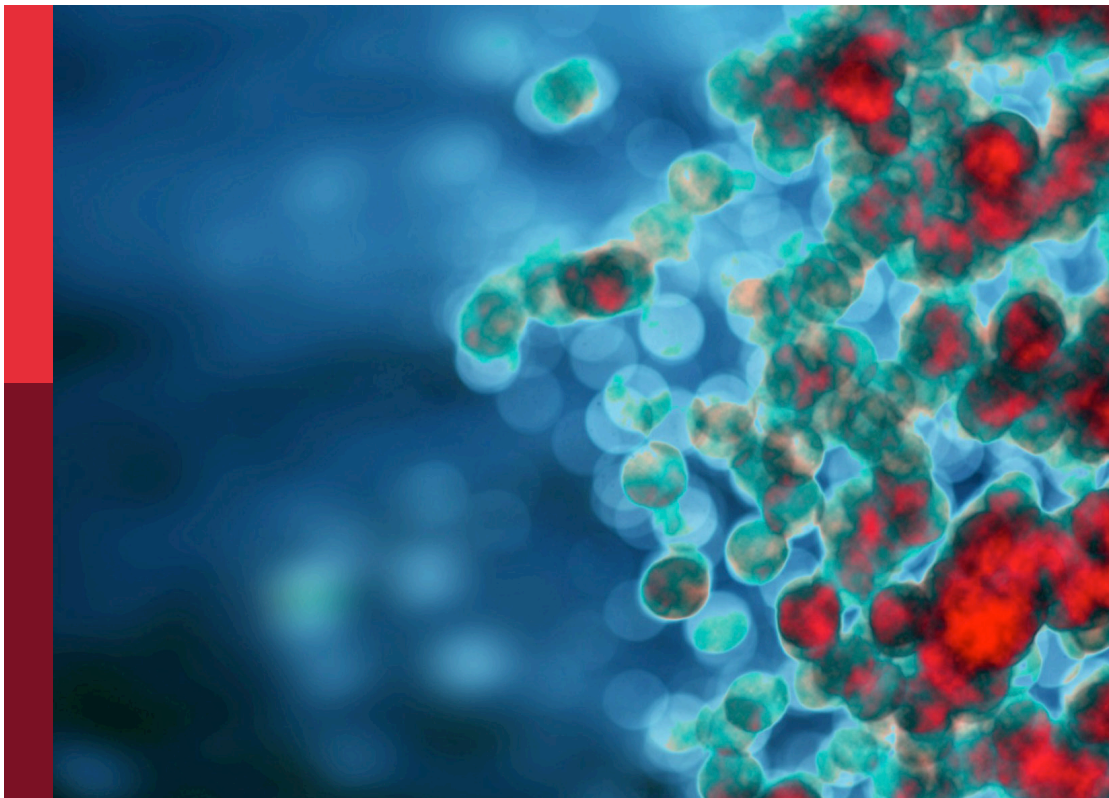
frontiers | Research Topics

Advances in mathematical and computational oncology, volume III

Edited by
George Bebis, Dinler Amaral Antunes, Ken Chen, Mohammad Kohandel, Kathleen Wilkie, Mamoru Kato and Jinzhuang Dou

Published in
Frontiers in Immunology
Frontiers in Oncology
Frontiers in Molecular Biosciences
Frontiers in Medicine
Frontiers in Artificial Intelligence



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3664-3
DOI 10.3389/978-2-8325-3664-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Advances in mathematical and computational oncology, volume III

Topic editors

George Bebis — University of Nevada, Reno, United States
Dinler Amaral Antunes — University of Houston, United States
Ken Chen — University of Texas MD Anderson Cancer Center, United States
Mohammad Kohandel — University of Waterloo, Canada
Kathleen Wilkie — Toronto Metropolitan University, Canada
Mamoru Kato — National Cancer Centre, Japan
Jinzhuang Dou — University of Texas MD Anderson Cancer Center, United States

Citation

Bebis, G., Antunes, D. A., Chen, K., Kohandel, M., Wilkie, K., Kato, M., Dou, J., eds. (2023). *Advances in mathematical and computational oncology, volume III*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3664-3

Table of contents

06 Editorial: Advances in mathematical and computational oncology, volume III
George Bebis, Mamoru Kato, Mohammad Kohandel, Kathleen Wilkie, Dinler A. Antunes, Ken Chen and Jinzhuang Dou

10 Tumor Microenvironment Characterization for Assessment of Recurrence and Survival Outcome in Gastric Cancer to Predict Chemotherapy and Immunotherapy Response
Yan Chen, Zepang Sun, Li Wan, Hongzhuan Chen, Tieju Xi and Yuming Jiang

23 Network Control Models With Personalized Genomics Data for Understanding Tumor Heterogeneity in Cancer
Jipeng Yan, Zhuo Hu, Zong-Wei Li, Shiren Sun and Wei-Feng Guo

33 DDOST Correlated with Malignancies and Immune Microenvironment in Gliomas
Xiaojing Chang, Jie Pan, Ruoyu Zhao, Tianfang Yan, Xinrui Wang, Cunle Guo, Yining Yang and Guohui Wang

46 Analysis of the Equilibrium Phase in Immune-Controlled Tumors Provides Hints for Designing Better Strategies for Cancer Treatment
Kevin Atsou, Sokchea Khou, Fabienne Anjuère, Véronique M. Braud and Thierry Goudon

71 Exploration of Novel Immunological Terms in Lung Cancer With Large Populations: Implications for Immunotherapy
Yuanshan Yao, Jing Wang, Fuzhi Yang and Wen Gao

85 A Signature Based on Costimulatory Molecules for the Assessment of Prognosis and Immune Characteristics in Patients With Stomach Adenocarcinoma
Bangjie Chen, Yong Yao, Deshen Mao, Conghan Li, Xingyu Wang, Shuyan Sheng, Lizhi Zhang, Xinyi Wang, Sanwei Chen, Wentao Xu, Jianyi Deng, Chenyu Sun, Qin Zhou, Scott Lowe, Rachel Bentley, Wei Shao and Haiwen Li

99 Oncogenic Role of miR-217 During Clear Cell Renal Carcinoma Progression
Jose María Zamora-Fuentes, Enrique Hernández-Lemus and Jesús Espinal-Enríquez

112 m6A-related metabolism molecular classification with distinct prognosis and immunotherapy response in soft tissue sarcoma
Zhen-Dong Huang, Yong-Cheng Fu, Shu-Yan Liu, Ya-Juan Mao, Yan Zhang, Chao Hu and Ren-Xiong Wei

126 Mathematical characterization of population dynamics in breast cancer cells treated with doxorubicin
Emily Y. Yang, Grant R. Howard, Amy Brock, Thomas E. Yankeelov and Guillermo Lorenzo

147 Construction and validation of nomograms based on the log odds of positive lymph nodes to predict the prognosis of lung neuroendocrine tumors
Suyu Wang, Juan Wei, Yibin Guo, Qiumeng Xu, Xin Lv, Yue Yu and Meiyun Liu

163 A novel cuproptosis-related molecular pattern and its tumor microenvironment characterization in colorectal cancer
Zhonglin Zhu, Qiuyan Zhao, Wang Song, Junyong Weng, Shanbao Li, Tianan Guo, Congcong Zhu and Ye Xu

179 Non-invasive measurement of tumor immune microenvironment and prediction of survival and chemotherapeutic benefits from ^{18}F fluorodeoxyglucose PET/CT images in gastric cancer
Junmeng Li, Chao Zhang, Huihui Guo, Shuang Li, Yang You, Peiming Zheng, Hongquan Zhang, Huanan Wang and Junwei Bai

191 Hydrophobicity identifies false positives and false negatives in peptide-MHC binding
Arnav Solanki, Marc Riedel, James Cornette, Julia Udell and George Vasmatzis

201 Assessing the efficacy of immunotherapy in lung squamous carcinoma using artificial intelligence neural network
Siqi Li, Wei Li, Tianyu Ma, Siyun Fu, Xiang Gao, Na Qin, Yuhua Wu, Xinyong Zhang, Jinghui Wang, Yuanming Pan and Zhidong Liu

213 Optimal predictive probability designs for randomized biomarker-guided oncology trials
Emily C. Zabor, Alexander M. Kaizer, Nathan A. Pennell and Brian P. Hobbs

224 SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth
Harsh Vardhan Jain, Kerri-Ann Norton, Bernardo Bianco Prado and Trachette L. Jackson

242 A new natural killer cell-specific gene signature predicting recurrence in colorectal cancer patients
Carolyn Shembrey, Momeneh Foroutan and Frédéric Hollande

257 Efficient feature extraction from highly sparse binary genotype data for cancer prognosis prediction using an auto-encoder
Junjie Shen, Huijun Li, Xinghao Yu, Lu Bai, Yongfei Dong, Jianping Cao, Ke Lu and Zaixiang Tang

269 Landmark mediation survival analysis using longitudinal surrogate
Jie Zhou, Xun Jiang, H. Amy Xia, Brian P. Hobbs and Peng Wei

283 A Boolean-based machine learning framework identifies predictive biomarkers of HSP90-targeted therapy response in prostate cancer
Sung-Young Shin, Margaret M. Centenera, Joshua T. Hodgson, Elizabeth V. Nguyen, Lisa M. Butler, Roger J. Daly and Lan K. Nguyen

299 Deep learning methods for drug response prediction in cancer: Predominant and emerging trends
Alexander Partin, Thomas S. Brettin, Yitan Zhu, Oleksandr Narykov, Austin Clyde, Jamie Overbeek and Rick L. Stevens

320 Construction of a lipid metabolism-related risk model for hepatocellular carcinoma by single cell and machine learning analysis
Lisha Mou, Zuhui Pu, Yongxiang Luo, Ryan Quan, Yunhu So and Hui Jiang

332 Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images
Alexander Partin, Thomas Brettin, Yitan Zhu, James M. Dolezal, Sara Kochanny, Alexander T. Pearson, Maulik Shukla, Yvonne A. Evrard, James H. Doroshow and Rick L. Stevens

343 Computational modeling of locoregional recurrence with spatial structure identifies tissue-specific carcinogenic profiles
Sharafudeen Dahiru Abubakar, Mitsuaki Takaki and Hiroshi Haeno

356 A computational model for the cancer field effect
Karl Deutscher, Thomas Hillen and Jay Newby

frontiers | Frontiers in Oncology

TYPE Editorial
PUBLISHED 25 September 2023
DOI 10.3389/fonc.2023.1282882

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Claudio Sette,
Catholic University of the Sacred Heart, Rome, Italy

*CORRESPONDENCE
George Bebis
bebis@cse.unr.edu

RECEIVED 25 August 2023
ACCEPTED 11 September 2023
PUBLISHED 25 September 2023

CITATION
Bebis G, Kato M, Kohandel M, Wilkie K, Antunes DA, Chen K and Dou J (2023) Editorial: Advances in mathematical and computational oncology, volume III. *Front. Oncol.* 13:1282882. doi: 10.3389/fonc.2023.1282882

COPYRIGHT
© 2023 Bebis, Kato, Kohandel, Wilkie, Antunes, Chen and Dou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Advances in mathematical and computational oncology, volume III

George Bebis[1]*, Mamoru Kato[2], Mohammad Kohandel[3], Kathleen Wilkie[4], Dinler A. Antunes[5], Ken Chen[6] and Jinzhuang Dou[6]

[1]Department of Computer Science and Engineering, University of Nevada, Reno, NV, United States, [2]Division of Bioinformatics, Research Institute, National Cancer Center Japan, Tokyo, Japan, [3]Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada, [4]Department of Mathematics, Ryerson University, Toronto, ON, Canada, [5]Department of Biology and Biochemistry, University of Houston, Houston, TX, United States, [6]Department of Bioinformatics and Computational Biology, Division of Basic Science Research, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

KEYWORDS

mathematical oncology, computational oncology, machine learning, mathematical modeling, computational modeling

Editorial on the Research Topic
Advances in mathematical and computational oncology, volume III

Despite significant advances in the understanding of the principal mechanisms leading to various cancer types, less progress has been made toward developing patient-specific treatments. Advanced mathematical and computational models could play a significant role in examining the most effective patient-specific therapies. Tumors, for example, undergo dynamic spatio-temporal changes, both during their progression and in response to therapies. Multiscale advanced mathematical and computational models could provide the tools to make therapeutic strategies adaptable enough and to address the emerging targets. Similarly, understanding the interrelationship amongst complex biological processes requires analyzing very large databases of cellular pathways. This Research Topic includes contributions to the state of the art and practice in mathematical and computational oncology addressing some of the challenges and difficulties in this field, as well as prototypes, systems, tools, and techniques.

Atsou et al. provided partial differential equations to represent interactions between immune and tumor cells. Introducing the equilibrium state between cancer cell proliferation and the elimination by immune cells, they invented a numerical method that determines the tumor size at the equilibrium from biological parameters. A sensitivity analysis was performed to find that the elimination rate of tumor cells by immune cells and their combinations with other parameters were the most influential factors. This suggests that the most effective strategy in immunotherapies is to act on the immune system rather than the tumor itself.

Chen et al. describes a study that investigated the tumor microenvironment (TME) characterization in gastric cancer (GC) patients and its association with recurrence, survival, and therapeutic response. The study used gene-expression data and clinical annotations from twelve cohorts of GC patients and evaluated the TME characteristics

using three computational algorithms. The study developed a TME-classifier, a TME-cluster, and a TME-based risk score to predict recurrence, survival, and response to chemotherapy and immunotherapy. It found that TME characterization was significantly associated with these outcomes and identified subgroups of patients who benefited from different treatments.

Yan et al. provide a timely to the field review on the application of network control methods in personalized cancer genomics. The authors have focused on methods that can be applied to one or few samples from an individual patient, for personalized medicine. Directions for future research are also highlighted providing a useful resource for researchers in the field.

RNA methylation (m6A) plays a significant role in numerous crucial physiological processes. Huang et al. analyze multi-omics data from 568 soft tissue sarcoma (STS) patients to investigate the relationship between m6A mRNA methylation, metabolic pathways, tumor microenvironment (TME), and patient prognosis. Using machine learning algorithms, they identify two distinct subtypes based on m6A-related metabolism and establish a scoring system (m6A-metabolic Score) to predict patient prognosis and immunotherapeutic responses. Additionally, they identify 11 m6A-related metabolic pathways associated with STS prognosis and demonstrate differences in TME characteristics and stemness features between the subtypes. This study provides insights into the potential for targeting m6A-related metabolism in STS treatment and guidance for personalized immunotherapy.

Chang et al. describes the role of DDOST (a protein involved in N-glycosylation) in gliomas, a type of brain tumor. The authors found that DDOST was overexpressed in gliomas and correlated with poor prognosis, tumor grade, IDH status, 1p19q status and MGMT methylation status. They showed that DDOST was associated with the immune microenvironment of gliomas and negatively related to tumor-infiltrating B cells and CD4+ T cells and positively related to CAFs and tumor-associated macrophages. They suggested that DDOST could be an important biomarker for diagnosing and treating gliomas.

Immunotherapy has provided new treatment options for cancer patients, with particular success against melanoma and lung cancer (1). However, response to treatment can vary dramatically, and there are no reliable biomarkers to predict response in a personalized manner. In this context, Yao et al. aimed at exploring novel immunological classifications associated with immunotherapy response through the Single-sample Gene Set Enrichment Analysis (ssGSEA) algorithm. They established a novel score to evaluate the immune-related risk (IRS). This IRS model was linked to prognosis and other tumor characteristics, predicting overall survival and immunotherapy response in non-small cell lung cancer patients.

Chen et al. examined the gene expression patterns of costimulatory molecule genes in patients with stomach adenocarcinoma (STAD) and developed a predictive signature to aid in therapy selection and outcome prediction. The authors conducted the first complete costimulatory molecular analysis in patients with STAD using 60 costimulatory family genes from prior research. They identified nine costimulatory molecular gene pairs (CMGPs) with prognostic value and developed a costimulatory molecule-related prognostic signature that performed well in an external dataset. The signature was proven to be a risk factor independent of the clinical characteristics. A further connection between the signature and immunotherapy response was discovered, suggesting that high-risk patients may have a better prognosis for immunotherapy.

The action of micro-RNAs (miRs) and their target genes during clear cell renal carcinoma (ccRC) progression was investigated by Zamora-Fuentes et al. To identify miRs that may have different roles during cancer progression, they developed a methodology for constructing miR-gene co-expression networks for each progression stage of ccRC, as well as for adjacent-normal renal tissue. Using these networks, they were able to observe those miR-gene interactions that are shared and unique for all the progression stages. The main finding of their study is that although miR-217 is differentially expressed in all contrasts, its targets were different depending on the ccRC stage.

There is a trend in implementing artificial intelligence methods on predicting cancer immunotherapy. Li et al. employed an auto-encoder workflow to extract features from binary genotype data for cancer prognosis prediction. The compressed transformation could better improve the model's original predictive performance and might avoid an overfitting problem due to the high dimensional data. Similarly, Shen et al. has demonstrated the efficiency of immunotherapy prediction based on artificial intelligence neural network. Their study enrolled 289 lung squamous carcinoma patients who received immunotherapy at Beijing Chest Hospital. With clinic features fed in the training model, the predicted disease control rate (DCR) can reach 0.95. In the work by Zhu et al. the researchers further characterize the tumor microenvironment (TME) of colorectal cancer based on cuproptosis-related molecular pattern. They derived the cuproptosis-related molecular patterns from 1,274 colorectal cancer samples. Such cuproptosis-related features are likely to strengthen our understanding of TME.

Zabor et al. proposes three randomized designs for early phase biomarker-guided oncology clinical trials. The designs use the optimal efficiency predictive probability method to monitor multiple biomarker subpopulations for futility. A simulation study results suggest that potentially smaller phase II trials can be designed using randomization and futility stopping to efficiently obtain more information about both the treatment and control groups prior to phase III study.

Yang et al. use mathematical modeling to explore the emergence of chemoresistance in a breast cancer cell line (MCF-7) treated with doxorubicin. Starting with logistic growth, they assume that a new subpopulation of irreversibly damaged cells is created after each dose. They fit their model to *in vitro* data and explore the dependence of model parameters on increasing drug dose, inter-treatment interval, and number of doses. This work suggests that longer delays between doses may promote chemoresistance.

A visual nomogram for cancer-specific survival and overall survival is proposed by Wang et al. Using lung neuroendocrine cancer SEER 18 data, they compute the log of the ratio of the number of positive lymph nodes to the number of dissected, non-positive lymph nodes. They then use multivariable Cox regression analysis to construct nomograms to predict survival, and show that

its prognostic ability is an improvement over the existing tumor-node-metastasis staging system.

Zhou et al. introduces a framework for conducting landmark mediation survival analyses that incorporate longitudinal assessment of tumor burden in clinical cancer trials. Zhou et al. compare the predictive performance of different tumor response characterizations, including conventional RECIST criteria and longitudinal tumor burden assessments using functional principal component (FPC) scores and integrated response measures. The framework is applied to two colorectal cancer trials, comparing survival prediction with and without longitudinal analysis. The findings show that longitudinal models utilizing FPC scores yield higher predictive accuracy, particularly when tumor burden exhibits U-shaped trends. In cases without U-shaped trends, binary objective response provides limited information. The study highlights the importance of incorporating longitudinal tumor burden data in treatment survival mediation analysis and suggests the practicality of FPC scores for capturing these patterns.

In recent years, natural killer (NK) cells have also emerged as an alternative to T-cell-based immunotherapy, as in cases of secondary resistance to checkpoint blockade therapy (2). NK cells are known to play a protective role in colorectal cancer (CRC), but most patients show limited intra-tumoral NK cell infiltration. In their new study, Shembrey et al. identified a new NK cell-specific gene signature to predict recurrence in colorectal cancer patients. By prioritising genes based on NK-specificity rather than expression level, they increased precision when determining the NK-specific contribution. High NK score was associated with several clinically useful molecular parameters, as well as with improved survival outcomes in CRC patients.

Developing noninvasive markers to assess the tumor microenvironment signature is of great significance in clinical practice. In the study by Li et al., the correlation between radiomics and immune microenvironment was determined using the preoperative PET/CT imaging data and the immunohistochemical results of the immune microenvironment of the pathological tissue after the operation. The main result of this study is that the radiomics score is related to the survival of patients and the benefit of adjuvant chemotherapy.

Some personalized forms of cancer immunotherapy rely on immunoinformatics tools, such as AI-based predictors of peptide-binding to Human Leukocyte Antigen (HLA) receptors (3). Although powerful, these methods can suffer from biases introduced by the limitations of available training sets, as recently demonstrated by Solanki et al. Their analysis showed that the allele-specific tool NetMHC-4.0 has a statistically significant bias towards predicting highly hydrophobic peptides as strong binders to two HLA alleles with differing hydrophobicity requirements. On the other hand, the pan-allele tool NetMHCpan-4.1 provides more reliable predictions, potentially due to the inclusion of immunopeptidomics on the training data.

Hepatocellular carcinoma (HCC) is a common and deadly cancer. Mou et al. focused on lipid metabolism-related genes differentially expressed between primary and metastatic HCC in single-cell RNA sequencing data. From these genes, they further selected 8 genes with a machine learning technique in TCGA bulk-cell RNA sequencing data and constructed a Cox regression model with the selected genes. Surprisingly, this model was validated with ICGC data. Clinical utility was shown in their monogram and decision curve analysis. This study represents an important step forward in the fight against the devastating disease.

Partin et al. presents a novel approach to improve drug response prediction in patient-derived xenografts (PDXs) using neural networks. The authors propose to combine drug descriptors, gene expressions and histology images as inputs for a multimodal neural network (MM-Net) and to augment the data by homogenizing drug representations and doubling drug-pair samples. They show that their method outperforms unimodal neural networks and baselines in terms of prediction performance. The article demonstrates the potential of data augmentation and multimodal learning for advancing cancer research with PDXs.

Agent-based models (ABMs) are effective tools for capturing intratumoral heterogeneity and integrating various scales of tumor dynamics. However, the computational costs of ABMs become expensive when simulating large cell populations, making parameter exploration and calibration challenging. Jain et al. present a novel method called Surrogate Modeling for Reconstructing Parameter Surfaces (SMoRe ParS), which uses surrogate models as intermediaries between ABM inputs and experimental data. The authors demonstrate the effectiveness of SMoRe ParS using an ABM of 3D vascular tumor growth as the computational model and utilized data obtained from tumor xenograft growth experiments as the real-world data. This application is an illustrative example of how SMoRe ParS can effectively connect the outputs of computational models with empirical data.

Deutscher et al. present a hybrid cellular automaton model describing the cancerization of an area due to carcinogen exposure. They focus on tobacco and alcohol exposure and implement a multi-layer perceptron to infer the resulting gene-related changes potentially leading to cancer development. With this model, they explore mutation and phenotypic evolution due to exposure, the effect of tumor excision, and the heterogeneity of clones in emerging tumors. Their results suggest that partial excision can lead to more aggressive recurrence and that tumors mainly form through polyclonality.

Partin et al. present a comprehensive review of machine learning methods for drug response prediction mainly from omics data, with strong attention to deep learning. This review extensively covers deep learning structures such as CNN, graph neural networks, and transformers, as well as learning schemes such as autoencoders, transfer learning, and multi-task learning. Additionally, the authors provide all other aspects of the method development, including data preparation, cancer models, measures of response, representations of drug compounds, cross-validation schemes, baseline models, and useful development practices. This review will be an essential guide for ones who develop and use machine learning for drug response prediction.

A machine learning methodology for the identification of multi-gene predictive biomarkers for targeted cancer drugs was developed by Shin et al. The methodology was applied to identify highly predictive biomarker panels for Hsp90-targeted treatment in prostate cancer from patient-derived proteomic data. The authors

were able to identify 5-protein panel biomarker that can predict cancer drug sensitivity. This is a promising approach, however, additional validation on different tumor types and cancer drugs is needed to further establish the usefulness of the method.

Local and regional recurrence of cancer after surgery is a major problem in cancer management, and the means to identify patients who are more likely to have a shorter recurrence are still lacking (4, 5). To address this issue, Abubakar et al. developed a mathematical model of cancer development using Moran and branching processes. The model is fit to historical data on disease-free survival in 27 difference cancer types and is used to estimate the cell turnover rate per month, relative fitness of pre-malignant cells, growth rate and death rate of cancer cells in each cancer type.

Author contributions

GB: Writing – original draft. MK: Writing – original draft. MK: Writing – original draft. KW: Writing – original draft. KC: Writing – original draft. DA: Writing – original draft. JD: Writing – original draft.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Montironi C, Pinedo OZ, Eldering E. Hematopoietic versus solid cancers and T cell dysfunction: looking for similarities and distinctions. *Cancers (Basel)* (2021) 13(2). doi: 10.3390/cancers13020284

2. Rodig SJ, Gusenleitner D, Jackson DG, Gjini E, Giobbie-Hurder A, Jin C, et al. MHC proteins confer differential sensitivity to CTLA-4 and PD-1 blockade in untreated metastatic melanoma. *Sci Trans Med Am Assoc Advancement Sci* (2018) 10:3342. doi: 10.1126/scitranslmed.aar3342

3. Nielsen M, Andreatta M, Peters B, Buus S. Immunoinformatics: predicting peptide-MHC binding. *Annu Rev Biomed Data Sci* (2020) 3:191–215. doi: 10.1146/annurev-biodatasci-021920-100259

4. Mahvi DA, Liu R, Grinstaff MW, Colson YL, Raut CP. Local cancer recurrence: the realities, challenges, and opportunities for new therapies. *CA Cancer J Clin* (2018) 68(6):488–505. doi: 10.3322/caac.21498

5. Karacz CM, Yan J, Zhu H, Gerber DE. Timing, sites, and correlates of lung cancer recurrence. *Clin Lung Cancer* (2020) 21(2):127–35 e3. doi: 10.1016/j.cllc.2019.12.001

frontiers | Frontiers in Immunology

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/fimmu.2022.890922

Tumor Microenvironment Characterization for Assessment of Recurrence and Survival Outcome in Gastric Cancer to Predict Chemotherapy and Immunotherapy Response

Yan Chen[1†], Zepang Sun[2†], Li Wan[1], Hongzhuan Chen[1], Tieju Xi[1] and Yuming Jiang[2,3*]*

[1] Shatou Community Health Service Center, Shenzhen Hospital of Integrated Traditional Chinese and Western Medicine, The Second People's Hospital of Bao'an Shenzhen (Group), Shenzhen Bao'an Shajing People's Hospital, Guangzhou Medical University, Shenzhen, China, [2] Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China, [3] Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA, United States

Background: The tumor microenvironment (TME) is crucial for tumor recurrence, prognosis, and therapeutic responses. We comprehensively investigated the TME characterization associated with relapse and survival outcomes of gastric cancer (GC) to predict chemotherapy and immunotherapy response.

Methods: A total of 2,456 GC patients with complete gene-expression data and clinical annotations from twelve cohorts were included. The TME characteristics were evaluated using three proposed computational algorithms. We then developed a TME-classifier, a TME-cluster, and a TME-based risk score for the assessment of tumor recurrence and prognosis in patients with GC to predict chemotherapy and immunotherapy response.

Results: Patients with tumor recurrence presented with inactive immunogenicity, namely, high infiltration of tumor-associated stromal cells, low infiltration of tumor-associated immunoactivated lymphocytes, high stromal score, and low immune score. The TME-classifier of 4 subtypes with distinct clinicopathology, genomic, and molecular characteristics was significantly associated with tumor recurrence (P = 0.002), disease-free survival (DFS, P <0.001), and overall survival (OS, P <0.001) adjusted by confounding variables in 1,193 stage I–III GC patients who underwent potential radical surgery. The TME cluster and TME-based risk score can also predict DFS (P <0.001) and OS (P <0.001). More importantly, we found that patients in the TMEclassifier-A, TMEclassifier-C, and TMEclassifier-D groups benefited from adjuvant chemotherapy, and patients in the TMEclassifier-B group without chemotherapy benefit responded best to pembrolizumab treatment (PD-1 inhibitor), followed by patients in the TMEclassifier-A, while patients in the C and D groups of the TMEclassifier responded poorly to immunotherapy.

OPEN ACCESS

Edited by:
Ken Chen,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Xianli Jiang,
University of Texas MD Anderson Cancer Center, United States
Merve Dede,
University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Yuming Jiang
jiangymbest@163.com
Yan Chen
494357044@qq.com

†These authors share first authorship

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: 07 March 2022
Accepted: 11 April 2022
Published: 29 April 2022

Citation:
Chen Y, Sun Z, Wan L, Chen H, Xi T and Jiang Y (2022) Tumor Microenvironment Characterization for Assessment of Recurrence and Survival Outcome in Gastric Cancer to Predict Chemotherapy and Immunotherapy Response. Front. Immunol. 13:890922. doi: 10.3389/fimmu.2022.890922

Conclusion: We determined that TME characterization is significantly associated with tumor recurrence and prognosis. The TME-classifier we proposed can guide individualized chemotherapy and immunotherapy decision-making.

Keywords: tumor microenvironment, gastric cancer, recurrence, prognosis, chemotherapy, immunotherapy, classifier (classification tool)

INTRODUCTION

Gastric cancer (GC) is the fifth most common malignancy and the third leading cause of cancer-related deaths in the world (1). Despite significant advances in early diagnosis and treatment, the 5-year overall survival (OS) rate for patients with local GC remains around 40% in western countries (2, 3). Furthermore, most patients with GC die of tumor recurrence or metastasis (1–3). It has been reported that about 40% of GC patients relapse within 2 years after surgery, which has often been found in peritoneal, hematogenous, and nodal metastases (4, 5). Moreover, the median survival time from recurrence to death is approximately 4–6 months (5, 6). Additionally, with early diagnosis of GC, radical surgery is being performed more frequently, which has resulted in a rapid increase in relapse. Worse yet, there are no effective therapies for GC recurrence (1–7). Therefore, the prediction of GC recurrence in clinical practice seems to be of great importance.

The widespread application of genomic sequencing technology in tumor studies has provided us with possibilities for us to dissect the potential characteristics of the tumor microenvironment (TME) in GC (8–10). The TME, composed of the extracellular matrix, signaling molecules, and non-tumor cells, is heterogeneous. A growing body of evidence has suggested that TME is crucial in tumor progression and therapeutic responses (11–14). For instance, different compositions of tumor-associated cells infiltrating, namely, cytotoxic T cells, follicular helper T cells, natural killer cells, dendritic cells, tumor-associated macrophages, and cancer-associated fibroblasts, in TME are associated with diverse clinical outcomes, chemotherapy benefits, and immunotherapy responses (12–16). Thus, a more in-depth understanding of TME is indispensable and urgent. However, to date, the comprehensive landscape of TME characteristics in tumor recurrence has not been elucidated.

Evaluation of tumor recurrence in GC presents a major challenge. Traditional imaging modalities, such as X-ray, computed tomography (CT), Positron Emission Tomography-Computed Tomography (PET-CT), or ultrasonography, yield unsatisfactory predictions of GC recurrence (17). Additionally, these techniques could only detect metastases after their occurrence, thus delaying treatments. Using liquid biopsy, such as circulating tumor DNA, and radiomics, are the emerging means for detecting recurrence (18, 19). However, both are technically demanding and prone to interference currently. Thus, assessment of GC recurrence-associated TME characteristics and identification of patients at high risk of relapse after definitive therapy have great significance. Increased surveillance and early intervention may improve the quality of life and survival for these patients.

The heterogeneous nature of GC results in diverse clinicopathological and molecular features, which generates greater challenges to individualized diagnosis and treatment (20). Although chemotherapy and immunotherapy have improved survival in some GC patients, they are associated with unavoidable side effects (4, 21). However, only 9% absolute benefit is observed from adjuvant chemotherapy compared with the surgery-only group, and only 10–26% objective response rate is achieved in GC patients treated with immune checkpoint blockade (ICB), such as anti-programmed cell death protein 1 (PD-1) and anti-programmed death-ligand 1 (PD-L1) (4, 21–23). A considerable number of patients do not benefit from these treatments, suggesting that they could be spared from excessive intervention. Therefore, it is rewarding to identify patients with a response or resistance to specific treatments before initiation.

Recently, several molecular classification systems have been proposed for individualized diagnosis and treatment based on whole-genome and transcriptome data (8, 24–26). The Asian Cancer Research Group (ACRG) divided GC into 4 subtypes: microsatellite instability (MSI), mesenchymal transition (EMT), microsatellite stability/the tumor protein 53-active (MSS/TP53$^+$), and microsatellite stability/the tumor protein 53-inactive (MSS/TP53$^-$) (24). Alternatively, the Cancer Genome Atlas (TCGA) proposed four molecular subtypes: Epstein–Barr virus (EBV)-positive, microsatellite instability (MSI), genomically stable (GS), and chromosomal instability (CIN) (25). Subsequent studies defined other classification models based on tumor-associated infiltrating cells or mesenchymal–epithelial phenotypes (8, 24). However, the association between TME and GC recurrence has not been thoroughly explored in these studies.

The widely used algorithms, namely, CIBERSORT, MCPcounter, and ESTIMATE, have enabled us to explore the relationship between TME and tumor recurrence, survival, and therapeutic responses from bulk gene expression data recently (27–29). However, there are some differences in the cell compositions and data types calculated using different algorithms. Thus, a combination of multiple algorithms can complement each other and strengthen the conclusion, which may provide a better characterization of the TME.

In this work, we systematically evaluated the cellular component and prognostic landscape of the TME associated with GC relapse using three proposed computational algorithms (27–29). We next proposed a TME classifier composed of 4 subtypes associated with tumor recurrence and then validated its prognostic value for disease-free survival (DFS) and overall survival (OS) in multiple independent cohorts with 2,411

patients. Moreover, a novel classification system was observed with distinct clinicopathology, molecular, genomic, and epigenetic characteristics. Furthermore, we developed a TME-based risk score to predict DFS and OS, and confirmed that a TME-cluster of 3 phenotypes could predict DFS. Importantly, the TME classifier of the 4 subtypes we proposed could predict chemotherapy and immunotherapy responses.

MATERIALS AND METHODS

Gastric Cancer Datasets and Clinical–Genomic Information

We conducted a systematic search for GC gene expression dataset, which were publicly accessible and had complete clinical annotations. Cases without survival information were excluded from further analysis. Totally, we achieved eleven cohorts of 2,411 patients with GC for this study: GSE62254/ACRG, GSE26253/SMC, GSE13861/YUHS, GSE26899/KUGH, GSE26901/KUCM, GSE15460/SGP, GSE28541/MDACC, GSE29272/TYB, GSE57308/CGH, GSE84437/KOREA, and TCGA-STAD. Raw data for the microarray datasets generated by Affymetrix or Illumina platform were screened from the Gene-Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/), and processed for background adjustment, quantile normalization, and final summarization by Perl software and limma packages. The corresponding clinical information was downloaded or manually registered from the item page in the GEO dataset website. For some series whose clinical data could not be obtained through the aforementioned methods, we retrieved the exact clinical information from the supplementary materials of relevant published papers (26, 30, 31). Level 4 gene-expression profile (FPKM normalized) and corresponding clinical data of The Cancer Genome Atlas (TCGA) were downloaded from the Genomic Data Commons (https://portal.gdc.cancer.gov/). Missing or updated clinical–genomic data was replenished from the UCSC Xena browser (GDC hub: https://gdc.xenahubs.net) and cBio Cancer Genomics Portal (cBioPortal: https://www.cbioportal.org/). DFS was defined as the time to recurrence at any site. OS was defined as the time to death from any cause. Among them, relapse information was recorded for six cohorts of ACRG, SMC, YUHS, KUGH, KUCM, and TCGA-STAD.

All patients except for the SGP cohort were reported with or without an operation note. Most patients were classified according to the American Joint Committee on Cancer (AJCC) 6th edition or 7th edition (32, 33). We redefined patients into the AJCC 8th edition based on the status of tumor-node-metastasis (TNM) available when necessary, with accuracy as the first criterion (34). Under these circumstances, we defined four meta-cohorts (A, B, C, and D) using cases with specific characteristics from the aforementioned eleven cohorts. The meta-cohort A (1,193 patients) included stage I–III patients with relapse records after potential radical surgery based on the AJCC 8th edition. The meta-cohort B (1,365 patients) included stage I–IV patients from six cohorts of ACRG, SMC, YUHS, KUGH, KUCM, and TCGA-STAD. The meta-cohort C (1,046 patients) included stage I–IV patients from five cohorts of SGP, MDACC, TYB, CGH, and Korea. The meta-cohort D (2,411 patients) consisted of meta-cohorts B and C. Adjuvant chemotherapy was mainly fluorouracil-based and was available in six cohorts: ACRG, SMC, YUHS, KUGH, KUCM, and MDACC.

Gastric Cancer Dataset With PD-1 Inhibition Treatment

The original paired gene sequence and corresponding clinical information of the PRJEB25780 cohort (pembrolizumab treatment) were downloaded from the European Bioinformatics Institute (EMBL-EBI) database (https://www.ebi.ac.uk/). The adapter and low-quality sequences were removed from the raw data using Trim Galore software. The quality of the samples after filtration was checked and adjusted by the FastQC software. Clean reads were compared with the human genome (HG38 version) using HISAT2 software, and then a read count of gene expression was generated using FeatureCounts software. Finally, the gene expression profile was normalized by the limma package.

Tumor Microenvironment Infiltrating Cells Dissecting

To quantify the composition of tumor-associated infiltrating cells in the GC samples, two proposed computational algorithms, CIBERSORT and Microenvironment Cell Populations-counter (MCPcounter), were conducted (27, 28). Based on the gene expression profiles, the CIBERSORT algorithm was employed to quantify the proportions of tumor-infiltrating immune cells using standard reference files at parameter settings of LM22 signature and 1,000 permutations (27). A total of 22 types of immune cells, namely, naive B and memory B cells, $CD8^+$ T cells, naive $CD4^+$ T cells, resting memory $CD4^+$ T cells, activated memory $CD4^+$ T cells, T follicular helper cells (Tfh), regulatory T cells (Tregs), resting natural killer (NK) cells, activated NK cells, M0 macrophages, M1 macrophages, M2 macrophages, resting dendritic cells (DC), activated DC, resting mast cells, activated mast cells, plasma cells, gamma delta T cells, monocytes, neutrophils, and eosinophils, were obtained through this algorithm (27). Additionally, the absolute abundance of ten kinds of immune-stromal associated cells, including two stromal cells (tumor-associated fibroblasts and endothelial cells) and eight immune cells (CD3 T cells, CD8 T cells, cytotoxic lymphocytes, B cell lineage, NK cells, monocytic lineage, myeloid dendritic cells, and neutrophils), was estimated by the MCPcounter algorithm (28).

Discovery and Validation of the TME-Classifier

The tumor-associated immune and stromal scores representing TME characterization were calculated based on the normalized gene-expression matrix using the ESTIMATE algorithm for each GC sample (29). Subsequently, for each dataset, the expression of the immune-stromal score was transformed into a z-score. We then classified patients into a high-

immune group and a low-immune group or a high-stroma group and a low-stroma group using a score of 0 as the cutoff, as described in our published paper (11). Furthermore, we developed a TME-classifier of 4 subtypes based on the aforementioned results: TMEclassifier-A (low-immune and low-stroma score), TMEclassifier-B (high-immune and low-stroma score), TMEclassifier-C (low-immune and high-stroma score), and TMEclassifier-D (high-immune and high-stroma score). Finally, the TME-classifier of 4 subtypes was validated in 12 independent cohorts and 4 meta-cohorts to assess GC recurrence, clinical–genomic characteristics, components of tumor-associated infiltrating cells, DFS, OS, chemotherapy, and immunotherapy responses.

TME-Cluster and TME-Based Risk Score Developing

To further dissect the association between TME characteristics and GC recurrence, we employed an unsupervised consensus clustering algorithm on 22 tumor-associated infiltrating immune cells and 2 tumor-associated infiltrating stromal cells, whose values had been standardized by Z-score and defined the robust subgroup of patients (35). Next, given the diverse prognostic value of the immune and stromal scores described previously, we integrated them into a comprehensive TME-based risk score. The TME risk score was calculated using the following equation: TME risk score = tumor-associated stromal score − tumor-associated immune score. Both the TME-cluster and TME-based risk scores were used to evaluate DFS and OS. Similarly, the TME-cluster of 3 subtypes was used to predict chemotherapy and immunotherapy responses.

Functional Enrichment Analysis

Gene annotation enrichment analysis was performed to gain an in-depth understanding of tumor microenvironment characteristics using the R package clusterProfiler (36). An adjusted P-value of <0.05 was identified as significant. Considering the quantity, quality, completeness, and representativeness of the datasets, gene set enrichment analyses (GSEA) were performed in the TCGA-STAD and KOREA cohorts.

Prediction of Chemotherapy and Immunotherapy Response

We next evaluated the predictive performance of our TME-classifier and TME-cluster on chemotherapy and immunotherapy responses in a meta-cohort of 903 GC with chemotherapy information and the PRJEB25780 cohort with anti-PD-1 treatment. Specifically, Kaplan–Meier curves for overall survival were performed to decide which subgroup of TME-classifier and TME-cluster could benefit from chemotherapy. Moreover, the rate of immunotherapy response, including an objective response rate (ORR), was calculated to be observed in which subgroups of TME-classifier and TME-cluster could benefit from immunotherapy. Meanwhile, the area under the receiver operating characteristic curve (AUC) was used to assess the predictive power.

Statistical Analysis

Continuous variables were compared among groups using the t-test, Mann–Whitney U test, or Kruskal–Wallis test. Enumeration data were compared among groups by the Chi-square or Fisher exact test. All statistical analyses were performed using R software (version 3.5.3), origin software (version 2019b) and SPSS statistical software (version 24.0). A two-sided of $P < 0.05$ was considered statistically significant.

RESULTS

Tumor Microenvironment Characterization Associated With Cancer Recurrence

A total of 2,456 patients from 12 cohorts were included in this study. The clinicopathological and treatment information is presented in **Table 1** and **Table S1**. After preprocessing, 1,193 stage I–III patients with relapse records after potential radical surgery were identified from 6 cohorts, which included ACRG (n = 257), SMC (n = 365), YUHS (n = 55), KUCM (n = 102), KUGH (n = 85), and TCGA-STAD (n = 329). These cohorts and an integrated meta-cohort were used to assess the association between the TME characteristics and tumor recurrence. On the whole, our data showed that patients without recurrence at the last follow-up had an active immune response and inactive immunosuppression (**Figures S1**, **S2**). Specifically, significantly higher infiltration of CD4 activated cells, NK activated cells, plasma cells, but significantly lower infiltration of M2 macrophages in patients without recurrence in cohorts of ACRG, TCGA-STAD, SMC, KUCM, and KUGH, separately (**Figure S3A**). Additionally, patients with recurrence exhibited a significantly higher abundance of tumor-associated stromal cells, including fibroblasts and endothelial cells, in all cohorts, especially ACRG, SMC, KUCM, and KUGH (**Figures S2**, **S3B**). Importantly, and believably, a higher stromal score and a lower immune score were found in patients with recurrence (**Figures S3C, D**).

TME-Classifier Can Predict Disease-Free Survival and Overall Survival Independent of the TNM Staging

Next, we developed the TME-classifier, which divided patients into 4 subtypes based on the low-high immune and low-high stromal scores (**Figures 1A, B**). **Table S2** contains the exact information of the TME-classifier for 11 cohorts. Most importantly of all, we demonstrated that the TMEclassifier was a robust prognostic biomarker for DFS and OS in multiple cohorts of ACRG, SMC, YUHS, KUGH, KUCM, TCGA, KOREA, SGP, CGH, MDACC, TYB, meta-cohort A, meta-cohort B, meta-cohort C, and meta-cohort D (**Figures 1C–J** and **2A–H**). Although statistical significance was not found in some cohorts due to a small sample or short follow-up period, there was a similar tendency in prognosis: the TMEclassifier-B showed the best prognosis, followed by the TMEclassifier-A, TMEclassifier-D, and then TMEclassifier-C. **Tables S3–13** contain a summary of these findings. More importantly, when

TABLE 1 | Clinicopathologic and treatment information of patients with gastric cancer in eleven public cohorts.

Variables	ACRG	SMC	YUHS	KUCM	KUGH	TCGA-STAD	SGP	MDACC	TYB	CGH	KOREA
No. of patients	300	432	59	109	93	372	248	40	126	199	433
Median age (range)	64 (24–86)	53 (23–74)	63 (32–83)	58 (28–74)	60 (36–83)	67 (35–90)	68 (23–92)	58 (33–78)	59 (23–71)	61 (34–81)	62 (27–86)
Male (%)	199 (66.3)	280 (64.8)	41 (69.5)	69 (63.3)	73 (78.5)	240 (64.5)	161 (64.9)	27 (67.5)	99 (78.6)	145 (72.9)	296 (68.4)
Location (%)											
Cardia	32 (10.7)	54 (12.5)	4 (6.8)	13 (11.9)	9 (9.7)	133 (35.8)	–	–	–	–	–
Body	107 (35.7)	139 (32.2)	28 (47.5)	36 (33.0)	29 (31.2)	89 (23.9)	–	–	–	–	–
Antrum	155 (51.7)	226 (52.3)	24 (40.7)	56 (51.4)	55 (59.1)	134 (36.0)	–	–	–	–	–
Whole	6 (2.0)	13 (3.0)	1 (1.7)	4 (3.7)	–	–	–	–	–	–	–
Lauren type (%)											
Intestinal	150 (50.0)	139 (32.2)	17 (28.8)	82 (75.2)	59 (63.4)	162 (43.5)	138 (55.6)	–	–	55 (27.6)	–
Diffuse	141 (47.0)	280 (64.8)	28 (47.5)	11 (10.1)	31 (33.3)	74 (19.9)	86 (34.7)	–	–	101 (50.8)	–
Mixed	9 (3.0)	13 (3.0)	12 (20.3)	5 (4.6)	2 (2.2)	–	22 (8.9)	–	–	43 (21.6)	–
T stage (%)											
T1	0	–	–	–	–	20 (5.4)	–	–	–	1 (0.5)	11 (2.5)
T2	188 (62.7)	–	–	–	–	80 (21.5)	–	–	–	25 (12.6)	38 (8.8)
T3	91 (30.3)	–	–	–	–	170 (45.7)	–	–	–	145 (72.9)	92 (21.2)
T4	21 (7.0)	–	–	–	–	98 (26.3)	–	–	–	28 (14.1)	292 (67.4)
N stage (%)											
N0	38 (12.7)	–	–	–	–	114 (30.6)	–	–	–	42 (21.1)	80 (18.5)
N1	131 (43.7)	–	–	–	–	101 (27.2)	–	–	–	72 (36.2)	188 (43.4)
N2	80 (26.7)	–	–	–	–	75 (20.2)	–	–	–	66 (33.2)	132 (30.5)
N3	51 (17.0)	–	–	–	–	75 (20.2)	–	–	–	19 (9.5)	33 (7.6)
M stage (%)											
M0	273 (91.0)	–	55 (93.2)	102 (93.6)	85 (91.4)	347 (93.3)	–	–	–	175 (87.9)	–
M1	27 (9.0)	–	4 (6.8)	7 (6.4)	7 (7.5)	23 (6.2)	–	–	–	24 (12.1)	–
TNM stage (%)											
I	30 (10.0)	68 (15.7)	12 (20.3)	40 (36.7)	11 (11.8)	52 (14.0)	42 (16.9)	1 (2.5)	5 (4.0)	9 (4.5)	–
II	97 (32.3)	167 (38.7)	11 (18.6)	18 (16.5)	18 (19.4)	112 (30.0)	42 (16.9)	6 (15.0)	5 (4.0)	31 (15.6)	–
III	96 (32.0)	130 (30.1)	24 (40.7)	36 (33.0)	27 (29.0)	155 (41.7)	91 (36.7)	12 (30.0)	108 (85.7)	104 (52.3)	–
IV	77 (25.7)	67 (15.5)	12 (20.3)	15 (13.8)	36 (38.7)	39 (10.5)	73 (29.4)	21 (52.5)	8 (6.3)	55 (27.6)	–
Potential radical surgery	Reported	Reported	Reported	Reported	Reported	Reported	Not reported	Reported	Reported	Reported	Reported
Chemotherapy (%)	144 (48.0)	432 (100)	45 (76.3)	39 (35.8)	67 (72.0)	–	–	40 (100)	–	–	–
Recurrence (%)	125 (41.7)	177 (41.0)	26 (44.1)	60 (55.0)	32 (34.4)	123 (33.1)	–	–	–	–	–
Median follow-up, mo.	80	84	91	94	49	22	75	85	70	51	116

A dash indicates that data are not available. The version of TNM staging system for ACRG, SMC, YUHS, KUCM, KUGH, TCGA, SGP, MDACC, TYB, CGH, and KOREA was the American Joint Committee on Cancer (AJCC) 6th or 7th edition.

integrated into a meta-cohort A, a meta-cohort B, a meta-cohort C, and a meta-cohort D, a more robust outcome for TMEclassifier to predict DFS (P <0.001) and OS (P <0.001) in univariate and multivariate Cox analyses was confirmed (**Table S14** and **Table 2**), as demonstrated by the Kaplan–Meier curves (**Figures 1**, **2**).

Tumor Microenvironment Landscape of the TME-Classifier

Based on the above phenomenon, we confirmed that TME characterization was closely associated with cancer recurrence and survival outcome. We then screened and analyzed the tumor-associated infiltrating immune-stromal cell landscape of the TME-classifier in 1,193 stage I–III patients with relapse information. Generally, compared with patients in the TMEclassifier-C + D group, we found that patients in the TMEclassifier-B group were characterized by a higher abundance of immunoactivating cells and a lower abundance of immunosuppressive cells. For example, a high abundance of $CD8^+$ T cells, T follicular helper cells, memory CD4 T-activated cells, M1 macrophages, cytotoxic lymphocytes, and NK cells, while a low abundance of M2 macrophages, regulatory T cells, and mast cells were observed in patients of the TMEclassifier-B group (**Figure 3A** and **Figure S4**). Additionally, compared with patients in the TMEclassifier-B + C + D group, we found that patients in the TMEclassifier-A group presented the lowest abundance of immune-stromal cells, especially calculated by the MCPcounter algorithm, and the lowest immune-stromal score calculated by the ESTIMATE algorithm, which may be associated with immune deficiency (**Figure 3A** and **Figure S4**). Furthermore, compared with the TMEclassifier-A + B group, patients in the TMEclassifier-C + D group showed more characteristics of immunosuppression (high abundance of fibroblasts and endothelial cells, and high stromal score), especially in the TMEclassifier-C group (**Figure 3A** and **Figure S4**). Moreover, the findings suggested a complete immunosuppression phenomenon in the TMEclassifier-C group, while there was immunosuppression and immune-activation phenomenon (high immune score and limited immunoactivating cells) in the TMEclassifier-D group (**Figure 3A** and **Figure S4**). These results may further explain the different survival outcomes in the 4 subtypes of the TME-classifier.

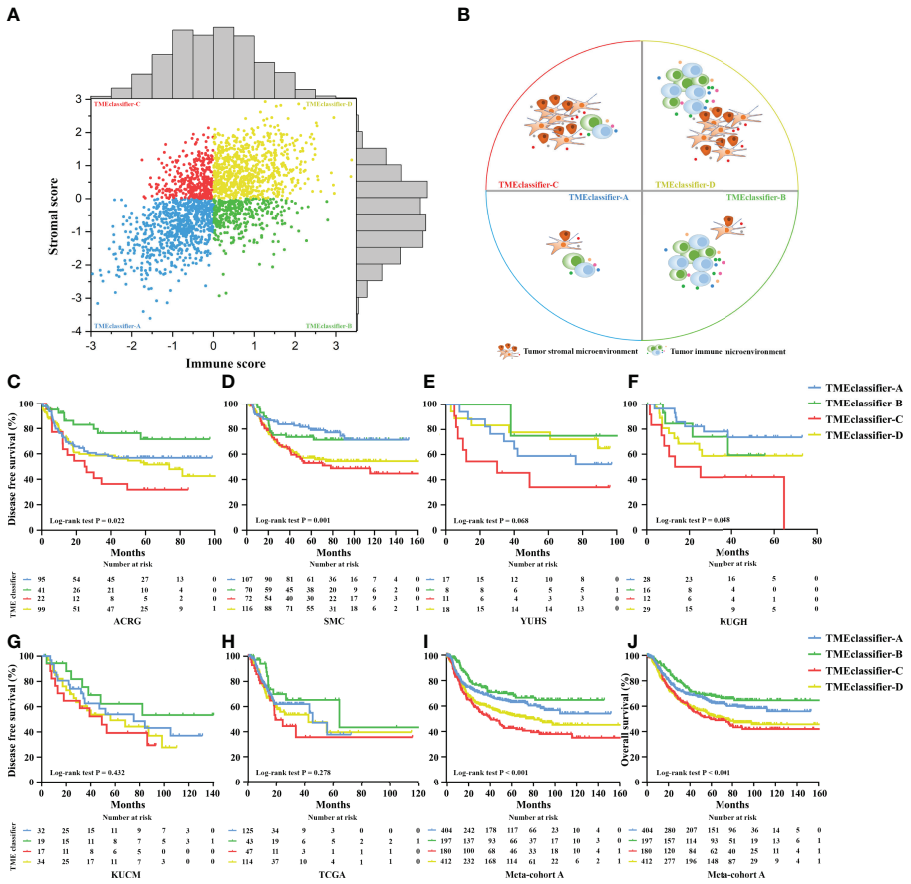


FIGURE 1 | Patients were divided into 4 subtypes of TME-classifier, based on the immune score and stromal score **(A, B)**. Kaplan–Meier curves for disease-free survival of patients with stage I–III gastric cancer after potential radical surgery based on TME-classifier in the ACRG **(C)**, SMC **(D)**, YUHS **(E)**, KUGH **(F)**, KUCM **(G)**, and TCGA **(H)** cohorts. Kaplan–Meier curves for disease-free survival **(I)**, and overall survival **(J)** of 1,193 patients with stage I–III gastric cancer after potential radical surgery based on TME-classifier in the meta-cohort A.

Assessing the Clinical, Genomic, and Molecular Characteristics Associated With the TME-Classifier

A higher proportion of earlier-stage GC patients was observed in the TMEclassifier-A + B group (**Figure 3B**). Moreover, the recurrence rate in the TMEclassifier-C + D group was significantly higher than that in the TMEclassifier-A + B group (**Figure 3C**). In particular, patients in the TMEclassifier-B group had the lowest recurrence rate. Similarly, we found that GC patients with cancer recurrence in 6 cohorts of ACRG, SMC, YUHS, KUCM, KUGH, and TCGA-STAD, had more patients in the TMEclassifier-C and TMEclassifier-D groups, but had fewer patients in the TMEclassifier-A and TMEclassifier-B groups (**Figure S5A**). Furthermore, when integrated into meta-cohort A, the number and percentage of TMEclassifier-A and TMEclassifier-B samples were still larger in patients without tumor recurrence (**Figures S5B, C**). In multivariable logistic regression analyses (**Table S15**), the TMEclassifier remained a significant predictor (P = 0.002) of GC recurrence after adjusting for clinicopathological factors. Likewise, the tumor location distribution among the 4 subtypes of TME-classifier was consistent (**Figure 3D**). Additionally, patients in the TMEclassifier-C + D group had a higher proportion of diffuse type GC (**Figure 3E**). Interestingly, the mutational load (mutation count and TMB) in the TMEclassifier-A + B group, especially group B, was significantly higher than that in the TMEclassifier-C + D group, showing an immunogenicity difference (**Figures 3F, G**). These statistical comparisons are presented in **Table S16**.

We also performed a comparison between our classification system and several existing molecular subtypes. Our findings showed that TMEcluster-A and -B groups had a higher degree of overlap with TMEclassifier-A and B groups, while almost all TMEcluster-C patients were observed in TMEclassifier-C and D groups (**Figure 3H** and **Figure S6A**). The EMT subtype of the ACRG classification with the worst prognosis had not been detected in patients of the TMEclassifier-A + B group, and patients in the TMEclassifier-B group presented the largest proportion of MSI and MSS/TP53^{+} subtype, followed by TMEclassifier-A (**Figure 3I** and **Figure S6B**). Additionally, we

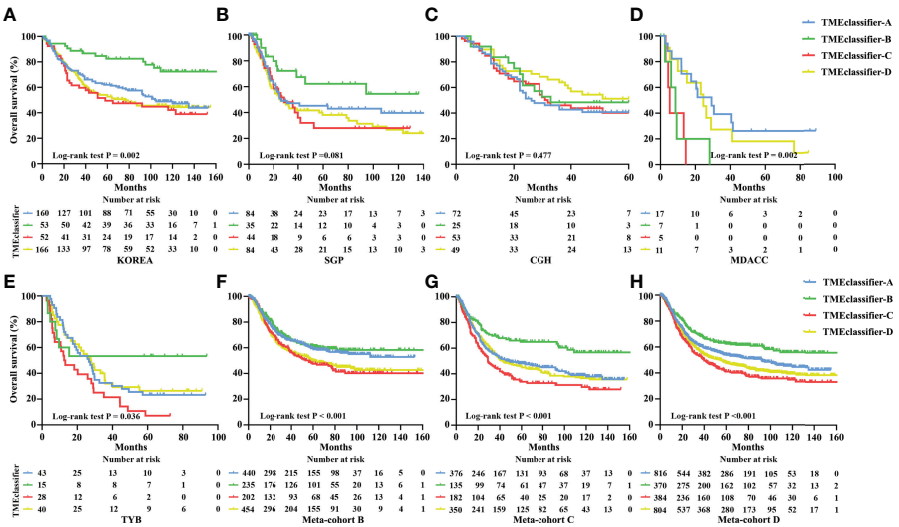


FIGURE 2 | Kaplan–Meier curves for overall survival of patients with stage I–IV gastric cancer based on TME-classifier in the KOREA **(A)**, SGP **(B)**, CGH **(C)**, MDACC **(D)**, and TYB **(E)** cohorts. Kaplan–Meier curves for overall survival **(F)** of 1,365 patients with stage I–IV gastric cancer based on TME-classifier in the meta-cohort B Kaplan–Meier curves for overall survival **(G)** of 1,046 patients with stage I–IV gastric cancer based on TME-classifier in the meta-cohort C Kaplan–Meier curves for overall survival **(H)** of 2,411 patients with stage I–IV gastric cancer based on TME-classifier in the meta-cohort D.

found significantly more patients with the EBV or MSI subtype of the TCGA classification in the TMEclassifier-B group, and a higher percentage of GS subtype was found in the TMEclassifier-C + D group (**Figure 3J** and **Figure S6C**). Meanwhile, there was almost no detection of any MP subtype in the TMEclassifier-A + B group, and the lowest proportion of EP cases was found in the TMEclassifier-C group (**Figure 3K** and **Figure S6D**). These results were highly consistent and significant, which confirmed the credibility and accuracy of our discovery.

TME Cluster and TME-Based Risk Score Are Markers for Prognosis

Furthermore, we developed a TME cluster and a TME-based risk score as per the aforementioned description. The optimal

TABLE 2 | Multivariable cox regression analyses for disease-free survival and overall survival in gastric cancer meta-cohort.

Variables	Disease-free survival		Overall survival	
	HR (95%CI)	P	HR (95%CI)	P
Age	1.010 (1.003–1.020)	0.047	1.008 (1.000–1.017)	0.062
Location				
Cardia	–	–	1	Reference
Body	–	–	1.043 (0.759–1.434)	0.793
Antrum	–	–	0.882 (0.645–1.205)	0.430
Whole	–	–	1.208 (0.659–2.215)	0.540
TNM stage				
I	1	Reference	1	Reference
II	3.100 (2.038-4.715)	<0.001	3.268 (2.153–4.960)	<0.001
III	5.687 (3.692–8.760)	<0.001	5.867 (3.815–9.022)	<0.001
IV	–	–	10.769 (6.848–16.935)	<0.001
Chemotherapy				
No	1	Reference	1	Reference
Yes	0.494 (0.386–0.633)	<0.001	0.446 (0.357–0.556)	<0.001
TMEclassifier				
A	1	Reference	1	Reference
B	0.795 (0.554–1.142)	0.214	0.808 (0.596–1.095)	0.169
C	1.697 (1.239–2.323)	0.001	1.463 (1.097–1.953)	0.010
D	1.392 (1.063–1.824)	0.016	1.300 (1.014–1.665)	0.038

CI, confidence interval; HR, hazard ratio. Patients for disease-free survival assessment had undergone potential radical surgery.

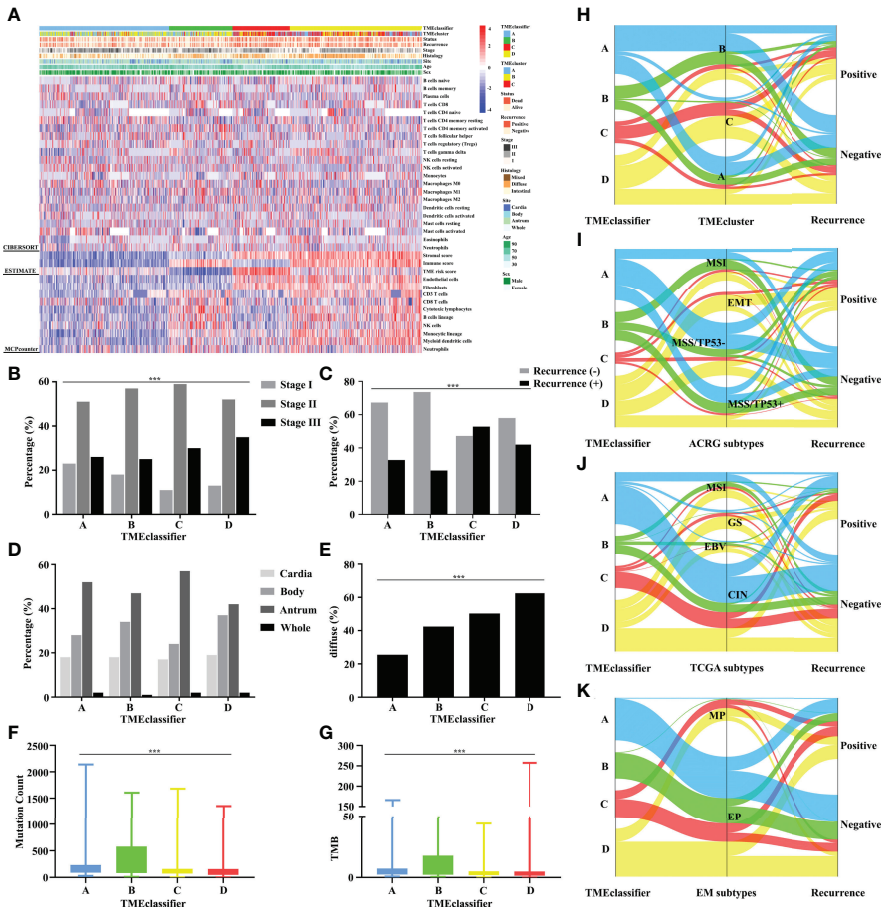


FIGURE 3 | Unsupervised clustering of TME cells for 1,193 patients. TME-classifier, TME cluster, survival status, cancer recurrence, stage, histologic subtype, tumor site, age, and sex are shown as annotations **(A)**. TME-classifier differences in the TNM staging **(B)**, cancer recurrence **(C)**, tumor site **(D)**, and histologic subtype **(E)**. Mutation count **(F)** and TMB score **(G)** stratified by TME-classifier. Alluvial diagram of TME-classifier in groups with different TME clusters **(H)**, ACRG subtypes **(I)**, TCGA subtypes **(J)**, and EM subtypes **(K)**. ***P<0.001.

number of clusters was found to be three, with maximal consensus within clusters and minimal ambiguity among clusters. When compared with the TMEcluster-C group, the TMEcluster-A group presented a higher abundance of immunoactivating cells and a lower abundance of immunosuppressive cells (immune-inflamed phenotype), while the TMEcluster-C group presented the highest abundance of stromal cells and the highest stromal score, which may be associated with an immune excluded phenotype. Moreover, the infiltration of both immune and stromal cells in the TMEcluster-B group was weak (immune-desert phenotype) (**Figure 4A** and **Figure S7**). We observed that the TME cluster can predict both DFS and OS in multiple cohorts and the meta-cohort as a supplement to the published results (**Figures 4B–G** and **Figure S8**). Additionally, patients were divided into low-risk and high-risk TME scores based on the median value. We found that patients with a high TME risk score had a poor prognosis, as shown by DFS and OS (**Figures 5A, B**).

Underlying Mechanism of Tumor Microenvironment Characterization

A comprehensive GSEA analysis was performed between the low and high TME risk scores to identify the underlying mechanisms. We observed significant differences in inflammation-related pathways, chemokine pathways, tumor progression-associated pathways, and metabolic pathways between the low and high TME risk score groups. Patients in the low TME risk score group showed a preferable immunogenicity and tumor suppressive response, which was consistent with a better survival outcome. These results were supported by data from the TCGA and KOREA cohorts (**Figures 5C–F**).

TME-Classifier Can Predict Chemotherapy and Immunotherapy Responses

Considering the important role of chemotherapeutic agents and immune checkpoint inhibitors in clinical application, we next explored patients whose subtype could benefit from chemotherapy

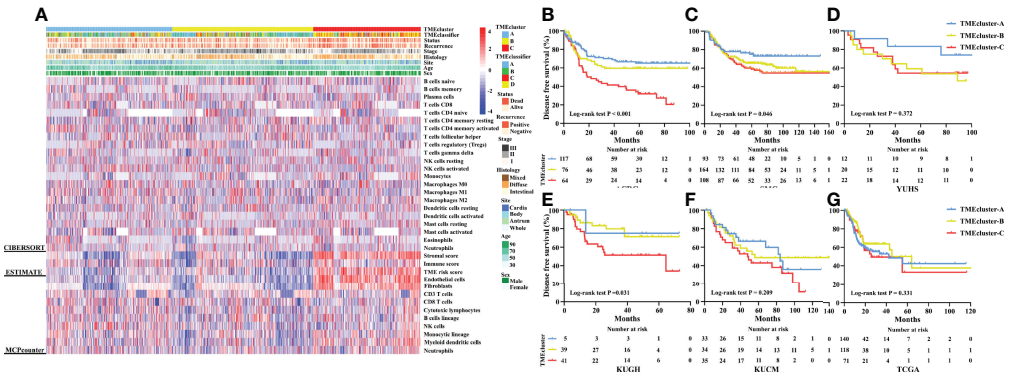


FIGURE 4 | Unsupervised clustering of TME cells for 1,193 patients. TME cluster, TME-classifier, survival status, cancer recurrence, stage, histologic subtype, tumor site, age, and sex are shown as annotations **(A)**. Kaplan–Meier curves for disease-free survival of patients with stage I–IV gastric cancer based on TME-classifier in the ACRG **(B)**, SMC **(C)**, YUHS **(D)**, KUGH **(E)**, KUCM **(F)**, and TCGA **(G)** cohorts.

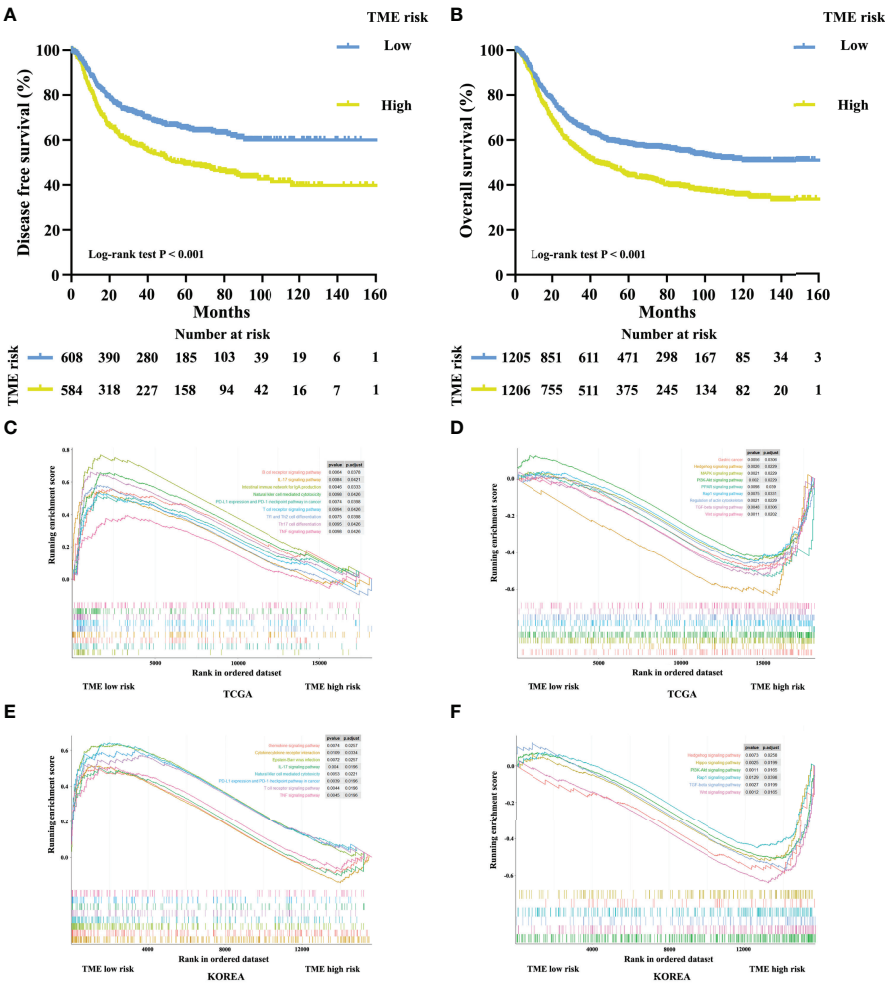


FIGURE 5 | Kaplan–Meier curves for disease-free survival **(A)** of 1,193 patients with stage I–III gastric cancer after potential radical surgery based on TME risk score. Kaplan–Meier curves for overall survival **(B)** of 2,411 patients with stage I–IV gastric cancer based on TME risk score. Enrichment plots for upregulation **(C, E)** and downregulation **(D, F)** pathways of TME low risk group in the TCGA and KOREA cohorts.

and immunotherapy based on our TME-classifier and TME-cluster. Adjuvant chemotherapy resulted in a significant survival benefit for patients in the TMEclassifier-A, TMEclassifier-C, and TMEclassifier-D groups (**Figures 6A, C, D**), but not for patients in the TMEclassifier-B group (**Figure 6B**). Furthermore, forty-five GC patients with complete transcriptome matrix and clinical information were used to evaluate the TMEclassifier characterization for further analysis (**Table S1**). Interestingly, we found that patients in the TMEclassifier-B group (ORR: 62.5%) without chemotherapy benefit responded best to pembrolizumab treatment (PD-1 inhibitor), followed by the TMEclassifier-A group (ORR: 26.7%), while patients in the TMEclassifier-C (ORR: 14.3%) and TMEclassifier-D (ORR: 13.3%) groups responded poorly to immunotherapy (**Figures 6E, F**) (P <0.05). We further observed that our TMEclassifier integrated with the TMB score, with an AUC of 0.773, could predict immunotherapy response well (**Figure 6G**). Additionally, patients in the TMEcluster-A group responded well to chemotherapy and immunotherapy, while patients in the TMEcluster-C group responded to chemotherapy, and patients in the TMEcluster-B group responded poorly to both treatments (**Figure S9**).

DISCUSSION

In this study, we drew the TME landscape of GC recurrence and confirmed that TME characterization was significantly associated with tumor recurrence. Inactive immune responses and active stromal responses were observed in patients with GC recurrence. Subsequently, we proposed a TMEclassifier of 4 subtypes with distinct clinicopathology, epigenetic and molecular characteristics from multiple cohorts of 2,411 patients with GC. The TMEclassifier yielded results that were comparable with the existing molecular classification systems. Importantly, the TMEclassifier remained a robust prognostic biomarker of DFS and OS when adjusted for other clinical factors. We also confirmed that a TME cluster with 3 subtypes can predict DFS as a supplement to the published study. Furthermore, we developed a TME-based risk score and observed that patients with a high TME risk score had a poor prognosis as shown by DFS and OS. More importantly, we found that patients in the TMEclassifier-A, TMEclassifier-C, and TMEclassifier-D groups benefited from adjuvant chemotherapy, and patients in the TMEclassifier-B group without chemotherapy benefit responded best to pembrolizumab treatment (PD-1 inhibitor), followed by the TMEclassifier-A, while patients in the C and D groups of the TMEclassifier responded poorly to immunotherapy.

We observed that the immune-stromal score on the borderline between the subgroups of the TME-classifier was approximate, which made it difficult to classify patients near the cut-off value, a phenomenon common in any clinical marker. To solve this challenge, future studies should integrate multimodal data, such as radiomics. Moreover, a single

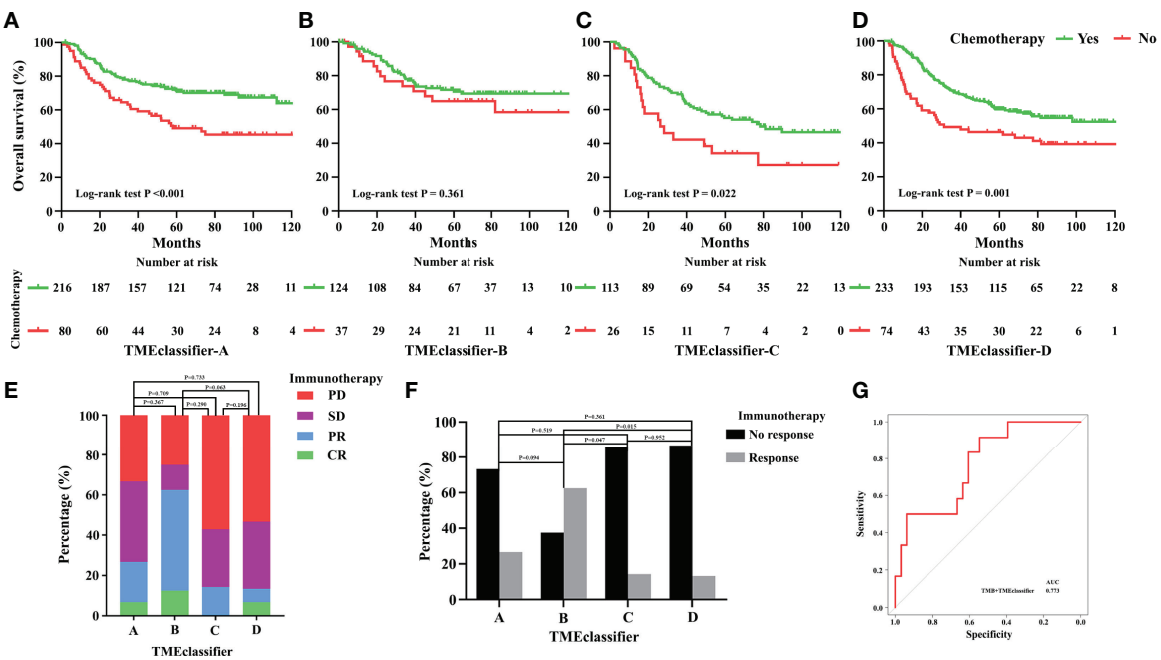


FIGURE 6 | Predictive relevance of the TME-classifier for the benefit of chemotherapy in stage I–III gastric cancer. Patients of TME-classifier **(A, C, D)** derived a significant survival benefit from adjuvant chemotherapy **(A, C, D)**. However, patients of TME-classifier B did not benefit from adjuvant chemotherapy **(B)**. Patients in TME-classifier B group respond best to pembrolizumab treatment (PD-1 inhibitor), followed by TMEclassifier-A, TME-classifier C, and TME-classifier D groups **(E, F)**. TMEclassifier integrated with TMB score predict immunotherapy response well **(G)**.

algorithm, such as CIBERSORT, may have limitations in distinguishing similar cell types. However, our results, which were based on three algorithms, were similar or almost identical and were validated in 13 cohorts of 2,456 patients with GC, which indicated that our conclusion was relatively correct and well-supported.

Recently, an increasing number of studies have indicated that TME is important in tumor progression and therapeutic responses (11–14, 37–39). Furthermore, our previous works have shown that immune signature is an independent predictor for survival and chemotherapeutic benefits in GC (11, 12, 37, 39). Thompson et al. proposed that increasing CD8 cell infiltration was correlated with impaired PFS and OS in GC (40). On the other hand, Grunberg et al. observed that cancer-associated fibroblasts promoted gastric cancer progression (14). Additionally, Sakamoto et al. indicated that tumor-associated macrophages (M2) promoted peritoneal dissemination of GC (13). Compared with previous studies, ours had several following strengths. Firstly, considering a paucity of study examining the tumor recurrence of GC associated with TME or describing one immune-stromal cell simply (13, 14), this study drew the cellular component and prognostic landscape of the TME associated with GC relapse and survival outcome systematically. Secondly, this study included the largest sample, with 2,456 patients from multiple cohorts as validation to date. Thirdly, unlike previous studies, our study combined three independent computational algorithms to confirm that TME characterization was closely related to cancer recurrence.

Cancer recurrence is a fatal complication that compromises the survival and quality of life of patients with GC after comprehensive therapy (41). In a retrospective review of 1,172 GC patients who underwent radical surgery, 42% of patients developed recurrence, and 79% relapsed within two years (5). The median time to death from the time of recurrence was 6 months (5, 41). Currently, the prediction and diagnosis of GC recurrence mainly depends on TNM staging, clinical signs, medical imaging, or even reoperation during follow-up, which may result in delayed diagnosis and treatment (5, 6, 41–43). Our work adds to a growing body of evidence supporting the crucial role of the TME in cancer recurrence. Based on the novel classification model, patients in the TMEclassifier-C and D groups with a high risk of tumor recurrence deserve intensified therapeutic regimens and active surveillance to prevent cancer relapse and improve survival outcomes.

Although chemotherapy and immunotherapy are widely used in clinical practice currently, not all patients with GC can benefit from these treatments (4, 21–23). Disparate clinical outcomes have been observed among patients of the same TNM stage who received similar treatments (44, 45), suggesting that such patients should be given individualized interventions. In this study, we observed that patients in the TMEclassifier-B group with the largest proportion of MSI could not benefit from chemotherapy but had the best response to anti-PD1 treatment. Similarly, several international trials, namely, the MAGIC trial (International Standard Randomized Controlled Trial Number [ISRCTN] 93793971) and the CLASSIC trial (ClincalTrials.gov identifier NCT00411229), reported that patients having tumors with high MSI did not benefit from perioperative or adjuvant chemotherapy (46–48). Moreover, patients with tumors with high MSI, EBV, and mutation burden were more likely to obtain durable responses to immunotherapy (49–51). Our results agree well with these reports. However, a considerable number of patients who were chemotherapy-resistant and showed a response to immunotherapy in the TMEclassifier-B group had MSS or GS tumors, which may be a challenge to the existing hypothesis and needs to be investigated further. Lastly, our results suggested that patients in the TMEclassifier-A, C, and D groups were sensitive to chemotherapy and insensitive to immunotherapy. This indicated that our novel classification system could identify patients who had no therapeutic benefits to avoid the side-effects of adjuvant treatments, and conversely, other patients would receive aggressive regimens and frequent surveillance to prevent cancer recurrences and improve survival outcomes. We also observed that the TME-cluster could predict the response to chemotherapy and immunotherapy. However, these findings were based on a small sample size and a long treatment duration, which may have limited the quality of the observation. Future research is needed to confirm these discoveries.

Despite these findings, this study still has several limitations. Firstly, the primary point is its retrospective nature. Secondly, although multiple cohorts of 2,411 patients are included in this study, validation from our center is lacking and patients are waiting for it. Thirdly, because the immune-stromal data were generated from gene expression profiles, further in-depth data from immunohistochemistry, cell, or animal experiments are required to validate the present findings. Fourthly, patients on the borderline between the two groups are classified into a particularly distinct group, which should be verified with multimodal data in the future. Fifthly, the sample of patients who underwent chemotherapy and immunotherapy is small, and thus a large validation cohort is needed.

In conclusion, we drew a TME landscape on GC recurrence and confirmed that TME characterization was significantly associated with tumor recurrence. We then proposed a TMEclassifier of 4 subtypes with distinct clinicopathology, genomic and molecular characteristics from multiple cohorts of 2,411 patients with GC. Importantly, the TMEclassifier remained a robust prognostic biomarker of DFS and OS when adjusted for clinical factors. We also confirmed that a TME cluster with 3 subtypes and a TME-based risk score can predict DFS and OS. More importantly, we found that patients in 4 subtypes of TMEclassifier had different responses to chemotherapy and immunotherapy.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

TCGA, GEO, and EMBL-EBI belong to public databases. The patients involved in the databases have obtained ethical approval. The studies were reviewed and approved by the Ethics Committee of Shenzhen Hospital of Integrated Chinese and Western Medicine and Nanfang Hospital of Southern Medical University. All procedures involving human participants were in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

All authors listed had made substantial contribution to this work. YJ and YC conceived and designed the study. ZS, LW, HC, and TX collected and collated the data. YC and ZS analyzed the data and wrote the manuscript together. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by grants from the Natural Science Foundation of Guangdong Province (2019A1515011445), the National Natural Science Foundation of China (82102156) and the Sanming Project of Medicine in Shenzhen (No. SZZYSM202106009).

ACKNOWLEDGMENTS

We acknowledge the TCGA, GEO, and EMBL-EBI databases for providing their platforms and contributors for uploading their meaningful datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.890922/full#supplementary-material

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2018) 68 (6):394–424. doi: 10.3322/caac.21492
2. Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, et al. Perioperative Chemotherapy Versus Surgery Alone for Resectable Gastroesophageal Cancer. *N Engl J Med* (2006) 355(1):11–20. doi: 10.1056/NEJMoa055531
3. Badgwell B. Multimodality Therapy of Localized Gastric Adenocarcinoma. *J Natl Compr Canc Netw* (2016) 14(10):1321–7. doi: 10.6004/jnccn.2016.0139
4. Noh SH, Park SR, Yang HK, Chung HC, Chung IJ, Kim SW, et al. Adjuvant Capecitabine Plus Oxaliplatin for Gastric Cancer After D2 Gastrectomy (CLASSIC): 5-Year Follow-Up of an Open-Label, Randomised Phase 3 Trial. *Lancet Oncol* (2014) 15(12):1389–96. doi: 10.1016/S1470-2045(14)70473-5
5. D'Angelica M, Gonen M, Brennan MF, Turnbull AD, Bains M, Karpeh MS. Patterns of Initial Recurrence in Completely Resected Gastric Adenocarcinoma. *Ann Surg* (2004) 240(5):808–16. doi: 10.1097/01.sla.0000143245.28656.15
6. Kikuchi H, Kamiya K, Hiramatsu Y, Miyazaki S, Yamamoto M, Ohta M, et al. Laparoscopic Narrow-Band Imaging for the Diagnosis of Peritoneal Metastasis in Gastric Cancer. *Ann Surg Oncol* (2014) 21(12):3954–62. doi: 10.1245/s10434-014-3781-8
7. Isobe Y, Nashimoto A, Akazawa K, Oda I, Hayashi K, Miyashiro I, et al. Gastric Cancer Treatment in Japan: 2008 Annual Report of the JGCA Nationwide Registry. *Gastric Cancer* (2011) 14(4):301–16. doi: 10.1007/s10120-011-0085-6
8. Zeng D, Li M, Zhou R, Zhang J, Sun H, Shi M, et al. Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. *Cancer Immunol Res* (2019) 7(5):737–50. doi: 10.1158/2326-6066.CIR-18-0436
9. Li B, Jiang Y, Li G, Fisher GAJr., Li R. Natural Killer Cell and Stroma Abundance are Independently Prognostic and Predict Gastric Cancer Chemotherapy Benefit. *JCI Insight* (2020) 5(9):1–15. doi: 10.1172/jci.insight.136570
10. Chen Y, Sun Z, Chen W, Liu C, Chai R, Ding J, et al. The Immune Subtypes and Landscape of Gastric Cancer and to Predict Based on the Whole-Slide Images Using Deep Learning. *Front Immunol* (2021) 12:685992. doi: 10.3389/fimmu.2021.685992
11. Jiang Y, Zhang Q, Hu Y, Li T, Yu J, Zhao L, et al. ImmunoScore Signature: A Prognostic and Predictive Tool in Gastric Cancer. *Ann Surg* (2018) 267 (3):504–13. doi: 10.1097/SLA.0000000000002116
12. Jiang Y, Xie J, Han Z, Liu W, Xi S, Huang L, et al. Immunomarker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit. *Clin Cancer Res* (2018) 24(22):5574–84. doi: 10.1158/1078-0432.CCR-18-0848
13. Sakamoto S, Kagawa S, Kuwada K, Ito A, Kajioka H, Kakiuchi Y, et al. Intraperitoneal Cancer-Immune Microenvironment Promotes Peritoneal Dissemination of Gastric Cancer. *Oncoimmunology* (2019) 8(12):e1671760. doi: 10.1080/2162402X.2019.1671760
14. Grunberg N, Pevsner-Fischer M, Goshen-Lago T, Diment J, Stein Y, Lavon H, et al. Cancer-Associated Fibroblasts Promote Aggressive Gastric Cancer Phenotypes *via* Heat Shock Factor 1-Mediated Secretion of Extracellular Vesicles. *Cancer Res* (2021) 81(7):1639–53. doi: 10.1158/0008-5472.CAN-20-2756
15. Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The Immune Contexture in Cancer Prognosis and Treatment. *Nat Rev Clin Oncol* (2017) 14(12):717–34. doi: 10.1038/nrclinonc.2017.101
16. Chen LJ, Zheng X, Shen YP, Zhu YB, Li Q, Chen J, et al. Higher Numbers of T-Bet(+) Intratumoral Lymphoid Cells Correlate With Better Survival in Gastric Cancer. *Cancer Immunol Immunother* (2013) 62(3):553–61. doi: 10.1007/s00262-012-1358-6
17. Wang Z, Chen JQ. Imaging in Assessing Hepatic and Peritoneal Metastases of Gastric Cancer: A Systematic Review. *BMC Gastroenterol* (2011) 11:19. doi: 10.1186/1471-230X-11-19
18. Maron SB, Chase LM, Lomnicki S, Kochanny S, Moore KL, Joshi SS, et al. Circulating Tumor DNA Sequencing Analysis of Gastroesophageal Adenocarcinoma. *Clin Cancer Res* (2019) 25(23):7098–112. doi: 10.1158/1078-0432.CCR-19-1704
19. Sun Z, Jiang Y, Chen C, Zheng H, Huang W, Xu B, et al. Radiomics Signature Based on Computed Tomography Images for the Preoperative Prediction of Lymph Node Metastasis at Individual Stations in Gastric Cancer: A Multicenter Study. *Radiother Oncol* (2021) 165:179–90. doi: 10.1016/j.radonc.2021.11.003
20. Ajani JA, Lee J, Sano T, Janjigian YY, Fan D, Song S. Gastric Adenocarcinoma. *Nat Rev Dis Primers* (2017) 3:17036. doi: 10.1038/nrdp.2017.36
21. Kim ST, Cristescu R, Bass AJ, Kim KM, Odegaard JI, Kim K, et al. Comprehensive Molecular Characterization of Clinical Responses to PD-1 Inhibition in Metastatic Gastric Cancer. *Nat Med* (2018) 24(9):1449–58. doi: 10.1038/s41591-018-0101-z

22. Kang YK, Boku N, Satoh T, Ryu MH, Chao Y, Kato K, et al. Nivolumab in Patients With Advanced Gastric or Gastro-Oesophageal Junction Cancer Refractory to, or Intolerant of, at Least Two Previous Chemotherapy Regimens (ONO-4538-12, ATTRACTION-2): A Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial. *Lancet* (2017) 390(10111):2461–71. doi: 10.1016/S0140-6736(17)31827-5
23. Shitara K, Ozguroglu M, Bang YJ, Di Bartolomeo M, Mandala M, Ryu MH, et al. Pembrolizumab Versus Paclitaxel for Previously Treated, Advanced Gastric or Gastro-Oesophageal Junction Cancer (KEYNOTE-061): A Randomised, Open-Label, Controlled, Phase 3 Trial. *Lancet* (2018) 392 (10142):123–33. doi: 10.1016/S0140-6736(18)31257-1
24. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular Analysis of Gastric Cancer Identifies Subtypes Associated With Distinct Clinical Outcomes. *Nat Med* (2015) 21(5):449–56. doi: 10.1038/nm.3850
25. Cancer Genome Atlas Research N. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* (2014) 513 (7517):202–9. doi: 10.1038/nature13480
26. Oh SC, Sohn BH, Cheong JH, Kim SB, Lee JE, Park KC, et al. Clinical and Genomic Landscape of Gastric Cancer With a Mesenchymal Phenotype. *Nat Commun* (2018) 9(1):1777. doi: 10.1038/s41467-018-04179-8
27. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust Enumeration of Cell Subsets From Tissue Expression Profiles. *Nat Methods* (2015) 12(5):453–7. doi: 10.1038/nmeth.3337
28. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression. *Genome Biol* (2016) 17 (1):218. doi: 10.1186/s13059-016-1070-5
29. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
30. Wang G, Hu N, Yang HH, Wang L, Su H, Wang C, et al. Comparison of Global Gene Expression of Gastric Cardia and Noncardia Cancers From a High-Risk Population in China. *PloS One* (2013) 8(5):e63826. doi: 10.1371/journal.pone.0063826
31. Qian Z, Zhu G, Tang L, Wang M, Zhang L, Fu J, et al. Whole Genome Gene Copy Number Profiling of Gastric Cancer Identifies PAK1 and KRAS Gene Amplification as Therapy Targets. *Genes Chromosomes Cancer* (2014) 53 (11):883–94. doi: 10.1002/gcc.22196
32. Frederick LG. *AJCC Cancer Staging Manual*. Chicago: Springer Science & Business Media (2002). MD ea.
33. Washington K. 7th Edition of the AJCC Cancer Staging Manual: Stomach. *Ann Surg Oncol* (2010) 17(12):3077–9. doi: 10.1245/s10434-010-1362-z
34. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to Build a Bridge From a Population-Based to a More "Personalized" Approach to Cancer Staging. *CA Cancer J Clin* (2017) 67(2):93–9. doi: 10.3322/caac.21388
35. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool With Confidence Assessments and Item Tracking. *Bioinformatics* (2010) 26 (12):1572–3. doi: 10.1093/bioinformatics/btq170
36. Yu G, Wang LG, Han Y, He QY. Clusterprofiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118
37. Jiang Y, Xie J, Huang W, Chen H, Xi S, Han Z, et al. Tumor Immune Microenvironment and Chemosensitivity Signature for Predicting Response to Chemotherapy in Gastric Cancer. *Cancer Immunol Res* (2019) 7(12):2065–73. doi: 10.1158/2326-6066.CIR-19-0311
38. Zeng D, Wu J, Luo H, Li Y, Xiao J, Peng J, et al. Tumor Microenvironment Evaluation Promotes Precise Checkpoint Immunotherapy of Advanced Gastric Cancer. *J Immunother Cancer* (2021) 9(8):1–15. doi: 10.1136/jitc-2021-002467
39. Lin Y, Pan X, Zhao L, Yang C, Zhang Z, Wang B, et al. Immune Cell Infiltration Signatures Identified Molecular Subtypes and Underlying Mechanisms in Gastric Cancer. *NPJ Genom Med* (2021) 6(1):83. doi: 10.1038/s41525-021-00249-x
40. Thompson ED, Zahurak M, Murphy A, Cornish T, Cuka N, Abdelfatah E, et al. Patterns of PD-L1 Expression and CD8 T Cell Infiltration in Gastric Adenocarcinomas and Associated Immune Stroma. *Gut* (2017) 66(5):794–801. doi: 10.1136/gutjnl-2015-310839
41. Yoo CH, Noh SH, Shin DW, Choi SH, Min JS. Recurrence Following Curative Resection for Gastric Carcinoma. *Br J Surg* (2000) 87(2):236–42. doi: 10.1046/j.1365-2168.2000.01360.x
42. Hasbahceci M, Akcakaya A, Guler B, Kunduz E, Malya FU, Muslumanoglu M. Use of Peritoneal Washing Cytology for the Detection of Free Peritoneal Cancer Cells Before and After Surgical Treatment of Gastric Adenocarcinoma. *J Cancer Res Ther* (2018) 14(6):1225–9. doi: 10.4103/0973-1482.184518
43. Chen Y, Zhou Q, Wang H, Zhuo W, Ding Y, Lu J, et al. Predicting Peritoneal Dissemination of Gastric Cancer in the Era of Precision Medicine: Molecular Characterization and Biomarkers. *Cancers (Basel)* (2020) 12(8):1225–29. doi: 10.3390/cancers12082236
44. Sasako M, Sakuramoto S, Katai H, Kinoshita T, Furukawa H, Yamaguchi T, et al. Five-Year Outcomes of a Randomized Phase III Trial Comparing Adjuvant Chemotherapy With S-1 Versus Surgery Alone in Stage II or III Gastric Cancer. *J Clin Oncol* (2011) 29(33):4387–93. doi: 10.1200/JCO.2011.36.5908
45. Joshi SS, Badgwell BD. Current Treatment and Recent Progress in Gastric Cancer. *CA Cancer J Clin* (2021) 71(3):264–79. doi: 10.3322/caac.21657
46. Pietrantonio F, Miceli R, Raimondi A, Kim YW, Kang WK, Langley RE, et al. Individual Patient Data Meta-Analysis of the Value of Microsatellite Instability As a Biomarker in Gastric Cancer. *J Clin Oncol* (2019) 37 (35):3392–+. doi: 10.1200/Jco.19.01124
47. Smyth EC, Wotherspoon A, Peckitt C, Gonzalez D, Hulkki-Wilson S, Eltahir Z, et al. Mismatch Repair Deficiency, Microsatellite Instability, and Survival An Exploratory Analysis of the Medical Research Council Adjuvant Gastric Infusional Chemotherapy (MAGIC) Trial. *JAMA Oncol* (2017) 3(9):1197–203. doi: 10.1001/jamaoncol.2016.6762
48. Choi YY, Kim H, Shin SJ, Kim H, Lee J, Yang HK, et al. Microsatellite Instability and Programmed Cell Death-Ligand 1 Expression in Stage II/III Gastric Cancer *Post Hoc* Analysis of the CLASSIC Randomized Controlled Study. *Ann Surg* (2019) 270(2):309–16. doi: 10.1097/Sla.0000000000002803
49. Janjigian YY, Sanchez-Vega F, Jonsson P, Chatila WK, Hechtman JF, Ku GY, et al. Genetic Predictors of Response to Systemic Therapy in Esophagogastric Cancer. *Cancer Discov* (2018) 8(1):49–58. doi: 10.1158/2159-8290.Cd-17-0787
50. Chao J, Fuchs CS, Shitara K, Tabernero J, Muro K, Van Cutsem E, et al. Pembrolizumab (Pembro) in Microsatellite Instability-High (MSI-H) Advanced Gastric/Gastroesophageal Junction (G/GEJ) Cancer by Line of Therapy. *J Clin Oncol* (2020) 38(4):430. doi: 10.1200/JCO.2020.38.4_suppl.430
51. Shitara K, Ozguroglu M, Bang YJ, Di Bartolomeo M, Mandal M, Ryu MH, et al. The Association of Tissue Tumor Mutational Burden (tTMB) Using the Foundation Medicine Genomic Platform With Efficacy of Pembrolizumab Versus Paclitaxel in Patients (Pts) With Gastric Cancer (GC) From KEYNOTE-061. *J Clin Oncol* (2020) 38(15):4537. doi: 10.1200/JCO.2020.38.15_suppl.4537

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chen, Sun, Wan, Chen, Xi and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Oncology

MINI REVIEW
published: 31 May 2022
doi: 10.3389/fonc.2022.891676

Network Control Models With Personalized Genomics Data for Understanding Tumor Heterogeneity in Cancer

Jipeng Yan[1], *Zhuo Hu*[2], *Zong-Wei Li*[2], *Shiren Sun*[1*] and *Wei-Feng Guo*[2,3*]

[1] Department of Nephrology, Xijing Hospital, The Fourth Military Medical University, Xi'an, China, [2] School of Electrical Engineering, Zhengzhou University, Zhengzhou, China, [3] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Avantika Lal,
National Centre for Biological Sciences, India
Fei Liu,
Baoji University of Arts and Sciences, China
Zhi-Ping Liu,
Shandong University, China

***Correspondence:**
Wei-Feng Guo
guowf@zzu.edu.cn
Shiren Sun
sunshiren@medmail.com.cn

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 08 March 2022
Accepted: 12 April 2022
Published: 31 May 2022

Citation:
Yan J, Hu Z, Li Z-W, Sun S and Guo W-F (2022) Network Control Models With Personalized Genomics Data for Understanding Tumor Heterogeneity in Cancer. Front. Oncol. 12:891676. doi: 10.3389/fonc.2022.891676

Due to rapid development of high-throughput sequencing and biotechnology, it has brought new opportunities and challenges in developing efficient computational methods for exploring personalized genomics data of cancer patients. Because of the high-dimension and small sample size characteristics of these personalized genomics data, it is difficult for excavating effective information by using traditional statistical methods. In the past few years, network control methods have been proposed to solve networked system with high-dimension and small sample size. Researchers have made progress in the design and optimization of network control principles. However, there are few studies comprehensively surveying network control methods to analyze the biomolecular network data of individual patients. To address this problem, here we comprehensively surveyed complex network control methods on personalized omics data for understanding tumor heterogeneity in precision medicine of individual patients with cancer.

Keywords: personalized omics, network control principles, tumor heterogeneity, precision medicine, cancer individual patients

INTRODUCTION

Increasing studies on cancer genomics data have revealed that individual heterogeneity of cancer patients is one of the main reasons for no substantive breakthrough in cancer treatment methods. With the recent development in high-throughput omics technology, data resources have become available for cancer research, such as genomic and transcriptomic data (1, 2). Personalized omics data of individual patients should be analyzed for understanding the tumor heterogeneity of cancer diseases. The key challenges are how to integrate multi-level omics data, such as genomic and transcriptomic data, for understanding the regulatory mechanism in individual patients, and how to identify cancer-related drug targets (3). Therefore, it is of great theoretical significance and clinical application value for designing computational methods through integration of the omics data of individual patients and screening for drug targets related to phenotype transitions of these patients.

Modern medical studies have shown that cancer is generally an outcome of the dysfunction of related dynamic systems. From the system biology perspective, cancer can be driven by the state transition of key driver genes, which can lead to the dysfunction of molecular networks (e.g., gene

regulation networks or signal transduction networks) that regulate molecular pathways and cellular processes. Moreover, the state of biomolecules (e.g., gene expression value) with complex dynamic characteristics in individual patients changes with time and environmental conditions (**Figure 1**). To understand the dynamics of individual patients, the regulatory mechanism of molecular networks needs to be understood from the perspective of network control theory. This theory considers the state variables of a high-dimensional dynamic system as a complex network and studies how to effectively control the state of driving variables through control signals in the system with optimal control objectives (such as minimum number of controllers or minimum energy), thus changing the network state to the desired stable state (4).

Although traditional control theory (5) has been studied extensively, it is not suitable for a biological network system with numerous nodes (e.g., genes). Network control methods provide the technology for analyzing biomolecule networks with complex dynamic characteristics and quantifying their ability to intervene the biomolecule system of individual patients through proper control signals (6, 7). Moreover, researchers have made progress about network control principles. However, few studies comprehensively surveyed network control methods to analyze the biomolecular network data of individual patients. Considering these facts, this study provided a comprehensive survey for complex network control methods on the multi-omics data of individual patients including methods for personalized gene interaction network construction, network control principles, driver gene prediction, and drug combination identification (**Figure 1**), which aims to reveal the molecular mechanism and regulation law of personalized biomolecular systems for the diagnosis, prevention, and treatment of individual patients.

DATASETS

With the development of cancer genomics technology, many data resources are available for understanding the cancer mechanism. In the past decade, a large amount of cancer genome data from large-scale cancer genomics projects facilitated the development of computational methods for mining personalized omics data of individual patients and understanding tumor heterogeneity in cancer precision medicine. Among these cancer genomics projects, The Cancer Genome Atlas, an important database for mining cancer omics data (2), has created a genomic panorama of different cancers. It currently contains 33 cancer types and more than 20,000 samples. The Cancer Cell Line Encyclopedia is a compilation of gene expression, chromosomal copy number, and massively parallel sequencing data from 1457 cell lines. It provides the pharmacological activities of 24 anticancer drugs in 504 cell lines (8, 9). Gene Expression Omnibus (GEO) is a public repository of functional genomics data currently storing approximately 23,002 public series submitted directly by 168,607 laboratories. This series comprises 4,851,647 samples derived from more than 1600 organisms (10).

BioGPS (11) is an online gene annotation database integrating 150 resources. It can query gene name information, chromosome location, gene function, transcript information,

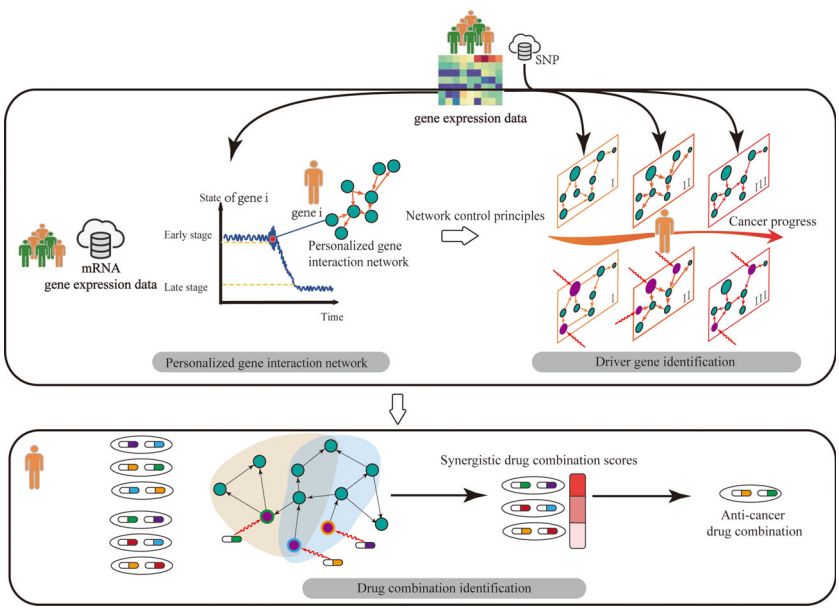


FIGURE 1 | Overview of our review. The contents of our review consist of three parts. Firstly, we summarized the works to construct personalized gene interaction network from genomics of individual patients. Then on the personalized gene interaction networks, we pointed out how to identify personalized driver gene by using network control tools. Finally, we described how to discover synergistic drug combinations by targeting personalized driver genes.

encoded protein information, and related protein names. However, this database cannot provide a detailed gene annotation list, and therefore, users find it difficult to annotate the genes for a large number of samples. The cancer gene census (CGC) data (12) offers a detailed list of driver genes that have been experimentally verified as cancer driver genes (13–15). The Network of Cancer Genes (NCG) (16) is a database that collects and annotates cancer genes from a large amount of cancer sequencing data. This database contains 2372 genes, including experimentally verified cancer driver genes.

CONSTRUCTION OF PERSONALIZED GENE INTERACTION NETWORK

A biological system is a complex dynamic multi-scale system involving different time, space, and functions. Cells contain genes that store information, proteins, and metabolites and perform biological functions for forming basic functional modules. A complex biological system is composed of multiple functional modules. In system biology, the key to constitute a biological system is not its components (e.g., genes, proteins, and small biological molecules), but their interactions with components having different properties. These interactions constitute the regulatory network controlling different biological functions.

The rich information can be obtained from the high-dimensional data of samples of individual patients. However, the individual genomics data of these patients are often limited and incomplete. Therefore, methods to ensure complete use of personalized genomics data for designing effective gene interaction network construction algorithms of individual patients must be developed. The personalized gene interaction network represents which gene pairs are involved in the disease development for each patient. Because the principles of the personalized network dynamics are hidden, it is important to reconstruct the personalized state transition networks with the personalized genetic data (e.g., expression profiles). It is a key challenge to unravel the dynamic nature of gene regulation during a biological process in systems biology.

Current gene interaction network construction methods, such as Gene Network Reconstruction tool (GNR) (17), dynamic cascaded method (DCM) (18), and Hotnet2 (19), use gene expression data of population cancer patients. Although these gene regulation networks can reflect the gene interaction mechanism of the disease, they cannot describe the gene interaction relationship of individual patients. Numerous single-sample gene interaction network construction methods have recently been proposed. Several common techniques including Single Sample Network construction method (SSN) (20), Paired Single Sample Network construction method (Paired-SSN) (21), Single Pearson Correlation Coefficient calculation method (SPCC) (22, 23), and Cell Specific Network construction method (CSN) (24), and Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) (25) were introduced as follows. In **Table S1 of Supplementary Tables**, we gave a summary of these methods including brief descriptions and input data for constructing personalized gene interaction network.

1) SSN

SSN is a statistical method to construct an individual-specific network based on statistical perturbation analysis of a single sample against a group of given control samples (20). For the SSN method, the co-expression network of the tumor sample network or normal sample network for each patient is constructed based on statistical perturbation analysis of one sample against a group of given reference samples (e.g., choosing the normal sample data of all of the patients as the reference data).

2) Paired-SSN

For the paired-SSN method (21), the co-expression network of the tumor sample network and normal sample network for each patient is firstly constructed in the same way as for the SSN method. Then, the personalized differential co-expression network between the normal sample network and tumor sample network can be constructed in which the edge will exist if the *P*-value of the gene pair is less than (greater than) 0.05 in the tumor network but greater than (less than) 0.05 in the normal network for their corresponding patient.

3) SPCC method

To overcome the difficulty in obtaining correlations or edges from one sample, the SPCC approach (22, 23) was developed by decomposing each PCC measurement into multiple additive elements that form a new vector embedding correlation-like information of two variables for one sample.

4) LIONESS

LIONESS does not rely upon differential analysis between the tumor sample and a group of normal samples, and it reconstructs the individual specific network in a population of tumor samples as the personalized gene state transition network for each tumor sample (25). LIONESS constructs the state transition network by calculating the edge statistical significance between all the tumor samples and the tumor samples without a given single sample.

5) CSN

The CSN method is derived from a theoretical model based on statistical dependency (24), which can be viewed as data transformation from the "unstable" gene expression data to the "stable" gene association data. CSN designs a statistic for gene pairs and can obtain the P-value corresponding to the edge between genes by the statistic.

We should note that conditional or partial sample-specific correlation network can be generally used to eliminate the indirect co-expressions between genes (26, 27). Furthermore, the reference reliable gene/protein interaction network are generally used to take overlapped edges from the original gene co-expression edges, forming the final personalized gene interaction networks for the above methods. However, the current single-sample gene regulation network construction

methods ignore temporal data of individual patients (28), and their accuracy and stability need improvement.

NETWORK CONTROL PRINCIPLES

A core concept in network science is to control and intervene on network dynamics (4). Network control methods have recently received extensive attention (29–36). Therefore, network control methods are better than the traditional control concept in revealing the dynamic characteristics of biological networks with a lot of noise in edge weight. In particular, considering the gene expression profiles in normal and tumor samples as the respective state of a given patient, network control tools aim to detect a small number of driver nodes by the input signals related with the state transition of individual patient depending on adequate knowledge of the network structure. The input signals may be oncogene activation signals such as gene mutation or metabolites changes in specific tissue. The "controllers" in network control problem for molecular networks mean the genetic or environment factors which produce the oncogene activation signals. As per our best understanding, current methods can be classified as directed and undirected network control methods.

i) For the control methods of directed networks, Liu et al. (37) studied the structural control of directed networks and proposed Maximum Matching Sets-based controllability methods by referring to the structural controllability theory of linear systems, which has greatly inspired the promotion of research of network control methods and applications. Although these network control tools have been applied to biomolecular systems, some interesting properties of biological systems have also been discovered. For example, driving mutant genes (38) and drug targets (39) were found in cancer datasets, and driving metabolites were detected in human liver metabolic networks (40). However, these tools only describe the linear dynamic behavior of the network and are not sufficient for completely characterizing the complex nonlinear dynamic system. Recently, a Feedback Vertex Set (FVS)-based control method based on the framework of feedback vertex set control theory (FC) that can be used to study network systems with nonlinear dynamics was proposed. However, for the FVS-based control method, not only the network structure needs to be known but also the functional form of the governing equation must satisfy certain properties (41, 42). Zanudo et al. applied FVS-based control to directed networks (43). By comparing FVS predictions with those of MMS-based controllability methods, they identified topological features underlying different observed phenomena.

ii) For the control methods of undirected networks, Yuan et al. proposed an accurate control method (44) that can identify the minimum set of driver nodes in the undirected networks. Because the precise control method only describes the linear dynamic behavior of biological networks, it cannot be used to accurately describe their nonlinear dynamic behavior. To overcome the aforementioned problems, some researchers proposed minimum dominating set (MDS)-based control methods (45). These methods, however, have a strong assumption on control signals, that is, these signals can independently control their neighborhood nodes. However, most controllers cannot satisfy the strong conditions; therefore, FVS-based methods of undirected networks (namely NCUA) based on the framework of feedback vertex set control theory (FC) have been proposed (46). Since most current methods are designed based on the time-invariant network system, temporal networks can accurately describe the characteristics of cancer omics data. Thus, more accurate network control methods need to be further developed to accurately understand the dynamic characteristics of the network.

To easier understand these network control methods, we gave the concept comparisons including the network types and targeted states and input and network dynamics between different network control method including MMS, MDS and DFVS and NCUA (**Table S2 of Supplementary Tables**). In **Figure 2**, we intuitively explained these methods. We summarized some key points of different structure network control methods as follow:

i) The MMS control methods investigate the controllability of directed structural networks with linear or local nonlinear dynamics through a minimum set of input nodes and they only give an incomplete view of the network control properties for a system with nonlinear dynamics.

ii) MDS control method studies the controllability of undirected networks by assuming that each driver node in the MDS model can control its associated edges independently in the undirected networks. Since MDS works with the strong assumption that the controllers can control its outgoing links independently, it requires higher costs in many kinds of networks which may underestimate the structural control analysis of undirected networks;

iii) NCUA and DFVS study the structural network control of undirected and direct networks respectively based on the framework of FC (42). Therefore NCUA and DFVS methods ultimately depict the structure-based network control of the large-scale system with nonlinear dynamics. Since FC assumes that the functional form of the governing equations must satisfy some continuous, dissipative, and decaying properties, DFVS and NCUA may be only suitable some specialized nonlinear systems.

DRIVER GENE PREDICTION

Using new methods, researchers have recently made some progress in predicting cancer driver genes of population cohorts. These methods are based on mutation frequency, machine learning, and complex network. (1) In gene mutation frequency-based methods, the mutation frequency of driver

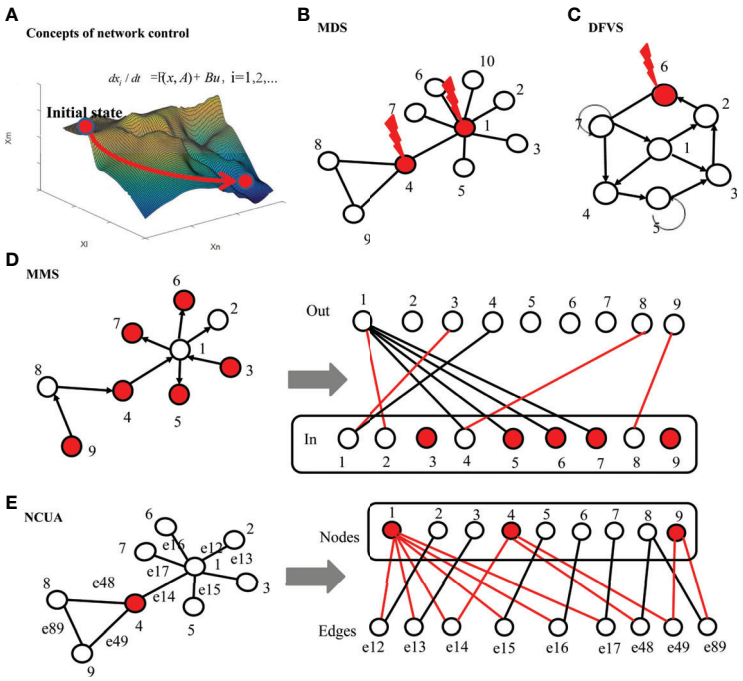


FIGURE 2 | The principles of different network control methods. **(A)** Concept demonstration of network control methods. Network control tools aim to detect a small number of driver nodes which form the input matrix and are injected by the input signals for driving the state transition of high dimension networked system depending on adequate knowledge of the network structure. **(B)** MDS based control methods. If the connected edges of MDS are removed, there will be no edges in the network. By assuming that the driving node can independently control all neighbor nodes, the minimum dominating set (MDS) in the undirected network is taken as the set of driver nodes, and the red node represents the minimum driving node. **(C)** DFVS based control methods. The red nodes represent the minimum set of feedback nodes (FVS), that is, if the connected edges of FVS are removed, there will be no loops in the network. For FVS based control methods, by controlling the nodes in FVS, the whole system can be transformed from one stable attractor to another attractor. **(D)** MMS based control methods. The directed network is transformed into a bipartite graph. For the bipartite graph, the upper side represents the out degree of the original network, while the bottom side represent the in degree of the original network nodes. If there is an edge from one node to another node in the original network, an edge connecting these two nodes is added to the bipartite graph. According to the maximum matching of bipartite graph, the maximum matching (i.e., red edges) can be obtained, and 6 unmatched nodes (i.e., red nodes) can be found in the bottom side of bipartite graph. By controlling these 6 nodes, the system structure can be completely controllable for MMS based control methods. **(E)** NCUA based control methods. Firstly, the original undirected network is transformed into a bipartite graph, in which the upper side represents the nodes of the original network and the bottom side represent the edges of the original network respectively. Then, the nodes covering the nodes on the bottom side (i.e., red nodes) are obtained in the bipartite graph and are considered as driver nodes for the NCUA method. The red edges represent the links between the driver nodes and the nodes of bottom side in the bipartite graph.

genes is usually assumed to be significantly higher than that of other genes (13, 47–51). However, due to the tumor heterogeneity, it is difficult to build a reliable background mutation model. In addition, these methods cannot be used to detect the low-mutated frequency and non-mutated cancer driver genes, because a part of driver genes is mutated at high frequencies (>20%), while most of cancer mutations occur at intermediate frequencies (2–20%) or even lower (52), and even many genes that play important roles in tumorigenesis are not altered on the DNA sequences, and these genes are dysregulated through various cellular mechanisms (53). (2) Machine learning-based methods (49, 51, 54–56) usually train the classifier (e.g., random forest, support vector machine) by extracting various kinds of features from different types of cancer data to predict new cancer driver genes. Although machine learning-based methods can effectively predict some cancer driver genes, because of the incomplete database, some key driver genes may be ignored, thus generating false-positive results. (3) Complex network-based methods usually assume that driver genes have obvious structural characteristics at the biological network level (19, 57, 58). Although these methods have been successfully used for detecting cancer driver genes, they are still limited to incomplete and unreliable interactions in biological network (59).

The aforementioned algorithms focus on how to identify driver genes in population cohorts but cannot be directly applied to the data of individual patients because of the following reasons. On the one hand, TCGA provides 33 cancer types with more than 10,000 samples and 3,000,000 mutation data, while the sample size for individual patients is typically very small. On the other hand, the functional characteristics of cancer mutations in population cohorts are different from those observed in individual cancer patient data with complex and unclear dynamics. Therefore, considering that network techniques such as random walk with restart (RWR) (60), network diffusion (19, 61), subnetwork

enrichment analysis (62), matrix completion (63) and network structure control (6, 64–66) to predict cancer driver genes at the biological network level by incorporating the knowledge of pathways, protein-protein interactions, can deal with high-dimensional data having a small sample size, researchers proposed some network algorithms for predicting personalized driver genes of individual patients (14, 15, 67). Although these algorithms can predict personalized driver genes with important biological functions, they do not consider dynamic changes in the structure of the personalized gene interaction network, thereby leading to false-positive results and affecting the accuracy of driver gene identification. Therefore, dynamic changes in the structure of the personalized gene interaction network must be considered for inferring the evolution trajectory of driver genes and accurately understanding the cancer development mechanism.

DRUG COMBINATION IDENTIFICATION

Computational methods for predicting combination drugs have recently attracted extensive attention (68). These methods can predict a large number of combination drugs with enhanced efficacy and reduced adverse effects (69), which are beneficial for providing efficient clinical treatments (70). At present, drug combination prediction methods mainly include complex network- and machine learning-based methods. (1) The complex network-based methods generally use some network optimization algorithms to predict drug combinations (71–73). However, their predictive performance relies on prior knowledge of drug targets and disease-related networks and is generally only suitable for a few specific diseases. (2) The machine learning-based methods use drug attribute information and cell line experimental data to predict combination drugs (74). The drug characteristics include chemical structure (75, 76), physical and chemical properties of the substructure and toxic modules (77), drug targets (78), and single drug dose response (79, 80). The cell line data include gene expression profile (81), transcriptome (76), pathway network (82), gene interaction network (83), microRNA expression and protein abundance (84), gene variation information, and copy number variation (78). The machine learning algorithms for drug combination prediction mainly include logistic regression (76), Bayesian network (85), random forest (86), multi-decision assemble (87), deep neural network (77), and deep residual neural network (88).

The existing methods ignore the heterogeneity of combination drugs among individual patients, and thus cannot predict effective drug combinations for individual patients. Therefore, an appropriate and effective drug combination prediction method needs to be designed by considering the information of individual patients. However, because personalized genomics data are generated from a small sample size and have a high dimension, these drug combination methods based on large samples cannot be used to accurately identify individual drug combinations from such data. Several studies have attempted to provide such recommendations through predictive models that can predict the efficacy of a drug for an input genomic profile. For instance, Sheng et al. (89) proposed an algorithm on a cell line (or a patient profile) based on similarity of the input drug and cell line to those in Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE). Recently, Drug Recommendations by Integrating Multiple Biomedical Databases (DruID) (90) utilized a Prescriptive Analytics framework based on Integer Programming on multiple public well known databases (91) for personalized drug recommendations. These methods require genomics data from cell lines and a list of genes with mutations, from a single patient, as input which ignore the personalized gene interaction for identifying personalized drug combinations. In our previous work, we used the network control theory to design a personalized drug combination prediction model, aiming to identify personalized synergistic drug combinations by targeting personalized driver genes of individual patients (6, 7). In fact, the model neglected the individual dynamic characteristics of drug activity and toxic concentration in drug combination therapy, and thus could not provide precise personalized drug combination. Therefore, a more effective personalized drug combination prediction model needs to be designed considering more information of individual patients' omics data.

FUTURE DIRECTIONS

Due to the complex dynamics of cancer data, some of the future directions for designing network control methods are as follows:

i)Designing Boolean network control methods. Although network control methods can analyze the dynamics of a biomolecular network, the positive and negative sign characteristics of interactions are currently ignored. For example, in a gene interaction network, the activation or inhibition regulatory relationship between genes is indicated as a connection with positive and negative signs (92). These signs of the gene interaction network become important when the network is controlled in a particular manner, whereas most current network control methods for personalized genome omics data only assume that their state of interactions are non-zero and do not make any assumptions about the sign of interactions. Therefore, designing Boolean network control methods considering the positive and negative sign characteristics of network interactions is an important future direction.

ii) Designing temporal network control methods. Driver genes influence the cell state through a combination of molecular interactions that may change dynamically during cancer progression (93, 94). At present, most network control methods are designed based on the static time-invariant network structure. However, the network structures of cancer patients differ at different cancer stages, which needs to be considered (95). Therefore, how to design reasonable temporal network control methods for inferring the

evolutionary trajectory of driver genes in cancer patients needs to be determined.

iii) Predicting biomarker for individual patients based on network observability. Individual early diagnosis has become essential in precision medicine, which has thus made biomarker prediction increasingly important in drug development (96–102). Therefore, more effective methods are required to describe transitions in cancer status and to identify more biologically significant biomarkers. Network observability is dual with network controllability for network with linear dynamics (103, 104). It focuses on how to select key sensor nodes in the network to reconstruct the state of the entire network. Therefore, developing effective biomarker prediction algorithms for individual patients based on network observability is another crucial future direction.

iv) Designing network control methods on personalized single-cell data. With the development of biological sequencing technology, single-cell data provides a powerful resource for revealing the gene interaction of a single cell and understanding the tumor heterogeneity of individual patients with cancer (105). Therefore, how to design effective network control methods on the single-cell data of individual patients to predict biomarkers, driver genes, and drug targets for such patients is another important research direction.

v) Designing deep learning techniques for network controllability. Over the past decade, deep learning has become a focal topic in artificial intelligence and machine learning (106, 107).In fact, deep learning techniques have been developed for predicting the controllability robustness according to the input network-adjacency matrices (108). However, there are no related works to apply deep learning techniques especially graph based deep learning techniques (107) for studying controllability of personalized gene interaction network. Therefore, how to utilize deep learning methods to analyze controllability of personalized gene network is another interesting and important topic in the future.

CONCLUSIONS

The genomic profiles of cancer patients are diverse and heterogeneous. These profiles are believed to be responsible for heterogeneity of drug response in cancer patients. The current main challenge in cancer precision medicine is to develop effective computational methods for finding personalized biomarkers, driver genes, and drug targets for individual patients. These personalized biomarkers, driver genes, and drug targets would help improve the outcome of cancer patients, especially those with drug resistance. As the number of samples of an individual patient is usually limited, the accuracy and reliability of statistical methods based on a large sample size (109, 110) will greatly be reduced for mining personalized omics data of individual patients. Therefore, considering the multi-omics data of individual patients, this study discusses cancer datasets, construction of gene regulation network, network structure control method, driver gene prediction, and drug combination prediction for individual patients in order to understand tumor heterogeneity in precision medicine.

AUTHOR CONTRIBUTIONS

W-FG developed the methodology and designed research. JY and W-FG wrote the original manuscript. JY, ZH, Z-WL, SS and W-FG revised the manuscript. SS and W-FG supervised the work, made critical revisions of the paper, and approved the submission of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This paper was supported by the National Natural Science Foundation of China [62002329 (W-FG)], and Key scientific and technological projects of Henan Province [2102310083 (W-FG)], and China postdoctoral foundation [2021M692915 (W-FG)], and Henan postdoctoral foundation [202002021(W-FG)], Research start-up funds for top doctors in Zhengzhou University [32211739 (W-FG)], and open Funds of State Key Laboratory of Oncology in South China [HN2021-01 (W-FG)]. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank BMCSCI (www.bmcscience.com) for its linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.891676/full#supplementary-material

REFERENCES

1. Levine DAC.G.A.R. Network. Integrated Genomic Characterization of Endometrial Carcinoma. *Nature* (2013) 497:67. doi: 10.1038/nature12113
2. Zhu Y, Qiu P, Ji Y. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. *Nat Methods* (2014) 11:599. doi: 10.1038/nmeth.2956
3. Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R, De Sano L, Mauri G, et al. Algorithmic Methods to Infer the Evolutionary Trajectories in Cancer Progression. *Proc Natl Acad Sci* (2016) 113:E4025–34. doi: 10.1073/pnas.1520213113
4. Liu Y-Y, Barabási A-L. Control Principles of Complex Systems. *Rev Mod Phys* (2016) 88:035006. doi: 10.1103/RevModPhys.88.035006
5. Kalman RE. Mathematical Description of Linear Dynamical Systems. *J Soc Ind Appl Math Ser A: Control* (1963) 1:152–92. doi: 10.1137/0301010
6. Guo WF, Zhang SW, Feng YH, Liang J, Zeng T, Chen L. Network Controllability-Based Algorithm to Target Personalized Driver Genes for Discovering Combinatorial Drugs of Individual Patients. *Nucleic Acids Res* (2021) 49:e37. doi: 10.1093/nar/gkaa1272
7. Guo W-F, Yu X, Shi Q-Q, Liang J, Zhang S-W, Zeng T. Performance Assessment of Sample-Specific Network Control Methods for Bulk and

Single-Cell Biological Data Analysis. *PloS Comput Biol* (2021) 17:e1008962. doi: 10.1371/journal.pcbi.1008962
8. Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic Agreement between Two Cancer Cell Line Data Sets. *Nature* (2015) 528:84. doi: 10.1038/nature15736
9. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* (2012) 483:603. doi: 10.1038/nature11003
10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Res* (2002) 30:207–10. doi: 10.1093/nar/30.1.207
11. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: An Extensible and Customizable Portal for Querying and Organizing Gene Annotation Resources. *Genome Biol* (2009) 10:R130. doi: 10.1186/gb-2009-10-11-r130
12. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A Census of Human Cancer Genes. *Nat Rev Cancer* (2004) 4:177. doi: 10.1038/nrc1299
13. Jia P, Zhao Z. VarWalker: Personalized Mutation Network Analysis of Putative Cancer Genes From Next-Generation Sequencing Data. *PloS Comput Biol* (2014) 10:e1003460. doi: 10.1371/journal.pcbi.1003460
14. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, et al. Patient-Specific Driver Gene Prediction and Risk Assessment Through Integrated Network Analysis of Cancer Omics Profiles. *Nucleic Acids Res* (2015) 43:e44–4. doi: 10.1093/nar/gku1393
15. Hou JP, Ma J. DawnRank: Discovering Personalized Driver Genes in Cancer. *Genome Med* (2014) 6:56. doi: 10.1186/s13073-014-0056-8
16. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): A Comprehensive Catalogue of Known and Candidate Cancer Genes From Cancer Sequencing Screens. *Genome Biol* (2019) 20:1. doi: 10.1186/s13059-018-1612-0
17. Wang Y, Joshi T, Zhang X-S, Xu D, Chen L. Inferring Gene Regulatory Networks From Multiple Microarray Datasets. *Bioinformatics* (2006) 22:2413–20. doi: 10.1093/bioinformatics/btl396
18. Zhu H, Rao RSP, Zeng T, Chen L. Reconstructing Dynamic Gene Regulatory Networks From Sample-Based Transcriptional Data. *Nucleic Acids Res* (2012) 40:10657–67. doi: 10.1093/nar/gks860
19. Leiserson MD, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations Across Pathways and Protein Complexes. *Nat Genet* (2015) 47:106–14. doi: 10.1038/ng.3168
20. Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized Characterization of Diseases Using Sample-Specific Networks. *Nucleic Acids Res* (2016) 44:e164–4. doi: 10.1093/nar/gkw772
21. Guo W-F, Zhang S-W, Wei ZG, Zeng T, Chen L. Constrained Target Controllability of Complex Networks. *J Stat Mech Theory Exp* (2017) 2017 (6):063402. doi: 10.1088/1742-5468/aa6de6
22. Yu X, Zeng T, Wang X, Li G, Chen L. Unravelling Personalized Dysfunctional Gene Network of Complex Diseases Based on Differential Network Model. *J Transl Med* (2015) 13:189–9. doi: 10.1186/s12967-015-0546-5
23. Zhang W, Zeng T, Liu X, Chen L. Diagnosing Phenotypes of Single-Sample Individuals by Edge Biomarkers. *J Mol Cell Biol* (2015) 7:231–41. doi: 10.1093/jmcb/mjv025
24. Dai H, Li L, Zeng T, Chen L. Cell-Specific Network Constructed by Single-Cell RNA Sequencing Data. *Nucleic Acids Res* (2019) 47:e62–2. doi: 10.1093/nar/gkz172
25. Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating Sample-Specific Regulatory Networks. *iScience* (2019) 14:226–40. doi: 10.1016/j.isci.2019.03.021
26. Huang Y, Chang X, Zhang Y, Chen L, Liu X. Disease Characterization Using a Partial Correlation-Based Sample-Specific Network. *Brief Bioinform* (2020) 22:6. doi: 10.1093/bib/bbaa062
27. Li L, Dai H, Fang Z, Chen L. C-CSN: Single-cell Rna Sequencing Data Analysis by Conditional Cell-Specific Network. *Genomics Proteomics Bioinformatics* (2021) 19(2):319–29. doi: 10.1101/2020.01.25.919829
28. van der Wijst MG, de Vries DH, Brugge H, Westra H-J, Franke L. An Integrative Approach for Building Personalized Gene Regulatory Networks for Precision Medicine. *Genome Med* (2018) 10:96. doi: 10.1186/s13073-018-0608-4
29. Barabási A-L. The Network Takeover. *Nat Phys* (2011) 8:14. doi: 10.1038/nphys2188
30. Lombardi A, Hörnquist M. Controllability Analysis of Networks. *Phys Rev E* (2007) 75:056110. doi: 10.1103/PhysRevE.75.056110
31. Gao J, Liu Y-Y, D'souza RM, Barabási A-L. Target Control of Complex Networks. *Nat Commun* (2014) 5:5415. doi: 10.1038/ncomms6415
32. Cornelius SP, Kath WL, Motter AE. Realistic Control of Network Dynamics. *Nat Commun* (2013) 4:1942. doi: 10.1038/ncomms2939
33. Wu F-X, Wu L, Wang J, Liu J, Chen L. Transittability of Complex Networks and its Applications to Regulatory Biomolecular Networks. *Sci Rep* (2014) 4:4819. doi: 10.1038/srep04819
34. Nacher JC, Akutsu T. Minimum Dominating Set-Based Methods for Analyzing Biological Networks. *Methods* (2016) 102:57–63. doi: 10.1016/j.ymeth.2015.12.017
35. Delpini D, Battiston S, Riccaboni M, Gabbi G, Pammolli F, Caldarelli G. Evolution of Controllability in Interbank Networks. *Sci Rep* (2013) 3:1626. doi: 10.1038/srep01626
36. Lin C-T. Structural Controllability. *IEEE Trans Automat Control* (1974) 19:201–8. doi: 10.1109/TAC.1974.1100557
37. Liu Y-Y, Slotine J-J, Barabási A-L. Controllability of Complex Networks. *Nature* (2011) 473:167. doi: 10.1038/nature10011
38. Dinstag G, Shamir R. PRODIGY: Personalized Prioritization of Driver Genes. *Bioinformatics* (2019) 36:6. doi: 10.1093/bioinformatics/btz815
39. Asgari Y, Salehzadeh-Yazdi A, Schreiber F, Masoudi-Nejad A. Controllability in Cancer Metabolic Networks According to Drug Targets as Driver Nodes. *PloS One* (2013) 8:e79397. doi: 10.1371/journal.pone.0079397
40. Liu X, Pan L. Detection of Driver Metabolites in the Human Liver Metabolic Network Using Structural Controllability Analysis. *BMC Syst Biol* (2014) 8:51. doi: 10.1186/1752-0509-8-51
41. Fiedler B, Mochizuki A, Kurosawa G, Saito D. Dynamics and Control at Feedback Vertex Sets. I: Informative and Determining Nodes in Regulatory Networks. *J Dyn Differ Equ* (2013) 25:563–604. doi: 10.1007/s10884-013-9312-7
42. Mochizuki A, Fiedler B, Kurosawa G, Saito D. Dynamics and Control at Feedback Vertex Sets. II: A Faithful Monitor to Determine the Diversity of Molecular Activities in Regulatory Networks. *J Theor Biol* (2013) 335:130–46. doi: 10.1016/j.jtbi.2013.06.009
43. Zañudo JGT, Yang G, Albert R. Structure-Based Control of Complex Networks with Nonlinear Dynamics. *Proc Natl Acad Sci* (2017) 114:7234–9. doi: 10.1073/pnas.1617387114
44. Yuan Z, Zhao C, Di Z, Wang W-X, Lai Y-C. Exact Controllability of Complex Networks. *Nat Commun* (2013) 4:2447. doi: 10.1038/ncomms3447
45. Nacher JC, Akutsu T. Dominating Scale-Free Networks with Variable Scaling Exponent: Heterogeneous Networks are Not Difficult to Control. *New J Phys* (2012) 14:073005. doi: 10.1088/1367-2630/14/7/073005
46. Guo WF, Zhang SW, Zeng T, Li Y, Gao JX, Chen LN. A Novel Network Control Model for Identifying Personalized Driver Genes in Cancer. *PloS Comput Biol* (2019) 15:27. doi: 10.1371/journal.pcbi.1007520
47. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: Identifying Mutational Significance in Cancer Genomes. *Genome Res* (2012) 22:1589–98. doi: 10.1101/gr.134635.111
48. Lawrence MS, Petar S, Paz P, Kryukov GV, Kristian C, Andrey S, et al. Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes. *Nature* (2013) 499:214–8. doi: 10.1038/nature12213
49. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: Toolkit for Detecting Biologically Important Single Nucleotide Mutations in Cancer. *Bioinformatics* (2011) 27:2147–8. doi: 10.1093/bioinformatics/btr357
50. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. *PloS One* (2013) 8:e77945. doi: 10.1371/journal.pone.0077945
51. Luo P, Ding Y, Lei X, Wu F-X. Deepdriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. *Front Genet* (2019) 10:13. doi: 10.3389/fgene.2019.00013

52. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and Saturation Analysis of Cancer Genes Across 21 Tumour Types. *Nature* (2014) 505:495–501. doi: 10.1038/nature12912
53. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LAJr., Kinzler KW. Cancer Genome Landscapes. *Science* (2013) 339:1546–58. doi: 10.1126/science.1235122
54. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the Evaluation of Cancer Driver Genes. *Proc Natl Acad Sci USA* (2016) 113:14330–5. doi: 10.1073/pnas.1616440113
55. Rogers MF, Gaunt TR, Campbell C. Prediction of Driver Variants in the Cancer Genome *Via* Machine Learning Methodologies. *Brief Bioinform* (2021) 22:bbaa250. doi: 10.1093/bib/bbaa250
56. Liu C, Dai Y, Yu K, Zhang ZK. Enhancing Cancer Driver Gene Prediction by Protein-Protein Interaction Network. *IEEE/ACM Trans Comput Biol Bioinform* (2021) 99:1–1. doi: 10.1109/TCBB.2021.3063532
57. Ciriello G, Cerami E, Sander C, Schultz N. Mutual Exclusivity Analysis Identifies Oncogenic Network Modules. *Genome Res* (2012) 22:398–406. doi: 10.1101/gr.125567.111
58. Zhang S-Y, Zhang S-W, Liu L, Meng J, Huang Y. M6a-Driver: Identifying Context-Specific mRNA m6A Methylation-Driven Gene Interaction Networks. *PloS Comput Biol* (2016) 12:e1005287. doi: 10.1371/journal.pcbi.1005287
59. Cheng FX, Zhao JF, Zhao ZM. Advances in Computational Approaches for Prioritizing Driver Mutations and Significantly Mutated Genes in Cancer Genomes. *Brief Bioinform* (2016) 17:642–56. doi: 10.1093/bib/bbv068
60. Zhou Y, Wang S, Yan H, Pang B, Zhang X, Pang L, et al. Identifying Key Somatic Copy Number Alterations Driving Dysregulation of Cancer Hallmarks in Lower-Grade Glioma. *Front Genet* (2021) 12:654736. doi: 10.3389/fgene.2021.654736
61. Cowen L, Ideker T, Raphael BJ, Sharan R. Network Propagation: A Universal Amplifier of Genetic Associations. *Nat Rev Genet* (2017) 18:551–62. doi: 10.1038/nrg.2017.38
62. Zhang D, Bin YN. Driversubnet: A Novel Algorithm for Identifying Cancer Driver Genes by Subnetwork Enrichment Analysis. *Front Genet* (2021) 11:10. doi: 10.3389/fgene.2020.607798
63. Zhang T, Zhang S-W, Li Y. Identifying Driver Genes for Individual Patients through Inductive Matrix Completion. *Bioinformatics* (2021) 37, 23:4477–84. doi: 10.1093/bioinformatics/btab477
64. Guo WF, Zhang SW, Liu LL, Liu F, Shi QQ, Zhang L, et al. Discovering Personalized Driver Mutation Profiles of Single Samples in Cancer by Network Control Strategy. *Bioinformatics* (2018) 34:1893–903. doi: 10.1093/bioinformatics/bty006
65. Guo WF, Zhang SW, Shi QQ, Zhang CM, Zeng T, Chen L. A Novel Algorithm for Finding Optimal Driver Nodes to Target Control Complex Networks and its Applications for Drug Targets Identification. *BMC Genomics* (2018) 19(S1):924. doi: 10.1186/s12864-017-4332-z
66. Guo WF, Zhang SW, Zeng T, Akutsu T, Chen LN. Network Control Principles for Identifying Personalized Driver Genes in Cancer. *Brief Bioinform* (2020) 21:1641–62. doi: 10.1093/bib/bbz089
67. Page L, Brin S, Motwani R, Winograd T. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford University, Stanford: Stanford Digital Libraries Working Paper (1998).
68. Huang L, Jiang Y, Chen Y. Predicting Drug Combination Index and Simulating the Network-Regulation Dynamics by Mathematical Modeling of Drug-Targeted EGFR-ERK Signaling Pathway. *Sci Rep* (2017) 7:40752. doi: 10.1038/srep40752
69. Chen D, Liu X, Yang Y, Yang H, Lu P. Systematic Synergy Modeling: Understanding Drug Synergy From a Systems Biology Perspective. *BMC Syst Biol* (2015) 9:56. doi: 10.1186/s12918-015-0202-y
70. Madani Tonekaboni SA, Soltan GL, Manem VS, Haibe-Kains B. Predictive Approaches for Drug Combination Discovery in Cancer. *Brief Bioinform* (2016) 19:2. doi: 10.1093/bib/bbw104
71. Kaifang P, Ying-Wooi W, Choi WT, Donehower LA, Jingchun S, Dhruv P, et al. Combinatorial Therapy Discovery Using Mixed Integer Linear Programming. *Bioinformatics* (2014) 30:1456. doi: 10.1093/bioinformatics/btu046
72. Lei H, Fuhai L, Jianting S, Xiaofeng X, Jinwen M, Ming Z, et al. DrugComboRanker: Drug Combination Discovery Based on Target Network Analysis. *Bioinformatics* (2014) 30:i228. doi: 10.1093/bioinformatics/btu278
73. Yadav B, Wennerberg K, Aittokallio T, Tang J. Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model. *Comput Struct Biotechnol J* (2015) 13:504–13. doi: 10.1016/j.csbj.2015.09.001
74. Ding P, Luo J, Liang C, Xiao Q, Cao B, Li G. Discovering Synergistic Drug Combination From a Computational Perspective. *Curr Top Med Chem* (2018) 18:965–74. doi: 10.2174/1568026618666180330141804
75. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G. Nllss: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. *PloS Comput Biol* (2016) 12:e1004975. doi: 10.1371/journal.pcbi.1004975
76. Liu Y, Zhao H. Predicting Synergistic Effects Between Compounds Through Their Structural Similarity and Effects on Transcriptomes. *Bioinformatics* (2016) 32:3782. doi: 10.1093/bioinformatics/btw509
77. Preuer K, Rpi L, Hochreiter S, Bender A, Bulusu KC, Klambauer G. Deepsynergy: Predicting Anti-Cancer Drug Synergy With Deep Learning. *Bioinformatics* (2018) 34:9. doi: 10.1093/bioinformatics/btx806
78. Jeon M, Kim S, Park S, Lee H, Kang J. *In Silico* Drug Combination Discovery for Personalized Cancer Therapy. *BMC Syst Biol* (2018) 12:16. doi: 10.1186/s12918-018-0546-1
79. Gayvert KM, Aly O, Platt J, Bosenberg MW, Stern DF, Elemento O. A Computational Approach for Identifying Synergistic Drug Combinations. *PloS Comput Biol* (2017) 13:e1005308. doi: 10.1371/journal.pcbi.1005308
80. Zhang M, Lee S, Yao B, Xiao G, Xu L, Xie Y. DIGREM: An Integrated Web-Based Platform for Detecting Effective Multi-Drug Combinations. *Bioinformatics* (2018) 35(10):1792–4. doi: 10.1093/bioinformatics/bty860
81. Hsu YC, Chiu YC, Chen Y, Hsiao TH, Chuang EY. A Simple Gene Set-Based Method Accurately Predicts the Synergy of Drug Pairs. *BMC Syst Biol* (2016) 10:66. doi: 10.1186/s12918-016-0310-3
82. Chen D, Zhang H, Lu P, Liu X, Cao H. Synergy Evaluation by a Pathway-Pathway Interaction Network: A New Way to Predict Drug Combination. *Mol Biosyst* (2015) 12:614–23. doi: 10.1039/C5MB00599J
83. Rabadan R, Kribelbauer J, Wang J. Network Propagation Reveals Novel Features Predicting Drug Response of Cancer Cell Lines. *Curr Bioinfrm* (2016) 11:2. doi: 10.2174/1574893611666160125222144
84. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep Learning for Computational Biology. *Mol Syst Biol* (2016) 12:878. doi: 10.15252/msb.20156651
85. Peng L, Chao H, Yingxue F, Jinan W, Ziyin W, Jinlong R, et al. Large-Scale Exploration and Analysis of Drug Combinations. *Bioinformatics* (2015) 31:2007. doi: 10.1093/bioinformatics/btv080
86. Li X, Xu Y, Cui H, Huang T, Wang D, Lian B, et al. Prediction of Synergistic Anti-Cancer Drug Combinations Based on Drug Target Network and Drug Induced Gene Expression Profiles. *Artif Intell Med* (2017) 83:35–43. doi: 10.1016/j.artmed.2017.05.008
87. Shi JY, Li JX, Gao K, Lei P, Yiu SM. Predicting Combinative Drug Pairs Towards Realistic Screening *Via* Integrating Heterogeneous Features. *BMC Bioinform* (2017) 18:409. doi: 10.1186/s12859-017-1818-2
88. Xia F, Shukla M, Brettin T, Garcia-Cardona C, Cohn J, Allen JE, et al. Predicting Tumor Cell Line Response to Drug Pairs With Deep Learning. *BMC Bioinform* (2018) 19:486. doi: 10.1186/s12859-018-2509-3
89. Sheng J, Li F, Wong STC. Optimal Drug Prediction from Personal Genomics Profiles. *J BioMed Health Inform* (2017) 19:1264–70. doi: 10.1109/JBHI.2015.2412522
90. Liany H, Jeyasekharan A, Rajan V. DruID: Personalized Drug Recommendations by Integrating Multiple Biomedical Databases for Cancer. *bioRxiv* (2021). 2021.04.11.439315. doi: 10.1101/2021.04.11.439315
91. Piñeiro-Yáñez E, Reboiro-Jato M, Gómez-López G, Perales-Patón J, Troulé K, Rodríguez JM, et al. PanDrugs: A Novel Method to Prioritize Anticancer Drug Treatments According to Individual Genomic Data. *Genome Med* (2018) 10:41. doi: 10.1186/s13073-018-0546-1
92. Choo SM, Ban B, Joo JI, Cho KH. The Phenotype Control Kernel of a Biomolecular Regulatory Network. *BMC Syst Biol* (2018) 12:49. doi: 10.1186/s12918-018-0576-8
93. Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R, De SL, Mauri G, et al. Algorithmic Methods to Infer the Evolutionary Trajectories in Cancer Progression. *Proc Natl Acad Sci USA* (2015) 113:e4025. doi: 10.1101/027359

94. Nussinov R, Jang H, Tsai C-J, Cheng F. Precision Medicine Review: Rare Driver Mutations and Their Biophysical Classification. *Biophys Rev* (2019) 11:5–19. doi: 10.1007/s12551-018-0496-2
95. Patil A, Kumagai Y, Liang K-c, Suzuki Y, Nakai K. Linking Transcriptional Changes Over Time in Stimulated Dendritic Cells to Identify Gene Networks Activated During the Innate Immune Response. *PloS Comput Biol* (2013) 9: e1003323. doi: 10.1371/journal.pcbi.1003323
96. Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting Early-Warning Signals for Sudden Deterioration of Complex Diseases by Dynamical Network Biomarkers. *Sci Rep* (2012) 2:342. doi: 10.1038/srep00342
97. Kazuyuki A, Liu R, Koizumi K, Liu X, Chen L. Dynamical Network Biomarkers: Theory and Applications. *Gene* (2021) 808:145997. doi: 10.1016/j.gene.2021.145997
98. Honda K, Ono M, Shitashige M, Masuda M, Kamita M, Miura N, et al. Proteomic Approaches to the Discovery of Cancer Biomarkers for Early Detection and Personalized Medicine. *Jpn J Clin Oncol* (2012) 43:103–9. doi: 10.1093/jjco/hys200
99. Li M, Zeng T, Liu R, Chen L. Detecting Tissue-Specific Early Warning Signals for Complex Diseases Based on Dynamical Network Biomarkers: Study of Type 2 Diabetes by Cross-Tissue Analysis. *Brief Bioinform* (2013) 15:229–43. doi: 10.1093/bib/bbt027
100. Xu J, Wu M, Zhu S, Lei J, Gao J. Detecting the Stable Point of Therapeutic Effect of Chronic Myeloid Leukemia Based on Dynamic Network Biomarkers. *BMC Bioinform* (2019) 20:202. doi: 10.1186/s12859-019-2738-0
101. Sun Y, Zhao H, Wu M, Xu J, Zhu S, Gao J. Identifying Critical States of Hepatocellular Carcinoma Based on Landscape Dynamic Network Biomarkers. *Comput Biol Chem* (2020) 85:107202. doi: 10.1016/j.compbiolchem.2020.107202
102. Zhang X, Xie R, Liu Z, Pan Y, Liu R, Chen P. Identifying Pre-Outbreak Signals of Hand, Foot and Mouth Disease Based on Landscape Dynamic Network Marker. *BMC Infect Dis* (2021) 21:6. doi: 10.1186/s12879-020-05709-w
103. Liu YY, Slotine JJ, Barabási AL. Observability of Complex Systems. *Proc Natl Acad Sci U S A* (2013) 110:2460–5. doi: 10.1073/pnas.1215508110
104. Montanari AN, Duan C, Aguirre LA, Motter AE. Functional Observability and Target State Estimation in Large-Scale Networks. *Proc Natl Acad Sci* (2022) 119:e2113750119. doi: 10.1073/pnas.2113750119
105. Ulirsch JC, Lareau CA, Bao EL, Ludwig LS, Guo MH, Benner C, et al. Interrogation of Human Hematopoiesis at Single-Cell and Single-Variant Resolution. *Nat Genet* (2019) 51:683–93. doi: 10.1038/s41588-019-0362-6
106. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *J R Soc Interface* (2018) 15:20170387. doi: 10.1098/rsif.2017.0387
107. Greener JG, Kandathil SM, Moffat L, Jones DT. A Guide to Machine Learning for Biologists. *J Nat Rev Mol Cell Biol* (2021) 23:40–55. doi: 10.1038/s41580-021-00407-0
108. Lou Y, He Y, Wang L, Chen G. Predicting Network Controllability Robustness: A Convolutional Neural Network Approach. *IEEE Trans Cybern* (2020) 1–12. doi: 10.1109/TCYB.2020.3013251
109. Harley JB, Alarcón-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, et al. Genetics, Genome-Wide Association Scan in Women With Systemic Lupus Erythematosus Identifies Susceptibility Variants in ITGAM, PXK, KIAA1542 and Other Loci. *Nat Genet* (2008) 40:204. doi: 10.1038/ng.81
110. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. *N Engl J Med* (2010) 363:166–76. doi: 10.1056/NEJMra0905980

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yan, Hu, Li, Sun and Guo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Immunology

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fimmu.2022.917014

DDOST Correlated with Malignancies and Immune Microenvironment in Gliomas

Xiaojing Chang[1†], Jie Pan[2†], Ruoyu Zhao[3†], Tianfang Yan[4], Xinrui Wang[3], Cunle Guo[3], Yining Yang[3] and Guohui Wang[3*]*

[1] Department of Radiotherapy, the Second Hospital of Hebei Medical University, Shijiazhuang, China, [2] Department of Pathology, Stanford University School of Medicine, Stanford, CA, United States, [3] Department of Radiotherapy, Tianjin First Center Hospital, Tianjin, China, [4] Department of Neurological Diagnosis and Restoration, Osaka University Graduate School of Medicine, Suita, Japan

OPEN ACCESS

Edited by:
Ken Chen,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Salvatore J. Coniglio,
Kean University, United States
Deepak Bhere,
University of South Carolina, United States

***Correspondence:**
Guohui Wang
wangguohuipku@163.com
Yining Yang
Yining_1624@163.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *10 April 2022*
Accepted: *20 May 2022*
Published: *23 June 2022*

Citation:
Chang X, Pan J, Zhao R, Yan T, Wang X, Guo C, Yang Y and Wang G (2022) DDOST Correlated with Malignancies and Immune Microenvironment in Gliomas. Front. Immunol. 13:917014. doi: 10.3389/fimmu.2022.917014

Among the most common types of brain tumor, gliomas are the most aggressive and have the poorest prognosis. Dolichyl-diphosphooligosaccharide protein glycosyltransferase non-catalytic subunit (DDOST) encodes a component of the oligosaccharide transferase complex and is related to the N-glycosylation of proteins. The role of DDOST in gliomas, however, is not yet known. First, we performed a pan cancer analysis of DDOST in the TCGA cohort. The expression of DDOST was compared between glioma and normal brain tissues in the GEO and Chinese Glioma Genome Atlas (CGGA) databases. In order to explore the role of DDOST in glioma, we analyze the impact of DDOST on the prognosis of glioma patients, with the CGGA 325 dataset as a test set and the CGGA 693 dataset as a validation set. Immunohistochemistry was performed on tissue microarrays to examine whether DDOST has an impact on glioma patient survival. Next, using single-cell sequencing analysis, GSEA, immune infiltration analysis, and mutation analysis, we explored how DDOST affected the glioma tumor microenvironment. Finally, we evaluated the clinical significance of DDOST for glioma treatment by constructing nomograms and decision curve analysis (DCA) curves. We found that DDOST was overexpressed in patients with high grade, IDH wild type, 1p19q non-codel and MGMT un-methylated, which was associated with poor prognosis. Patients with high levels of DDOST, regardless of their clinical characteristics, had a worse prognosis. Immunohistochemical analysis confirmed the results of the above bioinformatics analysis. Mechanistic analysis revealed that DDOST was closely associated with the glioma microenvironment and negatively related to tumor-infiltrating B cells and CD4+ T cells and positively related to CAFs and tumor-associated macrophages. In conclusion, these findings suggested that DDOST mediated the immunosuppressive microenvironment of gliomas and could be an important biomarker in diagnosing and treating gliomas.

Keywords: DDOST, glioma, progression, microenvironment, prognosis

INTRODUCTION

Gliomas are the most common malignant primary intracranial tumors that arise from glial or precursor cells, accounting for approximately 24.5% of all primary brain and other central nervous system (CNS) tumors and 80.9% of CNS malignancies (1). According to the 2021 fifth edition World Health Organization (WHO) classification, gliomas are categorized into four grades, WHO 1–4 (2). WHO 1 and WHO 2, with a relatively good OS, are defined as low-grade gliomas, and WHO 3 and WHO 4, which are related to a poor prognosis, are defined as high-grade gliomas. The 5-year survival rate has a large variation from 94.7% for pilocytic astrocytoma (WHO 1 grade) to 6.8% for glioblastoma (WHO 4 grade) (1). More recently, molecular biomarkers, such as isocitrate dehydrogenase (IDH), telomerase reverse transcriptase (TERT), and O-6-methylguanine-DNA methyltransferase (MGMT), have gained importance in providing diagnostic, prognostic, and therapeutic information (3). The leading change in the 2021 fifth edition WHO classification is advancing the role of molecular diagnostics, which could provide the most accurate classification of CNS neoplasms and guide clinical management and prognosis of gliomas (2). Thus, the investigation of biomarkers is a pivotal area in the research of gliomas.

Glycosylation, one of the important posttranslational modifications of proteins, is involved in various cell biological processes, and is closely related to many pathological processes, such as tumorigenesis and inflammatory responses (4). Glycosylation modifications are classified into O-glycosylation, N-glycosylation, C-glycosylation, and glycosylphosphatidylinositol anchor linkage based on the glycosylation site, among which O-glycosylation and N-glycosylation are the most common. Alterations in glycosylation could produce tumor-associated polysaccharides or glycoproteins that can serve as tumor-related markers and correlate with tumor development and prognosis (5). Dolichyl-diphosphooligosaccharide protein glycosyltransferase non-catalytic subunit (DDOST) encodes a component of the oligosaccharide transferase (OST) complex and is related to the N-glycosylation of proteins. Some researchers found that patients with skin squamous cell carcinoma with highly expressed DDOST showed a worse prognosis (6). Moreover, studies have shown that DDOST is an independent prognostic factor in patients with liver cancer, and its expression is positively correlated with the level of Th2 cell infiltration, but negatively correlated with the level of cytotoxic cell infiltration (7). However, no studies have been conducted on DDOST in gliomas.

In this project, we investigated the expression and significance of DDOST in gliomas by bioinformatics and immunohistochemical staining, and preliminarily discussed its impact on the glioma microenvironment. Furthermore, these results were verified by different databases.

MATERIALS AND METHODS

Data Acquisition

Pan cancer analysis data, TCGA Pan Cancer, were obtained from University of California, Santa Cruz (UCSC) (https://xenabrowser.net/). Microarray data of glioma expression profile, GSE4290 (8) and GSE50161 (9), were downloaded from the GEO database. From the TCGA database, gene expression and clinical information files of LGG and GBM were downloaded. Additionally, we analyzed the Chinese Glioma Genome Atlas (CGGA) to obtain RNA profiles and clinical features of gliomas and single-cell sequence information (10). All gene expression data were standardized and batch-processed using the limma R package. In the analyzed dataset, we retained only the gliomas with complete clinicopathological data and survival data, excluding patients with unknown or incomplete data. Single-cell sequencing data were normalized and quality controlled by the Seurat R package.

Pan Cancer Analysis

We performed differential analysis of DDOST on the pan cancer expression profile data downloaded from UCSC by Wilcoxon rank sum and signed rank tests. Then, the data were subjected to Cox proportional hazards regression models to determine the effects of DDOST on overall survival (OS).

DDOST Differential Expression in Glioma and Normal Brain Tissue

Firstly, we compared the expression of DDOST in gliomas and normal brain tissues by Wilcoxon rank sum test in CGGA, GSE4290, GSE50161, TCGA, and GTEx databases. Then, the CGGA 325 cohort was used as the test set and the CGGA 693 cohort was used as the validation set. We compared the expression of DDOST in gliomas with different clinical characteristics, such as age (age<42, age≥42), gender (female, male), grade (WHO 2, WHO 3, and WHO 4), IDH (mutant, wild type), 1p19q (codel, non-codel), and MGMT (methylated, un-methylated) status.

Prognostic Analysis

Kaplan–Meier survival analysis was used to determine the survival prognosis for gliomas. Patients were grouped into the "high" or "low" group based on the median value of DDOST expression. According to the above clinical characteristics, we analyzed the survival subgroups of glioma patients, including different age, gender, grade, and molecular characteristics. The CGGA 325 cohort was used as the test set and the CGGA 693 cohort was used as the validation set. The results were displayed by hazard ratio (HR), 95% confidence intervals (95% CI), and log-rank p-value (significant threshold <0.05).

Tissue Microarray and Immunohistochemistry

Shanghai Outdo Biotech Co., Ltd. (China) provided the tissue microarray applied for this research. The detailed procedure and scoring criteria of immunohistochemistry were referred to previously published articles from our team (11). A mouse monoclonal anti-DDOST antibody at a dilution of 1:100 was purchased from Santa Cruz (sc-74408). This semi-quantitative analysis was done by two independent assessors without prior knowledge of the patient outcome.

Single-Cell Analysis

Gene alteration information could be obtained from single-cell data in a more detailed way. The processed single-cell data containing 6,148 cells were downloaded from the CGGA database (12). The Seurat R software package was performed to reduce the dimensions of cells and generate t-SNE diagram for cell-type visualization. In order to study the role of DDOST in tumors, we compared the expression differences of DDOST between different cell types.

Database for Annotation, Visualization, and Integrated Discovery and Gene Set Enrichment Analysis

Differentially expressed genes (DEGs) between gliomas with high vs. low DDOST expression in the CGGA 325 cohort were identified using the limma package on R using $|\log FC0| > 1.5$ and FDR < 0.05 as cutoffs. To further explore its possible mechanism, we conducted GO and KEGG pathway analysis using DAVID 6.8 (https://david.abcc.ncifcrf.gov/) (13). To further explore the possible mechanism, the GSEA 4.0.2 software was used (14). A normalized enrichment score (NES) >1 and false discovery rate <0.05 were considered significant.

Evaluation of the Effect of DDOST on the Glioma Microenvironment

The fraction of immunocytes in gliomas was estimated using single-sample gene set enrichment analysis (ssGSEA) and CIBERSORT as previously published (11). Mass cytometry data from immune cells can be analyzed in a standardized manner using the Estimate the Proportion of Immune and Cancer cells (EPIC) web-based analytical and discovery platform (15).

Gliomas Were Clustered According to Immune-Related Genes by NMF

First, we download the list of immune-related genes from the ImmPort database (16). Non-negative matrix factorization (NMF) is an effective dimensionality reduction method that is widely used for molecular pattern recognition of high-dimensional genomic data and provides a powerful approach for class discovery. We extracted immune-related gene expression quantities from the CGGA 325 database and then performed NMF clustering analysis. The optimal number of clusters was calculated according to the values obtained from cophenetic.

Molecular Characteristic Specific for Low and High DDOST Groups

The TCGA database was used to collect single-nucleotide variants (SNVs) of LGG and GBM. Based on the expression values of DDOST, gliomas were divided into high and low groups. Wilcoxon rank sum test was used to compare gene mutations between the two patient groups. Waterfall plots were utilized to demonstrate the mutation type and mutation expression relationship with DDOST for the top 15 mutations.

Construction of a Predictive Nomogram

Nomograms referred to quantitative analysis plots representing the functional relationship between multiple variables with a cluster of mutually disjoint segments in plane coordinates. Clinical features such as age, gender, grade, and molecular characteristic of IDH, 1p19q, MGMT, and DDOST were used to build a nomogram. The probability of 1-, 3-, and 5-year OS of patients with glioma was determined using the rms R package. Decision curve analysis (DCA) curves assessed the value of DDOST for prognostic assessment.

RESULTS

Characteristics of Patients With Gliomas

We obtained gene expression data from 1,572 gliomas and 41 normal brain tissues by screening publicly available databases, as well as a single-cell sequencing dataset comprising 6,148 cells. Tissue microarray included 121 cases of glioma and 3 cases of normal brain tissue. The CGGA 325 cohort served as the test set and the CGGA 693 cohort served as the validation set. In **Table 1**, the clinical characteristics and molecular characteristics are listed.

Pan Cancer Analysis Revealed That DDOST Was Highly Expressed in a Subset of Tumors and Associated With Poor Prognosis

We calculated the expression difference between normal and tumor samples in each tumor *via* the R software, using unpaired Wilcoxon rank sum and signed rank tests for differential significance analysis. Results showed that significant upregulation of DDOST was observed in 19 tumors, such as GBM (tumor: 7.37 ± 0.46, normal: 6.15 ± 0.07, p = 2.0e-4) and CESC (tumor: 7.51 ± 0.53, normal: 6.95 ± 0.13, p = 0.03), while DDOST was downregulated in only 2 tumors, namely, THCA (tumor: 7.32 ± 0.47, normal: 7.69 ± 0.36, p = 3.6e-11) and KICH (tumor: 6.18 ± 0.56, normal: 6.97 ± 0.31, p = 2.1e-21) (**Figure 1A**). Cox proportional hazards regression models were performed to analyze the prognostic relationship between gene expression and prognosis within each tumor. High expression of DDOST was associated with poor prognosis in 11 tumor types (**Figure 1B**).

Gliomas With High DDOST Expression Showed Poor Prognosis

We further performed an expression analysis about DDOST in gliomas using sequencing analysis data from CGGA, GEO, TCGA, and GTEx databases. Results showed that DDOST expression was significantly higher than that in normal brain tissues (**Figures 2A–E**). In addition, the relationship between DDOST expression and different clinical features was also analyzed. Significantly higher expression of DDOST was identified in those clinical features known to confer worse prognosis, such as age ≥ 42, high grade, IDH wild type, 1p19q non-codel, and MGMT methylation (**Figure 3**). In the

TABLE 1 | Clinicopathological characteristics of glioma patients from the GEO, TCGA, and CGGA database and tissue microarray.

	GSE4290 (*n* = 176)	GSE50161 (*n* = 130)	TCGA (*n* = 592)	CGGA 325 (*n* = 286)	CGGA 693 (*n* = 429)	Tissue Microarray (*n* = 124)
Age						
<42	NA	NA	241	124	197	44
≥42	NA	NA	346	162	232	77
Gender						
Female	NA	NA	246	110	191	35
Male	NA	NA	341	176	238	86
Normal Tissue	23	13	5	NA	NA	3
Tumor Tissue	153	117	587	286	429	121
Grade						
II	45	NA	211	86	100	37
III	31	NA	234	68	164	40
IV	77	NA	142	132	165	47
IDH Status						
Wild type	NA	NA	219	136	193	NA
Mutation	NA	NA	368	150	236	NA
1p/19q						
Codel	NA	NA	149	57	90	NA
Non-codel	NA	NA	438	229	339	NA
MGMT						
Methylated	NA	NA	NA	147	251	NA
Un-methylated	NA	NA	NA	139	178	NA
Status						
Dead	NA	NA	173	202	278	47
Alive	NA	NA	414	84	151	77

NA, Not available.

CGGA 693 cohort, which was used as an external validation dataset, we reached the same conclusion (**Supplementary 2**). Next, we investigated the impact of DDOST on survival of gliomas using the CGGA 325 and CGGA 693 datasets. The high/low expression of DDOST was determined by the median value. Kaplan–Meier survival analysis revealed that patients with high DDOST expression had significantly worse OS than those with low expression (HR = 5.11, 95% CI: 3.76–6.95, $p \leq 0.001$, **Figure 4A**). Stratification analysis was carried out on DDOST to assess its impact on glioma prognosis. Patients with glioma were divided into different groups based on different clinical characteristics for age (age<42, age≥42), gender (female, male), grade (WHO 2, WHO 3, and WHO 4), IDH (mutant, wild type), 1p19q (codel, non-codel), and MGMT (methylated, un-methylated) status. Survival analysis showed that patients with high expression of DDOST had a poor prognosis in various subgroups (**Figures 4B–N**). However, in the 1p19q codel group, we observed only patients with low expression of DDOST. We obtained similar results from the external validation set (**Supplementary 3**). Immunohistochemistry of tissue microarray was used to analyze and confirm the relationship between DDOST expression and prognosis. As shown in **Figures 5A–C**, DDOST expression was significantly positively correlated with WHO grade, with a strong positivity in WHO 4, a moderate positivity in WHO 3, and a weak positivity in WHO 2. Brain tissue from normal individuals had no evidence of DDOST expression (**Figure 5D**). Survival analysis showed that the OS was significantly poorer in cases with high DDOST expression (HR = 5.01, 95% CI: 2.23–11.26, $p < 0.001$, **Figure 5E**). All of the above results strongly suggest that DDOST can accurately predict the prognosis of glioma patients.

DDOST Was Heterogenous in the Tumor Immune Microenvironment

In 6,148 cells from 13 gliomas, cells were mainly divided into 5 types, namely, astrocytes, macrophages, epithelial cells, monocytes, and T cells (**Figure 6A**). Based on single-cell analysis, the expression of DDOST was significantly different in different clusters (**Figure 6B**). DDOST expression was increased in some clusters, especially in astrocytes, macrophages, and monocytes (**Figure 6C**).

DDOST Was Associated With Cancer-Promoting Pathways

There were 1,279 DEGs, of which 613 were upregulated and 666 were downregulated (**Figure 7A** and **Supplementary 4**). DAVID analysis was conducted on the above differential genes to identify functional enrichments. The results showed that the differential genes were significantly enriched in neuroactive ligand–receptor interaction, cAMP signaling pathway, calcium signaling pathway, epithelial–mesenchymal transition, KRAS, inflammatory response, and other pathways (**Figures 7B, C**). Enrichment results of GSEA did not depend on differential genes, and their results together with those of DAVID could better predict the underlying mechanisms of DDOST. Our GSEA analysis identified pathways associated with glucose metabolism, and oncogenic and tumor microenvironment (TME) as being enriched (**Supplementary Figures 3A, B**).

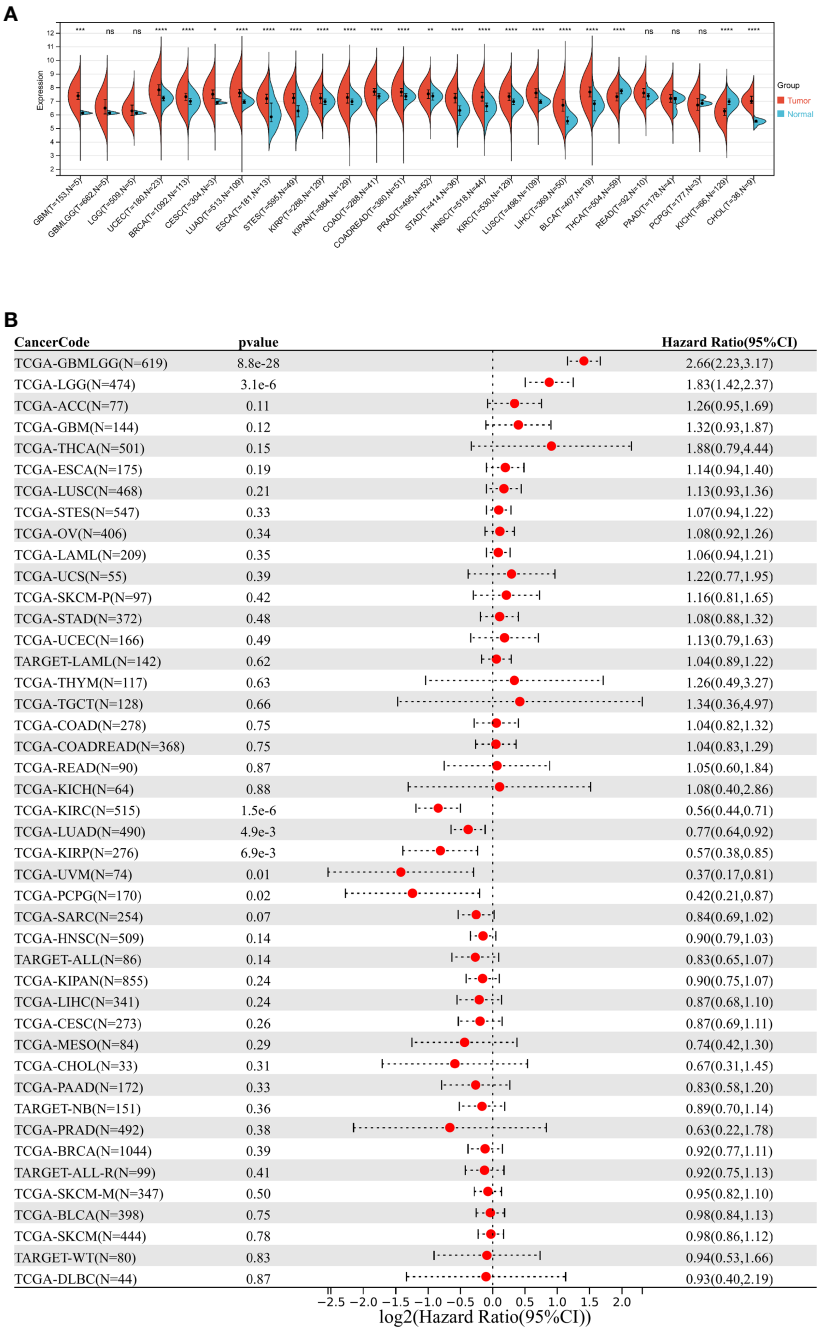


FIGURE 1 | Pan cancer analysis of DDOST expression. **(A)** Analysis of DDOST expression in 33 tumors on TCGA. *, **, ***, **** indicate $p < 0.05$, $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively; ns, not significant (Wilcoxon test). **(B)** Risk plot of correlation between DDOST levels and OS.

The Expression of DDOST Was Related to the Infiltration of Immune Cells in the Tumor Microenvironment of Gliomas

The infiltration of immune cells was assessed using ssGSEA. The immune cell infiltration status of the patients was used to categorize them into low and high immunity groups. Results showed that the expression of DDOST in the group with more immune cell infiltration was significantly higher than that with lower immune cell infiltration (**Figure 8A**). The expression level of DDOST was negatively correlated with tumor purity ($r = -0.333$, $p < 0.001$, **Figure 8B**). However, it was positively correlated with stromal score ($r = 0.318$, $p < 0.001$, **Figure 8C**), ESTIMATE score ($r = 0.333$, $p < 0.001$, **Figure 8D**), and immune score ($r = 0.322$, $p < 0.001$, **Figure 8E**). Furthermore, we analyzed the relationship between DDOST and infiltrating B cells, CAFs,

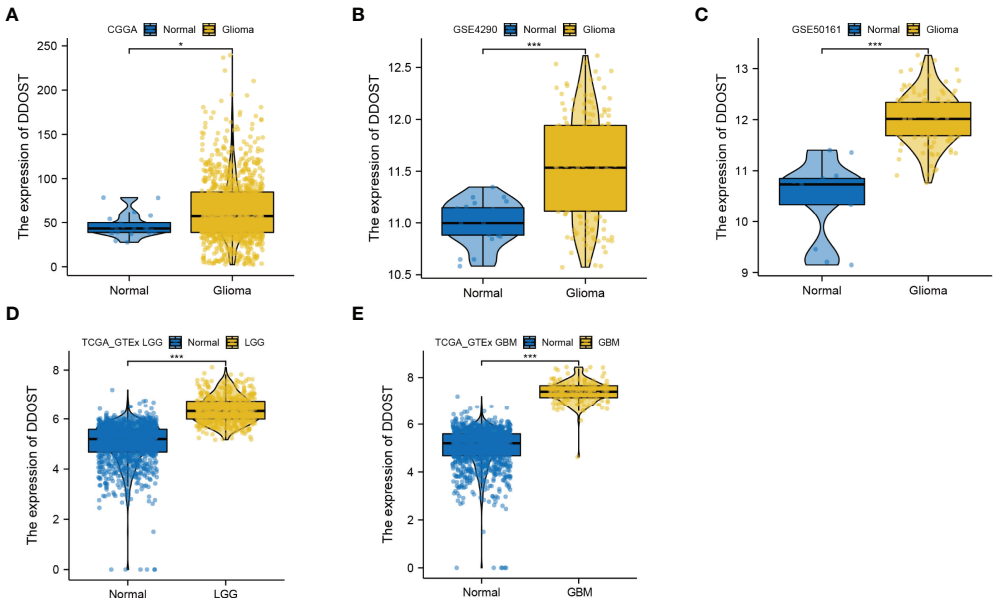


FIGURE 2 | Expression difference of DDOST between glioma and normal tissue. **(A)** The expression of DDOST was different between glioma and normal brain tissue in the CGGA cohort. **(B, C)** In the GSE4290 and GSE50161 cohort, DDOST expression was significantly higher in glioma patients than in normal brain. **(D, E)** In the TCGA and GTEx, the expression of DDOST in LGG and GBM was significantly higher than that in normal tissue. *$p < 0.05$, ***$p < 0.001$.

CD4+ T cells, CD8+ T cells, macrophages, NK cells, and other cells through EPIC. The expression of DDOST was negatively correlated with the infiltrating B cells (r = -0.587, p < 0.001) and CD4+ T cells (r = -0.498, $p < 0.001$). However, it was positively correlated with the infiltrating CAFs (r = 0.589, $p < 0.001$) and macrophages (r = 0.463, $p < 0.001$). However, there was no significant correlation between the expression of DDOST and NK cells and other cells. Scatter plots of the association of DDOST with these immune cells are shown in **Figures 9A–G**. Then, we obtained 2,483 immune-related genes from the website (**Supplementary Table 3**). NMF clustering

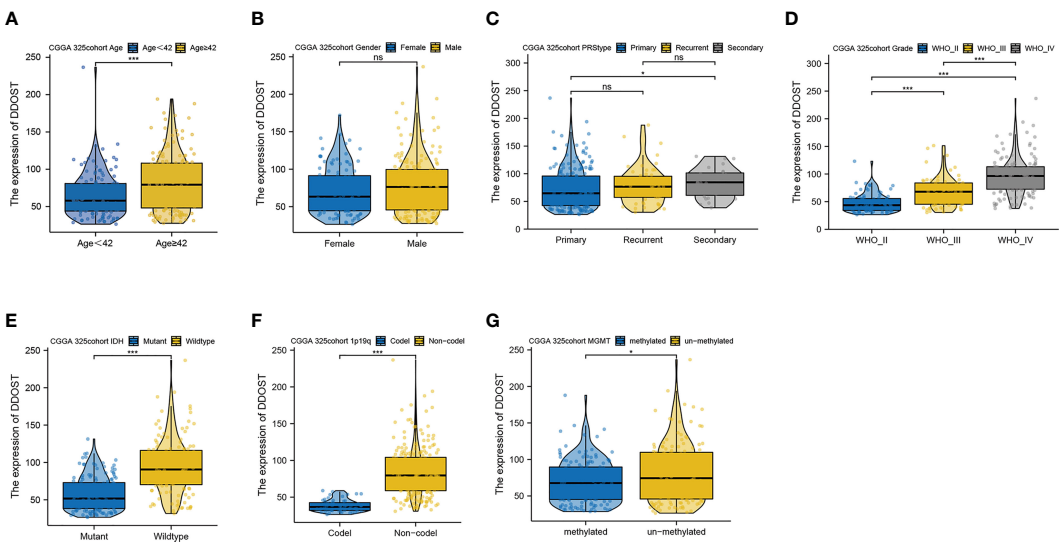


FIGURE 3 | Expression difference of DDOST between different clinical characters in patients with glioma in the CGGA 325 cohort. The expression of DDOST in different age **(A)**, gender **(B)**, PRS type **(C)**, grade **(D)**, IDH **(E)**, 1p19q **(F)**, and MGMT status **(G)**. *$p < 0.05$, ***$p < 0.001$. ns, not significant.

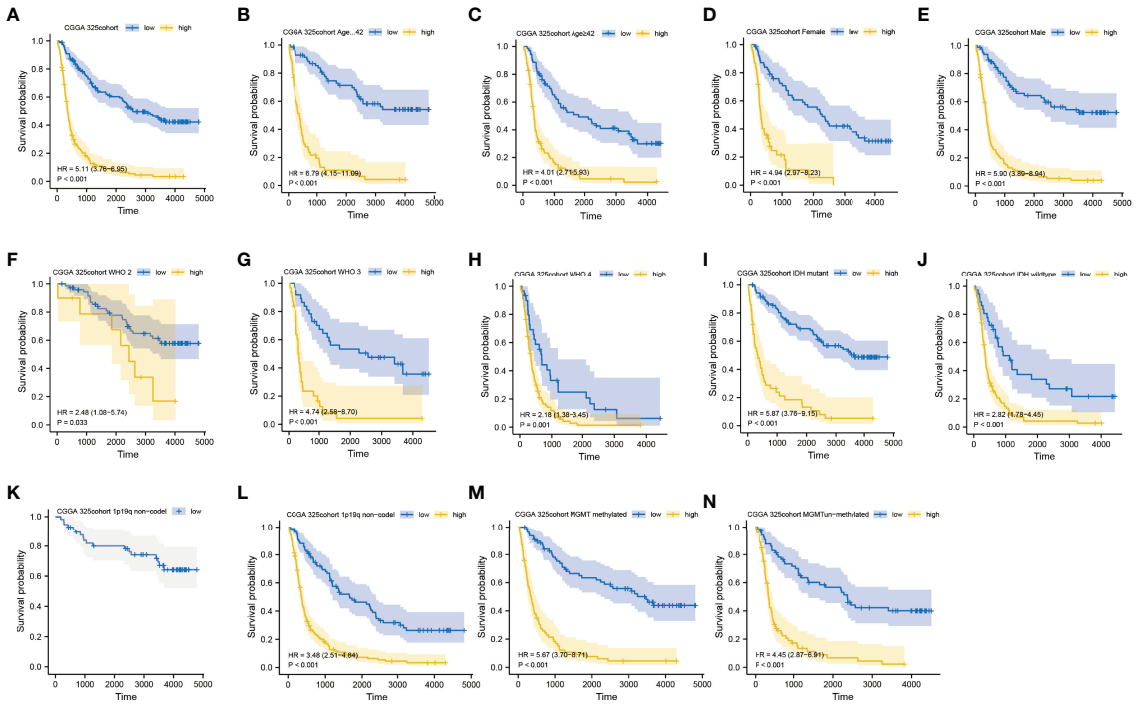


FIGURE 4 | Prediction of outcome of the DDOST in stratified patients in the CGGA 325 dataset. Survival curve was used to analyze OS in the low- and high-DDOST groups in CGGA 325 set **(A)**. Survival analysis of the signature in patients stratified by age **(B, C)**, gender **(D, E)**, grade **(F, H)**, IDH **(I, J)**, 1p19q status **(K, L)**, and MGMT promoter **(M, N)**.

analysis was performed according to the expression of these genes. According to the values of the cophenetic, we clustered gliomas into 2 categories, Cluster 1 and Cluster 2 (**Figure 10A**). Survival analysis showed that patients in group C2 had a significantly worse survival than those in group C1 (**Figure 10B**). Interestingly, we also found that the expression of DDOST was significantly higher in group C2 than in group C1 (**Figure 10C**).

Different DDOST Expression Exhibited Distinct Molecular Features

We detected 509 and 153 samples containing mutations in LGG and GBM. According to the expression of DDOST, gliomas were divided into high- and low-expression groups. Waterfall plots show the differences of the first 15 mutation genes. We found that IDH mutations exist in 80.6% of LGG, and high expression

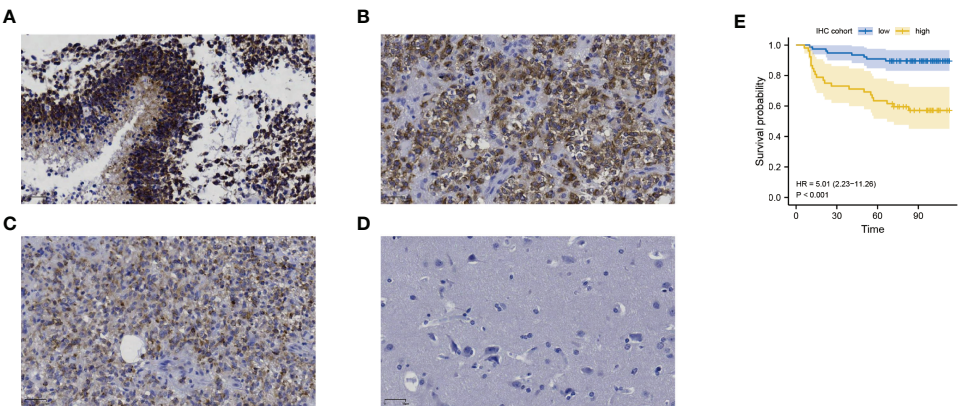


FIGURE 5 | The expression of DDOST in gliomas and its prognostic significance were analyzed by immunohistochemistry. **(A–C)** showed that DDOST is strongly, moderately, and weakly positive in gliomas, respectively. **(D)** shows the expression of DDOST in normal brain tissue. High DDOST expression in glioma was related to poor OS **(E)**.

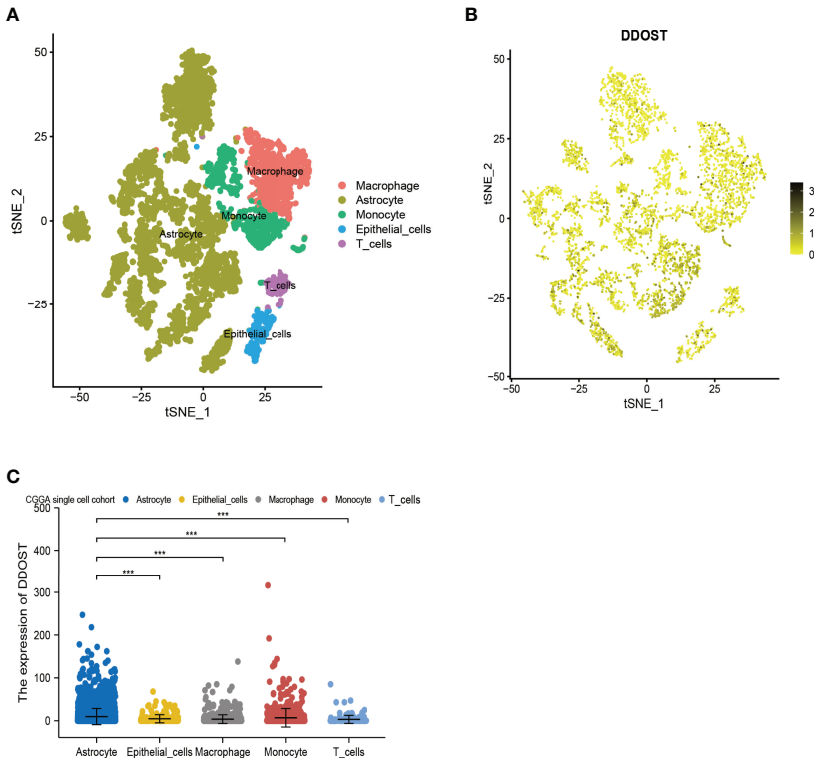


FIGURE 6 | The expression of DDOST in glioma by single-cell analysis. **(A)** By dimensionality reduction analysis of CGGA single-cell data, 6,148 cells were divided into astrocytes, epithelial cells, macrophages, monocytes, and T cells. **(B)** Scatter plot of DDOST distribution in glioma. **(C)** The scatter plot shows the expression of DDOST in astrocytes, epithelial cells, macrophages, monocytes, and T cells. ***$p < 0.001$.

of DDOST had less IDH mutation. However, gene alteration of CIC and FUBP1 in the low-expression DDOST group was significantly higher than that in the high-expression group (**Supplementary Figure 4A**). In GBM, TP53 had the highest rate of gene mutation, but the difference between the two groups was not significant. In the low-expression group, the mutation rate of PTEN was higher (**Supplementary Figure 4B**).

DDOST Could Better Improve the Predictive Model

Nomograms are powerful tools that can be used to determine individuals' risk in a clinical setting based on the integration of multiple factors (17). A nomogram incorporating DNA repair genes, age, PRS, grade, IDH, 1p19q, and MGMT status was created to predict the probability of 1-, 3-, and 5-year OS. As illustrated in **Figure 11A**, a score was assigned to each of the above features. We predicted patient 1-, 3-, and 5-year survival by calculating the sum of the scores for each glioma patient. For example, a patient with an overall score of 393 would have a 5-year survival of 19.7%, a 3-year survival of 37%, and a 1-year survival of 76.5%. According to calibration curves, actual and predicted survival were very close (**Figure 11B**). The DCA curve showed that DDOST combined with other clinical traits can bring benefits to predict the prognosis of glioma patients (**Figure 11C**).

DISCUSSION

Malignancies of the brain account for 1.6% of new cancers and 2.5% of deaths annually worldwide (18). Gliomas account for 80% of brain malignancies. To date, surgery remains to be the first-line treatment for gliomas, and maximal safe resection is the mainstay and the most effective therapeutic option. Conventional radiotherapy combined with chemotherapy based on temozolomide and PCV (procarbazine, lomustine, and vincristine) after maximal safe resection is the standard adjuvant treatment in high-grade glioma and some WHO 2 gliomas, and tumor treating fields (TTFields) are newly recommended for glioblastoma according to the 2021 NCCN guideline (https://www.nccn.org/professionals/physician_gls/pdf/cns.pdf). Unfortunately, the prognosis is still unsatisfactory, especially for glioblastoma, which has a 5-year survival rate of 3%. Evidence shows that genetic alterations are involved in the development of gliomas (3). The leading change in the 2021 fifth edition WHO classification is the recommendation of molecular diagnostics, which could provide the most accurate classification of CNS neoplasms and guide clinical management and prognosis of gliomas. For example, the presence of combined 1p/19q loss and/or IDH1 mutations predicts a favorable prognosis, and could get benefit from PCV chemotherapy.

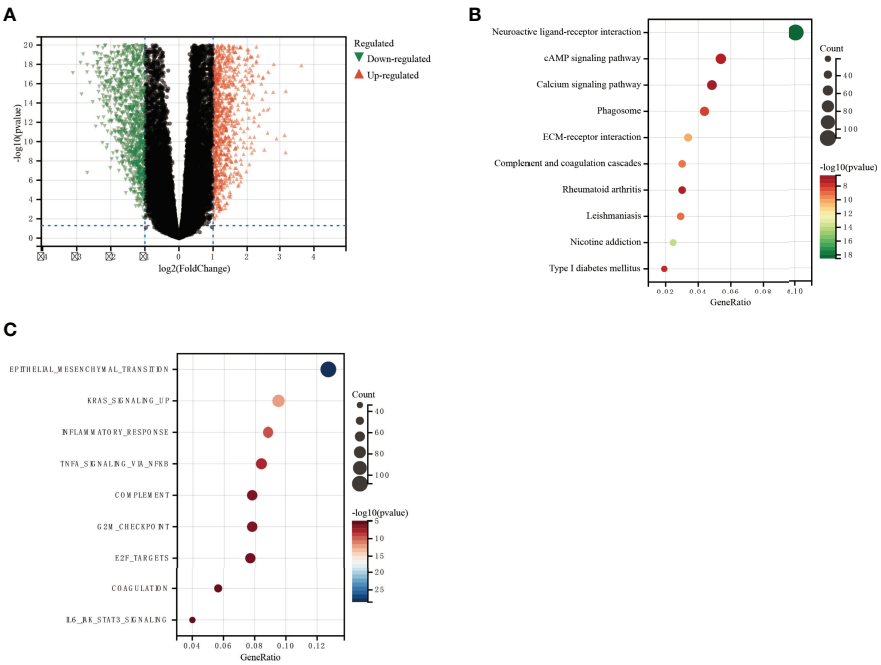


FIGURE 7 | GO and KEGG functional enrichment analysis of the role of DDOST in glioma. **(A)** The volcano map shows the differential genes between high- and low-DDOST groups by the CGGA 325 cohort **(A)**. KEGG **(B)** and GO **(C)** were used to analyze the relevant mechanisms.

A glycosylated protein is an important form of protein modification in eukaryotes, occurring in more than 50% of all proteins (19). Moreover, evidence shows that the glycosylation of proteins is closely related to the biological behavior of tumors. In particular, OST-mediated protein glycosylation is a key factor of signal transduction, protein folding, and degradation. Some studies have also shown that it is closely related to the immune escape of tumor cells (20). A subunit of the OST complex, DDOST, plays a vital role in N-glycosylation (21). It has been shown that DDOST also plays an important role in tumors and can serve as a marker for HCC and is closely associated with the TME (7). Numerous studies on glycomics and glycoproteomics have established a strong link between glycosylation and gliomas (22, 23). In this study, we firstly analyzed the expression and significance of DDOST in pan cancer. Results showed that DDOST expression was upregulated in most tumors compared with the corresponding normal tissues including gliomas, and tumors that expressed high levels of DDOST displayed poor prognosis. IHC analysis of tissue microarray confirmed our data, and DDOST expression was significantly positively correlated with WHO grade, with a

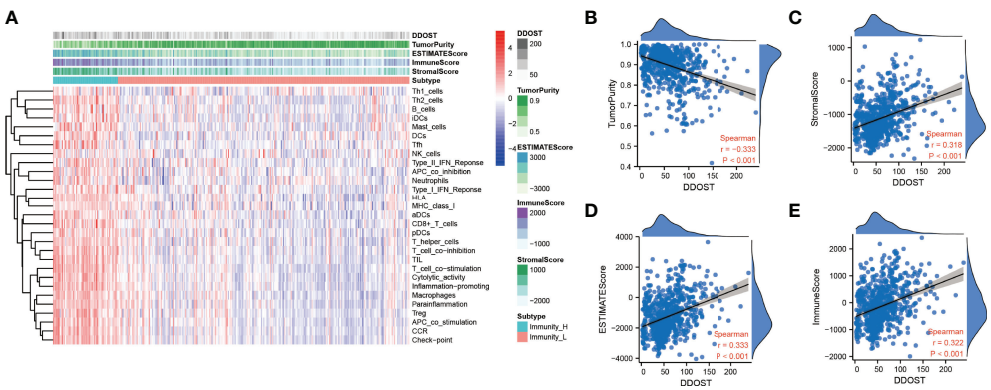


FIGURE 8 | Immune infiltration patterns of low- and high-DDOST analyzed by ssGSEA methods in glioma from the CGGA dataset. **(A)** Heatmap revealing the scores of immune cells in low and high immunities. **(B–E)** Scatter plot showing the correlation between DDOST and tumor purity, stromal, ESTIMATE, and immune scores.

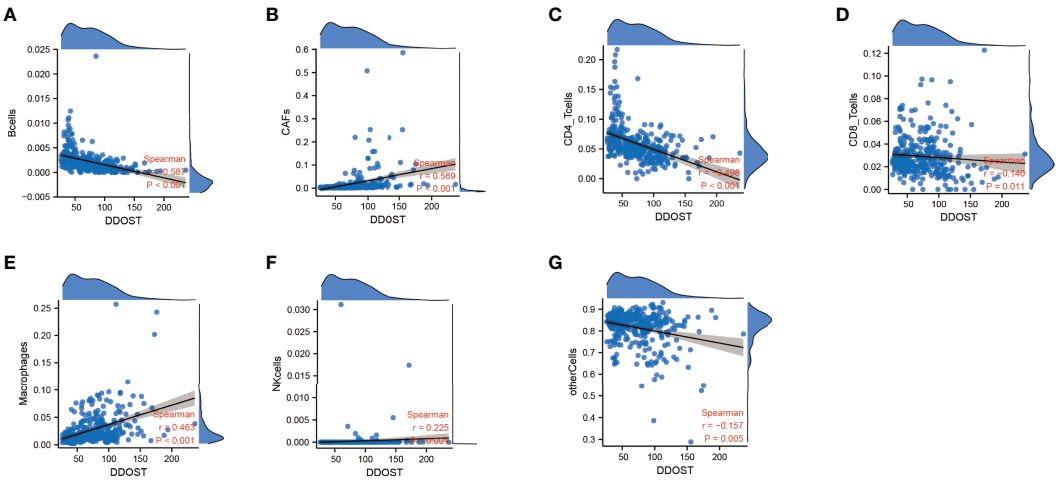


FIGURE 9 | The relationship between the expression of DDOST and immune cells were analyzed by EPIC. The scatter plot shows the correlation between DDOST and B cells **(A)**, CAFs **(B)**, CD4+ T cells **(C)**, CD8+ T cells **(D)**, macrophages **(E)**, NK cells **(F)**, and other cells **(G)**.

strong positivity in WHO 4, a moderate positivity in WHO 3, and a weak positivity in WHO 2, and negative in brain tissue from normal individuals. We further found that DDOST was overexpressed in patients with worse molecular features, which predicted poor survival, such as IDH wild type, 1p19q non-codel, and MGMT un-methylated. Patients with high expression of DDOST had worse prognosis than those with lower levels in gliomas. These results were in accordance with a previous study by Shapanis et al., who discovered that in cervical and oropharyngeal cancer, an increased expression of ANXA5 and DDOST was associated with a shorter time to metastasis and decreased survival (6). It suggests that DDOST may be a candidate oncogene and an independent factor for predicting poor prognosis in gliomas.

It is widely known that tumors consist not only of cancerous cells but also of non-cancerous cells containing a significantly altered surrounding stroma including immune cells, endothelial cells, fibroblasts, and molecules produced and released by them,

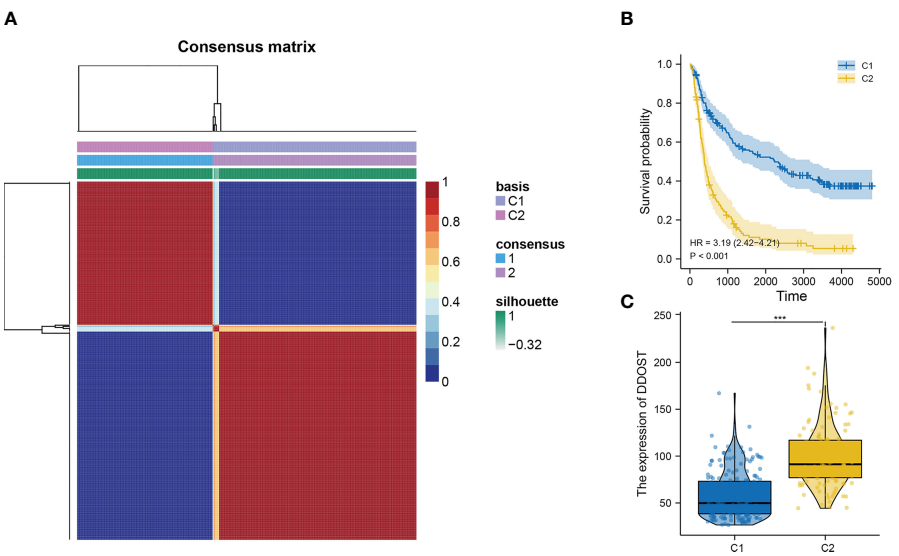


FIGURE 10 | The relationship between DDOST and tumor immune infiltration was analyzed by NMF cluster analysis of glioma patients with immune-related genes. **(A)** According to the cophenetic value and the expression of immune-related genes, glioma patients were divided into two clusters, C1 and C2. **(B)** The survival curve was used to analyze the survival difference between C1 and C2 groups. **(C)** The expression differences of DDOST in C1 and C2 clusters were compared. ***$p < 0.001$.

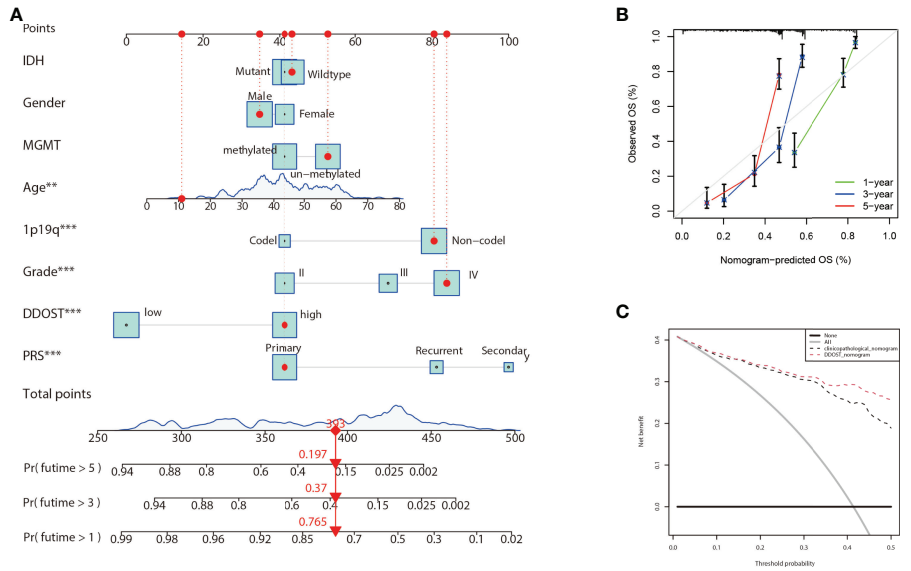


FIGURE 11 | Nomogram for the prediction of prognostic probabilities in the CGGA dataset. **(A)** The nomogram for the prediction of OS was developed using the CGGA dataset. **(B)** The calibration plots for predicting 1-, 3-, and 5-year survival. **(C)** Decision curve analysis for the DDOST nomogram and the clinicopathological nomogram to estimate the OS.

which is called the TME. Evidence shows that the tumor cells could interact with TME by releasing extracellular signals, promoting tumor angiogenesis, and inducing immunity tolerance to influence tumorigenesis, tumor progression, therapeutic response, and clinical outcome (24). Now, TME has become a therapeutic target in cancers. For example, anti-VEGF and EGFR agents could induce tumor vascular normalization, and immunotherapy drugs including anti-PD-1 (programmed death-1) and anti-PD-L1 could remodel the immuno-microenvironment by regulating the immune cells to treat tumors (25).

Gliomas are a group of immunosuppressive tumors with a complicated TME. Although immunotherapy has achieved remarkable survival benefits in multiple cancers, due to the blood–brain barrier (BBB), the effectiveness of target and immunotherapy drugs is dramatically reduced in gliomas. Furthermore, gliomas have a high heterogeneity, and show differences in grade of malignancy and molecule characteristics. Researching the molecular biomarkers and unique immunological status of gliomas has a profound effect on cancer diagnosis and therapeutic strategies. New studies have found that some biomarkers, such as ITGA-5 and Tenascin-C, could predict glioma prognosis, and are closely related to the immunosuppressive microenvironment, particularly for the clinical application of immunotherapy in glioma (26, 27). New studies also reported that DDOST could interact with the microenvironment of gliomas and was negatively related to tumor-infiltrating B cells and CD4+ T cells and positively related to CAFs and tumor-associated macrophages (TAMs). B lymphocytes form the major immune cell population in the TME and contribute to tumor progression (28). In recent studies, however, it has been found that the density of B and Th1 cells was closely correlated to improving patient survival (29). Our data showed that higher levels of DDOST expression were associated with fewer tumor-infiltrating B cells and CD4+ T cells in glioma, which meant that less tumor-infiltrating B cells were unable to activate a powerful antitumor immune response and induce immune escape, which, in turn, led to a poorer prognosis. CD4+ T cells play an anti-tumor role in the TME. Our results showed that DDOST expression was inversely correlated with CD4+ T cells and further illustrated that DDOST may contribute to the immunosuppressive microenvironment of gliomas, which affected the effective of immunotherapy drugs. Many studies have demonstrated that CAFs are not individual cells that surround tumors, but rather interact with cancer cells to promote tumor growth and survival and maintain their malignant propensity (30). TAMs are a key component of the cancer microenvironment, and contribute to tumor growth and progression (31). Based on our immunoinfiltration analysis, we concluded that DDOST was associated with an immunosuppressive microenvironment in gliomas. Lastly, based on the above data, we constructed a nomogram incorporating DNA repair genes, age, PRS, grade, IDH, 1p19q, and MGMT status to predict the probability of 1-, 3-, and 5-year OS. One-, 3-, and 5-year OS could be predicted by calculating the sum of the scores for each glioma patient. For example, a patient with an overall score of 393 would have a 5-year survival of 19.7%, a 3-year survival of 37%, and a 1-year survival of 76.5%. The DCA curve showed that DDOST combined with other clinical traits can bring benefits to predict the prognosis of glioma patients.

These results were mainly based on the data of bioinformatics; therefore, large-sample immunohistochemistry data, Western blot, and RT-PCR of glioma tissues are urgently

needed to confirm the above results. We will also further detect and confirm these results in *in vivo* and *in vitro* experiments in glioma.

In the present study, we found that DDOST could be used as an accurate predictor for patients with gliomas. The immune infiltration analysis showed that DDOST was closely related to the TME and might mediate the suppressive immune microenvironment of glioma. Finally, DDOST and clinical features were used to construct a nomogram that could accurately assess the prognosis of patients. Our results preliminarily suggest that DDOST could serve as a potential target for the treatment of glioma patients, and an independent factor for predicting prognosis in gliomas, but further detailed mechanisms need to be explored in future research.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YY and GW were responsible for the overall design of this study. XC and RZ were responsible for manuscript writing as well as data analysis. JP mainly performed immunohistochemistry and scoring. TY and XW were responsible for data collection and R code writing. CG assisted GW with manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Special pharmaceutical research project of Hebei Pharmaceutical Association Hospital (2HC2022012).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.917014/full#supplementary-material

Supplementary Figure 1 | Expression difference of DDOST between different clinical characters in patients with glioma in the CGGA 693 cohort. The expression of DDOST in different age **(A)**, gender **(B)**, PRS type **(C)**, grade **(D)**, IDH **(E)**, 1p19q **(F)**, and MGMT status **(G)**.

Supplementary Figure 2 | Prediction of outcome of the DDOST in stratified patients in the CGGA 693 dataset. Survival curve was used to analyze OS in the low- and high-DDOST groups in CGGA 325 set(A). Survival analysis of the signature in patients stratified by age **(B, C)**, gender **(D, E)**, grade **(F–H)**, IDH **(I, J)**, 1p19q status **(K, L)**, and MGMT promoter **(M, N)**.

Supplementary Figure 3 | The role of DDOST in glioma was analyzed by GSEA. GO**(A)** and hallmark**(B)** gene sets were performed to explore the mechanism of DDOST in glioma.

Supplementary Figure 4 | Waterfall diagram shows the relationship between DDOST expression and gene mutation in LGG **(A)** and GBM **(B)**.

Supplementary Table 1 | Clinicopathological characteristics of glioma patients from the GEO, TCGA, CGGA database and tissue microarray.

Supplementary Table 2 | The differential genes between high- and low-DDOST groups by CGGA 325 cohort.

Supplementary Table 3 | Immune-related genes.

REFERENCES

1. Ostrom QT, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014-2018. *Neuro Oncol* (2021) 23(12 Suppl 2):iii1–1iii105. doi: 10.1093/neuonc/noab200
2. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary. *Neuro Oncol* (2021) 23(8):1231–51. doi: 10.1093/neuonc/noab106
3. Yang Z, Ling F, Ruan S, Hu J, Tang M, Sun X, et al. Clinical and Prognostic Implications of 1p/19q, IDH, BRAF, MGMT Promoter, and TERT Promoter Alterations, and Expression of Ki-67 and P53 in Human Gliomas. *Cancer Manag Res* (2021) 13:8755–65. doi: 10.2147/CMAR.S336213
4. Schjoldager KT, Narimatsu Y, Joshi HJ, Clausen H. Global View of Human Protein Glycosylation Pathways and Functions. *Nat Rev Mol Cell Biol* (2020) 21(12):729–49. doi: 10.1038/s41580-020-00294-x
5. Oliveira-Ferrer L, Legler K, Milde-Langosch K. Role of Protein Glycosylation in Cancer Metastasis. *Semin Cancer Biol* (2017) 44:141–52. doi: 10.1016/j.semcancer.2017.03.002
6. Shapanis A, Lai C, Smith S, Coltart G, Sommerlad M, Schofield J, et al. Identification of Proteins Associated With Development of Metastasis From Cutaneous Squamous Cell Carcinomas (cSCCs) *via* Proteomic Analysis of Primary cSCCs. *Br J Dermatol* (2021) 184(4):709–21. doi: 10.1111/bjd.19485
7. Zhu C, Xiao H, Jiang X, Tong R, Guan J. Prognostic Biomarker DDOST and Its Correlation With Immune Infiltrates in Hepatocellular Carcinoma. *Front Genet* (2021) 12:819520. doi: 10.3389/fgene.2021.819520
8. Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, Pastorino S, et al. Neuronal and Glioma-Derived Stem Cell Factor Induces Angiogenesis Within the Brain. *Cancer Cell* (2006) 9(4):287–300. doi: 10.1016/j.ccr.2006.03.003
9. Griesinger AM, Birks DK, Donson AM, Amani V, Hoffman LM, Waziri A, et al. Characterization of Distinct Immunophenotypes Across Pediatric Brain Tumor Types. *J Immunol* (2013) 191(9):4880–8. doi: 10.4049/jimmunol.1301966

10. Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, et al. Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource With Functional Genomic Data From Chinese Glioma Patients. *Genomics Proteomics Bioinf* (2021) 19(1):1–12. doi: 10.1016/j.gpb.2020.10.005
11. Wang G, Li H, Pan J, Yan T, Zhou H, Han X, et al. Upregulated Expression of Cancer-Derived Immunoglobulin G Is Associated With Progression in Glioma. *Front Oncol* (2021) 11:758856. doi: 10.3389/fonc.2021.758856
12. Yu K, Hu Y, Wu F, Guo Q, Qian Z, Hu W, et al. Surveying Brain Tumor Heterogeneity by Single-Cell RNA-Sequencing of Multi-Sector Biopsies. *Natl Sci Rev* (2020) 7(8):1306–18. doi: 10.1093/nsr/nwaa099
13. Huang da W, Sherman BT, Lempicki RA. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat Protoc* (2009) 4(1):44–57. doi: 10.1038/nprot.2008.211
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc Natl Acad Sci U S A* (2005) 102(43):15545–50. doi: 10.1073/pnas.0506580102
15. Yeo JG, Wasser M, Kumar P, Pan L, Poh SL, Ally F, et al. The Extended Polydimensional Immunome Characterization (EPIC) Web-Based Reference and Discovery Tool for Cytometry Data. *Nat Biotechnol* (2020) 38(6):679–84. doi: 10.1038/s41587-020-0532-1
16. Chaussabel D, Baldwin N. Democratizing Systems Immunology With Modular Transcriptional Repertoire Analyses. *Nat Rev Immunol* (2014) 14 (4):271–80. doi: 10.1038/nri3642
17. Iasonos A, Schrag D, Raj GV, Panageas KS. How to Build and Interpret a Nomogram for Cancer Prognosis. *J Clin Oncol* (2008) 26(8):1364–70. doi: 10.1200/JCO.2007.12.9791
18. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
19. Eichler J. Protein Glycosylation. *Curr Biol* (2019) 29(7):R229–229R231. doi: 10.1016/j.cub.2019.01.003
20. Mikolajczyk K, Kaczmarek R, Czerwinski M. How Glycosylation Affects Glycosylation: The Role of N-Glycans in Glycosyltransferase Activity. *Glycobiology* (2020) 30(12):941–69. doi: 10.1093/glycob/cwaa041
21. Jones MA, Ng BG, Bhide S, Chin E, Rhodenizer D, He P, et al. DDOST Mutations Identified by Whole-Exome Sequencing are Implicated in Congenital Disorders of Glycosylation. *Am J Hum Genet* (2012) 90(2):363–8. doi: 10.1016/j.ajhg.2011.12.024
22. Hu Y, Mayampurath A, Khan S, Cohen JK, Mechref Y, Volchenboum SL. N-Linked Glycan Profiling in Neuroblastoma Cell Lines. *J Proteome Res* (2015) 14(5):2074–81. doi: 10.1021/pr5011718
23. Lemjabbar-Alaoui H, McKinney A, Yang YW, Tran VM, Phillips JJ. Glycosylation Alterations in Lung and Brain Cancer. *Adv Cancer Res* (2015) 126:305–44. doi: 10.1016/bs.acr.2014.11.007
24. Hinshaw DC, Shevde LA. The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res* (2019) 79(18):4557–66. doi: 10.1158/0008-5472.CAN-18-3962
25. Xiao Y, Yu D. Tumor Microenvironment as a Therapeutic Target in Cancer. *Pharmacol Ther* (2021) 221:107753. doi: 10.1016/j.pharmthera.2020.107753
26. Li S, Zhang N, Liu S, Zhang H, Liu J, Qi Y, et al. ITGA5 Is a Novel Oncogenic Biomarker and Correlates With Tumor Immune Microenvironment in Gliomas. *Front Oncol* (2022) 12:844144. doi: 10.3389/fonc.2022.844144
27. Zhang P, Liu G, Hu J, Chen S, Wang B, Peng P, et al. Tenascin-C can Serve as an Indicator for the Immunosuppressive Microenvironment of Diffuse Low-Grade Gliomas. *Front Immunol* (2022) 13:824586. doi: 10.3389/fimmu.2022.824586
28. Affara NI, Ruffell B, Medler TR, Gunderson AJ, Johansson M, Bornstein S, et al. B Cells Regulate Macrophage Phenotype and Response to Chemotherapy in Squamous Carcinomas. *Cancer Cell* (2014) 25(6):809–21. doi: 10.1016/j.ccr.2014.04.026
29. Lu Y, Zhao Q, Liao JY, Song E, Xia Q, Pan J, et al. Complement Signals Determine Opposite Effects of B Cells in Chemotherapy-Induced Immunity. *Cell* (2020) 180(6):1081–97.e24. doi: 10.1016/j.cell.2020.02.015
30. Chen Y, McAndrews KM, Kalluri R. Clinical and Therapeutic Relevance of Cancer-Associated Fibroblasts. *Nat Rev Clin Oncol* (2021) 18(12):792–804. doi: 10.1038/s41571-021-00546-5
31. Mantovani A, Marchesi F, Malesci A, Laghi L, Allavena P. Tumour-Associated Macrophages as Treatment Targets in Oncology. *Nat Rev Clin Oncol* (2017) 14(7):399–416. doi: 10.1038/nrclinonc.2016.217

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chang, Pan, Zhao, Yan, Wang, Guo, Yang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 27 June 2022
doi: 10.3389/fonc.2022.878827

Check for updates

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Maria Ramirez,
University of Texas MD Anderson Cancer Center, United States
Caterina AM La Porta,
University of Milan, Italy

***Correspondence:**
Véronique M. Braud
braud@ipmc.cnrs.fr
Thierry Goudon
thierry.goudon@inria.fr

†Present address:
Sokchea Khou,
Department of Cell, Developmental & Cancer Biology, Oregon, Health & Science University, Portland, OR, United States

‡These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Oncology

Received: 18 February 2022
Accepted: 02 May 2022
Published: 27 June 2022

Citation:
Atsou K, Khou S, Anjuère F, Braud VM and Goudon T (2022) Analysis of the Equilibrium Phase in Immune-Controlled Tumors Provides Hints for Designing Better Strategies for Cancer Treatment. Front. Oncol. 12:878827. doi: 10.3389/fonc.2022.878827

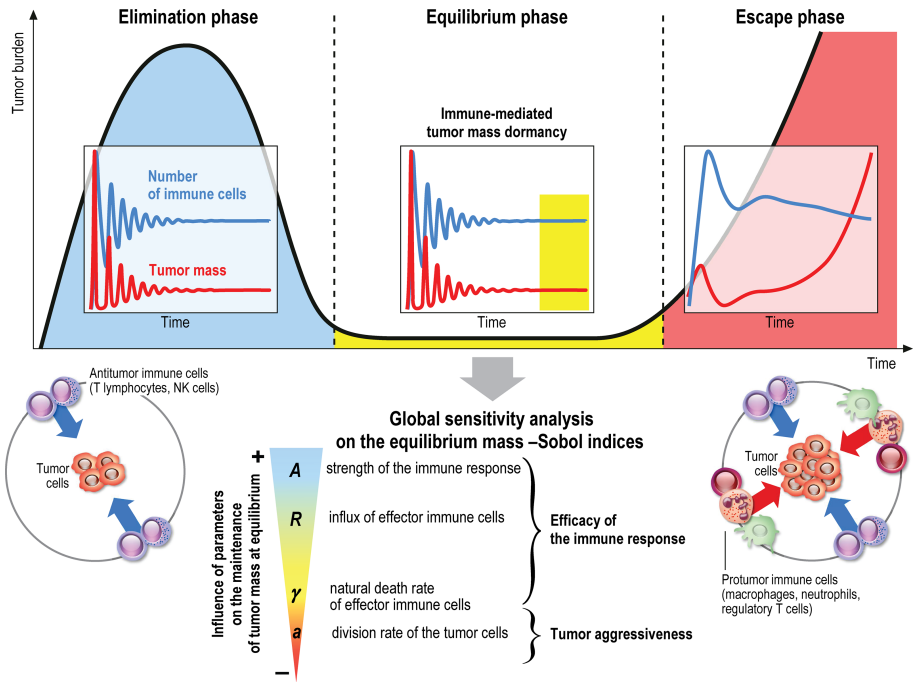
Analysis of the Equilibrium Phase in Immune-Controlled Tumors Provides Hints for Designing Better Strategies for Cancer Treatment

Kevin Atsou[1], Sokchea Khou[2†], Fabienne Anjuère[1], Véronique M. Braud[2‡] and Thierry Goudon[1*‡]*

[1] Université Côte d'Azur, Inria, CNRS, LJAD, Nice, France, [2] Université Côte d'Azur, CNRS, Institut de Pharmacologie Moléculaire et Cellulaire UMR 7275, Valbonne, France

When it comes to improving cancer therapies, one challenge is to identify key biological parameters that prevent immune escape and maintain an equilibrium state characterized by a stable subclinical tumor mass, controlled by the immune cells. Based on a space and size structured partial differential equation model, we developed numerical methods that allow us to predict the shape of the equilibrium at low cost, without running simulations of the initial-boundary value problem. In turn, the computation of the equilibrium state allowed us to apply global sensitivity analysis methods that assess which and how parameters influence the residual tumor mass. This analysis reveals that the elimination rate of tumor cells by immune cells far exceeds the influence of the other parameters on the equilibrium size of the tumor. Moreover, combining parameters that sustain and strengthen the antitumor immune response also proves more efficient at maintaining the tumor in a long-lasting equilibrium state. Applied to the biological parameters that define each type of cancer, such numerical investigations can provide hints for the design and optimization of cancer treatments.

Keywords: cancer, mathematical oncology, equilibrium phase, immunotherapy, drug response



GRAPHICAL ABSTRACT |

1 INTRODUCTION

The immune system plays a major role in the control of tumor growth. This has led to the concept of immune surveillance and cancer immunoediting composed of three phases (1–3): the elimination, when tumors are rapidly eradicated by the immune system, the equilibrium, a latency period when tumors can survive but remain on a controlled state, and the escape, the final outgrowth of tumors that have outstripped immunological restraints. In this later phase, immune suppression is prevailing and immune cells are also subverted to promote tumor growth. Numerous cancer immunotherapy strategies have been designed and assessed to counteract immune suppression and restore effective and durable elimination of tumors (4–8). They show improved efficacy over conventional anticancer treatments but only a minority of patients respond. The challenge to face now is to identify key biological parameters which will convert a fatal outcome into a chronic, manageable state, the durable maintenance of cancer in a viable equilibrium phase controlled by immunity. Reaching such immune-mediated tumor mass dormancy is indeed the first key step for successful control of tumor growth and a goal for immunotherapy (9). The equilibrium state is however difficult to apprehend experimentally because the tumor mass at equilibrium is below detectable limits (3). Mathematical modeling of the tumor-immune system interactions offers useful information about the features of the equilibrium phase during primary tumor development, and such tools could be used to guide the design of optimal anticancer therapies (10–13).

We previously (10) introduced a specific multiscale mathematical model based on partial differential equations (PDE), intended to describe the earliest stages of tumor-immune system interactions. We conjecture that the space heterogeneities of the distribution of active and resting immune cells, which are subjected to several interaction mechanisms with the tumor cells, plays a critical role in the efficiency of the immune response, and the ability in reaching the equilibrium phase. This, in turn, motivates the appeal to PDEs descriptions and can complete the already established modeling based on ordinary differential systems, on which there exists a wide literature, see for instance (11, 14–19) Extension to the PDE framework has permitted to bring out the role of space organisation (20–23). The reader can find further details and references about the mathematical modeling of tumor-immune system interactions, based on different viewpoints and addressing several issues of the efficacy of the immune response, in the reviews (24–29). The original model developed in (10) thus accounts for both the growth of the tumor, by natural cell growth and cell divisions, and the displacement of the immune cells towards the tumor, by means of activation processes and chemotaxis effects. The most notable finding from (10) was that an equilibrium state, with residual tumor and active immune cells, can be observed. Moreover, mathematical analysis provides a basis for the explanation of the formation of the equilibrium. How the biological parameters shape this equilibrium is the main question investigated in the present article. Indeed, the equilibrium can be mathematically interpreted by means of an eigenproblem coupled to a stationary

diffusion equation with constraint. This observation permits us to develop an efficient numerical strategy to determine *a priori* the shape of the equilibrium — namely, the size distribution of the tumor cells and the residual tumor mass — for a given set of biological tumor and immune cell parameters. Consequently, the equilibrium state can be computed at low numerical cost since we can avoid the resolution of the evolution problem on a long time range. The use of this simple and fast algorithm allows us to address the question of the sensitivity of the residual mass to the parameters and to discuss the impact of treatments. This information can be decisive to design clinical studies and choose therapeutic strategies that will revert to an equilibrium phase. Our work therefore provides hints for cancer treatment management.

1.1 Quick Guide to Equations: A Coupled PDE Model for Tumor-Immune System Interactions

The modeling approach imposes to select a few phenomena, considered as the leading effects for the situation under consideration; other effects are just roughly described by tuning some parameters or are simply disregarded. Choices for designing the mathematical model are also dictated by the difficulty in attributing numerical values to the parameters of the equations, due to a lack of experimental measurements: the poor knowledge of driving quantities leads to keep a description as simple as possible, with a reduced number of unknown parameters. The principles of the modeling adopted in (10), summarized by **Figure 1**, led to couple an evolution equation for the size-distribution of the tumor cells, and a convection-diffusion equation for the activated immune cells. The two-way coupling arises from the death term induced by the action of the immune cells on the tumor cells, and by the activation and the attraction of immune cells towards the tumor, which are determined by the total mass of the tumor. The model is intended to describe the earliest stages of the tumor formation, when the size of the tumor is relatively small. The tumor is located at the center of a domain Ω (there is no displacement of the tumor). The model distinguishes two distinct and independent length scales: the size of the tumor cells, described by the variable $z \geq 0$, is considered as "infinitely small" compared to the scale of displacement of the immune cells, described by the space variable $x \in \Omega$.

The unknowns are

- The size density of tumor cells $(t, z) \mapsto n(t, z)$ so that the integral $\int_a^b z \quad n(t, z) \quad \mathrm{d}z$ gives the volume of the tumor occupied at time t by cells having their size z in the interval (a, b);
- The concentration of activated immune cells which are fighting against the tumor $(t, x) \mapsto c(t, x)$;
- The concentration of chemical signal that attracts the immune cells towards the tumor microenvironment $(t, x) \mapsto \phi(t, x)$.

The specific biological assumptions made to construct the model are fully described in (10). **Figure 2** offers an overview of the interaction mechanisms embodied in the equations and of the role of the parameters of the model.

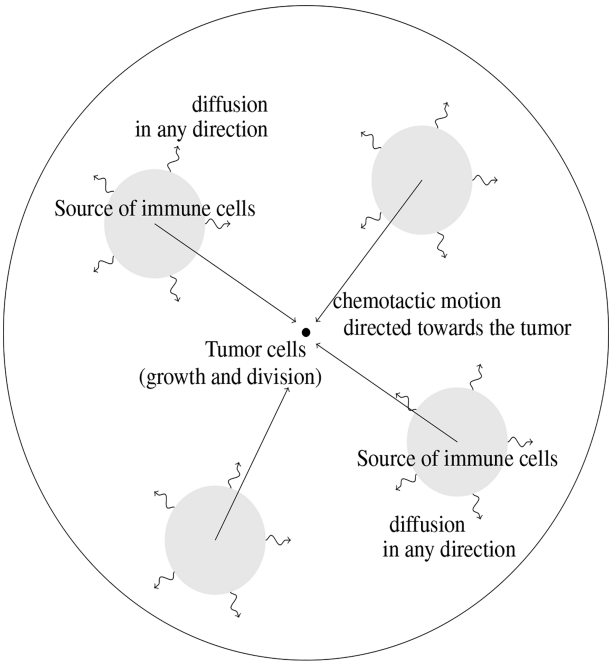


FIGURE 1 | Schematic view of the geometry of the mathematical model. The tumor cells are located at the center of the domain where they are subjected to growth and division mechanisms. Immune cells are activated from baths of resting cells; their motion is driven by diffusion combined to a convection field, due to chemotactic mechanisms and directed towards the tumor.

Immune cells, once activated from a bath of resting cells, are subjected to natural diffusion and to a chemotactic drift, induced by the presence of the tumor. The strength of this drift, as well as the activation of immune cells, directly depends on the total mass of the tumor, proportional to the quantity

$$\mu_1(t) = \int_0^\infty z n(t,z)\, \mathrm{d}z\,.$$

The immune system-tumor competition is described by the following system of PDEs

$$\partial_t\, n + \partial_z(Vn) = Q(n) - m(n,c), \tag{1a}$$

$$\partial_t\, c + \nabla_x \cdot (c\chi\, \nabla_x\, \phi - D\, \nabla_x\, c) = \mu_1 R - \gamma c, \tag{1b}$$

$$-\mathcal{K}\Delta_x \phi = \mu_1 \left(\sigma(x) - \frac{1}{|\Omega|} \int_\Omega \sigma(y)\, \mathrm{d}y \right), \tag{1c}$$

$$n(t,0) = 0,\ \ c\,|_{\partial\Omega} = 0,\ \ \mathcal{K}\, \nabla_x\, \phi \cdot \nu\,|_{\partial\Omega} = 0, \tag{1d}$$

$$n(t=0,z) = n_0(z),\ \ c(t=0,x) = c_0(x)\,. \tag{1e}$$

The features of the growth-division dynamics for the tumor cells (1a) are embodied into the (possibly size-dependent)

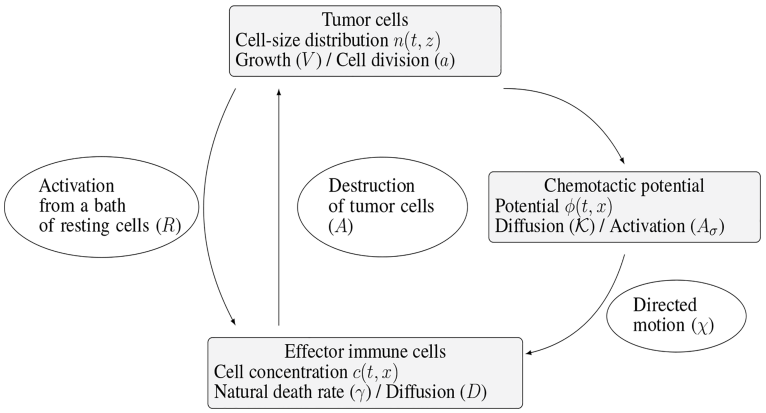


FIGURE 2 | Schematic view of the interaction mechanisms described by the system (1a)-(1e).

growth rate $z \mapsto V(z) \geq 0$ and the cell division operator $Q(n)$. We refer the reader to (30–37) for further details on this evolution equation (with $m(n, c) = 0$) for cell growth and division, and its application to cancer modeling. What is crucial for modeling purposes is the principle that cell-division does not change the total mass: the operator Q satisfies $\int_0^\infty zQ(n)\,\mathrm{d}z = 0$. However, the total number of cells in the tumor increases since $\int_0^\infty Q(n)\,\mathrm{d}z \geq 0$ (we refer the reader to (10) and Appendix A for further details). In what follows, we restrict to the mere symmetric binary division operator

$$Q(n)(t,z) = a(z)(4n(t,2z) - n(t,z)), \tag{2}$$

with $z \mapsto a(z) \geq 0$ the division rate. It simply describes the situation where cells are cut into two cells having half the size of the original cell. Further relevant examples of division operators can be found in (32) (see Appendix A). The specific case where the division rate a in (2) is a positive constant makes the model simpler, and is often used. It is however likely relevant to incorporate more complex behaviors through the size-dependence; for instance divisions can be prohibited below a certain size threshold. Similarly, it can be convenient to assume that the growth rate V is a positive constant, but more intricate laws can take into account some important phenomena. For instance, logistic or Gompertz law can incorporate size limitation effects, and roughly describe difficulties in accessing nutrients or necrotic effects (38–40); a detailed study of growth laws can be found in (41). As mentioned above, though, using such complex laws, also raises the issue of determining more parameters. The boundary condition for n in (1d) means that no tumor cells are created with size 0.

Despite the fact that there exists several types of immune cells – at least T-cells and NK cells – fighting against the tumor, they are all described here through the single concentration c. It also means that coefficients of the equation – the death rate $\gamma > 0$, the chemotactic strength $\chi > 0$, and the diffusion coefficient D – correspond to an averaged behavior of all these cells. By the way, working with a constant diffusion coefficient $D > 0$ is again a simplification, neglecting the architecture of the tumor environment, which might induce directional effects. The effector immune cells that effectively fight against the tumor, are activated from a "reservoir" of resting cells, described in the right hand side of (1b) by $(t,x) \mapsto R(t,x)$. This given function, possibly time and space dependent, stands for the space distribution of the influx rate of activated effector immune cells. It takes into account the sources of resting immune cells that can be activated in the tumor microenvironment or in the draining lymph nodes into cells fighting the tumor. At early stages of tumor growth, the rate of the activation process is supposed to be directly proportional to the tumor mass μ_1. Again, more complex activation law, for instance based on Michaelis-Menten kinetics can incorporate relevant limitation mechanisms. The Dirichlet boundary condition for c in (1d) means that the immune cells far from the tumor are non-activated. Immune cells are directed towards the tumor by a chemo-attractive potential ϕ, induced by the presence of the tumor cells. Through (1c), the strength of the signal is proportional to the total mass of the tumor, and it is shaped by a form function $x \mapsto \sigma(x)$ which will be a function peaked at the tumor location. The potential is thus defined by the diffusion equation (1c), that involves a positive coefficient $K > 0$ (that could be matrix valued), and the Neumann boundary condition in (1d), where ν stands for the unit outward normal vector on $\partial\Omega$. Finally, the activated immune cells are able to destroy tumor cells, as described by the death term in (1a)

$$m(c,n)(t,z) = \underbrace{\int_\Omega \delta(y)c(t,y)\mathrm{d}y}_{:=\mu_c(t)} \times n(t,z), \tag{3}$$

where $\delta \geq 0$ is another form function, also peaked in the vicinity of the tumor. For the numerical experiments, we shall work with the Gaussian profiles

$$\delta(x) = \frac{A}{\theta\sqrt{2\pi}} \exp\left(-\frac{|x|^2}{2\theta^2}\right),\ \sigma(x) = \frac{A_\sigma}{\theta_\sigma\sqrt{2\pi}} \exp\left(-\frac{|x|^2}{2\theta_\sigma^2}\right), \quad (4)$$

where the positive parameters A,A_σ and θ,θ_σ can be used to tune the amplitude and spreading of these functions, and thus the strength and radius of influence of the related phenomena. We refer the reader to (10) for further details and comments about the model. Note that this model neglects the possible additional protumoral effects that can take place and are crucial to swing to the escape phase. Such protumor effects can have different forms: they can directly enhance the tumor growth, and make antitumor immune cells exhausted, a state where they are hyporesponsive and cannot kill the tumor, see (42) on these issues. Remarkably, the model (1a)-(1e) is able to reproduce equilibrium phases where the tumor growth is controlled by the immune response.

2 MATERIALS AND METHODS

2.1 Development of Numerical Methods Predicting Parameters of the Equilibrium in Immune-Controlled Tumors

According to (2, 3, 9), the equilibrium phase corresponds to a long-lasting period of immune-mediated latency, also known as tumor mass dormancy, prior to the emergence of clinically detectable malignant disease, with a residual tumor which has not be fully destroyed by the immune system, maintained under the control of immunity. The simulations of the initial-boundary value problem (1a)-(1e) performed in (10) revealed that such a behavior can be reproduced by the model. Here, we wish to study the features of the equilibrium phase in immune-controlled tumors and, in particular, we want to predict, for given biological parameters (see Section 2.2 below), the total mass of the residual tumor and its size distribution. To this end, we developed specific numerical procedures based on the mathematical interpretation of the equilibrium.

2.1.1 Equilibrium States

The definition of the equilibrium relies on the following arguments. When disregarding the immune response, the cell-division equation

$$\partial_t n + \partial_z(Vn) = Q(n)\,. \quad (5)$$

admits a positive eigenstate, which drives the large time behavior of the solution. To be more specific, there exists λ>0 and a non negative function $z \geq 0 \mapsto \bar{N}(z)$ satisfying

$$\begin{cases} \partial_z(V\bar{N}) - Q(\bar{N}) + \lambda\bar{N} = 0 \quad \text{for} \quad z \geq 0 \\ \bar{N}(0) = 0,\ \bar{N}(z) > 0 \quad \text{for z} > 0,\ \int_0^{+\infty} \bar{N}(z) \quad \mathrm{d}z = 1. \end{cases} \quad (6)$$

The existence-uniqueness of the eigenpair $(\lambda, \bar{N})$ can be found in (32, 34). Furthermore, when the tumor does not interact with the immune system, the large time behavior is precisely driven by the eigenpair: the solution of (5) behaves like

$$n(t,z) \underset{t\to\infty}{\sim} \mu_0 e^{\lambda t}\bar{N}(z)$$

where μ_0 >0 is a constant determined by the initial condition, see (33, 34). Consequently, in the immune-free case, the tumor population grows exponentially fast, with a rate λ>0 , and, as time becomes large, its size repartition obeys a certain profile $\bar{N}$. In the specific case where V is constant and Q is the binary division operator (2), with a constant division rate a, we simply have λ=a and the profile $\bar{N}$ is explicitly known (43, 44). However, for general growth rates and division kernels the solution should be determined by numerical approximations; we are going to detail a numerical procedure to effectively compute the pair $(\lambda, \bar{N})$.

Coming back to the coupled model (1a)-(1e), we infer that the equilibrium phase corresponds to the situation where the death rate – the integral of the immune cells concentration with weight δ, denoted as $\bar{\mu}_c$ in (3) – precisely counterbalances the natural exponential growth of the tumor cell population. In other words, at equilibrium we expect that

- The size distribution of tumor cells is proportional to the eigenstate $\mu_0\bar{N}(z)$. The proportionnality factor is related to the total mass by the relation $\mu_1 = \mu_0 \int_0^\infty z\bar{N}(z)\,\mathrm{d}z$.
- The concentration of immune cells is defined by the stationary equation

$$\gamma C - \nabla_x \cdot (D\,\nabla_x C) + \mu_1\,\nabla_x \cdot(\chi C\,\nabla_x \Phi) = \mu_1 R,\ C\,|_{\partial\Omega=0} = 0, \quad (7)$$

- Where Φ is the solution of

$$-\mathcal{K}\Delta_x\Phi = \sigma - \frac{1}{|\Omega|}\int_\Omega \sigma(y) \quad \mathrm{d}y,$$

- Endowed with the homogeneous Neumann boundary condition, together with the constraint

$$\int_\Omega \delta(x)C(x) \quad \mathrm{d}x = \lambda\,. \quad (12)$$

This can be interpreted as an implicit definition of the total mass μ_1 to be the value such that the solution of the boundary value problem (7) satisfies (8): in other words, it defines implicitly the mass of the residual tumor μ_1 to be the value such that the solution of the stationary boundary value problem for C defines a death rate that exactly compensates the exponential growth rate of the growth division equation. The existence of an equilibrium state defined in this way is rigorously justified in (10, Theorem 2).

Theorem 2.1. Let $x \mapsto R(\ x\ ) \in L^2\ (\ \Omega\ )$ be a non negative function. If λ>0 is small enough, there exists a unique $\mu_1\ (\ \lambda\ )$>0 such that the solution $C_{\mu_1}\ (\ \lambda\ )$ of the stationary equation (7) satisfies (8).

Theorem 2.1 requires a smallness assumption; for (2) with constant growth rate V and division rate a, this is a smallness assumption on a. Numerical experiments have shown different large time behaviors for the initial-boundary value problem (1a)-(1e) (an example will be presented later on):

- When the source term R is space-homogeneous, the expected behavior seems to be very robust. The immune cell concentration tends to fulfill the constraint $\bar{\mu}_c(t) \sim \lambda$ as time becomes large, and the size repartition of tumor cells tends to the eigenfunction $\bar{N}$. The total mass μ_1 tends to a constant; however the asymptotic value cannot be predicted easily.
- When R has space variations, the asymptotic behavior seems to be much more sensitive to the parameters of the model, in particular to the aggressiveness of the tumor (characterized by the cell division rate a). On short time scale of simulations, we observe alternance of growth and remission phases, and the damping to the equilibrium could be very slow.

These observations bring out the complementary roles of different type of cytotoxic cells (45). The NK cells could be seen as a space-homogenous source of immune cells, immediately available to fight against the tumor, at the early stage of tumor growth. In contrast, T-cells need an efficient priming which occurs in the draining lymph nodes, and their sources is therefore non-homogeneously distributed. Eventually, NK and $CD8^+$ T-cells cooperate to the anti-tumor immune response.

Numerical experiments thus show that the model (1a)–(1e) is able to reproduce, in the long-time range, cancer-persistent equilibrium, but the features of the equilibrium, and its ability to establish, are highly sensitive to the parameters. To discuss this issue further, we focus here on the mass at equilibrium considered as a critical quantity that evaluates the efficacy of the immune response. Indeed, it is known that a tumor gains in malignancy when its mass reaches certain thresholds (45, 46). The smaller the tumor mass at equilibrium, the better the vital prognosis of the patient. In doing so, we do not consider transient states and time necessary for the equilibrium to establish. The interest of the interpretation of the equilibrium by means of an eigenproblem relies on the fact that the equilibrium state can be determined a priori, at least through numerical simulations, without running the initial boundary value problem over long time ranges: given a set of biological parameters it can be obtained by solving the eigenvalue problem for $(\lambda, \bar{N})$ and the constrained stationary drift-diffusion equation for C, see **Figure 3**. In turn, since the equilibrium state can be computed at low numerical cost, a wide range of parameters can be considered and the role of the parameters can be investigated in details. The determination, on numerical grounds, of the equilibrium state relies on a two-step process, as schematised in **Figure 3**. First, we compute the normalized eigenstate of the tumor cell equation, second, we find the tumor mass which

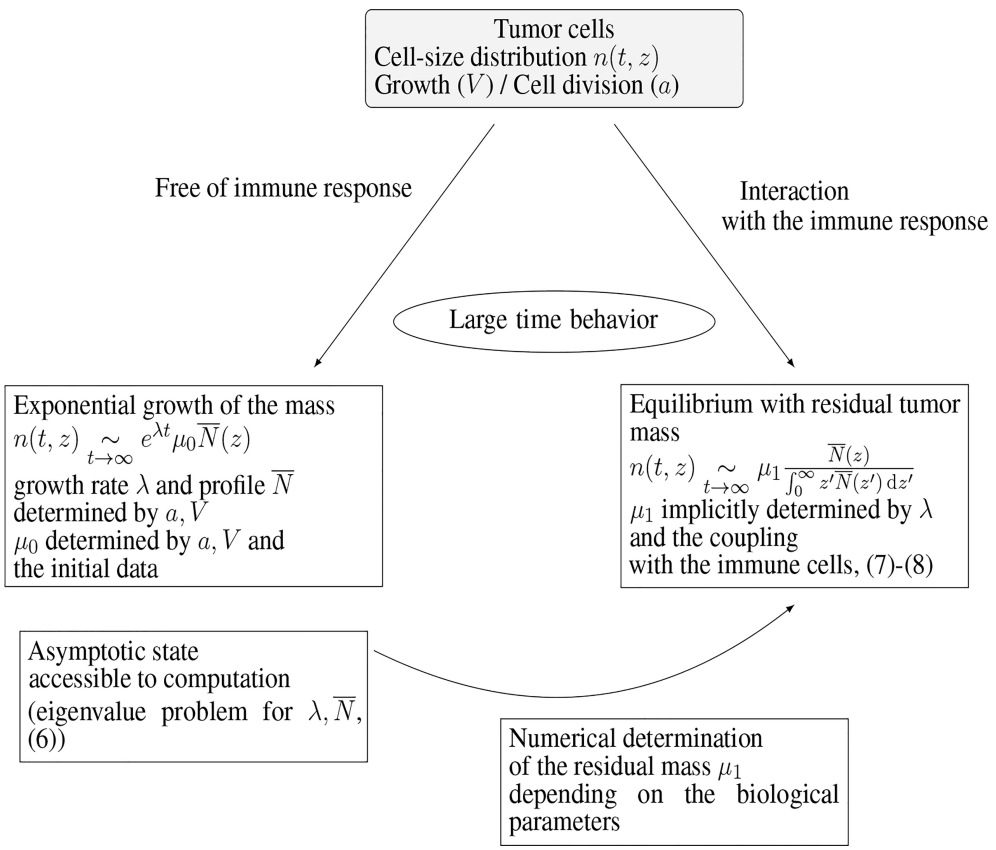


FIGURE 3 | Connection of the equilibrium state with the eigenstate of the growth-division equation, and interpretation of the residual tumor mass.

makes the coupled death rate fit with the eigenvalue. To this end, we have developed a specific numerical approach.

2.1.2 The Eigen-Elements of the Growth-Division Equation

The numerical procedure is fully detailed and analyzed in Appendix A; it is inspired from the spectral analysis of the equation: λ is found as the leading eigenvalue of a conveniently shifted version of the growth-division operator. In practice, we work with a problem where the size variable is both truncated and discretized. Hence, the problem recasts as finding the leading eigenvalue of a shifted version of the underlying matrix, which can be addressed by using the inverse power method [(47), Section 1.2.5]. We refer the reader to (48, 49) for a thorough analysis of the approximation of eigenproblems for differential and integral operators, which provides a rigorous basis to this approach. It is also important to check a priori, based on the analysis of the equation (32), how large the shift should be, and that it remains independent on the numerical parameters. As already mentioned, for some specific division and growth rates, the eigenpair $(\lambda, \bar{N})$ is explicitly known, see (32). We used these formula to validate the ability of the algorithm to find the expected values and profiles.

2.1.3 Computation of the Equilibrium Mass

Having at hand the eigenvalue λ, we go back to the convection-diffusion equation (7) and the constraint (8) that determine implicitly the total mass μ_1 of the residual tumor. For a given value of μ_1, we numerically solve (7) by using a finite volume scheme, see (10, Appendix C). Then, we use the dichotomy algorithm to fit the constraint:

- The chemo-attractive potential Φ is computed once for all.
- Pick two reference values $0 < \mu_a < \mu_b$; the mass we are searching for is expected to belong to the interval (μ_a, μ_b).
- Set $\mu_1 = \frac{\mu_a+\mu_b}{2}$ and compute the associated solution C_{μ_1} of (7) (the subscript emphasizes the dependence with respect to μ_1). Evaluate the discrete version of the quantity $I=\int \delta C_{\mu_1} \, dx-\lambda$
- If $I < 0$, then replace μ_a by μ_1, otherwise replace μ_b by μ_1.
- We stop the algorithm when the relative error $\frac{\mu_b-\mu_a}{\mu_a} < \varepsilon$ is small enough.

It is also possible to design an algorithm based on the Newton method. However, this approach is much more numerically demanding (it requires to solve more convection-diffusion equations) and does not provide better results.

2.2 Identification of Biological Parameters

In order to go beyond the qualitative discussion of (10), the model should be challenged with biological data. The PDE system (1a)-(1e) is governed by the set of parameters collected in **Table 1**. Most parameter values were retrieved from previously published experimental results and we propose an estimation of the remaining parameters *R, a, V* based on the experimental study performed in (59) where the development of chemically-induced cutaneous squamous cell carcinoma (cSCC) is investigated.

Calibrating the parameters of the equations is an issue due to the lack of direct measurements, and the fact that experimental data are obtained at the price of the sacrifice of mice. Consequently, beyond the cost of the experiments, it also means that a time evolution of the quantities of interest is usually not affordable. Therefore, a specific procedure should be developed in order to estimate the parameters from the experimental data points. Since the informations on the parameters are quite poor, we restrict to the case where the coefficients *a, V, R* are constant, which is also a reasonable assumption when dealing with the earliest stages of the tumor development. In order to identify the parameters, we shall use a degraded version of the equations.

Neglecting the immune response, the tumor growth is driven by (5). As explained above, this leads to an exponential growth of the tumor mass, see (32–34, 44). Let $t \mapsto \mu_0(t) = \int_0^\infty n(t,z) \quad dz$, the total number of tumor cells, and $t \mapsto \mu_1(t) = \int_0^\infty zn(t,z) \quad dz$.

Integrating (9) with respect to size variable, with integration by parts, and bearing in mind that the cell division operator is mass preserving, we thus get

$$\frac{d}{dt}\mu_0 = a\mu_0, \ \frac{d}{dt}\mu_1 = V\mu_0 . \tag{9}$$

Next, assuming space homogeneity of the immune cells concentration and neglecting the displacement and the natural death rate of the immune cells, the immune cells concentration is driven by

TABLE 1 | Key model parameters and their biophysical meaning.

Symbol	Description	Value and unit	References
X	chemotactic coefficient	8.64×10^{1}–$8.64\times10^{6} mm^2 \cdot mmol^{-1} \cdot day^{-1}$	(Macrophages) (50)
D	natural space diffusion coef. of the cytotoxic effector cells population	8.64×10^{-5}–$10^{-3} mm^2 \cdot day^{-1}$	(CD8$^+$ T-cells) (23, 51)
R	the normal rate of influx of effector immune cells	$\log(R) \sim \mathcal{N}(\log(2.2\times10^{-6}), 0.84)(\frac{cell_n \cdot mm^{-3}}{cell_n \cdot \mu m^3} \cdot day^{-1})$	estimated
γ	natural death rate of the tumor antigen-specific cytotoxic effector cells	2×10^{-2}–$1\ day^{-1}$	(14, 20, 22, 52)
A	strength of the immune response	$2 - 57.6 \quad cell_n^{-1} \cdot day^{-1}$	(53–56)
K	diffusion coefficient for the attractive potential ϕ	10^{-2}–$1\ mm^2 \cdot day^{-1}$	(23, 57)
A_σ	strength of the chemical signal induced by each tumor cell	$5\cdot10^{-17}$–$0.625\times10^{-16} mmol \cdot^{-1} \mu m^3 \cdot day^{-1}$	(58)
a	division rate of the tumor cells	$\log(a) \sim N(\log(0.12), 0.2)(a \text{ in } day^{-1})$	estimated
V	growth rate of the tumor cells	$\log(V) \sim N(\log(816.33), 0.51)(V \text{ in } \mu m^3 \cdot day^{-1})$	estimated

$$\frac{d}{dt}c = R\mu_1 \tag{10}$$

Based on this simplified dynamics, reduced to (9)-(10), we used the Nonlinear Mixed Effects Modeling (NMEM) in order to estimate the parameters *a, V, R* from the experimental data. Let N denote the number of mice within the population and $Y_i^{(k)} = \{y_{i1}^{(k)}, \cdots, y_{in^i}^{(k)}\}$ the vector of longitudinal measurements for the ith mouse: $y_{ij}^{(k)}$ is a typical observation of the mouse i for a given measurement type $k \in \{0,1,2\}$ (with (0, 1, 2) referring to (μ_0 ,μ_1 , c) respectively) at time t_{ij}^k for $i \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, N\}$. We suppose that the statistics of the measurements obeys, for $k \in \{0, 1, 2\}$, $j \in \{1, \ldots n_i^k\}$, $i \in \{1, \ldots, N\}$

$$y_{ij}^{(k)} = f^{(k)}\left(t_{ij}^k; \theta_i^k\right) + e_{ij}^{(k)}, \tag{11}$$

where $f^{(k)}(t_{ij}^k; \theta_i^k)$ is the evaluation of the model at time t_{ij}^k, $\theta_i^k \in \mathbb{R}^p$ is the vector of the parameters describing the individual i and $e_{ij}^{(k)}$ the residual error model. The inter-individual variability is described by the combination of fixed effects , which, by definition, are constant within the population and along time, and random effects η_i^k which explain the inter-individual variability among the mice. The positivity of the parameters is ensured by assuming that the individual parameters follow a log-normal distribution. In other words, the random effects are normally distributed with mean zero and a variance-covariance matrix W. For instance W=diag(ω^0 ,ω^1 ,ω^2) where the ω^k's stand for the variance of the parameters *a, V, R*. Therefore, we have

$$\log\ \theta_i^k = \log\ (\theta_{pop}^k) + \eta_i^k,\ \eta_i^k \sim \mathcal{N}\left(0, \omega^k\right) \tag{12}$$

for $k \in \{0,1,2\}$. The error model is assumed to be proportional to the model evaluation and is defined as follows:

$$e_{ij}^{(k)} = \left(b^{(k)} f^{(k)}\left(t_{ij}^k; \theta_i^k\right)\right)\epsilon_{ij} \tag{13}$$

Where $\epsilon_{ij} \sim N(0,1)$ represents the statistical model residual errors and $b^{(k)}$ is the proportionality factor measuring the relative amplitude of the errors.

2.2.1 Estimation of the Model Parameters

According to the experimental procedure in (59), 5×10^5 mSCC38 were injected to each mouse at time $t_0 = 0$. Therefore we fixed the initial number of tumor cells to $\mu_0(0) = 5\times 10^5$ *cells*. Assuming that each tumor cell is spherically shaped with a radius 15 μ m, we set $\mu_1(0) = 7.1 mm^3$. The initial concentration of immune cells is fixed to $c_0 = 0$: we suppose that initially there is no effector immune cells (or at least it means that the initial concentration of activated immune cells is negligible compared to the concentration of resting cells). Some data points were censored due to the sacrifice of the individual for flow cytometry cell counting. The censored data points have been handled by Limit Of Quantification (LOQ) censoring (60). Let I_{ij}^k be the finite or infinite censoring interval for mouse i, measurement k and time t_{ij}^k and

$$\mathbb{P}(y_{ij}^{(k)} \in I_{ij}^k \mid \theta_i^k) = \int_{I_{ij}^k} p_{y_{ij}^{(k)} \mid \theta_i^k}(x \mid \theta_i^k)\quad \mathrm{d}x,$$

where $p_{y_{ij}^{(k)} \mid \theta_i^k}$ is the conditional distribution of $y_{ij}^{(k)}$ given θ_i^k. Let us collect in a vector α=(a_{pop},V_{pop}, R_{pop}, ω_a, ω_V, ω_R, b_a, b_V, b_R) the parameters of the model; they are estimated by maximizing the observed likelihood function

$$\begin{aligned}\mathcal{L}(\alpha, y) = \prod_{k=0}^{2}\prod_{i=1}^{N}\prod_{j=1}^{n_i^k}\int p\left(y_{ij}^{(k)} \mid \theta_i^k\right)^{1_{\left\{y_{ij}^{(k)} \in I_{ij}^k\right\}}} \\ \times \mathbb{P}\left(y_{ij}^{(k)} \in I_{ij}^k \mid \theta_i^k\right)^{1_{\left\{y_{ij}^{(k)} \in I_{ij}^k\right\}}} p\left(\theta_i^k; \alpha\right)\quad \mathrm{d}\theta_i^k .\end{aligned} \tag{14}$$

To this end, we used the Stochastic Approximation of the Expectation Maximization algorithm (SAEM) implemented in the MONOLIX R API (61). Furthermore, the individual parameter estimators $\hat{\theta}_i^k$ are computed in MONOLIX (61) by means of the Empirical Bayes Estimate (EBE) of θ_i^k which corresponds to the mode of the conditional distribution $p(\theta_i^k \mid y_i^k; \hat{\alpha})$ (where $\hat{\alpha}$ corresponds to estimated parameters).

A preliminary estimation procedure indicates a significant correlation between the parameters a and R (t-test p-value 2.6×10^{-6}) . Hence, introducing this correlation into the variance covariance matrix of the random effects by setting covar (a,R)=$\rho_{aR}\ \omega_a\ \omega_R$, where ρ_{aR} represents the correlation coefficient between a and R, enhances the goodness of fit. The estimated value of ρ_{aR} is 0.8 with a relative standard error of 13%. The parameters in a were estimated with reasonable standard errors (computed using the stochastic approximation) and relative standard errors (max (*R.S.E*) = 30.6 and min (*R.S.E.*) = 3) which indicate that the model parameters are identifiable. The ShapiroWilk test reinforces the normality hypotheses on the random effects $\eta_i^{(k)}$ (the p-values for η_a ,η_V and η_R are respectively 0.83, 0.61, 0.2). Pictures indicating the fits are provided in **Figure 4**, and detailed parameter estimates are given in **Table 2**.

2.3 Materials

2.3.1 Mice

FVB/N wild-type (WT) mice (Charles River Laboratories, St Germain Nuelles, France) were bred and housed in specific-pathogen-free conditions. Experiments were performed using 6-7 week-old female FVB/N, in compliance with institutional guidelines and have been approved by the regional committee for animal experimentation (reference MESR 2016112515599520; CIEPAL, Nice Côte d'Azur, France).

2.3.2 *In Vivo* Tumor Growth

mSCC38 tumor cell line was established from DMBA/PMA induced sSCCs and maintained in DMEM (Gibco-ThermoFisher Scientific, Courtaboeuf, France) supplemented with 10% heat-inactivated fetal bovine serum (FBS) (GE Healthcare, Chicago, Illinois, USA) penicillin (100 *U/ml*) and

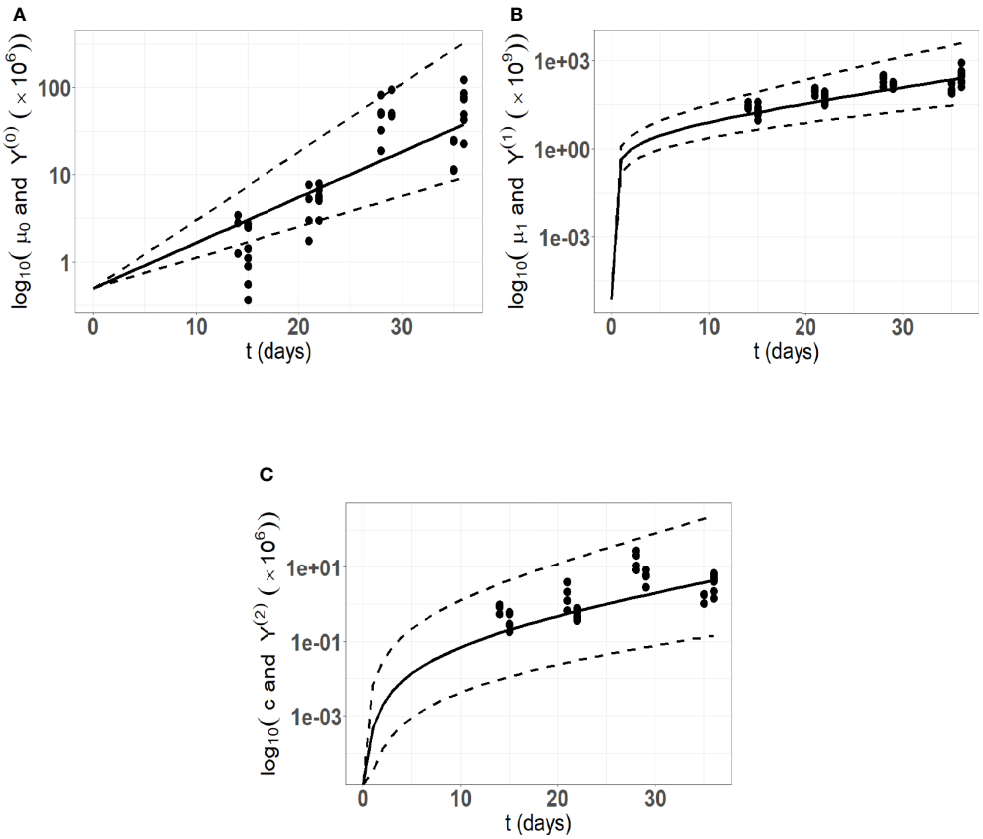


FIGURE 4 | Model fitting to the *in vivo* experimental cSCC tumor growth data. Here, we are using 34 data points from an *in vivo* experimental cutaneous squamous cell carcinoma (cSCC) tumor growth mouse model (59). **(A)** Number of tumor cells kinetics; **(B)** Tumor volume kinetics (μm^3); **(C)** Concentration of immune cells kinetics. The solid lines represent the model prediction using the mean estimated parameters, the dashed lines represent the model predictions using the 5th and 95th percentiles of the parameters distribution.

streptomycin (100 $\mu g/ml$) (Gibco-ThermoFisher Scientific, Courtaboeuf, France). 5×10^5 mSCC38 were intradermally injected in anesthetized mice after dorsal skin shaving. Tumor volume was measured manually using a ruler and calculated according to the ellipsoid formula: Volume=Length(mm)×Width(mm)×Height(mm)×$\pi/6$.

2.3.3 Tissue Preparation and Cell Count

mSCC38 were excised and enzymatically treated twice with collagenase IV (1mg/ml) (Sigma-Aldrich, St Quentin Fallavier, France), and DNase I (0.2 mg/ml) (Roche Diagnostic, Meylan, France) for 20 minutes at 37°C . Total cell count was obtained on a Casy cell counter (Ovni Life Science, Bremen, Germany).

TABLE 2 | Estimated value of the parameters with their Standard Error (S.E.) and Relative Standard Error (R.S.E).

parameters	value	S.E	R.S.E (%)
a_{pop}	0.12	0.0041	3
V_{pop}	816.33	92.59	11
R_{pop}	2.2×10^{-6}	3.6×10^{-7}	16
ω_a	0.20	0.027	13.5
ω_V	0.51	0.075	15
ω_R	0.84	0.11	13
b_a	0.37	0.041	11
b_V	0.17	0.052	31
b_R	0.18	0.056	30
ρ_{aR}	0.8	0.1	13

Immune cell count was determined from flow cytometry analysis. Briefly, cell suspensions were incubated with anti-CD16/32 (2.4G2) to block Fc receptors and stained with anti-CD45 (30-F11)-BV510 antibody and the 7-Aminoactinomycin D (7-AAD) to identify live immune cells (BD Biosciences, Le Pont de Claix, France). Samples were acquired on a BD LSR Fortessa and analyzed with DIVA V8 and FlowJo V10 software (BD Biosciences, Le Pont de Claix, France).

2.3.4 Mathematical and Statistical Analysis

Computations were realized in Python and we made use of dedicated libraries, in particular the gmsh library for the computational domain mesh generation, the packages optimize (for the optimization methods using the Levenberg-Marquard mean square algorithm; similar results have been obtained with the CMA-ES algorithm of the library cma) from the library scipy, the MONOLIX R API and application for the model calibration to the experimental data (61), the library Pygpc for the generalized Polynomial Chaos approximation (62) and the library Salib for the sensitivity analysis (63).

3 RESULTS

3.1 Validation of the Method

For all the simulations discussed here, we adopt the same framework as in (10): the tumor is located at the origin of the computational domain Ω, which is the two-dimensional unit disk. Otherwise explicitly stated, we work with the lower bound of the parameters collected in **Table 1**. When necessary, the initial values of the unknowns are respectively $\mu_0\ (0) = 1\ cell_n$, ue $\mu_1\ (0) = 14137.2\ \mu m^3$, $c(0,x) = 0$.

To start with, we perform a simulation of the initial-boundary value problem (1a)-(1e). **Figure 5** illustrates how the equilibrium establishes in time: as time becomes large, the effective concentration of active immune cells, that is denoted

$$\overline{\mu_c}(t) = \int_\Omega \delta(x) c(t,x)\, \mathrm{d}x$$

tends to the eigenvalue of the cell-division equation, the total mass $\mu_1(t)$ tends to a constant and the size distribution of tumor cells takes the profile of the corresponding eigenstate. This result has been obtained by setting $(a,V,R) = (0.072, 713.61, 1.74\times 10^{-7})$. We observe a non symmetric shape of the size distribution of tumor cells, peaked about a diameter of 23 μm, which is consistent with observational data reporting the mean size distribution of cancer cells (64, 65).

For the simplest model of growth-division with a and V constant, we know an expression of the eigenstate $(\lambda, \bar{N})$; however, we do not know an explicit evaluation of the residual mass. Nevertheless, we can compare the results of the inverse power-dichotomy procedure that predicts the residual mass, to the large time simulations as performed in (10). Let μ_1^f be the asymptotic value of the total mass given by the large time simulation of the initial-boundary value problem (and checking that the variation of the total mass has become negligible) and let μ_1^{pd} be the mass predicted by the power-dichotomy procedure. We set

$$E_{\mu_1} = \frac{\left|\mu_1^f - \mu_1^{pd}\right|}{\mu_1^f}.$$

The results for several cell division rates a are collected in **Table 3**: the numerical procedures finds the same equilibrium mass as the resolution of the evolution problem, which is a further validation of the method.

Further validation concerning the ability in finding the leading eigenstate are presented in Appendix A. The method has been successfully employed to predict equilibrium state when dealing with complex growth rate and division operator in (42).

3.2 Numerical Simulations Show How Parameters Influence Equilibrium

The numerical methods were next used to assess how the parameters influence the equilibrium. In particular, we wish to assess the evolution of the tumor mass at equilibrium according to immune response and tumor growth parameters.

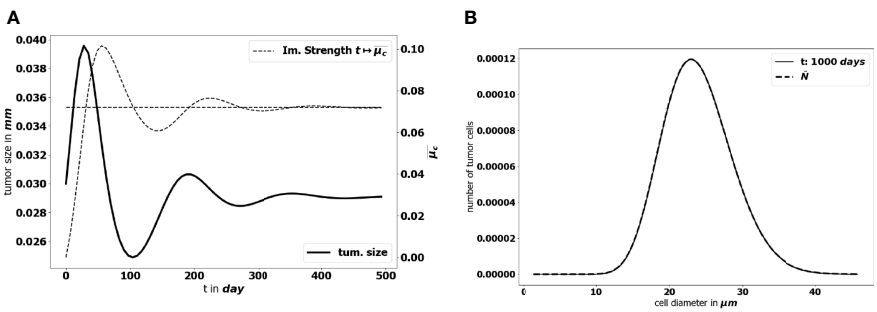


FIGURE 5 | **(A)** Time evolution of the diameter of the tumor (bold black line) and concentration of active immune cells (dotted gray line). The predicted asymptotic value for the latter is represented by the horizontal dotted line. **(B)** Comparison of the tumor cell-size distribution at t = 1000 *days* with the positive eigenstate of the cell division equation (x-axis: size of the tumor cells, y-axis: number of tumor cells at the final time). For this simulation $\Omega = \{\, \| x \| \leq 1 \,\}$, the data are given by the lower bound of the parameters collected in **Table 1** and $(a,V,R) = (0.072, 713.61, 1.74\times 10^{-7})$.

TABLE 3 | Comparison of the large time tumor mass and the predicted tumor mass for several values of a.

a	μ_1^f (mm^3) at final time $T = 500$	μ_1^{pd} (mm^3)	E_{μ_1}
0.103	$7.67271875 \times 10^{-5}$	$7.67271872 \times 10^{-5}$	4.10×10^{-9}
0.15	$1.11701535 \times 10^{-4}$	$1.11701543 \times 10^{-4}$	7.97×10^{-8}
0.20	$1.48924575 \times 10^{-4}$	$1.48924641 \times 10^{-4}$	4.40×10^{-7}
0.3	$2.23420663 \times 10^{-4}$	$2.23420562 \times 10^{-4}$	4.53×10^{-7}
0.351	$2.61368442 \times 10^{-4}$	$2.61367974 \times 10^{-4}$	1.80×10^{-6}

For the numerical simulations presented here, we thus work on the eigenproblem (6) and on the constrained system (7)-(8). Unless precisely stated, the immune response parameters are fixed to the lower bounds in **Table 1**. The tumor growth parameters are set to = 0.1 day^{-1}, $V = 713.61\ \mu m^3\ day^{-1}$ and $R = 1.74 \times 10^{-7} \frac{cell_n \cdot mm^{-3}}{cell_\mu \cdot \mu m^3} \cdot day^{-1}$.

The main features of the solutions follow the observations made in (10), which were performed with arbitrarily chosen values for the parameters. We observe that $\bar{\mu}_c(t) = \int_\Omega \delta(y) c(t,y)\, \mathrm{d}y$ tends to the division rate a, which in this case corresponds to the leading eigenvalue of the cell-division equation. It is important to note that the predicted diameter of the tumor at equilibrium — see **Figure 5** — is significantly below modern clinical PET scanners resolution limit, which could detect tumors with a diameter larger than $7mm$ (66). This is consistent with the standard expectations about the equilibrium phase (3), but, of course, it makes difficult further comparison of the prediction with data.

The aggressiveness of the tumor is characterized by the division rate, the variations of which impact the size of the tumor at equilibrium: the larger a, the larger the residual tumor, see **Figure 6A**. Increasing the immune strength A increases the efficacy of the immune response, reducing the size of the residual tumor see **Figure 6B**. Similarly, increasing the mean rate of influx of effector immune cells in the tumor microenvironment R, decreases the tumor size at equilibrium, see **Figure 6C**. On the contrary, increasing the death rate of the immune cells γ reduces the efficacy of the immune response and increases the equilibrium tumor size see **Figure 6D**.

Moreover, as mentioned above, not only the parameters determine the equilibrium mass, but they also impact how the equilibrium establishes. **Figures 7A–C** shows what happens by making the tumor cell division rate a vary. There are more oscillations along time, with larger amplitude, as a increases. Similar observations can be made when reducing the strength of the immune system A (likely out of its realistic range), see **Figures 7D–F**. The smaller A, the weaker the damping of the

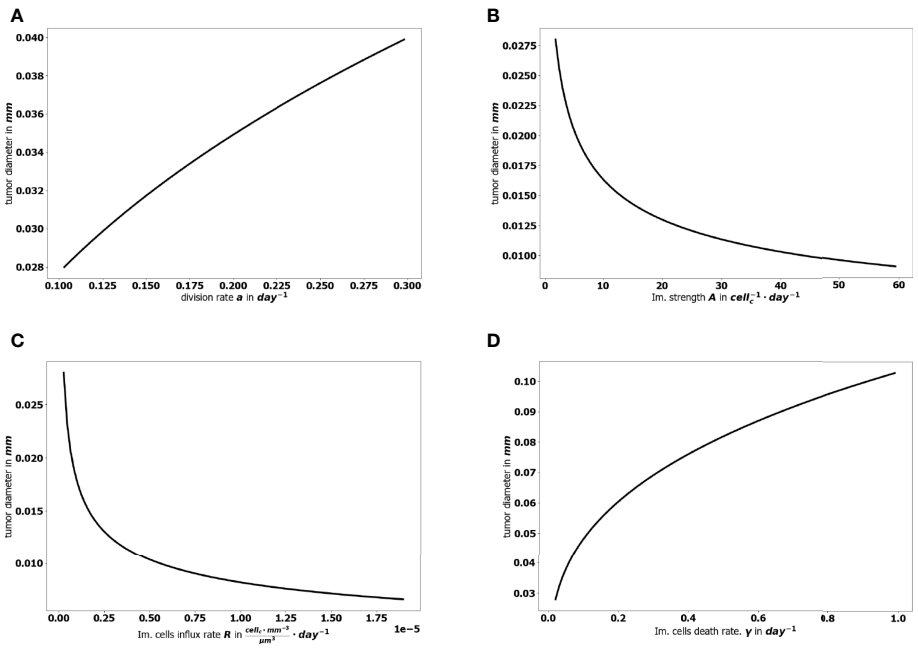


FIGURE 6 | Evolution of the tumor diameter at equilibrium, with respect to **(A)** the division rate of tumor cells a, **(B)** the strength of the effector immune cells A, **(C)** the influx rate of effector immune cells R, **(D)** the natural death rate γ of the effector cells.

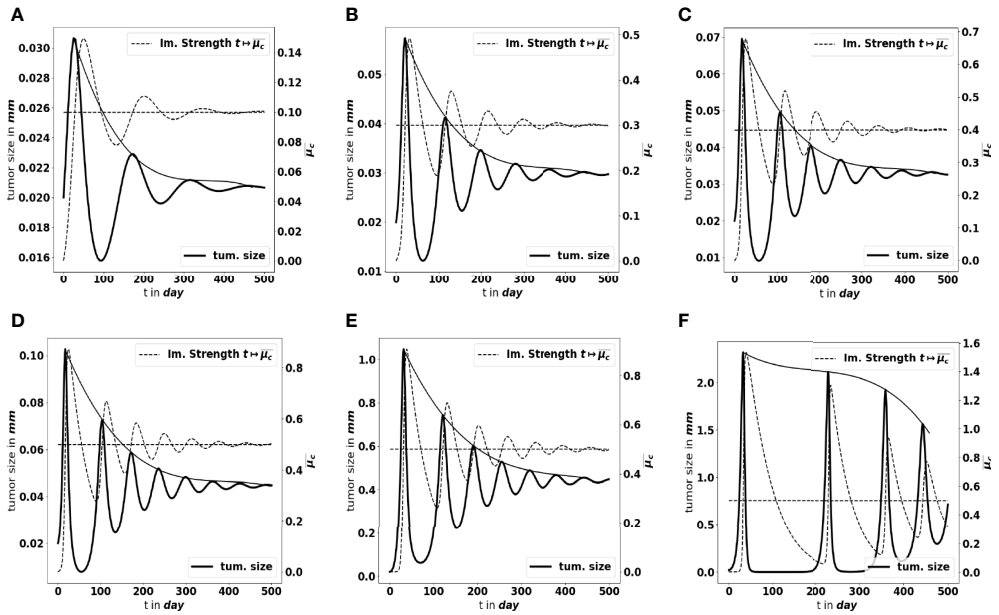


FIGURE 7 | Large-time simulation of the PDE system: evolution of the tumor diameter (bold black line, left axis), and of the concentration of immune cells $\bar{\mu}_c$ (dotted grey line, right axis), for several values of the division rate a: **(A)** $a = 0.1\ day^{-1}$, **(B)** $a = 0.3\ day^{-1}$, **(C)** $a = 0.4\ day^{-1}$ and for several values of the immune strength A: **(D)** $A = 1 \quad cell_c^{-1} \cdot day^{-1}$, **(E)** $A = 1 \cdot 10^{-3} \quad cell_c^{-1} \cdot day^{-1}$, **(F)** $A = 1 \cdot 10^{-5} cell_c^{-1} \cdot day^{-1}$. The horizontal dotted line represents the predicted asymptotic value for $\bar{\mu}_c$. The solid line gives the envelope of the oscillations, indicating a polynomial damping rate. The equilibrium needs more time to establish as the strength of the immune system decreases.

oscillations and the longer the periods. We notice that the decay of the maximal tumor radius holds at a polynomial rate. In extreme situations, either the damping is very strong and the equilibrium establishes oscillation-free or the equilibrium does not establish on reasonable observation times, and the evolution can be confounded with a periodic alternance of growing and remission phases. Such scenario illustrates that the relevance of the equilibrium can be questionable depending on the value of the parameters. In what follows, we focus on the details of the equilibrium itself, rather than on the transient states.

3.3 Global Sensitivity Analysis on the Equilibrium Mass Identifies the Key Parameters to Target in Cancer Therapy

Since the equilibrium state can be computed for a reduced numerical cost (it takes about 1/4 of a second on a standard laptop), we can perform a large number of simulations, sampling the range of the parameters. This allows us to discuss in further details the influence of the parameters on the residual mass and, by means of a global sensitivity analysis, to make a hierarchy appear according to the influence of the parameters on this criterion. Ultimately, this study can help in proposing treatments that target the most influential parameters.

Details on the applied methods for the sensitivity analysis can be found in Appendix B. Among the parameters, we distinguish:

- The tumor cell division rate a which drives the tumor aggressiveness,
- The efficacy of the immune system, governed by the mean influx rate of activated effector immune cells R, the strength of the immune response A, the chemotactic sensitivity χ, the death rate γ of the immune cells, and the strength of the chemical signal induced by each tumor cell A_σ
- Environmental parameters such as the diffusion coefficients D (for the immune cells) and K (for the chemokine concentration).

We assume that the input parameters except a and R are independent random variables. Due to the lack of knowledge on the specific distribution of most of the parameters, the most suitable probability distribution is the one which maximizes the continuous entropy (67), more precisely, the uniform distribution over the ranges defined in **Table 1**. Therefore, the uncertainty in the parameter values is represented by uniform distributions for the parameters ($A,\chi,D,A_\sigma,\gamma,K$) and by log-normal distributions for the parameters a and R. In what follows, the total mass at equilibrium, μ_1, given by the power-dichotomy algorithm, is seen as a function of the uncertain parameters:

$$\mu_1 = f(a, A, R, \chi, D, A_\sigma, \gamma, K) \tag{15}$$

To measure how the total variance of the output μ_1 of the algorithm is influenced by some subsets $i_1 \cdots i_p$ of the input parameters $i_1 \cdots i_p$ ($k \geq p$ being the number of uncertain input parameters), we compute the so-called Sobol's sensitivity indices. The total effect of a specific input parameter i is evaluated by the total sensitivity index $S_T^{(i)}$, the sum of the sensitivity indices which contain the parameter i. (Details on the computed Sobol indices

can be found in Appendix B). The computation of these indices is usually based on a Monte Carlo (MC) method [see (68, 69)] which requires a large number of evaluations of the model due to its slow convergence rate ($O(1/\sqrt{N})$ where N is the size of the experimental sample). To reduce the number of model evaluations, we use instead the so-called generalized Polynomial Chaos (gPC) method [see (70)]. The backbone of the method is based on building a surrogate of the original model by decomposing the quantity of interest on a basis of orthonormal polynomials depending on the distribution of the uncertain input parameters $\theta(\omega)=(a,A,R,\chi,D,A_\sigma,\gamma,K)$, where ω represents an element of the set of possible outcomes. Further details on the method can be found in (71). For uniform distributions, the most suitable orthonomal polynomial basis is the Legendre polynomials. The analysis of the distribution of μ_1 after a suitable sampling of the parameters space indicates that μ_1 follows a log-normal distribution. This distribution is not uniquely determined by its moments (the Hamburger moment problem) and consequently cannot be expanded in a gPC [see (72)]. Based on this observation, to obtain a better convergence in the mean square sense, we apply the gPC algorithm on the natural logarithm of the output μ_1. Typically, $\ln(\mu_1)$ is decomposed as follows:

$$\ln(\mu_1(\omega)) = \sum_{\alpha \in J_{k,p}} q_\alpha L_\alpha(\theta(\omega)) + \epsilon, \tag{16}$$

where ϵ corresponds to the approximation error, $\mathcal{J}_{k,p} = \{\alpha \in \mathbb{N}^k : \sum_{i=1}^k \alpha_i \leq p\}$ and p represents the highest degree of the expansion. Hence, the dimension of the polynomial basis is given by $\frac{(k+p)!}{k!\,p!}$. We reduce the number of model evaluations to 642 runs by constraining also the parameters interaction order to 2. For our purpose, a degree $p = 5$ gives a better fit (see **Figures 8A, B** to the original model and the goodness of fit of the gPC algorithm is measured by a Leave One Out Cross Validation (LOOCV) technique (73). The resulting LOO error indicates 0.4% prediction error. The Sobol's sensitivity indices are then computed from the exponential of the surrogate model (16) by using Monte Carlo simulations combined with a careful space-filling sampling of the parameters space [see (68, 74)]. For the computations, a sample with N=1.8× 10^6 points has been used in order to get stable second order Sobol indices. Indeed, the sensitivity indices that are needed to discriminate the impact of the input parameters are the first and total Sobol' sensitivity indices. Here, the analysis revealed a significant difference between some first order Sobol' indices and their corresponding total Sobol indices, which indicated the importance of computing also the second order Sobol' indices.

It is important to stress that the obtained results, and the associated conclusions, could be highly dependent on the range of the parameter values. This observation makes the measurement/estimation of the parameters a crucial issue which can be dependent on the type of cancer analyzed.

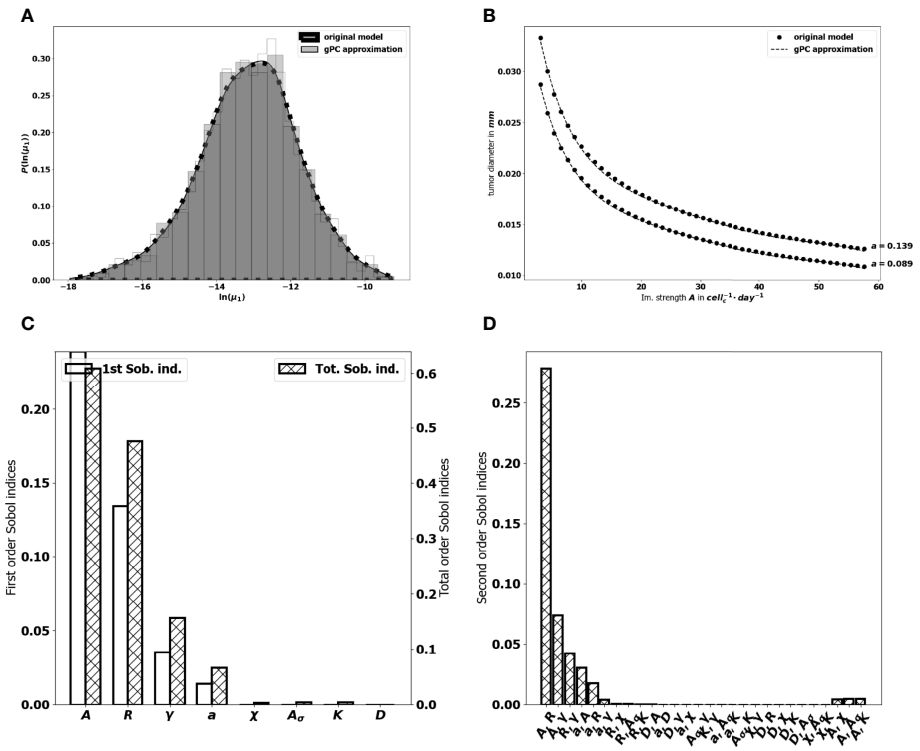


FIGURE 8 | **(A)** comparison between the pdf of ln (μ_1) from the gPC approximation and the pdf from the original model. **(B)** Comparison between the value of μ_1 generated by the power-dichotomy algorithm and the gPC approximation. **(C)** First (empty, left scale) and total (dashed, right scale) order Sobol indices for μ_1. **(D)** Second order Sobol indices for μ_1.

3.3.1 Efficacy of the Immune Response

The first order Sobol indices represented in **Figure 8C** indicate that the parameters which impact the most the variability of the immune-controlled tumor mass at equilibrium are:

- The strength of the lethal action of the immune cells on the tumor cells A, by far the most influential, and three additional parameters
- The influx rate of activated effector immune cells into the tumor microenvironment R.
- The natural death rate γ of the effector immune cells,
- And the division rate a of the tumor cells.

This result is consistent with the observations made from the numerical experiments above and in (10), showing a prominent role of the immune response which can be enhanced by increasing either A or R, and decreasing γ. That A is the most influential parameter is not that surprising but it is remarkable how far its importance exceeds that of the other parameters. It is also puzzling to see that the chemotactic sensitivity χ, like the strength of the chemical signal induced by each tumor cell A_σ, the space diffusion coefficient of the effector immune cells D and the diffusion coefficient of the chemokines K , have a negligible influence on the immune-controlled tumor mass, see **Figure 8C**, whether individually or in combination with other parameters. This result is consistent with the necessity for immune cells to be able to effectively kill the tumor cells once they reach the tumor site. The second order Sobol' indices indicate that the leading interactions are the pairs (A, R), (A,γ), (R, γ), (a,A), (a, R) and (a, γ). Accordingly, in order to enhance the immune response, an efficient strategy can be to act simultaneously on the immune strength A together with the influx rate of activated immune effector cells R. Increasing such influx into the tumor microenvironment by enhancing the activation/recruitment processes leading to the conversion of naive immune cells into activated immune cells potentiate anti-tumor immune responses. Besides, the natural death rate γ of the effector immune cells combined to A and R have an impact, as well as A combined with the division rate of the tumor cells, a.

3.3.2 The Tumor Aggressiveness

The tumor aggressiveness is mainly described by the cell division rate a. The first order Sobol indice indicates that a influences significantly the tumor mass at equilibrium, and we observe that the total Sobol index of a is higher than the individual one. This indicates that this parameter has strong interactions with the others. By taking a look at **Figure 8D** we remark that a interacts significantly with the parameters A, R, γ. However, the most significant interaction is the one with A. This suggests that combining therapies targeting tumor and immune cells should be more efficient at maintaining immune-mediated tumor mass dormancy (75).

3.3.3 Towards Optimized Treatments

Because equilibrium state can be computed for a reduced numerical cost, it allows a large number of simulation to be performed in a minimal time, so that an extensive sampling of the range of the parameters can be tested. The flexibility of the numerical simulations provides valuable tools to assess the

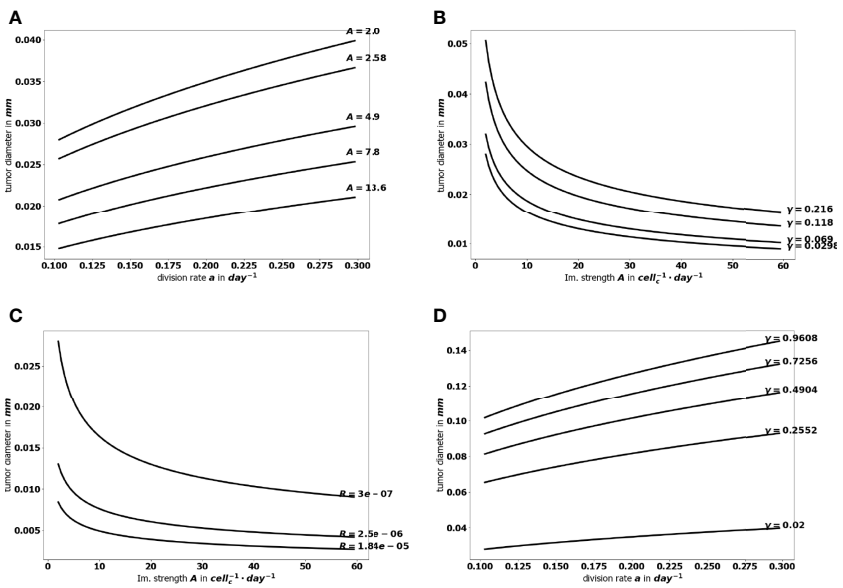


FIGURE 9 | Evolution of the tumor diameter at equilibrium, **(A)** with respect to the division rate a for several values of the immune strength A, **(B)** with respect to the immune strength A for several values of the death rate γ, **(C)** with respect to the immune strength A for several values of the influx rate of effector immune cells R+, and **(D)** with respect to the division rate a for several values of the death rate γ.

efficiency of a variety of therapeutic strategies and select those that sustain a viable equilibrium and prevent cancer relapses after a surgery or a treatment. **Figure 9** illustrates how the equilibrium mass is impacted when combining variations of two parameters, namely the immune strength A combined to the tumor cell division rate a, the mean rate of influx of effector immune cells R or the death rate of effector immune cells γ; and the tumor cell division rate a with the death rate γ. Interestingly, a reduction of the tumor mass at equilibrium can be obtained significantly more easily by acting on two parameters than on a single one. For instance, reducing the tumor cell division rate a from 0.35 to 0.1 cannot reduce the diameter of the tumor below.025 mm, with $A = 1$; while the final size is always smaller when $A = 3.95$. This observation highlights the interest of combined treatments having such complementary actions. The interest is two-fold: either smaller residual tumors can be obtained by pairing two actions, or the same final tumor size can be obtained with a combined treatment having less toxicity than a mono-therapy.

4 DISCUSSION

Controlling parameters that maintain immune-mediated tumor mass dormancy is key to the successful development of future cancer therapies. To understand how equilibrium establishes and how it is influenced by immune, environmental and tumor-related parameters, we evaluate the tumor mass which tends to a constant at equilibrium. In this study, we make use of the space and size structured mathematical model developed in (10) to provide innovative, efficient methods to predict, at low numerical cost, the residual tumor mass at equilibrium. By means of numerical simulations and global sensitivity analysis, we identify the elimination rate A of tumor cells by immune cells as the most influential factor. Therefore, the most efficient therapeutic strategy is to act primarily on the immune system rather than on the tumor itself. We also demonstrate the need to develop combined cancer treatments, boosting the immune capacity to kill tumor cells (increase A), the conversion into efficient immune cells (increase R), reducing natural death rate of effector immune cells (decrease γ) and reducing the ability of tumor cells to divide (decrease a). The combination of such approaches definitely outperforms the performances of a single action; it permits to maintain the tumor in a long-lasting equilibrium state, far below measurement capabilities.

Generally, therapeutic strategies are designed to target preformed, macroscopic cancers. Indeed, patients are diagnosed once their tumor is established and measurable, thus at the escape phase of the cancer immunoediting process (1). The goal of successful treatments is to revert to the equilibrium phase and ultimately to tumor elimination. Experimental evidence and clinical observations indicate that such equilibrium exists but it is difficult to study and measure, the residual tumor mass being below detection limits (1, 2, 3). It is regarded as "a immune-mediated tumor mass dormancy" when the rate of cancer cell proliferation matches their rate of elimination by immune cells. In human, cancer recurrence after therapy and long periods of remission or detection of low number of tumor cells in remission phases are suggestive of such equilibrium phase. Mathematical models can also be used to provide evidence of such state. The system of partial differential equations proposed in (10) is precisely intended to describe the earliest stages of immune control of tumor growth. Remarkably, while being in the most favorable condition, only taking into account the cytotoxic effector immune cells and no immunosuppressive mechanisms, the model reproduces the formation of an equilibrium phase with maintenance of residual tumor cells rather than their complete elimination. Besides suggesting that elimination may be difficult to reach, this finding also brings out the role of leading parameters that shape the equilibrium features and opens new perspectives to elaborate cancer therapy strategies that reach this state of equilibrium.

To decipher tumor-immune system dynamics leading to equilibrium state, we have developed here computational tools. The total mass of the tumor is a critical criterion of the equilibrium and was used to predict parameters that contribute the most to the establishment of the equilibrium. By means of global sensitivity analysis, we identified one leading parameter, A, and three others, R, γ and a that affect the most the variability of the immune-controlled tumor mass; A, R and γ are related to immune cells, and a to tumor cells. Moreover, the influence of the leading parameters is significantly increased when they are paired. This observation supports the development of combined therapeutic treatments which would be more efficient at reducing tumor growth and toxicity. Because the pairs (A, R), (A, γ), (R, γ), (A, a), (a, R) and (a, γ) are the most influential, we predict that a combination of drugs enhancing antitumor immune responses with drugs diminishing tumor aggressiveness will be the most efficient. This is consistent with the clinical benefit obtained when chemotherapies reducing the tumor cell division rate a are combined with immunotherapies increasing A and R (75), The parameter A which governs the efficacy of the immune system to eliminate tumor cells, is the most influential. This finding is consistent with the observation that "hot" tumors infiltrated with immune cells have better prognostic than "cold" tumors (76) and that the immune cells with the strongest positive impact on patient's survival are the cytotoxic $CD8^+$ T cells (77). It is also in line with the success of ICP which revert immune tolerance triggered by chronic activation and upregulation of exhaustion markers on effector T and NK cells, thus not only increasing the parameter A but also R (78). The leading role of the parameter A is also demonstrated by experimental studies and clinical trials, such as adoptive transfer of CAR-T and CAR-NK cells engineered to attack cancer cells, immunomodulating antibody therapies or cancer vaccines which boost the antitumor immune response (75, 79–81). Finally, our finding that the parameter γ is highly influential is confirmed by the administration of cytokines that stimulate and increase effector T and NK cell survival which are efficient at controlling tumor growth (81). Thus,

altogether, these experimental and clinical data validate the numerical method.

Interestingly, besides the dominant role of the parameter A, only two additional parameters related to immune cells R, γ seem to have an influence on the tumor mass at equilibrium. These data predict that to enhance the immune response, it is more efficient to increase the rate of influx and conversion of naive immune cells into effector cells (parameter R) or to increase the lifespan of immune effectors (parameter γ) than to increase chemotaxis as a whole (parameters χ, A_a, K). The lack of influence of chemotaxis emphasizes that the localization of immune cells within tumors is necessary but not sufficient. Indeed, the leading influence of the parameters A, R, γ stresses the importance of having functional immune cells infiltrating tumors. Overcoming immune suppression is therefore highly relevant in therapeutic strategies.

Targeting Immune-mediated tumor mass dormancy is gaining more and more attention, having been linked to recurrence and metastasis (9, 82). The persistence of undetectable tumor cells after primary tumor resection at the primary site but also their spreading to metastatic niches are major causes of treatment failure. Thus, developing strategies to maintain an equilibrium between these tumor cells and the immune response is crucial. Interestingly, a recent study demonstrated a role of the NK cell reservoir in blocking the reawakening of dormant tumor cells (83). The mechanisms involve IL-15 that drives NK cell proliferation and IFN- γ secreted by NK. Therapies boosting NK cell activity like IL-15 superagonists, or engineered NK cell engagers are therefore promising strategies to sustain NK cell-mediated maintenance of tumor dormancy (83, 84).

It is appropriate to finally comment on the limitations of this work and provide new avenues for future research. Firstly, the analysis focuses on the asymptotic state, taking full advantage of its mathematical interpretation which makes it easily computable. However, the transient states and the rate at which the equilibrium becomes observable are simply disregarded, while they are certainly essential for assessing the biological relevance of the equilibrium state. Further analysis is therefore needed in order to understand how the parameters of the model influence the trend to equilibrium. Secondly, the modeling approach is facing contradictory requests: on the one hand, the lack of knowledge on the parameters motivates working with a reduced set of equations, at the cost of considering an "averaged" behavior (say for instance between different types of immune cells); on the other hand, it might be important to keep under consideration many relevant and competing effects of cellular interactions. These issues can be addressed with a better access to biological data and through the development of dedicated methods of parameter identification. This is of course even more important when describing the effects of treatments. Thirdly, the present analysis is limited to an idealized situation in which many important effects have been overlooked. In particular, the immune response can also promote the tumor growth. Considering such immune actions leads to a much more complex dynamical behavior and the possible establishment of an escape phase, as shown in (42). Finally geometrical aspects and heterogeneity are poorly addressed and restrict the relevance of the description to the earliest stages of the tumor development. More complex models, with a full space structuration, should be elaborated in order to obtain a more accurate description of the tumor microenvironment.

5 CONCLUSION

In conclusion, clinical trials have been undertaken quite often on assumptions from acquired knowledge on tumor development and immune responses to cancer cells, but without tools to help the decision-making. The numerical methods developed here provide valuable hints for the design and the optimization of antitumor therapies. The approach is in agreement with published experimental findings and clinical evidence. By adapting the range of the parameters to the biological values, one can more precisely adapt the therapeutic strategies to specific types of tumors. We thus conclude that mathematical modelling combined with numerical validation provide valuable information that could contribute to better stratify the patients eligible for treatments and consequently save time and lives. In addition, it could also help to decrease the burden of treatment cost providing hints on optimized therapeutic strategies.

6 COMPUTATION OF THE EIGEN-ELEMENTS OF THE GROWTH-DIVISION EQUATION

The binary division operator (2) is a very specific case, and for applications it is relevant to deal with more general expressions. Namely, we have

$$Q(n)(t,z) = -a(z)n(t,z) + \int_z^\infty a(z')k(z \mid z')n(t,z')\,\mathrm{d}z'. \quad (17)$$

In (21), $a(z')$ is the frequency of division of cells having size z', and $k(z|z')$ gives the size-distribution that results from the division of a tumor cell with size z'. What is crucial for modeling purposes is the requirement

$$\int_0^z z' k(z' \mid z)\,\mathrm{d}z' = z,$$

which is related to the principle that cell-division does not change the total mass

$$\int_0^\infty zQ(n)\,\mathrm{d}z = 0.$$

We refer the reader to (32) for examples of such cell-division operators and the analysis of the eigenvalue problem (6) under

quite general assumptions of the growth rate V, the frequency a and the kernel k. Our numerical method can handle such general coefficients.

It is important to bear in mind the main arguments of the proof of the existence-uniqueness of the eigenpair $(\lambda, \bar{N})$ for the growth-division equation. Namely, for Λ large enough we consider the *shifted* operator

$$T_\Lambda N = \Lambda N + \partial_z(VN) + aN - \int_z^\infty a(z')k(z \mid z')N(z')\,\mathrm{d}z'.$$

Then, we check that the operator S_Λ which associates to a function f the solution n of $T_\Lambda n = f$ fulfills the requirements of the Krein-Rutman theorem (roughly speaking, positivity and compactness), see (85). Accordingly, the quantity of interest λ is related to the leading eigenvalue of S_Λ. In fact, this reasoning should be applied to a somehow truncated and regularized version of the operator, and the conclusion needs further compactness arguments; nevertheless this is the essence of the proof. In terms of numerical method, this suggests to appeal to the inverse power algorithm, applied to a discretized version of the equation. However, we need to define appropriately the shift parameter Λ. As far as the continuous problem is considered, Λ can be estimated by the parameters of the model (32), but it is critical for practical issues to check whether or not this condition is impacted by the discretization procedure. This information will be used to apply the inverse power method to the discretized and shifted version of the problem.

6.1 Analysis of the Discrete Problem

The computational domain for the size variable is the interval [0, R] where R is chosen large enough: due to the division processes, we expect that the support of the solution remains essentially on a bounded interval, and the cut-off should not perturb too much the solution. In what follows, the size step $h = z_{i+1} - z_i$ is assumed to be constant. The discrete unknowns N_i, with $i \in \{1, \dots, I\}$ and $h = R/I$, are intended to approximate $N(z_i)$ where $z_i = ih$. The integral that defines the gain term of the division operator is approximated by a simple quadrature rule. For the operator (2) the kernel involves Dirac masses which can be approached by peaked Gaussian. We introduce the operator $T_\Lambda^h : \mathbb{R}^I \to \mathbb{R}^I$ defined by

$$\begin{cases} (T_\Lambda^h N)_i = F_i - F_{i-1} + h(\Lambda + a_i)N_i \\ -h^2 \sum_{j=i}^{I} a(z_j)k(z_i \mid z_j)N_j, \\ N_1 = 0 \end{cases} \tag{18}$$

where $F_i = V_{i+1/2} N_i$ represents the convective numerical flux on the grid point $z_{i+1/2} = (i+1/2)h$, $i \in \{1, \dots, I\}$. This definition takes into account that the growth rate is non negative, and applies the upwinding principles. Note that the step size h should be small enough to capture the division of small cells, if any. The following statement provides the *a priori* estimate which allows us to determine the shift for the discrete problem.

Theorem 1.1. We suppose that

i. $z \mapsto V(z)$ is a continuous function which lies in L^∞ and it is bounded from below by a positive constant,
ii. $h\sum_{j=1}^{I} a(z_j)k(z_i \mid z_j)$ remains bounded uniformly with respect to h,
iii. for any $i \in \{1,\dots,I-1\}$, there exists $j \in \{1,\dots,I-1\}$ such that $a(z_j)k(z_i \mid z_j) > 0$,
iv. there exists $Z_0 \in (0, \infty)$ such that, setting $\bar{\mathcal{N}}(z) = h\sum_{j=2}^{I} k(z_j \mid z)$, we have $a(z)(\bar{\mathcal{N}}(z) - 1) \geq \nu_0 > 0$ for any $z \geq Z_0$.

Let

$$\Lambda > \frac{\|V\|_{L^\infty}}{\min_{j\in\{1,\dots,I\}} |V_{j+1/2}|} \max_{k\in\{1,\dots,I\}} \Big(h\sum_{j=k}^{I} a_j k(z_k \mid z_j)\Big) - \min_{j\in\{1,\dots,I\}} |a_j|, \tag{23}$$

and we suppose that $R > Z_0$ is large enough. Then, T_Λ^h is invertible and there exists a pair $\mu > 0$, $N \in \mathbb{R}^I$ with positive components, such that $Ker((T_\Lambda^h)^{-1} - \mu) = Span\{N\}$. Moreover $\lambda = \Lambda - \frac{1}{\mu} > 0$.

Note that the sum that defines $\bar{\mathcal{N}}(z)$ is actually reduced over the indices such that $jh \leq z$; this quantity is interpreted as the expected number of cells produced from the division of a cell with size z so that the forth assumption is quite natural.

Proof. Let $f \in \mathbb{R}^I$. We consider the equation

$$T_\Lambda^h N = f.$$

We denote $N = S_\Lambda^h f$ the solution. We are going to show that S_Λ^h is well defined and satisfies the assumptions of the Perron-Frobenius theorem, see e. g. (47, Theorem 1.37 & Corollary 1.39) or (86, Chapter 5).

It is convenient to introduce the change of unknown $U_i = N_i V_{i+1/2}$, $\forall i \in \{1, \cdots, I\}$. The problem recasts as

$$\begin{cases} (\tilde{T}_\Lambda^h U)_i = h\frac{f_i}{V_{i+1/2}}, \text{with} \\ (\tilde{T}_\Lambda^h U)_i = U_i - U_{i-1} + h\frac{\Lambda + a_i}{V_{i+1/2}} U_i \\ \quad -h^2 \sum_{j=i}^{I} \frac{a_j}{V_{j+1/2}} k(z_i \mid z_j)U_j, \\ U_1 = 0. \end{cases} \tag{20}$$

The solution is interpreted as the fixed point of the mapping

$$\xi \mapsto U = A^h \xi$$

where U is given by $U_{1=0}$ and

$$U_i = U_{i-1} + h^2 \sum_{j=i}^{I} \frac{a_j}{V_{j+1/2}} k(z_i \mid z_j)\xi_j + h\frac{f_i}{V_{i+1/2}}.$$

We are going to show that A^h is a contraction: $\|A^h \xi\|_{\ell^\infty} \leq k\|\xi\|_{\ell^\infty}$ for some $k < 1$. Multiplying (20) by sign (U_i), we obtain

$$\left(1+h\frac{\Lambda+a_i}{V_i}\right)\operatorname{sign}(U_i)U_i=\left(1+h\frac{\Lambda+a_i}{V_i}\right)|U_i|$$

$$=\operatorname{sign}(U_i)U_{i-1}+h^2\sum_{j=1}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)\operatorname{sign}(U_i)\xi_j$$

$$\le |U_{i-1}|+h^2\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)|\xi_j|\,.$$

We multiply this by the weight $\prod_{l=1}^{i-1}[1+h\frac{\Lambda+a_l}{V_{l+1/2}}]$, where all factors are ≥1 . We get

$$|U_i|\prod_{l=1}^{i}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]$$

$$\le |U_{i-1}|\prod_{l=1}^{i-1}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]$$

$$+h^2\prod_{l=1}^{i}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)|\xi_j|\,.$$

Then, summing over $i\in\{2,\dots,m\}$ yields

$$|U_m|\prod_{l=1}^{m}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]$$

$$\le |U_1|\left[1+h\frac{\Lambda+a_1}{V_{3/2}}\right]$$

$$+h^2\sum_{i=2}^{m}\prod_{l=1}^{i}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)|\xi_j|$$

where actually $U_1=0$. It follows that

$$|U_m|\le h^2\sum_{i=2}^{m}\prod_{l=i}^{m}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]^{-1}\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)|\xi_j|$$

$$\le\frac{h^2\,\|\,\xi\,\|_{\ell^\infty}}{\min\limits_{j\in\{1,\dots,I\}}V_{j+1/2}}\sum_{i=2}^{m}\prod_{l=i}^{m}\left[1+h\frac{\Lambda+a_l}{V_{l+1/2}}\right]^{-1}\sum_{j=i}^{I}a_jk(z_i\,|\,z_j)$$

$$\le\frac{h^2\,\|\,\xi\,\|_{\ell^\infty}}{\min\limits_{j\in\{1,\dots,I\}}V_{j+1/2}}\|\sum_{j=i}^{I}a_jk(z_i\,|\,z_j)\,\|_{\ell^\infty}$$

$$\sum_{i=2}^{m}\left[1+h\frac{\Lambda+\min\limits_{l\in\{1,\dots,I\}}a_l}{\|\,V\,\|_{L^\infty}}\right]^{i-m+1}$$

$$\le\frac{h\,\|\,\xi\,\|_{\ell^\infty}}{\min\limits_{j\in\{1,\dots,I\}}V_{j+1/2}}\|\sum_{j=i}^{I}a_jk(z_i\,|\,z_j)\,\|_{\ell_\infty}$$

$$\left[\frac{\Lambda+\min\limits_{l\in\{1,\dots,I\}}a_l}{\|\,V\,\|_{L^\infty}}\right]^{-1}$$

Therefore, A^h is a contraction provided (19) holds. This estimate is similar to the condition obtained for the continuous problem, see (32, Proof of Theorem 2, Appendix B); the discretization does not introduce further constraints.

We are now going to show that T_Λ^h is a M-matrix when (19) holds. Let $f\in\mathbb{R}^I\setminus\{0\}$ with non negative components. Let $U\in\mathbb{R}^I$ satisfy $(\tilde{T}_\Lambda^hU)_i=h\frac{f_i}{V_{i+1/2}}$. Let i_0 be the index such that $U_{i_0}=\min\{U_i\,,\,i\in\{2,\dots,I\}\}$. We have

$$U_{i_0}\left(1+h\frac{\Lambda+a_{i_0}}{V_{i_0+1/2}}\right)$$

$$=U_{i_0-1}+h^2\sum_{j=i_0}^{I}\frac{a_j}{V_{j+1/2}}k(z_{i_0}\,|\,z_j)U_j+h\frac{f_{i_0}}{V_{i_0+1/2}}\qquad(21)$$

$$\ge U_{i_0}\left(1+h^2\sum_{j=i_0}^{I}\frac{a_j}{V_{j+1/2}}k(z_{i_0}\,|\,z_j)\right)+h\frac{f_{i_0}}{V_{i_0+1/2}}$$

Since $f_{i_0}\ge0$, we get

$$U_{i_0}\left(\frac{\Lambda+a_{i_0}}{V_{i_0+1/2}}-h\sum_{j=i_0}^{I}\frac{a_j}{V_{j+1/2}}k(z_{i_0}\,|\,z_j)\right)\ge0$$

$$\underbrace{\qquad\qquad}_{>0\text{by}(19)}$$

which tells us that $U_{i_0}\ge0$. Suppose $U_{i_0}=0$ for some $i_0>1$. Coming back to (21), we deduce that U_{i_0-1} vanishes too, and so on and so forth, we obtain $U_1=\cdots U_{i_0}=0$. Finally, we use the irreductibility assumption iii): we can find $j_0>i_0$ such that $\frac{a_{j_0}}{V_{j_0+1/2}}k(z_{i_0}\,|\,z_{j_0})>0$ and (21) implies $\frac{a_{j_0}}{V_{j_0+1/2}}k(z_{i_0}\,|\,z_{j_0})U_{j_0}=0$, so that $U_{j_0}=0$. We deduce that $U=0$, which contradicts $f\neq0$. Therefore the components of U are positive, but U_1.

We conclude by applying the Perron-Froebenius theorem to $(T_\Lambda^h)^{-1}$, (86). It remains to prove that $\lambda=\Lambda-\frac{1}{\mu}$ is positive, with μ the spectral radius of $(T_\Lambda^h)^{-1}$. To this end, we make use of assumption iv). We set $Z_0=i_0h$. We argue by contradiction, supposing that $\lambda=\Lambda-1/\mu<0$. We consider the eigenvector with positive components and normalized by the condition $h\sum_{i=1}^{I}U_i=1$. We have

$$(\tilde{T}_0^hU)i=U_i-U_{i-1}+\frac{a_i}{V_{i+1/2}}hU_i$$

$$-h^2\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)U_j=-\lambda U_i\ge0$$

It follows that, for $m\ge i_0$,

$$U_m\ge-h\sum_{i=2}^{m}\frac{a_i}{V_{i+1/2}}U_i+-h^2\sum_{i=2}^{m}\sum_{j=i}^{I}\frac{a_j}{V_{j+1/2}}k(z_i\,|\,z_j)U_j$$

$$\ge-h\sum_{i=2}^{m}\frac{a_i}{V_{i+1/2}}U_i+h\sum_{j=2}^{m}\left(h\sum_{i=2}^{j}k(z_i\,|\,z_j)\right)\frac{a_j}{V_{j+1/2}}U_j$$

$$\ge-h\sum_{i=2}^{m}\frac{a_i}{V_{i+1/2}}U_i+h\sum_{j=2}^{m}\bar{\mathcal{N}}(z_j)\frac{a_j}{V_{j+1/2}}U_j$$

$$\ge h\sum_{i=2}^{m}\left(\bar{\mathcal{N}}(z_i)-1\right)\frac{a_i}{V_{i+1/2}}U_i$$

$$\ge h\sum_{i=i_0}^{m}\left(\bar{\mathcal{N}}(z_i)-1\right)\frac{a_i}{V_{i+1/2}}U_i\ge\frac{v_0}{\|\,V\,\|_{L^\infty}}h\sum_{i=i_0}^{m}U_i\,.$$

It implies

$$1 = h\sum_{m=1}^{I} U_m \geq h\sum_{m=i_0}^{I} U_m \geq h(I-i_0)\frac{v_0}{\| V \|_{L^\infty}} h\sum_{i=i_0}^{m} U_i\,.$$

We arrive at

$$1 \geq (R - Z_0)\frac{v_0}{\| V \|_{L^\infty}},$$

a contradiction when R is chosen large enough (but how large R should be does not depend on h). Therefore, we conclude that $\lambda > 0$.

6.2 Numerical Approximation of (λ, N)

We compute (an approximation of) the eigenpair (λ, N) by using the inverse power method which finds the eigenvalue of $(T_\Lambda^h)^{-1}$ with largest modulus:

- We pick Λ verifying (19).
- We compute once for all the LU decomposition of the matrix T_Λ^h.
- We choose a threshold $0 < \varepsilon \ll 1$.
- We start from a random vector $N^{(0)}$ and we construct the iterations

$$LUq^{(k+1)} = N^{(k)}$$

$$N^{(k+1)} = \frac{q^{(k+1)}}{\| q^{(k+1)} \|}$$

until the relative error $\frac{\| N^{(k+1)} - N^{(k)} \|}{\| N^{(k)} \|} \leq \varepsilon$ is small enough. Then, given the last iterate $N^{(K)}$, we set $LUq = N^{(K)}, \tilde{\mu} = \frac{q.N^{(K)}}{N^{(K)}.N^{(K)}}$, and $\tilde{\lambda} = \Lambda - 1/\tilde{\mu}$.

This approach relies on the ability to approximate correctly the eigenpair of the growth-fragmentation operator. In particular, it is important to preserve the algebraic multiplicity. This issue is quite subtle and it is known that the pointwise convergence of the operator is not enough to guarantee the convergence of the eigenelements and the consistency of the invariant subspaces, see (48) for relevant examples. This question has been thoroughly investigated in (48, 49) which introduced a suitable notion of stability. It turns out that one needs a uniform convergence of the operators. Namely, here, we should check that $\| (T_\Lambda^I)^{-1} - (T_\Lambda)^{-1} \| \to 0$ as $I \to \infty$. In the present framework, a difficulty relies on the fact that the size variable lies in an unbounded domain, which prevents for using usual compactness arguments. For this reason, we introduce a truncated version of the problem, which has also to be suitably regularized. Let us denote by $T_\Lambda^{R,\varepsilon}$ the corresponding operator, where ε represents the regularization parameter. This truncated and regularized perator appeared already in (32). Indeed, we know from (32) that $\| T_\Lambda^{R,\varepsilon} - T_\Lambda \| \to 0$ as $R \to \infty$ and $\varepsilon \to 0$, hence, this implies that $\| (T_\Lambda^{R,\varepsilon})^{-1} - (T_\Lambda)^{-1} \| \to 0$ as $R \to \infty$ and $\varepsilon \to 0$ by continuity of the map $\Pi: T_\Lambda \mapsto (T_\Lambda)^{-1}$. Moreover, $(T_\Lambda^{R,\varepsilon})^{-1}$ is well-defined, continuous and compact, see (32, Appendix. B). The discrete operators $(T_\Lambda^I)^{-1}$ converge pointwise to $(T_\Lambda^{R,\varepsilon})^{-1}$, and the compactness of $(T_\Lambda^{R,\varepsilon})^{-1}$ ensures that the discrete operator converges uniformly to $(T_\Lambda^{R,\varepsilon})^{-1}$, for $0 < R < \varepsilon$ and $0 < \varepsilon < 1$ fixed (see (49) for more details on this fact). Following (49), we deduce that the numerical eigenelements (λ^I, N^I) converges to $(\lambda^{R,\varepsilon}, N^{R,\varepsilon})$, the eigenelements of $(T_\Lambda^{R,\varepsilon})^{-1}$, while preserving their algebraic multiplicity. Finally the uniform convergence $\| (T_\Lambda^{R,\varepsilon})^{-1} - (T_\Lambda)^{-1} \| \to 0$ as $R \to \infty$ and $\in \to 0$ ensures the convergence of $(\lambda^{R,\varepsilon}, N^{R,\varepsilon})$, to (λ, N) (32).

6.3 Numerical Results

For some specific fragmentation kernels and growth rates, the eigenpair $(\lambda, \bar{N})$ is explicitly known, see (32). We can use these formula to check that the algorithm is able to find the expected values and profiles. To this end, we introduce the relative errors

$$E_\lambda^h = \frac{\left|\lambda - \tilde{\lambda}\right|}{\tilde{\lambda}} \quad \text{and} \quad E_V^h = h\sum_{i=1}^{I}\left|N_i^{(K)} - N(ih)\right|$$

where $N^{(K)}$ and N are both normalized by $h\sum_{i=1}^{I} N_i^{(K)} = h\sum_{i=1}^{I} N(ih) = 1$.

Mitosis fragmentation kernel. We start with the binary division kernel:

$$k(z\,|\,z') = \delta_{z'=2z} \tag{22}$$

The associated division operator is described by (2). We assume that a and V are constant. In this specific case the eigenpair is given by

$$\lambda = a,\ N(z) = N\sum_{n=0}^{\infty}(-1)^n \alpha_n \exp\left(-2^{n+1}\frac{a}{V}z\right), \tag{23}$$

with $N > 0$ an appropriate normalizing constant and $(\alpha_n)_{n\in\mathbb{N}}$ is the sequence defined by the recursion

$$\alpha_0 = 1,\ \alpha_n = \frac{2}{2^n - 1}\alpha_{n-1}.$$

In practice we shall use a truncated version of the series that defines N. For the numerical tests, we use the parameters collected in **Table 4**.

With this threshold ε, the approached eigenpair is reached in 43 iterations, independently of the size step. **Figure 10** represents the evolution of the error E_V^h as a function of h in a log-log scale, see also **Table 5**: $N^{(K)}$ approaches N at order 1. The rate improves when using a quadrature rule with a better accuracy. For this test, the approximation of the eigenvalue is already accurate with a coarse grid; it is simply driven by the threshold ε and E_L^h does not significantly change with h.

Uniform fragmentation. The Uniform fragmentation kernel is defined by:

$$k(z\,|\,z') = \frac{1}{z'}1_{0\leq z\leq z'}\,.$$

We apply the algorithm for the following two cases:

TABLE 4 | Data for the numerical tests: binary division kernel.

a	*V*	*R*	ϵ
4	0.6	5	10^{-6}

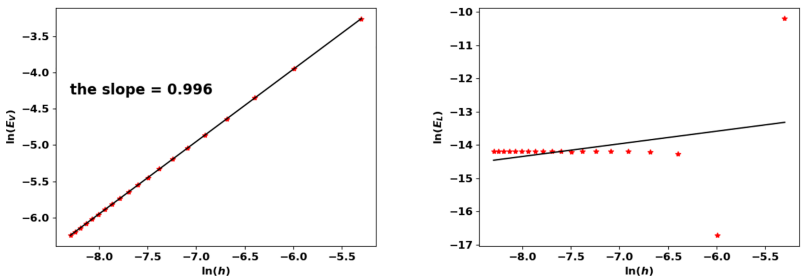


FIGURE 10 | Binary division kernel: convergence rates of $(\lambda^{(K)}, N^{(K)})$ with respect to h.

TABLE 5 | Binary division kernel: errors for several number of grid points.

Number of cells	E_λ	E_V
1000	3.73×10^{-5}	3.83×10^{-2}
2000	5.68×10^{-8}	1.93×10^{-2}
4000	6.77×10^{-7}	9.69×10^{-3}
8000	6.84×10^{-7}	4.85×10^{-3}

TABLE 6 | Uniform fragmentation, ex. 1: errors for several number of grid points.

Number of cells	E_λ	E_V
1000	1.30×10^{-2}	8.89×10^{-3}
2000	6.43×10^{-3}	4.50×10^{-3}
4000	3.23×10^{-3}	2.24×10^{-3}
8000	1.62×10^{-3}	1.13×10^{-3}

$V(z) = V_0$ and $a(z) = a_0 z$. We have $\lambda = \sqrt{a_0 V_0}$ and

$$N(z) = 2\sqrt{\frac{a_0}{V_0}}\left(Z + \frac{Z^2}{2}\right)\exp\left(-Z - \frac{Z^2}{2}\right).$$

We still use the values in **Table 4** (especially, $a_0 = a$ and $V_0 = V$). The approximated eigenpair is obtained in 84 iterations and, as in the previous test, it does not change with the size step. In this case, both the eigenvalue and the eigenfunction are approached at order 1, see **Table 6** and **Figure 11**.

$V(z) = V_0 z$ and $a(z) = a_0 z^n$ with $n \in \mathbb{N} \setminus \{0\}$. The eigenpair is defined by the following formula:

$$n = 1 \quad \lambda = V_0 \qquad N(z) = \frac{a_0}{V_0}\exp\left(-\frac{a_0}{V_0}z\right)$$

$$n = 2 \quad \lambda = V_0 \qquad N(z) = \frac{2a_0}{\pi V_0}\exp\left(-\frac{a_0}{2V_0}z^2\right)$$

$$n \quad \lambda = V_0 \; N(z) = \left(\frac{a_0}{nV_0}\right)^{\frac{1}{n}}\frac{n}{\Gamma\left(\frac{1}{n}\right)}\exp\left(-\frac{a_0}{nV_0}z^n\right)$$

Note that the growth rate V vanishes and Theorem 1.1 does not apply as such. Nonetheless, the algorithm works well and still captures the eigenpair. We perform the test for $n = 1$ and $n = 2$

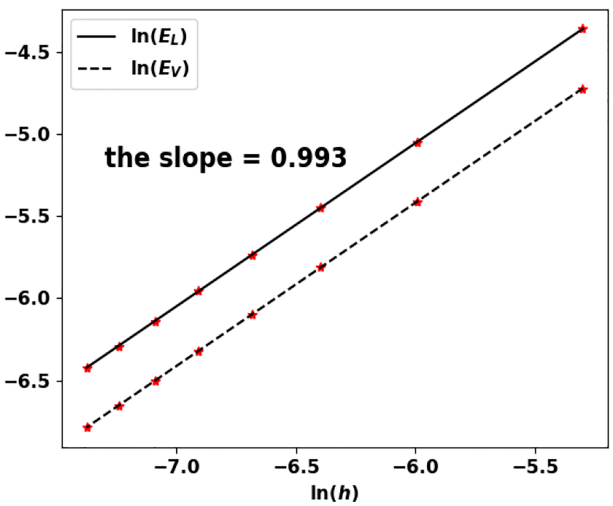


FIGURE 11 | Uniform fragmentation, ex. 1: rate of convergence to the exact eigenpair with respect to h.

and the results are recorded in **Table 7**; **Figure 12** and **Table 8**; **Figure 13**, respectively.

7 SENSITIVITY ANALYSIS ON THE EQUILIBRIUM MASS

Having an efficient procedure to predict the residual mass of the equilibrium phase also opens perspectives to discuss the influence of the parameters. This can provide useful hints for the design and the optimization of anti-tumor therapies. We address this issue by performing a global sensitivity analysis on the immune-controlled tumor mass. Sensitivity analysis also provides information on the quantification of uncertainty in the model output with respect to the uncertainties in the input parameters. We remind the reader that the equilbrium mass is seen as a function of the parameters in **Table 1**:

$$\mu_1 = f(a, A, p, \chi, D, \gamma). \tag{24}$$

We consider that the input parameters are independent random variables uniformly distributed in an interval $[\,x_1, x_2\,] \subset (\,0, \infty\,)$:

$$M = (a, A, p, \chi, D, \gamma) \quad \text{with} \quad M_i \sim U(x_1, x_2)\,. \tag{25}$$

The pillar of the Sobol sensitivity analysis is the decomposition of f into 2^n - 1 summands of increasing dimensions:

$$\begin{aligned} f(M) &= f_0 + \sum_{i=1}^{n} f_i(M_i) \\ &+ \sum_{1 \le i < j \le n} f_{ij}\left(M_i, M_j\right) + \cdots + f_{1\,\ldots\,n}(M_1, \cdots, M_n), \end{aligned} \tag{26}$$

Where

$$\frac{1}{x_2 - x_1} \int_{[x_1, x_2]} f_{i_1 \cdots i_p}\left(M_{i_1 \cdots i_p}\right) \mathrm{d}M_{i_k} = 0 \text{ for } \quad k \in \{1, \ldots, p\}, \tag{27}$$

$$f_0 = \frac{1}{(x_2 - x_1)^n} \int_{[x_1, x_2]^n} f(M) \quad \mathrm{d}M, \tag{28}$$

$$\int_{[x_1, x_2]^n} f_{i_1 \cdots i_p}\left(M_{i_1 \cdots i_p}\right) f_{j_1 \cdots j_p}\left(M_{j_1 \cdots j_p}\right) \quad \mathrm{d}M = 0, \tag{29}$$

and $M_{i_1 \cdots i_p} = (\,M_{i_1}, \cdots M_{i_p}\,)$. The existence and uniqueness of the above decomposition has been proven in (69), given f a square integrable function. Owing to the orthogonality condition (29), the total variance of f reads:

TABLE 7 | Uniform fragmentation, ex. 2, case n = 1: errors for different number of cells.

Number of cells	E_λ	E_V
1000	4.70×10^{-2}	2×10^{-2}
2000	2.43×10^{-2}	1.06×10^{-2}
4000	1.25×10^{-2}	5.5×10^{-3}
8000	6.39×10^{-3}	2.81×10^{-3}

$$\mathcal{V} = \mathrm{Var}(f(M)) = \frac{1}{(x_2 - x_1)^n} \int_{[x_1, x_2]^n} f(M)^2 \quad \mathrm{d}M - f_0^2\,. \tag{30}$$

Given (26), V can be decomposed as follows:

$$\mathcal{V} = \sum_{i=1}^{n} \mathcal{V}_i + \sum_{1 \le i < j \le n} \mathcal{V}_{ij} + \cdots + \mathcal{V}_{1\,\ldots\,n}, \tag{31}$$

where the terms $V_{\,i_1 \cdots i_p}$, called partial variances read:

$$\mathcal{V}_{i_1 \cdots i_p} = \frac{1}{(x_2 - x_1)^n} \int_{[x_1, x_2]^n} f^2_{i_1 \cdots i_p} \quad \mathrm{d}M_{i_1} \cdots \quad \mathrm{d}M_{i_p}\,. \tag{32}$$

Following the description in (69), the Sobol' sensitivity indices are defined as follows:

$$S_{i_1 \cdots i_p} = \frac{\mathcal{V}_{i_1 \cdots i_p}}{V} \tag{33}$$

They verify

$$\sum_{i=1}^{n} S_i + \sum_{1 \le i < j \le n} S_{ij} + \cdots + S_{1\,\ldots\,n} = 1. \tag{34}$$

Each index $S_{\,i_1 \cdots i_p}$ measures how the total variance of f is affected by uncertainties in the set of input parameters $i_1 \cdots i_p$. An equivalent definition of the above indices is given by [see (68)]:

$$\mathcal{V}_i = \mathrm{Var}(\mathbb{E}(Y \mid M_i)),\ \mathcal{V}_{ij} = \mathrm{Var}(\mathbb{E}(Y \mid M_i, M_j)) - \mathcal{V}_i - \mathcal{V}_j, \ldots \tag{35}$$

The total effect of a specific input parameter i is evaluated by the so-called total sensitivity index $S_T^{(i)}$, the sum of the sensitivity indices which contain i:

$$S_T^{(i)} = \sum_{C_i} S_{i_1 \cdots i_p} \tag{36}$$

where $C_i = \{\,(\,i_1 \cdots i_p\,) : \exists m \in \{\,1, \ldots, p\,\},\ i_m = i\,\}$. In practice, the sensitivity indices that are needed to discriminate the impact of the parameters are the first, second and total Sobol' sensitivity indices. The above indices are computed using Monte Carlo simulations combined with a careful sampling of the parameters space in order to reduce the computational load and the number of model evaluations. For this purpose, the following estimators can be derived using two different N samples A and B, see (68, 74),

$$\widehat{f_0} = \frac{1}{N} \sum_{l=1}^{N} f(M_l), \tag{37}$$

$$\hat{\mathcal{V}} = \frac{1}{N} \sum_{l=1}^{N} f^2(M_l) - \hat{f}_0^2, \tag{38}$$

$$\hat{\mathcal{V}}_i = \frac{1}{N} \sum_{l=1}^{N} f\left(M^{(A)}_{(-i)l}, M^{(A)}_{il}\right) f\left(M^{(B)}_{(-i)l}, M^{(A)}_{il}\right) - \hat{f}_0^2, \tag{39}$$

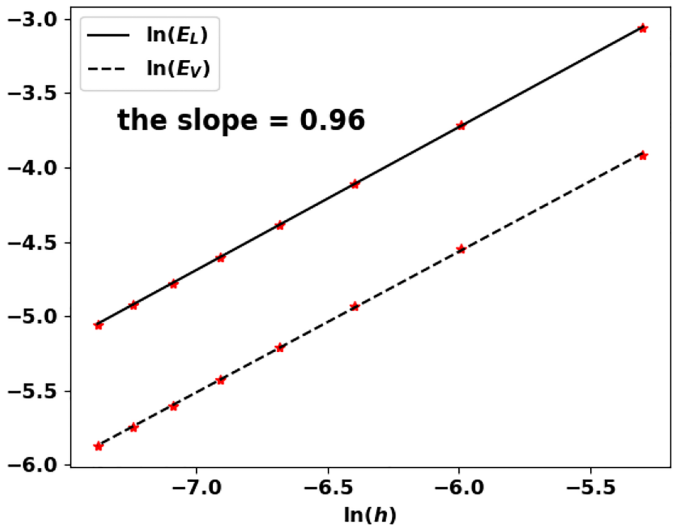


FIGURE 12 | Uniform fragmentation, ex. 2 case $n = 1$: rate of convergence to the exact eigenpair with respect to h.

TABLE 8 | Uniform fragmentation, ex. 2, case $n = 2$: errors for different number of cells.

Number of cells	E_λ	E_V
1000	2.39×10^{-2}	8.81×10^{-2}
2000	1.23×10^{-3}	4.53×10^{-3}
4000	6.41×10^{-3}	2.35×10^{-3}
8000	3.41×10^{-3}	1.24×10^{-3}

$$\hat{\mathcal{V}}_{ij} = \frac{1}{N}\sum_{l=1}^{N} f\left(M^{(A)}_{-(i,j)l}, M^{(A)}_{il}, M^{(A)}_{jl}\right) f\left(M^{(B)}_{-(i,j)l}, M^{(A)}_{il}, M^{(A)}_{jl}\right) - \hat{f}_0^2 - \hat{\mathcal{V}}_i - \hat{\mathcal{V}}_j \tag{40}$$

Here the notation $M_{-(i_1 \cdots i_p)l}$ stands for the l-th sample line where we get rid of the points corresponding to the indices $i_1, \cdots, i_p$. The total sensitivity (87) is given by:

$$S_{T_i} = 1 - S_{-i} \tag{41}$$

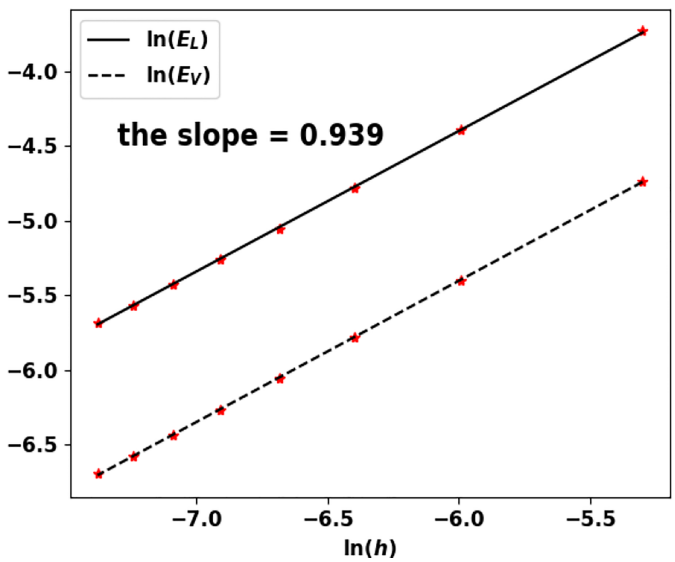


FIGURE 13 | Uniform fragmentation, ex. 2: rate of convergence to the exact eigenpair with respect to h.

where S_{-i} is the sum of all the sensitivity indices that do not contain the index i. Hence, the total sensitivity index estimator reads:

$$\hat{S}_{T_i} = 1 - \frac{\hat{\mathcal{V}}_{-i}}{\hat{\mathcal{V}}} \tag{42}$$

Where

$$\hat{\mathcal{V}}_{-i} = \frac{1}{N}\sum_{l=1}^{N} f\left(M_{(-i)l}^{(A)}, M_{il}^{(A)}\right) f\left(M_{(-i)l}^{(A)}, M_{il}^{(B)}\right) - \hat{f}_0^2 .$$

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/atsoukevin93/tumorgrowth.

AUTHOR CONTRIBUTIONS

Conception and design: KA, VB, and TG; Development of methodology: KA, VB, and TG; Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): FA, VB, and SK; Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): KA, VB, and TG; Writing, review, and/or revision of the manuscript: KA, FA, VB, and TG; Study supervision: FA, VB, and TG. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the French Government (National Research Agency, ANR) through the "Investments for the Future" programs LABEX SIGNALIFE ANR-11-LABX-0028 and IDEX UCAJedi ANR-15-IDEX-01.

ACKNOWLEDGMENTS

The authors acknowledge the support of UCAncer, an incentive Université Côte d'Azur network, which has permitted and encouraged this collaboration. We thank the IPMC's animal house and Imaging/Flow cytometry core facilities for technical support, and Franck Aguila for the artwork of the graphical abstract.

REFERENCES

1. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer Immunoediting: From Immunosurveillance to Tumor Escape. *Nat Immunol* (2002) 3:991–8. doi: 10.1038/ni1102-991
2. Dunn GP, Old LJ, Schreiber RD. The Immunobiology Review of Cancer Immunosurveillance and Immunoediting. *Immunity* (2004) 21:137–48. doi: 10.1016/j.immuni.2004.07.017
3. Koebel CM, Vermi W, Swann JB, Zerafa N, Rodig SJ, Old LJ, et al. Adaptive Immunity Maintains Occult Cancer in an Equilibrium State. *Nature* (2007) 450:903–8. doi: 10.1038/nature06309
4. Boon T, Coulie PG, den Eynde BJV, van der Bruggen P. Human *T*-Cell Responses Against Melanoma. *Annu Rev Immunol* (2006) 24:175–208. doi: 10.1146/annurev.immunol.24.021605.090733
5. Dranoff G. Cytokines in Cancer Pathogenesis and Cancer Therapy. *Nat Rev Cancer* (2004) 4:11–22. doi: 10.1038/nrc1252
6. Rabinovich GA, Gabrilovich D, Sotomayor EM. Immunosuppressive Strategies That are Mediated by Tumor Cells. *Ann Rev Immunol* (2007) 25:267–96. doi: 10.1146/annurev.immunol.25.022106.141609
7. Smyth MJ, Godfrey DI, Trapani JA. A Fresh Look at Tumor Immunosurveillance and Immunotherapy. *Nat Immunol* (2001) 2:293–9. doi: 10.1088/0031-9155/57/2/R1
8. Whiteside TL. Immune Suppression in Cancer: Effects on Immune Cells, Mechanisms and Future Therapeutic Intervention. *Semin Cancer Biol* (2006) 16:3–15. doi: 10.1016/j.semcancer.2005.07.008
9. Phan TG, Croucher PI. The Dormant Cancer Cell Life Cycle. *Nature* (2020) 20:398–411. doi: 10.1038/s41568-020-0263-0
10. Atsou K, Anjuère F, Braud VM, Goudon T. A Size and Space Structured Model Describing Interactions of Tumor Cells With Immune Cells Reveals Cancer Persistent Equilibrium States in Tumorigenesis. *J Theor Biol* (2020) 490:110163. doi: 10.1016/j.jtbi.2020.110163
11. Kirschner DE, Panetta JC. Modeling Immunotherapy of the Tumor-Immune Interaction. *J Math Biol* (1998) 37:235–52. doi: 10.1007/s002850050127
12. Konstorum A, Vella AT, Adler AJ, Laubenbacher RC. Addressing Current Challenges in Cancer Immunotherapy With Mathematical and Computational Modelling. *J R Soc Interface* (2017) 14:20170150. doi: 10.1098/rsif.2017.0150
13. Lai X, Carson WE, Stiff A, Friedman A, Wesolowski R, Duggan M. Modeling Combination Therapy for Breast Cancer With BET and Immune Checkpoint Inhibitors. *Proc Nat Acad Sc* (2018) 115:5534–9. doi: 10.1073/pnas.1721559115
14. de Pillis LG, Radunskaya AE, Wiseman CL. A Validated Mathematical Model of Cell-Mediated Immune Response to Tumor Growth. *Cancer Res* (2005) 65:7950–8. doi: 10.1158/0008-5472.CAN-05-0564
15. de Pillis L, Radunskaya A. The Dynamics of an Optimally Controlled Tumor Model: A Case Study. *Math Comput Model* (2003) 37:1221–44. doi: 10.1016/S0895-7177(03)00133-X
16. Kuznetsov V, Knott G. Modelling Tumor Regrowth and Immunotherapy. *Math Comput Model* (2001) 33:1275–87. doi: 10.1016/S0895-7177(00)00314-9
17. Li H, Wang S, Xu F. Dynamical Analysis of Tumor-Immune-Help T Cells System. *Int J Biomath* (2019) 12:1950075. doi: 10.1142/S179352451950075X
18. Robertson-Tessi M, El-Kareh A, Goriely A. A Mathematical Model of Tumor–Immune Interactions. *J Theor Biol* (2012) 294:56 – 73. doi: 10.1016/j.jtbi.2011.10.027
19. Wilkie KP, Hahnfeldt P. Modeling the Dichotomy of the Immune Response to Cancer: Cytotoxic Effects and Tumor-Promoting Inflammation. *Bull Math Biol* (2017) 79:1426–48. doi: 10.1007/s11538-017-0291-4
20. Chaplain MA, Kuznetsov VA, James ZH, Davidson F, Stepanova LA. Spatio-Temporal Dynamics of the Immune System Response to Cancer. In: MA Horn, S Gieri and GF Webb, editors. *Mathematical Models in Medical and Health Science*. Vanderbilt, TN, USA: Vanderbilt University Press (1998). International Conference on Mathematical Models in Medical and Health Sciences.
21. Eftimie R. Investigation Into the Role of Macrophages Heterogeneity on Solid Tumour Aggregations. *Math Biosc* (2020) 322:108325. doi: 10.1016/j.mbs.2020.108325
22. Kuznetsov VA, Makalkin IA, Taylor MA, Perelson AS. Nonlinear Dynamics of Immunogenic Tumors: Parameter Estimation and Global Bifurcation Analysis. *Bull Math Biol* (1994) 56:295–321. doi: 10.1007/BF02460644

23. Matzavinos A, Chaplain MAJ, Kuznetsov VA. Mathematical Modelling of the Spatio-Temporal Response of Cytotoxic *T*-Lymphocytes to a Solid Tumour. *Math Med Biol* (2004) 21:1–34. doi: 10.1093/imammb/21.1.1
24. de Pillis LG, Mallet DG, Radunskaya AE. Spatial Tumor-Immune Modeling. *Comput Math Methods Med* (2006) 7:159–76. doi: 10.1080/10273660600968978
25. Eftimie R, Bramson JL, Earn DJD. Interactions Between the Immune System and Cancer: A Brief Review of non-Spatial Mathematical Models. *Bull Math Biol* (2011) 73:2–32. doi: 10.1007/s11538-010-9526-3
26. Eftimie R, Gillard JJ, Cantrell DA. Mathematical Models for Immunology: Current State of the Art and Future Research Directions. *Bull Math Biol* (2017) 78:2091–134. doi: 10.1007/s11538-016-0214-9
27. Eladdadi A, Kim P, Mallet D. *Mathematical Models of Tumor-Immune System Dynamics, Springer Proceedings in Math. & Statistics* Vol. 107. SwitzerlandSpringer (2014).
28. Mahlbacher GE, Reihmer KC, Frieboesa HB. Mathematical Modeling of Tumor-Immune Cell Interactions. *J Theor Biol* (2019) 469:47–60. doi: 10.1016/j.jtbi.2019.03.002
29. Roose T, Chapman SJ, Maini PK. Mathematical Models of Avascular Tumor Growth. *SIAM Rev* (2007) 49:179–208. doi: 10.1137/S0036144504446291
30. Bellouquid A, Delitala M. Mathematical Methods and Tools of Kinetic Theory Towards Modelling Complex Biological Systems. *Math Mod Meth Appl Sci* (2005) 15:1639–66. doi: 10.1142/S0218202505000923
31. Bekkal Brikci F, Clairambault J, Ribba B, Perthame B. An Age-and-Cyclin-Structured Cell Population Model for Healthy and Tumoral Tissues. *J Math Biol* (2008) 57:91–110. doi: 10.1007/s00285-007-0147-x
32. Doumic-Jauffret M, Gabriel P. Eigenelements of a General Aggregation-Fragmentation Model. *Math Models Methods Appl Sci* (2010) 20:757–83. doi: 10.1142/S021820251000443X
33. Michel P, Mischler S, Perthame B. General Relative Entropy Inequality: An Illustration on Growth Models. *J Math Pures Appl* (2005) 84:1235–60. doi: 10.1016/j.matpur.2005.04.001
34. Michel P. Existence of a Solution to the Cell Division Eigenproblem. *Models Math Meth App Sci* (2006) 16:1125–53. doi: 10.1142/S0218202506001480
35. Michel P. *Principe D'entropie Généralisée Et Dynamique De Populations Structurées. Ph.D. Thesis, Paris Dauphine*. (2005).
36. Perthame B. Transport Equations in Biology. *Front Math (Birkhauser)* (2007). doi: 10.1007/978-3-7643-7842-4
37. Preziosi L. *Modeling Cancer Growth*. Chapman Hall: CRC-Press (2003).
38. Alzahrani EO, Kuang Y. Nutrient Limitations as an Explanation of Gompertzian Tumor Growth. *Disc Cont Dyn Syst B* (2016) 21:357–72. doi: 10.3934/dcdsb.2016.21.357
39. Gyllenberg M, Webb GF. A Nonlinear Structured Population Model of Tumor Growth With Quiescence. *J.Math Biol* (1990) 28:671–94. doi: 10.1007/BF00160231
40. Iwata K, Kawasaki K, Shigesada N. A Dynamical Model for the Growth and Size Distribution of Multiple Metastatic Tumors. *J Theor Biol* (2000) 203:177–86. doi: 10.1006/jtbi.2000.1075
41. Benzekry S, Lamont C, Beheshti A, Tracz A, Ebos J, Hlatky L, et al. Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth. *PloS Comput Biol* (2014) 10:e1003800. doi: 10.1371/journal.pcbi.1003800
42. Atsou K, Anjuère F, Braud VM, Goudon T. A Size and Space Structured Model of Tumor Growth Describes a Key Role for Protumor Immune Cells in Breaking Equilibrium States in Tumorigenesis. *PloS One* (2021), 0259291. doi: 10.1371/journal.pone.0259291
43. Baccelli F, McDonald D, Reynier J. A Mean Field Model for Multiple TCP Connections Through a Buffer Implementing RED. *Perform Eval* (2002) 49:77–97. doi: 10.1016/S0166-5316(02)00136-0
44. Perthame B, Ryzhik L. Exponential Decay for the Fragmentation or Cell-Division Equation. *J Differ Eq* (2005) 210:155–77. doi: 10.1016/j.jde.2004.10.018
45. Faget J, Groeneveld S, Boivin G, Sankar M, Zangger N, Garcia M, et al. Neutrophils and Snail Orchestrate the Establishment of a Pro-Tumor Microenvironment in Lung Cancer. *Cell Rep* (2017) 21:3190–204. doi: 10.1016/j.celrep.2017.11.052
46. Li CY, Shan S, Huang Q, Braun RD, Lanzen J, Hu K, et al. Initial Stages of Tumor Cell-Induced Angiogenesis: Evaluation *via* Skin Window Chambers in Rodent Models. *J Nat Cancer Inst* (2000) 92:143–7. doi: 10.1093/jnci/92.2.143
47. Goudon T. Mathematics for Modeling and Scientific Computing. *Math Stat (Wiley-ISTE)* (2016) 27(3):665–679. doi: 10.1002/9781119371120
48. Chatelin F. Convergence of Approximation Methods to Compute Eigenelements of Linear Operations. *SIAM J Numer Anal* (1973) 10:939–48. doi: 10.1137/0710080
49. Chatelin F. The Spectral Approximation of Linear Operators With Applications to the Computation of Eigenelements of Differential and Integral Operators. *SIAM Rev* (1981) 23:495–522. doi: 10.1137/1023099
50. Farrell BE, Daniele RP, Lauffenburger DA. Quantitative Relationships Between Single-Cell and Cell-Population Model Parameters for Chemosensory Migration Responses of Alveolar Macrophages to C5a. *Cell Motil* (1990) 16:279–93. doi: 10.1002/cm.970160407
51. Friedman A, Hao W. The Role of Exosomes in Pancreatic Cancer Microenvironment. *Bull Math Biol* (2018) 80:1111–33. doi: 10.1007/s11538-017-0254-9
52. Yates A, Callard R. Cell Death and the Maintenance of Immunological Memory. *Disc Cont Dyn Syst.-B* (2001) 1:43–59. doi: 10.3934/dcdsb.2001.1.43
53. Beck RJ, Slagter M, Beltman JB. Contact-Dependent Killing by CytotoxicTLymphocytes is Insufficient for EL4 Tumor Regression *In Vivo*. *Cancer Res* (2019) 79:3406–16. doi: 10.1158/0008-5472.CAN-18-3147
54. Cazaux M, Grandjean CL, Lemaître F, Garcia Z, Beck RJ, Milo I, et al. Single-Cell Imaging of CAR *T*-Cell Activity *In Vivo* Reveals Extensive Functional and Anatomical Heterogeneity. *J Exp Med* (2019) 216:1038–49. doi: 10.1084/jem.20182375
55. Hanson HL, Donermeyer DL, Ikeda H, White JM, Shankaran V, Old LJ, et al. Eradication of Established Tumors by CD8+ T-Cell Adoptive Immonutherapy. *Cell Press* (2000) 13:265–76. doi: 10.1016/s1074-7613(00)00026-1
56. Nolz JC, Hill AB. Strength in Numbers: Visualizing CTL-Mediated Killing *In Vivo*. *Immunity* (2016) 44:207–8. doi: 10.1016/j.immuni.2016.01.026
57. Kwok CS, Cole SE, Liao SK. Uptake Kinetics of Monoclonal Antibodies by Human Malignant Melanoma Multicell Spheroids. *Cancer Res* (1988) 48:1856–63.
58. Cairns CM, Gordon JR, Li F, Baca-Estrada ME, Moyana T, Xiang J. Lymphotactin Expression by Engineered Myeloma Cells Drives Tumor Regression: Mediation by CD4+ and CD8+ T-cells and neutrophils expressing XCR1 receptor. *J Immun* (2001) 167:57–65. doi: 10.4049/jimmunol.167.1.57
59. Khou S, Popa A, Luci C, Bihl F, Meghraoui-Kheddar A, Bourdely P, et al. Tumor-Associated Neutrophils Dampen Adaptive Immunity and Promote Cutaneous Squamous Cell Carcinoma Development. *Cancers* (2020) 10-12: E1860. doi: 10.3390/cancers12071860
60. Samson A, Lavielle M, Mentré F. Extension of the SAEM Algorithm to Left-Censored Data in Nonlinear Mixed-Effects Model: Application to HIV Dynamics Model. *Comput Stat Data Anal* (2006) 51:1562–74. doi: 10.1016/j.csda.2006.05.007
61. *[Dataset] Monolix Version 2020R1*. Antony, france: Lixoft sas (2020).
62. Weise K, Poßner L, Müller E, Gast R, Knösche TR. Pygpc: A Sensitivity and Uncertainty Analysis Toolbox for Python. *SoftwareX* (2020) 11:100450. doi: 10.1016/j.softx.2020.100450
63. Herman J, Usher W. SALib: An Open-Source Python Library for Sensitivity Analysis. *J Open Source Softw* (2017) 2:joss00097. doi: 10.21105/joss.00097
64. Shashni B, Ariyasu S, Takeda R, Suzuki T, Shiina S, Akimoto K, et al. Size-Based Differentiation of Cancer and Normal Cells by a Particle Size Analyzer Assisted by a Cell-Recognition Pc Software. *Biol Pharm Bull* (2018) 41:487–503. doi: 10.1248/bpb.b17-00776
65. Moore GE, Sandberg AA, Watne AL. The Comparative Size and Structure of Tumor Cells and Clumps in the Blood, Bone Marrow, and Tumor Imprints. *Cancer* (1960) 13:111–7. doi: 10.1002/1097-0142(196001/02)13:1<111::aid-cncr2820130121>3.0.co;2-y
66. Erdi YE. Limits of Tumor Detectability in Nuclear Medicine and PET. *Mol Imaging Radionuclid Ther* (2012) 21:23–8. doi: 10.4274/Mirt.138
67. Harenberg D, Marelli S, Sudret B, Winschel V. Uncertainty Quantification and Global Sensitivity Analysis for Economic Models. *Quanti Econom* (2019) 10:1–41. doi: 10.3982/qe866

68. Saltelli A. Sensitivity Analysis of Model Output. An Investigation of New Techniques. *Comput Stat Data Anal* (1993) 15:211–38. doi: 10.1016/0167-9473(93)90193-W
69. Sobol' I. Sensitivity Estimates for Nonlinear Mathematical Models. *Math Modell Comput* (1993) 1:407–14.
70. Crestaux T, Le Maître O, Martinez JM. Polynomial Chaos Expansion for Sensitivity Analysis. *Reliabil Eng Syst Saf* (2009) 94:1161–72. doi: 10.1016/j.ress.2008.10.008
71. Sudret B. Global Sensitivity Analysis Using Polynomial Chaos Expansions. *Reliabil Eng Syst Saf* (2008) 93:964–79. doi: 10.1016/j.ress.2007.04.002
72. Ernst OG, Mugler A, Starkloff HJ, Ullmann E. On the Convergence of Generalized Polynomial Chaos Expansions. *ESAIM: Math Model Numeric Anal* (2012) 46:317–39. doi: 10.1051/m2an/2011045
73. Le Gratiet L, Marelli S, Sudret B. Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes. In: R Ghanem, D Higdon and H Owhadi, editors. *Handbook of Uncertainty Quantification*. Cham: Springer International Publishing (2017). p. 1289–325. doi: 10.1007/978-3-319-12385-1_38
74. Saltelli A. Making Best Use of Model Evaluations to Compute Sensitivity Indices. *Comput Phys Commun* (2002) 145:280 – 297. doi: 10.1016/S0010-4655(02)00280-1
75. Bailly C, Thuru X, Quesnel B. Combined Cytotoxic Chemotherapy and Immunotherapy of Cancer: Modern Times. *NAR Cancer* (2020) 2:1–20. doi: 10.1093/narcan/zcaa002
76. Galon J, Bruni D. Approaches to Treat Immune Hot, Altered and Cold Tumours With Combination Immunotherapies. *Nat Rev Drug Discovery* (2019) 18:197–218. doi: 10.1038/s41573-018-0007-y
77. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science* (2006) 313:1960–4. doi: 10.1126/science.1129139
78. Sharma P, Allison J. The Future of Immune Checkpoint Therapy. *Science* (2015) 348:56–61. doi: 10.1126/science.aaa8172
79. Champiat S, Tselikas L, Farhane S, Raoult T, Texier M, Lanoy E, et al. Intratumoral Immunotherapy: From Trial Design to Clinical Practice. *Clin Cancer Res* (2020) 27(3):665–79. doi: 10.1158/1078-0432.CCR-20-0473
80. Kaufman HL, Atkins MB, Subedi P, Wu J, Chambers J, Mattingly TJ, et al. The Promise of Immuno-Oncology: Implications for Defining the Value of Cancer Treatment. *J ImmunoTher Cancer* (2019) 7:129–40. doi: 10.1186/s40425-019-0594-0
81. Shekarian T, Valsesia-Wittmann S, Caux C, Marabelle A. Paradigm Shift in Oncology: Targeting the Immune System Rather Than Cancer Cells. *Mutagenesis* (2015) 30:205–11. doi: 10.1093/mutage/geu073
82. Wang H, Wang S, Huang M, Liang X, Tang Y, Tang Y. Targeting Immune-Mediated Dormancy: A Promising Treatment of Cancer. *Front Oncol* (2019) 9:498:1–9. doi: 10.3389/fonc.2019.00498
83. Correia AL, Guimaraes JC, der Maur PA, Silva DD, Trefny MP, Okamoto R, et al. Hepatic Stellate Cells Suppress NK Cell-Sustained Breast Cancer Dormancy. *Nature* (2021) 594:566–92. doi: 10.1038/s41586-021-03614-z
84. Lopes N, Vivier E. Natural Killer Cells Lull Tumours Into Dormancy. *Nature* (2021) 594:501–2. doi: 10.1038/d41586-021-01381-5
85. Krein M, Rutman M. Linear Operator Leaving Invariant a Cone in a Banach Space. *Amer Math Soc Transl* (1962) 10:199–325.
86. Serre D. Matrices: Theory and Applications. In: *Graduate Texts in Math*, vol. 216. New-York, Dordrecht, Heidelberg, London: Springer (2002).
87. Homma T, Saltelli A. Importance Measures in Global Sensitivity Analysis of Nonlinear Models. *Reliabil Eng Syst Saf* (1996) 52:1 – 17. doi: 10.1016/0951-8320(96)00002-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Atsou, Khou, Anjuère, Braud and Goudon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Immunology

ORIGINAL RESEARCH
published: 30 June 2022
doi: 10.3389/fimmu.2022.924498

Exploration of Novel Immunological Terms in Lung Cancer With Large Populations: Implications for Immunotherapy

*Yuanshan Yao†, Jing Wang†, Fuzhi Yang† and Wen Gao**

Department of Thoracic Surgery, Shanghai Key Laboratory of Clinical Geriatric Medicine, HuaDong Hospital Affiliated to Fudan University, Shanghai, China

Background: Ideal biomarkers to predict the response to immunotherapy in lung cancer are still lacking. Therefore, there is a need to explore effective biomarkers in large populations.

Objective: The objective of this study is to explore novel immunological classifications that are associated with immunotherapy response through the ssGSEA algorithm.

Methods: Six independent lung cancer cohorts were collected for analysis including The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), and the EMBL-EBI database. The ssGSEA algorithm was performed to extract immune terms. Then, TCGA samples were involved as a training group and other cohorts were used as a validation group. After LASSO and Cox regression, prognostic associated immune terms were extracted and an immune-related risk score (IRS) signature was constructed. Furthermore, the association between IRS signature and clinical data, genome features, stemness indices analysis, tumor immune microenvironment, immunotherapy efficiency, and targeted therapy response was also investigated.

Results: A total of 1,997 samples were enrolled in this study including six large lung cancer cohorts. Fifty-four immune terms were calculated through the ssGSEA algorithm in combined cohorts. Then, a nine-immune-term risk score model named IRS signature was established to predict the prognosis in combined cohorts. We classified patients into high-risk and low-risk subgroups according to the cutoff point. Subsequently, analysis of clinical data and genome features indicated that the patients in the high-IRS group tend to have advanced clinical features (clinical stage and T classification), as well as a higher level of copy number variation burden, higher tumor burden mutation, and higher tumor stemness indices. Immune landscape analysis demonstrated that high-IRS groups exhibited lower immune cell infiltration and immune-suppressive state. More importantly, the predicted result of the Tumor Immune Dysfunction and Exclusion analysis showed that high-IRS groups might be more insensitive to immunotherapy. Meanwhile, we have also identified that high-IRS groups were associated with better efficiency of several targeted drugs.

OPEN ACCESS

Edited by:
Dinler Amaral Antunes, University of Houston, United States

Reviewed by:
Jin-Zhou Xu, Huazhong University of Science and Technology, China
Mustafa Elhag, University of Seychelles, Seychelles

****Correspondence:***
Wen Gao
gaowen1232021@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *20 April 2022*
Accepted: *08 June 2022*
Published: *30 June 2022*

Citation:
Yao Y, Wang J, Yang F and Gao W (2022) Exploration of Novel Immunological Terms in Lung Cancer With Large Populations: Implications for Immunotherapy. Front. Immunol. 13:924498. doi: 10.3389/fimmu.2022.924498

Conclusion: To summarize, we identified a novel IRS model based on nine immune terms, which was quantified by the ssGSEA algorithm. This model had good efficacy in predicting overall survival and immunotherapy response in non-small cell lung cancer patients, which might be an underlying biomarker.

Keywords: NSCLC, immune terms, IRS signature, prognosis, immunotherapy

INTRODUCTION

Nowadays, immunotherapy has significantly improved the prognosis of advanced non-small cell lung cancer (NSCLC) (1). Although PD1/L1 inhibitors have changed the treatment landscape of NSCLC, the optimal biomarker to predict the clinical response of immunotherapy is still lacking (2). PD-L1 expression level and tumor mutational burden (TMB) in tumor specimens are two reliable biomarkers in clinical practice so far (3). However, a proportion of PD-L1-negative patients can also benefit from PD1/L1 inhibitors. Thus, it is urgent to explore other effective biomarkers with greater accuracy.

Recently, there are various biomarkers to predict immunotherapy response to lung cancer patients derived from the tumor immune microenvironment, molecular alterations, and serum indexes (4). For example, tumor-infiltrating lymphocytes (T cells and B cells) and the spatial location of these immune cells were already validated to be associated with the benefit of immunotherapy (5, 6). However, controversy still exists. Colt et al. demonstrated that penetration of immune cells into the cancer cells represented a better prognosis compared to those in tumor stroma (7). Nevertheless, Paul et al. revealed that immune cells that were located at the tumor margin exhibited a better response to immunotherapy than those in the stroma or intra-tumor (8). In addition, the prognosis of patients who received immune checkpoint inhibitors was associated with a non-invasive source of pre-treatment serum NLR (neutrophil-to-lymphocyte ratio) (9, 10). Currently, there is still no single satisfactory biomarker to evaluate the efficacy of immunotherapy. Incorporating multiple methods may be the best way (11).

In recent years, immune-related genes have also attracted attention. T cell-inflamed gene expression profile, which consisted of 18 IFN-r associated genes, was also explored and confirmed to exert its clinical prediction value in various types of malignancies (12). Johnson et al. demonstrated that patients with higher HLA-DR obtained better clinical response, prognosis, and progression-free survival compared with lower HLA-DR patients (13). Nevertheless, little is known about the correlation between immune terms and immunotherapy response in NSCLC patients.

Meanwhile, two atypical patterns of treatment responses are particularly correlated with immunotherapy including pseudoprogression and hyperprogression, which occupy a certain amount of patients accepting immunotherapy (14). The underlying molecular mechanism remains unknown so far. In clinical applications even using the somewhat improved immune RECIST (iRECIST) criteria, it is probable that the treatment effect of several patients is misjudged, which could lead to loss of optimal treatment opportunities (15). Patients with improved or stable clinical symptoms may predict effective immunotherapy. Future work is warranted to refine imaging response criteria and explore potential biomarkers that can make treatment recommendations more clear and standardized.

In this study, we collected the expression data from the publicly accessible dataset with 1,997 NSCLC patients. Then, 54 immune terms were extracted through single-sample gene set enrichment analysis (ssGSEA). We identified nine immune terms with prognostic values and constructed an immune-related risk score (IRS) signature. Moreover, the IRS signature can be used as a biomarker to predict the efficacy of immunotherapy. Meanwhile, several targeted drugs were potential candidates for targeting this IRS signature. We hope that our study can provide a reference for the treatment of NSCLC.

MATERIALS AND METHODS

Data Collection and Preprocessing

Six cohorts with complete clinical information and available expression matrix data were collected (TCGA, GSE37745, GSE50081, GSE68465, GSE73403, and E-GEOD-30219) after a comprehensive search in this study. TCGA-LUAD data and TCGA-LUSC data were retrieved from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). In detail, the expression profile data from TCGA were originally in "FPKM" form and then converted to "TPM" form for higher comparability with microarray data. E-GEOD-30219 data were retrieved from the ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/). In addition, four individual Gene Expression Omnibus (GEO) datasets were identified from the GEO dataset (https://www.ncbi.nlm.nih.gov/geo/), including GSE37745 (GPL570), GSE50081 (GPL570), GSE68465 (GPL96), and GSE73403 (GPL6480). Three packages in the R environment, namely, sva, cluster, and oompaBase, were used to combine data and reduce the likelihood of batch effects and magnitude harmonization. The codes used were uploaded in the figshare website (https://figshare.com/articles/software/Code/19995041).

Single-Sample Gene Set Enrichment Analysis and Pathway Enrichment Analysis

The R package "clusterProfiler" was used to perform gene set enrichment analysis (GSEA), Gene Oncology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. The R

package GSVA was used to conduct ssGSEA to evaluate the enrichment scores of 54 immune terms. The Hallmark gene set (MSigDB) was selected as a reference set to explore the difference in the oncogenetic pathways. Immune-related features used for ssGSEA quantification were collected from Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/panimmune) and previous studies, which was available at figshare (https://figshare.com/articles/dataset/Untitled_Item/19368023).

Calculation of Immune-Related Risk Score Signature for the Combined Patient Cohort

After removing batch effects, TCGA-LUAD and TCGA-LUSC cohorts were selected as the training cohort and other datasets were used as the validation cohort (GSE37745, GSE50081, GSE68465, GSE73403, and E-GEOD-30219). For the training cohort, univariate Cox analysis was performed to identify prognosis-related immune terms with the criteria p-value <0.05. Then, L1-penalized (LASSO) estimation across 1,000 iterations was used for dimension reduction. Immune terms with a frequency higher than 50 times were then selected for multivariate Cox regression analysis and IRS calculation using the following formula: Immune risk score = term1*coef1 + term2*coef2 + term3*coef3 + … + termN*coefN. R packages "SimDesign" and "tdROC" were used to get the best cutoff value of risk score in training and validation cohorts. The R packages "survival" and "survivalROC" were used to assess the prognostic value of the signature through Kaplan–Meier survival curve and the ROC curve.

Features of Tumor Genomics Between Two IRS Groups

TMB means the number of mutations per megabase (mt/Mb). Somatic mutation data retrieved from the cBioPortal website (http://www.cbioportal.org/datasets, Lung Adenocarcinoma/Lung Squamous Cell Carcinoma, TCGA, PanCancer Atlas) was used to calculate TMB. The R package "maftools" was used to analyze significantly mutated genes with p-value < 0.05 between the two IRS groups and the interaction effect of gene mutations. GISTIC_2.0 (https://cloud.genepattern.org) was applied for copy number variation (CNV) analysis. Based on the output files from GISTIC_2.0, copy number gain burden and loss burden at the focal and arm levels were calculated. One-class logistic regression machine learning (OCLR) algorithm was used to quantify the tumor stemness index, including mRNAsi, mDNAsi, EREG-mRNAsi, and EREG-mDNAsi, whose process is available on https://bioinformaticsfmrp.github.io/PanCanStem_Web/. ESTIMATEScore, StromalScore, and ImmuneScore of each sample were calculated using the R package "ESTIMATE". A higher score represents a more considerable amount of the corresponding lower tumor purity and higher stromal and immune cells in TMB.

Prediction of Immunotherapeutic and Chemotherapeutic Response

The Tumor Immune Dysfunction and Exclusion (TIDE, http://tide.dfci.harvard.edu) algorithm was used to predict patient immunotherapy response. Each patient with a TIDE score <−0.2 was defined as a Responder, and a patient with a TIDE score >0.2 was defined as a No-responder. Subclass mapping algorithm (https://cloud.genepattern.org/) was used to assess similarities in response to immunotherapies between the 47 patients who responded to immunotherapies and two lung cancer patients from the IRS group. The response to chemotherapy drugs of patients was predicted based on the public pharmacogenomics database Genomics of Drug Sensitivity in Cancer (GDSC, https://www.cancerrxgene.org/), and the R package "pRRophetic" was used to estimate the half-maximal inhibitory concentration (IC_{50}) of twelve commonly used chemo drugs for predicting the sensitivity of chemotherapy drugs.

Statistical Analysis

All statistical analyses were performed with R version 4.0.2 and its appropriate packages. All statistics were two-sided and statistical significance was defined as p-value < 0.05. Data were analyzed with standard statistical tests as appropriate. Spearman correlation was used to estimate the correlations between continuous variables. Independent sample t-test was used to compare continuous variables with normal distribution. Wilcoxon rank-sum test was used to compare continuous variables with skewed distribution.

RESULTS

Immune Terms' Quantification

The flowchart of the whole study is shown in **Figure 1**. After a comprehensive search of the public database, six individual NSCLC cohorts met our criteria and were finally included in our analysis (TCGA, GSE73403, GSE68465, GSE50081, GSE37745, and E-GEOD-30219). A significant batch difference was observed between these cohorts (**Figure 2A**, comp1: 75.7% variance, comp2: 3.9% variance). Then, the sva package was used to remove the batch effect of these six NSCLC cohorts (**Figure 2B**, comp1: 8.9% variance, comp2: 6.2% variance). The combined expression profile was then used for the 54 immune terms' quantification based on the ssGSEA algorithm, which is shown in **Figure 2C**.

Identification of the Prognosis-Related Immune Terms

To further identify the immune terms associated with patient prognosis, we first performed a univariate Cox analysis with the threshold of $p < 0.05$. Then, iterative LASSO regression was used for high-frequency features selection and 18 prognosis-related terms were retained (**Figures 3A, B**). Nine immune terms were selected for prognosis model construction based on the multivariate Cox analysis, including B-cell receptor score, CD8 T cells, IFNG score, IL13 score, Inflammation promoting, Lymphs PCA, SERUM RESPONSE UP, Tfh cells, and Tgd cells (**Figure 3C**). Among these, IFNG score, SERUM RESPONSE UP, CD8 T cells, and IL13 score were risk factors, yet Tgd cells, Inflammation promoting, Lymphs PCA, Tfh cell,

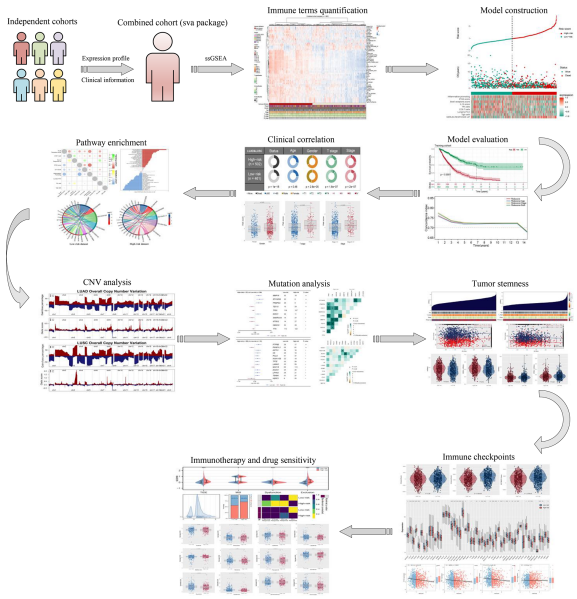


FIGURE 1 | The flowchart of the whole study was shown in **Figure 1**.

and B-cell receptor score were protective factors (**Figure 3D** and **Figure S1A**). The model IRS was calculated with the following formula: "IRS = B cell receptor score * −0.115 + CD8 T cell * 0.148 + SERUM RESPONSE UP * 0.225 + IFNG score * 0.265 + IL13 score * 0.142 + Inflammation promoting * −0.249 + Lymphs PCA * −0.135 + Tfh cells * −0.128 + Tgd cells * −0.161". SimDesign and the tdROC package were used to calculate the optimum cutoff values, and the result showed that the 0.98 and 1.38 were the best cutoff of training cohort and validation cohort, respectively (**Figures S1B,C**). In the training cohort, more death cases could be observed in the high-risk group (**Figure 3E**). The ROC curve indicated a satisfactory prediction efficiency of our model on patients' OS (**Figure 3F**, 3-year AUC: 0.781, 5-year AUC: 0.780, 8-year AUC: 0.767). Moreover, the Kaplan–Meier survival curve showed that the patients in the high-risk group have a worse prognosis (**Figure 3G**). Meanwhile, time-dependent AUC showed that the risk score combined clinical features tend to have a better prediction efficiency than risk score (**Figure 3H**). The same trend and great prediction efficiency were also observed in the validation cohort (**Figures 3I-L**). Univariate and multivariate Cox analysis showed that the IRS is a risk factor independent of other clinical features (**Figure S4**).

Clinical Correlation and Biological Role of the IRS Model

We further explored the clinical correlation and biological role of the IRS model to explain its prognosis effect. The result showed that patients in the high-risk group might have adverse clinical features, including T classification and clinical stage (**Figures 4A-D**). Interestingly, the male patients might have a higher IRS than the female patients, indicating their underlying immune microenvironment difference (**Figure 4B**). The co-expression relationship of nine model immune terms was visualized as a correlation coefficient heatmap to further explore their interactions (**Figure 4E**). GSEA demonstrated that in the high-risk group, the pathway of bile acid metabolism, KRAS signaling, fatty acid metabolism, xenobiotic metabolism, interferon-gamma response, inflammatory response, IL2-STAT5 signaling, and IL6-JAK-STAT3 signaling were activated, yet the hypoxia, P53 pathway, and DNA repair were downregulated (**Figure 4F** and **Figure S2**). GO and KEGG enrichment analysis showed that in the low-risk group, the terms post-translational protein modification (GO:0043687), protein–lipid complex remodeling (GO:0034368), endoplasmic reticulum lumen (GO:0005788), blood microparticle (GO:0072562), lipase inhibitor activity (GO:0055102), immunoglobulin binding (GO:0019865), cholesterol metabolism (hsa04979), and fat digestion and absorption (hsa04975) were enriched (**Figure 4G**). In the high-risk group, the terms cornification (GO:00702680), antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844), intermediate filament (GO:0005882), intermediate filament cytoskeleton (GO:0045111), G protein-coupled receptor binding (GO:0001664), *Staphylococcus aureus* infection (hsa05150), and Wnt signaling pathway (has04310) were enriched (**Figure 4H**).

The Correlation Between Genomic Features and IRS in NSCLC

The overview of gene mutation of TCGA-NSCLC is shown in **Figure S3**. Then, the copy number profile in TCGA-LUAD and TCGA-LUSC patients, including gain/loss percentage and gistic score, was characterized (**Figures 5A–D**). The result showed that high-risk patients had a higher degree of genomic instability

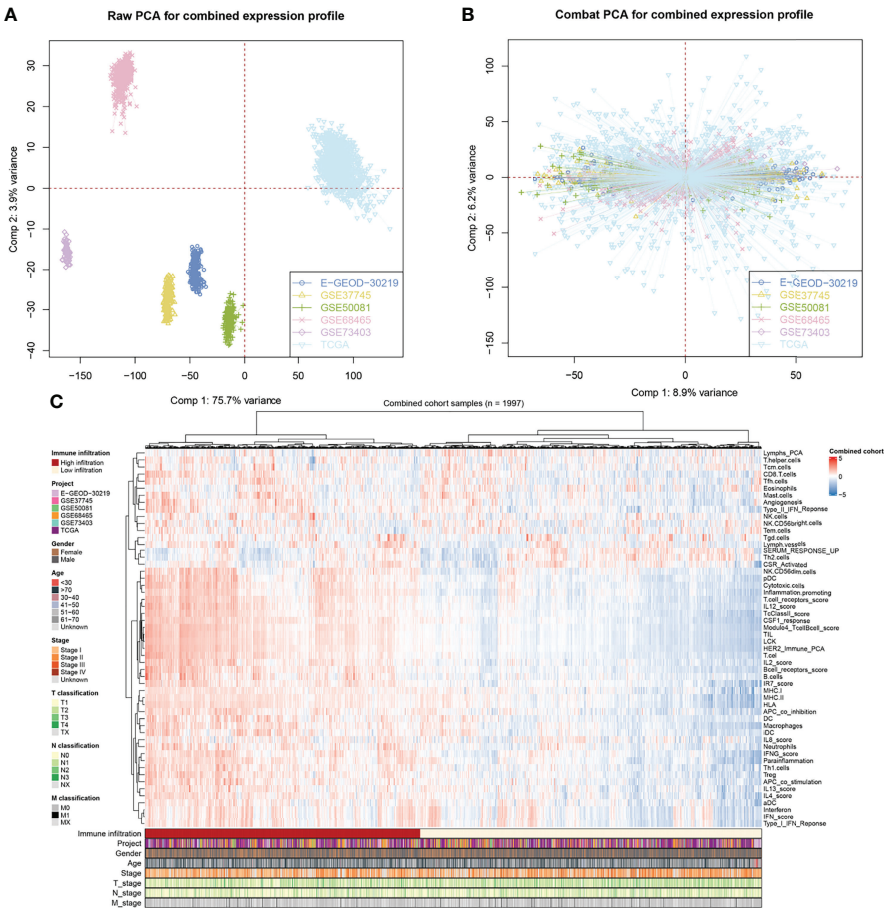


FIGURE 2 | Combination of NSCLC cohort and quantification of immune terms. **(A, B)** The sva package was used for reduce the batch difference of individual cohort, including TCGA, GSE73403, GSE68465, GSE50081, GSE37745 and E-GEOD-30219; **(C)** ssGSEA algorithm was used to quantify the 53 immune terms based on the transcript data.

(**Figures 5E-H**). In detail, CNV burden analysis showed that high-risk patients had a higher level of focal and broad CNV burden in both gain and loss aspects (**Figures 5F-H**). The overview of the tumor mutation landscape of NSCLC is shown in **Figure 6A**. Moreover, higher TMB was also observed in the high-risk group (**Figure 6B**, $p = 0.021$). Furthermore, considering the intrinsic biological differences, we separately explored the mutant gene in TCGA-LUAD and TCGA-LUSC cohorts. The result showed that in the TCGA-LUAD cohort, TEX15, TP53, SMARCA4, NTRK3, and TTN were the most common mutated genes in the high-risk group, while the MMP16, ZFYVE26, FRMPD4, DIDO1, and OBSCN were the opposite (**Figure 6C**). In the TCGA-LUSC cohort, PKHD1L1, MUC16, and HERC1 were the most common mutated genes in the high-risk group, while the PTPRB, ASTN1, C6, PCLO, KIAA1109, TPTE, LAMA2, DCDC1, LPHN3, and TENM1 were the opposite (**Figure 6E**). The co-mutation relationship between these genes was then explored for the underlying interaction (**Figure 6D**, LUAD; **Figure 6F**, LUSC). Tumor stemness plays an important role in cancer progression. Therefore, we quantified the tumor stemness in the RNA and DNA level (**Figures 7A, B**). The result showed that risk score has a significantly positive correlation with mRNAsi and EREG-mRNAsi, but a negative correlation with mDNAsi and EREG-mDNAsi (**Figure 7C**, mRNAsi, $r = 0.21$, $p < 0.001$, EREG-mRNAsi, $r = 0.11$, $p < 0.001$; **Figure 7D**, mDNAsi, $r = -0.13$, $p < 0.001$, EREG-mDNAsi, $r = -0.09$, $p < 0.001$). Meanwhile, we found a higher mRNAsi and EREG-mRNAsi level in the high-risk group (**Figures 7E, F**). In addition, a lower mDNAsi and EREG-mDNAsi level was also observed in high-risk patients (**Figures 7G, H**).

The IRS Model Is Closely Associated With Immunotherapy Response

The ESTIMATE algorithm was used to quantify the tumor microenvironment score, including ESTIMATEScore, StromalScore, and ImmuneScore. No significant difference in StromalScore was found in high- and low-risk patients (**Figure 8A**). However, high-risk patients tend to have a lower ImmuneScore and ESTIMATEScore compared with low-risk patients (**Figures 8B, C**). Recently, immune checkpoints have

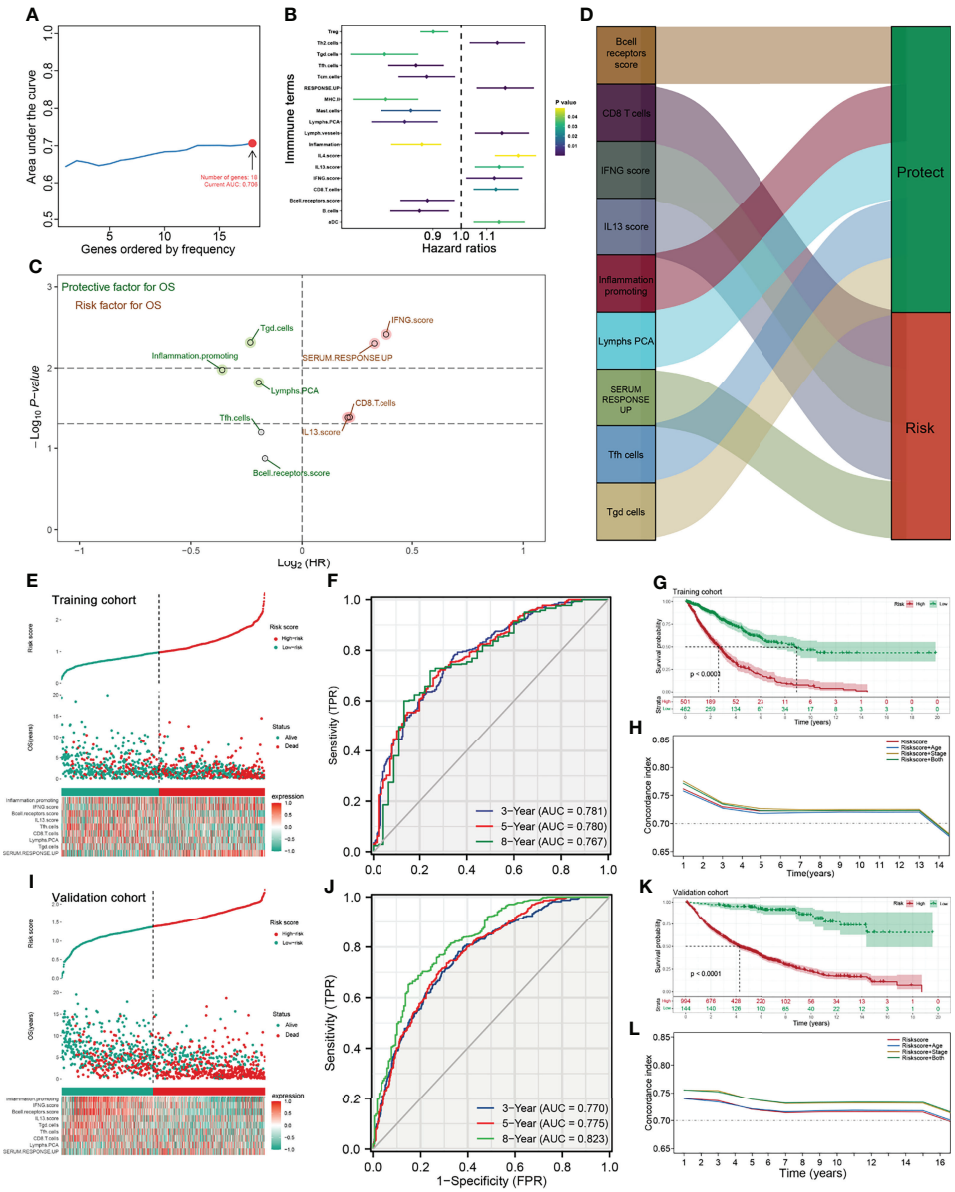


FIGURE 3 | Identification the immune terms significantly associated with patients prognosis. **(A)** Iterative LASSO regression was used for high-frequency features selection and 18 prognosis-related terms were retained; **(B)** Univariate Cox analysis was performed to identified prognosis-related immune terms with the threshold of P < 0.05; **(C, D)** Multivariate Cox analysis was performed to identify immune terms for prognosis-model construction; **(E)** Risk plot showed a higher percentage of dead cases in the high IRS group (training cohort); **(F)** ROC curve showed a satisfactory prediction efficiency of prognosis model; **(G)** Kaplan-Meier survival curve showed that high IRS patients might have a worse OS; **(H)** Time-dependent ROC curve showed that IRS signature has stable predictive ability of OS in a different time (training cohort); **(I–L)** The risk plot, ROC curve, Kaplan-Meier survival curve and Time-dependent ROC curve in validation cohort.

gained a lot of interest among the scientific community. Therefore, we performed a correlation analysis between our IRS model and multiple checkpoint modules. The result showed a significant difference in several immune checkpoint modules between high- and low-IRS patients (**Figure 8D**). CTLA-4, PD-1, PD-L1, and PD-L2 were the immune checkpoint modules with great attention. We found that high-risk patients might have a higher CTLA-4, PD-1, and PD-L1 level, indicating the underlying immunotherapy response difference between low- and high-risk patients (**Figures 8E-H**). TIDE analysis was performed to explore the effect of IRS on immunotherapy response. The result showed that the patients in the high-risk group might have a higher TIDE score (**Figure 9A**). The patients with TIDE score < 0 were defined as immunotherapy responders, while patients with TIDE score > 0 were defined as immunotherapy non-responders (**Figure 9B**). We found that the high-risk group had a higher percentage of immunotherapy non-responders (**Figure 9C**). Based on the submap algorithm obtained from GenePattern, we found that

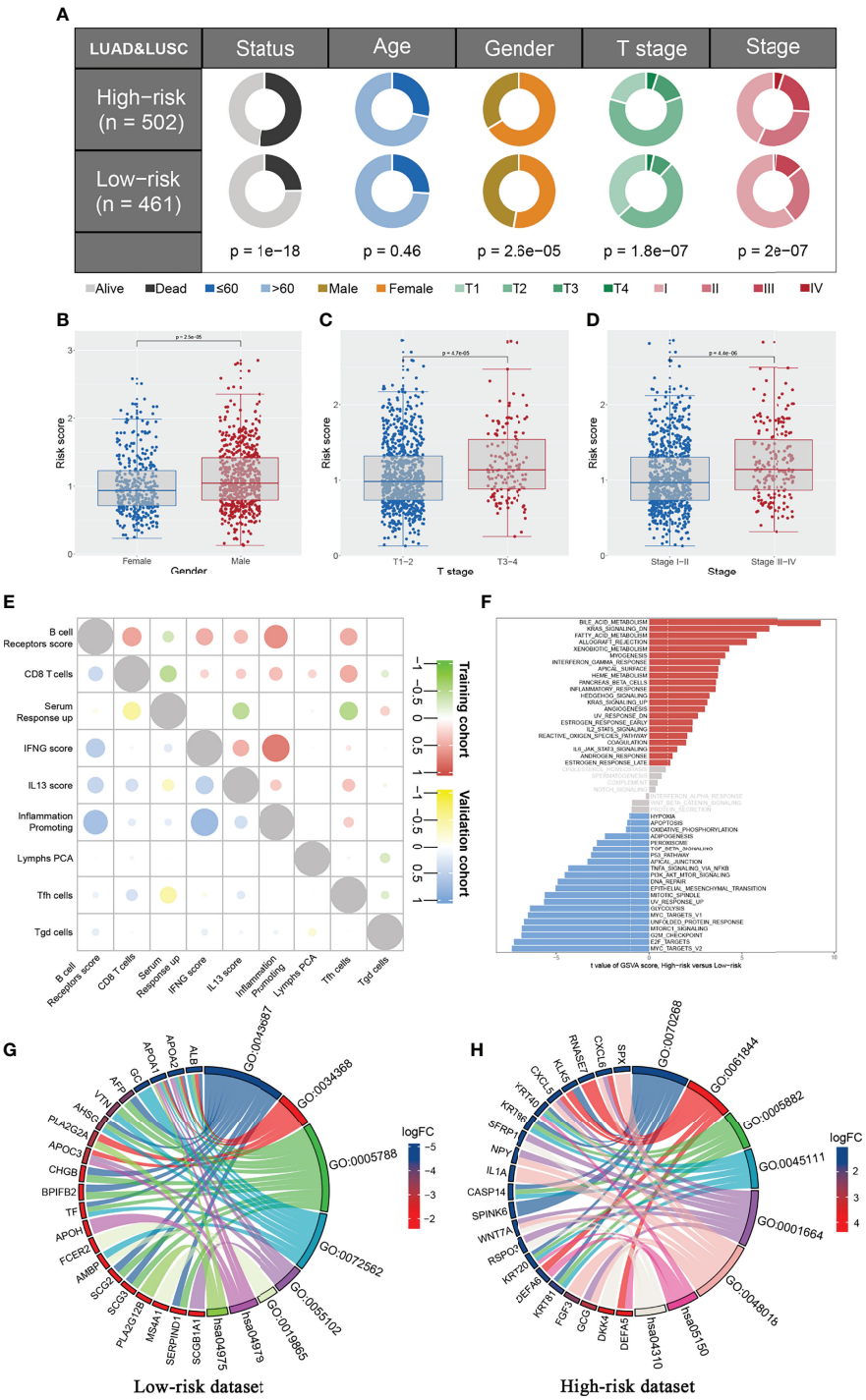


FIGURE 4 | Clinical correlation and biological effect of IRS model. **(A)** The patients in high risk group have a higher percentage of more aggressive clinical features compared with the low risk patients, including T stage and clinical stage; **(B–D)** The IRS level between different group (Male vs. Female, T3-4 vs. T1-2, Stage III-IV vs. Stage I-II); **(E)** Co-expression immune terms of model immune terms; **(F)** GSEA analysis was performed to explore the biological pathway difference between high and low IRS group; **(G, H)** GO and KEGG pathway enrichment in low and high risk patients.

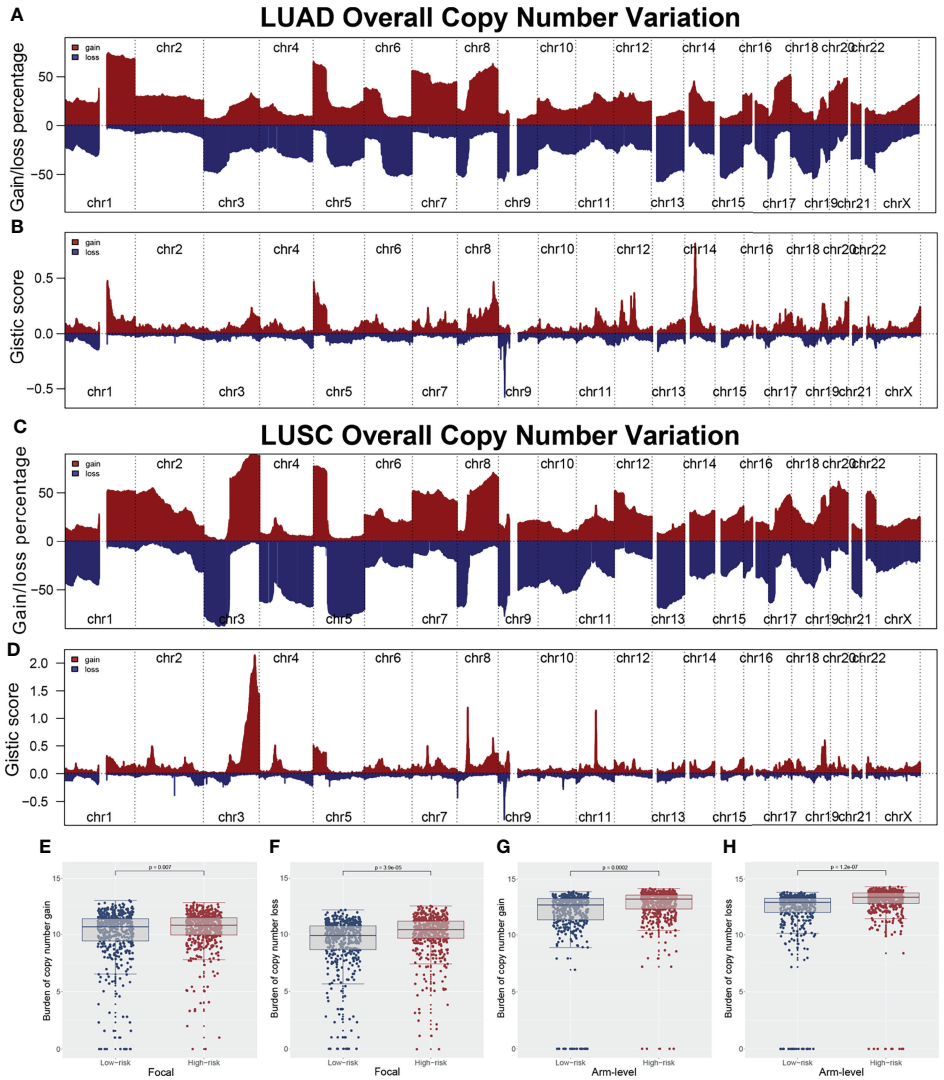


FIGURE 5 | Copy number variation of TCGA-LUAD and TCGA-LUSC. **(A, B)** Copy number profiles (gain and loss) and gistic score of TCGA-LUAD; **(C, D)** Copy number profiles (gain and loss) and gistic score of TCGA-LUSC; **(E, F)** The copy number burden in focal level between low and high risk patients; **(G, H)** The copy number burden in arm level between low and high risk patients.

patients in the low-risk group might be more sensitive to PD-1 immunotherapy (**Figure 9D**). Next, drug sensitivity analysis was performed to explore the underlying influence of IRS on common chemotherapy drugs (**Figures 9E–P**). The result showed that high-risk patients might be more sensitive to bosutinib, lapatinib, nilotinib, pazopanib, sunitinib, tipifarnib, temsirolimus, and vorinostat, yet resistant to metformin.

DISCUSSION

In recent years, PD-1/L1 inhibitors have changed the treatment landscape of advanced lung cancer patients. However, the relatively lower response rate and higher immune-related adverse occurrence rate are hardly satisfactory (16). Thus, it is necessary to explore another precise predictive biomarker for immunotherapy treatment. With the development of bioinformatics analysis, it is feasible for researchers to find another reliable biomarker in large populations.

In our study, we first enrolled 1,997 samples from six independent lung cancer cohorts that contained RNA-seq and clinical data. Then, the ssGSEA algorithm was performed to probe 54 immune terms in a combined expression profile. Furthermore, TCGA-LUAD combined with TCGA-LUSC samples were involved as a training group and other cohorts were classified as a validation group. Subsequently, nine prognostic-related immune terms were explored through univariate Cox regression analysis, LASSO analysis, and multivariate Cox regression analysis. We

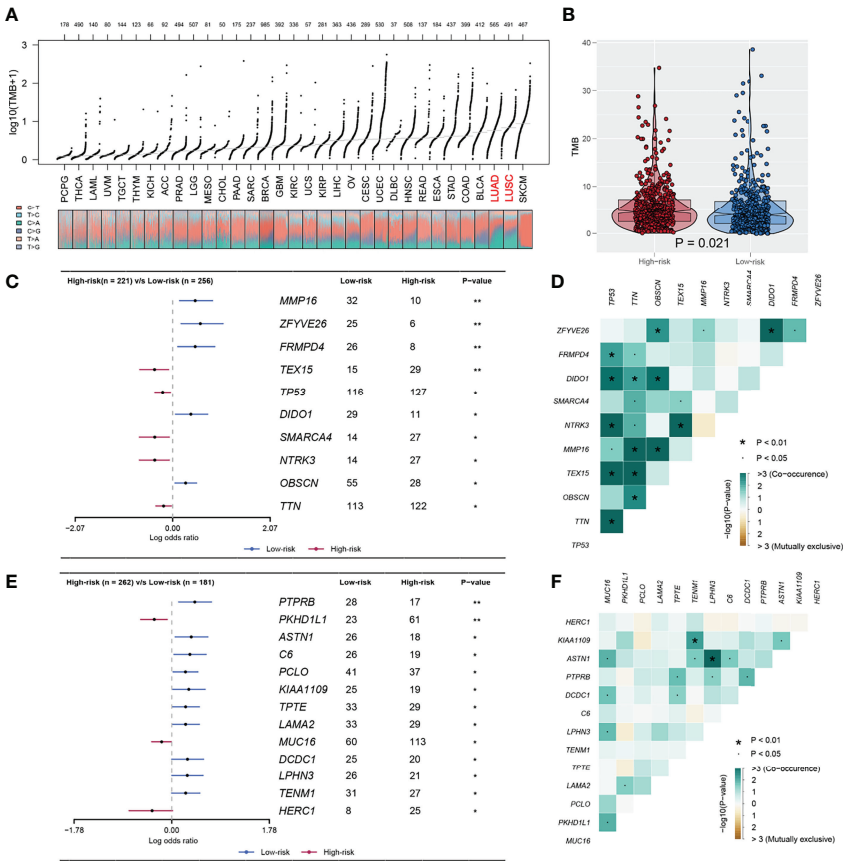


FIGURE 6 | The TMB difference between high and low IRS group. **(A)** The overview of TMB in TCGA-pancancer; **(B)** A higher TMB level was observed in high risk patients compared with the low risk patients; **(C, D)** The mutated gene in TCGA-LUAD between high and low risk with $P < 0.05$; **(E, F)** The mutated gene in TCGA-LUSC between high and low risk with $P < 0.05$. *$P < 0.05$, **$P < 0.01$.

calculated a prognosis-related immune term signature called IRS based on these nine immune terms. According to the cutoff point, all patients were classified as high-risk and low-risk groups. Our study showed that these immune terms were significantly or marginally significantly associated with the prognosis in NSCLC patients. In addition, high IRS demonstrated aggressive clinical characteristics, worse prognosis, higher level of CNV burden, and higher tumor burden mutation and tumor stemness indices. In addition, the IRS model might effectively predict the response to immunotherapy and targeted therapy in NSCLC patients.

In our study, to enhance trustworthiness and credibility, we collected a total of 1,997 samples from six large lung cancer cohorts. Nine prognosis-associated immune terms were extracted after LASSO and Cox regression analysis. Meanwhile, these nine immune terms that included five protective factors and four risk factors were significantly or marginally significantly associated with the prognosis of lung cancer patients, implying a role in lung cancer progression. For example, Kalli et al. and Papageorgis et al. both indicated that IL13 could promote metastasis of breast cancer cells to lung tissue (17, 18). Moreover, IFN-γ and CD8+T cells serving as risk factors to prognosis could have surprised us because CD8+T cells and IFN-γ secretion were always thought as anti-tumor immune surveillance. We postulated that effector phenotype and exhausted phenotype CD8+T cells both counted through ssGSEA may be the major causes. Exhausted phenotype CD8+T cells represented a dysfunctional state, which negatively regulated the function of tumor cell killing (19). In parallel, IFN-γ might enhance endogenous PD-L1 expression and boost tumor metastasis in the tumor immune microenvironment (20). Therefore, it is critically important to facilitate the transformation towards anti-tumor phenotype (21). Recent studies also showed that intratumor-infiltrating B cells inhibited the early-stage lung cancer progression and predicted better immunotherapy response (22, 23). Moreover, the concordance index indicated that the combination of tumor stage and IRS signature could be a strong predictor of overall survival. These nine immune terms provided essential foundation to further explore the underlying tumor microenvironment and genomics information in NSCLC samples.

Our study showed that this novel model had good prognostic value in both training and validation groups. Then, we found

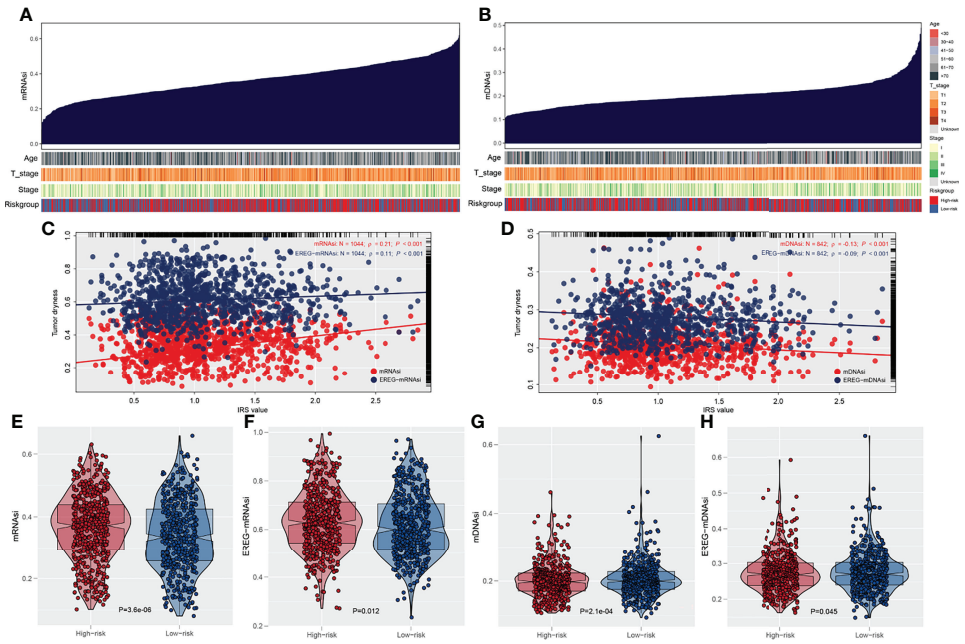


FIGURE 7 | Association of tumor stemness with IRS. **(A)** Quantification of tumor stemness in RNA level; **(B)** Quantification of tumor stemness in DNA level; **(C, D)** The association between IRS and tumor stemness; **(E, F)** A higher degree tumor stemness in RNA level was observed in high risk patients; **(G, H)** A lower degree tumor stemness in RNA level was observed in high risk patients.

that high IRS was correlated with advanced TNM stage, suggesting possible immune suppression. Pathway enrichment analysis displayed significant upregulation of bile acid metabolism, K-RAS signal, and fatty acid metabolism in high-IRS groups. Previous studies already showed that bile acid metabolism and fatty acid metabolism contributed to the tumor invasion (24, 25). Oncogenic K-RAS signaling that has been studied extensively promotes tumor progression in several cancers. Reck et al. indicated that K-RAS gene mutation was correlated with an immunosuppressive landscape through recruiting myeloid-derived suppressor cells (26). This could explain the different immunotherapy response between high- and low-IRS groups. P53 is a classical tumor-suppressive gene and downregulation of P53 was already verified to facilitate cancer progression in various cancers (27). GO and KEGG analysis demonstrated that programmed cell death signaling, G protein-coupled receptor signaling, humoral immune response, and WNT signaling pathway were mostly enriched pathways in high-IRS groups. Shen et al. reported that G protein-coupled estrogen receptor facilitated lung cancer cell metastasis through the NOTCH pathway (28). Our study indicated that high-IRS groups had worse clinical outcomes in combination with the abnormal activation of the mentioned pathways.

In addition, the high-IRS group demonstrated higher copy number alteration burden in the focal and arm level compared with the low-IRS group. This indicated that high-risk patients had a higher degree of genomic instability. It is generally accepted that genetic phenotypic differences caused by copy number alteration can facilitate tumor initiation and progression (29). Patients with a high level of CNV tend to be associated with decreased levels of immune cell infiltration landscape and elevated tumor proliferation ability (30). Meanwhile, CNVs were verified to be correlated with the prognosis of multiple cancers and the outcome of immunotherapy (31–33). A recent study showed that copy number alteration that can lead to genomic instability was indispensable in the transition from lung carcinoma *in situ* to invasive lung carcinoma (34). Gene mutation analysis revealed that patients with a high risk score possessed a higher TMB, indicating different potential response to immunotherapy. Our study was consistent with Liu et al., which showed that low TMB and low CNV were associated with better survival. In parallel, Liu et al. also revealed that those patients with high TMB and high CNV were relatively immunotherapy resistant (35). In addition, the probability of SMARCA4, NTRK3, and TEX15 mutation frequency was higher in high-risk LUAD groups. Concepcion et al. recently demonstrated that the SMARCA4 mutation type could decrease the chromatin accessibility and promote the progression of LUAD (36). NTRK was a rare genetic mutation in lung cancer and NTRK fusion inhibitor had already been used in clinical application (37). At the same time, there was a higher incidence of HERC1 and PKHD1L1 mutation in the LUSC group. The role of HERC1 and PKHD1L1 in LUSC was not reported and needed further validation. Furthermore, cancer stemness results showed that risk score has a positive correlation with mRNAsi and EREG-mRNAsi, but a weakly negative correlation with mDNAsi

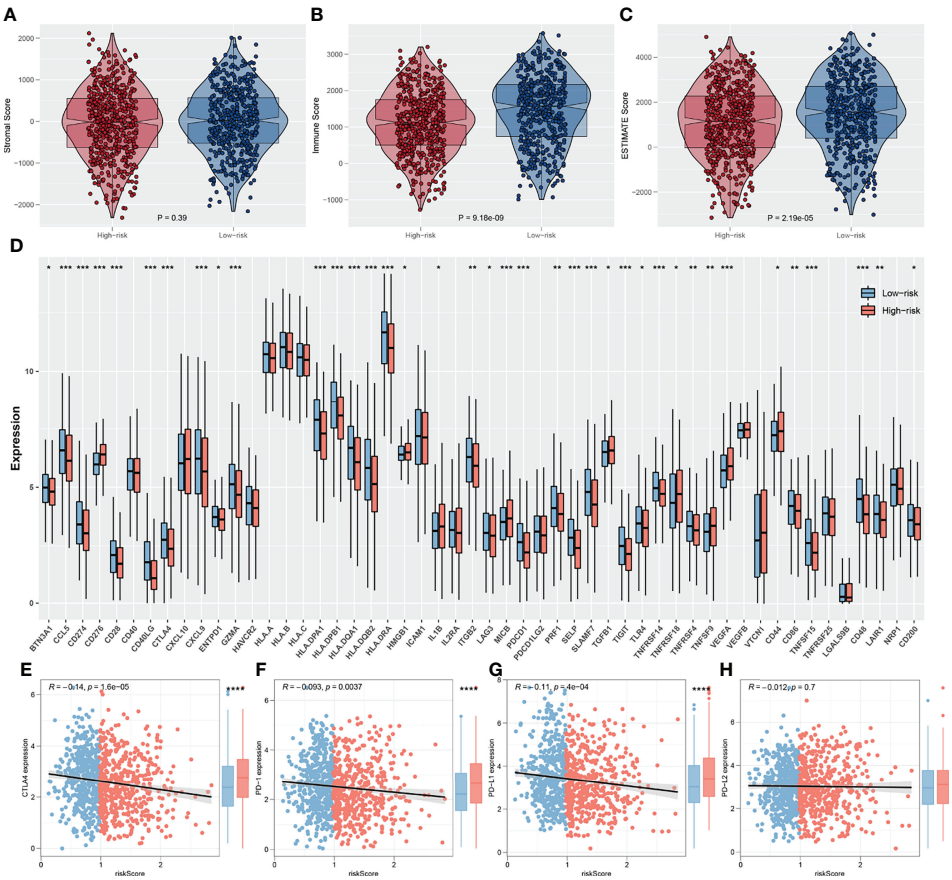


FIGURE 8 | Exploration of the association of immune checkpoint genes with IRS. **(A–C)** The stromalscore, immunescore and estimatescore difference between high and low IRS group; **(D)** Significant differences were observed in multiple immune checkpoint genes between high and low IRS groups; **(E–H)** The difference of serval important immune checkpoint in high and low IRS patients. *P < 0.05, **P < 0.01, ***P < 0.001.

and EREG-mDNAsi (mRNAsi, $r = 0.21$, EREG-mRNAsi, $r = 0.11$; mDNAsi, $r = -0.13$, EREG-mDNAsi, $r = -0.09$). In the meantime, because mRNAsi was reflective of gene transcription expression, the high-IRS group exhibited higher tumor stemness index compared with the low-risk group (38). The study of Malta et al. showed that a higher tumor stemness index was highly correlated with tumor progression and lower PD-L1 expression level, which was shown in high-risk samples (39).

More importantly, high-IRS group patients had lower immune cell infiltration compared with the low-IRS group through the ESTIMATE algorithm (40). Our study also showed that the high-IRS group harbored a relative low PD-1, PD-L1, and CTLA-4 expression level, which usually means decreased immunotherapy efficacy. Then, results of TIDE and Subclass mapping algorithm showed that the patients with high IRS were not likely to benefit from immunotherapy. This conclusion was consistent with the previous finding. In addition, bosutinib, lapatinib, nilotinib, pazopanib, sunitinib, tipifarnib, temsirolimus, and vorinostat were all potential molecular targeted drugs for the treatment of high-IRS patients. Most of these drugs were already demonstrated to have an effect on lung cancer cells in recent years. Tan and his partners showed that bosutinib could repress the KRAS mutant lung cancer cells (41). Pazopanib blocked tumor growth and reduced the metastatic sites in lung cancer mouse models (42). Metformin acting as a traditional glucose-lowering agent could not inhibit the progression of lung cancer alone (43). Thus, further development of these agents is still warranted.

Nevertheless, some inevitable limitations should be noticed in our analysis. Firstly, this study was a retrospective analysis and simply using bioinformatics analysis were the largest weaknesses. Furthermore, a prospective sequencing data cohort is needed to support our IRS signature. Secondly, our study consisted only of a few Asian individuals available, which accounted for a small proportion of all cohorts. This situation might limit the clinical application of this IRS signature in China. Thirdly, the open-access data used for analysis were all at the mRNA level, and not at the protein level, which hardly reflects the real situation of lung cancer tissue.

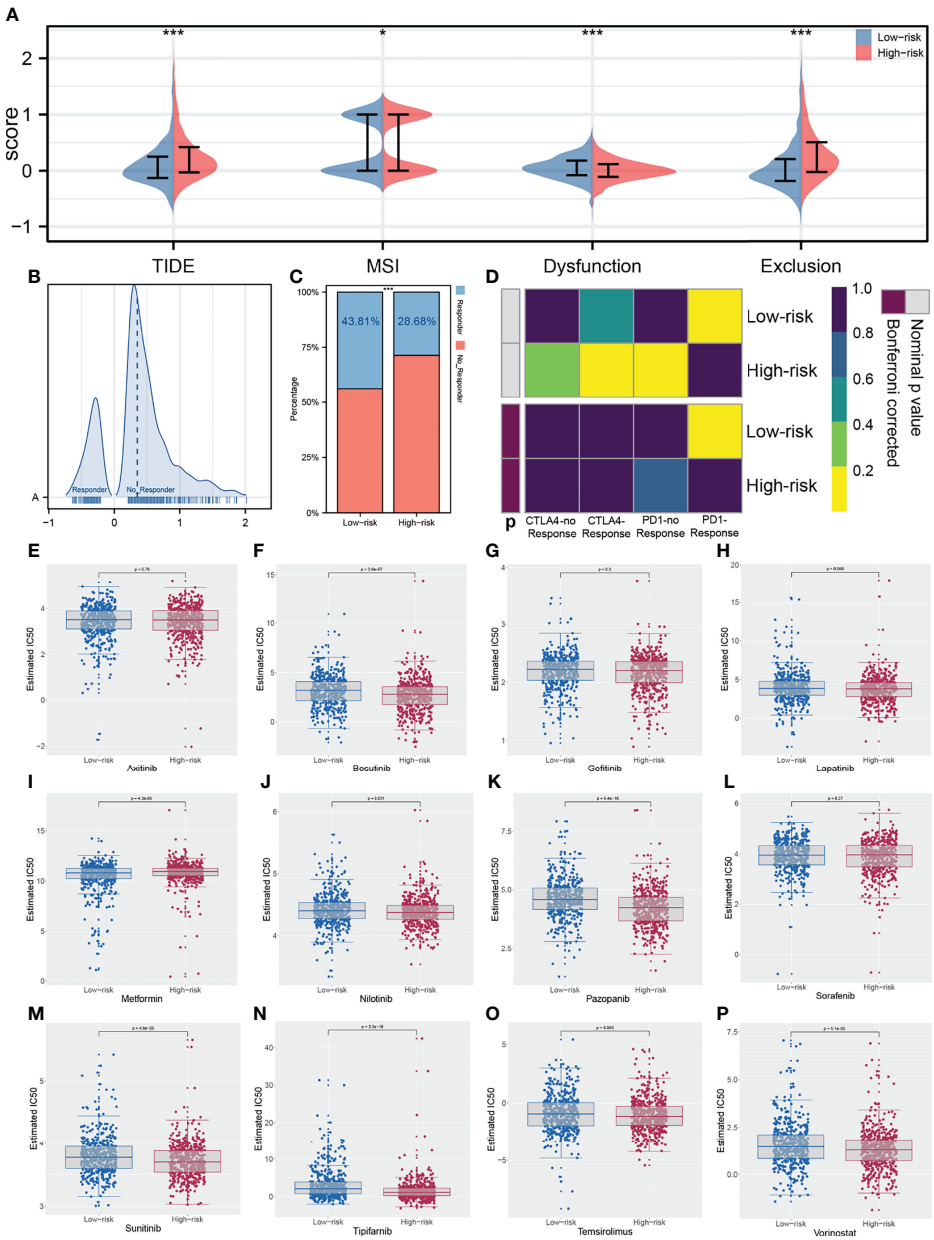


FIGURE 9 | Correlation of IRS immunotherapy response rate and drug sensitivity. **(A)** TIDE analysis was performed in enrolled patients; **(B)** The patients with TIDE score < 0 were defined as immunotherapy responders, while > 0 were defined as immunotherapy non-responders; **(C)** The high risk group had a higher percentage of immunotherapy non-responders; **(D)** The low risk group might more sensitive to PD-1 immunotherapy; **(E–P)** Drug sensitivity analysis was performed to explore the underlying influence of IRS on common chemotherapy drugs. *P < 0.05, ***P < 0.001.

CONCLUSION

According to the prognosis-related immune terms quantified through ssGSEA, we established IRS signature to evaluate the outcome of NSCLC patients and respond to immunotherapy. Moreover, results indicated that high-IRS patients are inclined to bear increased genome instability, cancer CNV, and stemness indices, which mainly accounted for a dismal prognosis and a relatively low efficiency of immunotherapy. In addition, some targeted agents were identified including bosutinib, lapatinib, nilotinib, pazopanib, sunitinib, tipifarnib, temsirolimus, and vorinostat, which might reverse this adversity.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

Conception and design: YY. Administrative support: JW. Provision of study materials: FY. Data analysis and interpretation: WG. Manuscript writing: All authors. Final approval of manuscript: All authors.

FUNDING

This study was supported by the Shanghai Key Laboratory of Clinical Geriatric Medicine, Shanghai Municipal Key Clinical Specialty (shslczdzk2801) and the Lung Cancer Diagnosis and Treatment Center (H1382).

ACKNOWLEDGMENTS

We thank all colleagues involved in the study for their contributions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.924498/full#supplementary-material

Supplementary Figure 1 | Kaplan-Meier survival curve of model immune terms, and the best cutoff of IRS in training and validation cohort. **(A)** Kaplan-Meier survival curve of model immune terms; **(B)** The best cutoff of IRS in training and validation cohort.

Supplementary Figure 2 | GSEA enrichment analysis of IRS. **(A)** GSEA enrichment analysis of IRS model with all the GSEA gene set; **(B)** Representative pathway of each gene set.

Supplementary Figure 3 | The overview of gene mutation of TCGA-NSCLC.

Supplementary Figure 4 | Univariate and multivariate Cox regression analysis of IRS and clinical features. **(A)** Univariate Cox regression analysis of IRS and clinical features; **(B)** Multivariate Cox regression analysis of IRS and clinical features.

REFERENCES

1. Hong T, Sun T, Zhang M, Liu X, Yuan Y, Dolo PR, et al. Surgical Perspective in Neoadjuvant Chemoimmunotherapy for Stage Ii-Iii Non-Small Cell Lung Cancer. *Thorac Cancer* (2021) 12(20):2796–802. doi: 10.1111/1759-7714.14127
2. Brozos-Vázquez EM, Díaz-Peña R, García-González J, León-Mateos L, Mondelo-Macía P, Peña-Chilet M, et al. Immunotherapy in Nonsmall-Cell Lung Cancer: Current Status and Future Prospects for Liquid Biopsy. *Cancer Immunol Immunother* (2021) 70(5):1177–88. doi: 10.1007/s00262-020-02752-z
3. Shi Y, Lei Y, Liu L, Zhang S, Wang W, Zhao J, et al. Integration of Comprehensive Genomic Profiling, Tumor Mutational Burden, and Pd-L1 Expression to Identify Novel Biomarkers of Immunotherapy in Non-Small Cell Lung Cancer. *Cancer Med* (2021) 10(7):2216–31. doi: 10.1002/cam4.3649
4. Kubo T, Shinkawa T, Kikuchi Y, Murata K, Kanaseki T, Tsukahara T, et al. Fundamental and Essential Knowledge for Pathologists Engaged in the Research and Practice of Immune Checkpoint Inhibitor-Based Cancer Immunotherapy. *Front Oncol* (2021) 11:679095. doi: 10.3389/fonc.2021.679095
5. Hashemi S, Fransen MF, Niemeijer A, Ben Taleb N, Houda I, Veltman J, et al. Surprising Impact of Stromal Til's on Immunotherapy Efficacy in a Real-World Lung Cancer Study. *Lung Cancer* (2021) 153:81–9. doi: 10.1016/j.lungcan.2021.01.013
6. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immuno-Oncology Biomarkers Working Group: Part 2: Tils in Melanoma, Gastrointestinal Tract Carcinomas, Non-Small Cell Lung Carcinoma and Mesothelioma, Endometrial and Ovarian Carcinomas, Squamous Cell Carcinoma of the Head and Neck, Genitourinary Carcinomas, and Primary Brain Tumors. *Adv Anatomic Pathol* (2017) 24(6):311–35. doi: 10.1097/pap.0000000000000161
7. Egelston CA, Avalos C, Tu TY, Rosario A, Wang R, Solomon S, et al. Resident Memory Cd8+ T Cells Within Cancer Islands Mediate Survival in Breast Cancer Patients. *JCI Insight* (2019) 4(19):e130000. doi: 10.1172/jci.insight.130000
8. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, et al. Pd-1 Blockade Induces Responses by Inhibiting Adaptive Immune Resistance. *Nature* (2014) 515(7528):568–71. doi: 10.1038/nature13954
9. Capone M, Giannarelli D, Mallardo D, Madonna G, Festino L, Grimaldi AM, et al. Baseline Neutrophil-To-Lymphocyte Ratio (Nlr) and Derived Nlr Could Predict Overall Survival in Patients With Advanced Melanoma Treated With Nivolumab. *J Immunother Cancer* (2018) 6(1):74. doi: 10.1186/s40425-018-0383-1
10. Valero C, Lee M, Hoen D, Weiss K, Kelly DW, Adusumilli PS, et al. Pretreatment Neutrophil-To-Lymphocyte Ratio and Mutational Burden as Biomarkers of Tumor Response to Immune Checkpoint Inhibitors. *Nat Commun* (2021) 12(1):729. doi: 10.1038/s41467-021-20935-9
11. Gibney GT, Weiner LM, Atkins MB. Predictive Biomarkers for Checkpoint Inhibitor-Based Immunotherapy. *Lancet Oncol* (2016) 17(12):e542–e51. doi: 10.1016/s1470-2045(16)30406-5
12. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. Ifn-Gamma-Related Mrna Profile Predicts Clinical Response to Pd-1 Blockade. *J Clin Invest* (2017) 127(8):2930–40. doi: 10.1172/JCI91190
13. Johnson DB, Estrada MV, Salgado R, Sanchez V, Doxie DB, Opalenik SR, et al. Melanoma-Specific Mhc-Ii Expression Represents a Tumour-Autonomous Phenotype and Predicts Response to Anti-Pd-1/Pd-L1 Therapy. *Nat Commun* (2016) 7:10582. doi: 10.1038/ncomms10582
14. Zhou L, Zhang M, Li R, Xue J, Lu Y. Pseudoprogression and Hyperprogression in Lung Cancer: A Comprehensive Review of Literature. *J Cancer Res Clin Oncol* (2020) 146(12):3269–79. doi: 10.1007/s00432-020-03360-1
15. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. Irecist: Guidelines for Response Criteria for Use in Trials Testing Immunotherapeutics. *Lancet Oncol* (2017) 18(3):e143–e52. doi: 10.1016/s1470-2045(17)30074-8
16. Song Y, Yan S, Fan W, Zhang M, Liu W, Lu H, et al. Identification and Validation of the Immune Subtypes of Lung Adenocarcinoma: Implications for Immunotherapy. *Front Cell Dev Biol* (2020) 8:550. doi: 10.3389/fcell.2020.00550
17. Kalli M, Mpekris F, Wong CK, Panagi M, Ozturk S, Thiagalingam S, et al. Activin a Signaling Regulates Il13rα2 Expression to Promote Breast Cancer Metastasis. *Front Oncol* (2019) 9:32. doi: 10.3389/fonc.2019.00032
18. Papageorgis P, Ozturk S, Lambert AW, Neophytou CM, Tzatsos A, Wong CK, et al. Targeting Il13ralpha2 Activates Stat6-Tp63 Pathway to Suppress Breast

Cancer Lung Metastasis. *Breast Cancer Res: BCR* (2015) 17(1):98. doi: 10.1186/s13058-015-0607-y
19. Kallies A, Zehn D, Utzschneider DT. Precursor Exhausted T Cells: Key to Successful Immunotherapy? *Nat Rev Immunol* (2020) 20(2):128–36. doi: 10.1038/s41577-019-0223-7
20. Abiko K, Matsumura N, Hamanishi J, Horikawa N, Murakami R, Yamaguchi K, et al. Ifn-Γ From Lymphocytes Induces Pd-L1 Expression and Promotes Progression of Ovarian Cancer. *Br J Cancer* (2015) 112(9):1501–9. doi: 10.1038/bjc.2015.101
21. Kurtulus S, Madi A, Escobar G, Klapholz M, Nyman J, Christian E, et al. Checkpoint Blockade Immunotherapy Induces Dynamic Changes in Pd-1(-) Cd8(+) Tumor-Infiltrating T Cells. *Immunity* (2019) 50(1):181–94 e6. doi: 10.1016/j.immuni.2018.11.014
22. Chen J, Tan Y, Sun F, Hou L, Zhang C, Ge T, et al. Single-Cell Transcriptome and Antigen-Immunoglobin Analysis Reveals the Diversity of B Cells in Non-Small Cell Lung Cancer. *Genome Biol* (2020) 21(1):152. doi: 10.1186/s13059-020-02064-6
23. Song P, Li W, Wu X, Qian Z, Ying J, Gao S, et al. Integrated Analysis of Single-Cell and Bulk Rna-Sequencing Identifies a Signature Based on B Cell Marker Genes to Predict Prognosis and Immunotherapy Response in Lung Adenocarcinoma. *Cancer Immunol Immunother* (2022). doi: 10.1007/s00262-022-03143-2
24. Nie M, Yao K, Zhu X, Chen N, Xiao N, Wang Y, et al. Evolutionary Metabolic Landscape From Preneoplasia to Invasive Lung Adenocarcinoma. *Nat Commun* (2021) 12(1):6479. doi: 10.1038/s41467-021-26685-y
25. Svensson RU, Parker SJ, Eichner LJ, Kolar MJ, Wallace M, Brun SN, et al. Inhibition of Acetyl-Coa Carboxylase Suppresses Fatty Acid Synthesis and Tumor Growth of Non-Small-Cell Lung Cancer in Preclinical Models. *Nat Med* (2016) 22(10):1108–19. doi: 10.1038/nm.4181
26. Reck M, Carbone DP, Garassino M, Barlesi F. Targeting Kras in Non-Small-Cell Lung Cancer: Recent Progress and New Approaches. *Ann Oncol* (2021) 32(9):1101–10. doi: 10.1016/j.annonc.2021.06.001
27. Cho HC, Huang Y, Hung JT, Hung TH, Cheng KC, Liu YH, et al. Puf-A Promotes Cancer Progression by Interacting With Nucleophosmin in Nucleolus. *Oncogene* (2022) 41(8):1155–65. doi: 10.1038/s41388-021-02138-0
28. Shen Y, Li C, Zhou L, Huang JA. G Protein-Coupled Oestrogen Receptor Promotes Cell Growth of Non-Small Cell Lung Cancer Cells *Via* Yap1/Qki/Circnotch1/M6a Methylated Notch1 Signalling. *J Cell Mol Med* (2021) 25(1):284–96. doi: 10.1111/jcmm.15997
29. Christinat Y, Chaskar P, Clement S, Ho L, Charrier M, McKee T, et al. Automated Detection of Arm-Level Alterations for Individual Cancer Patients in the Clinical Setting. *J Mol Diagn* (2021) 23(12):1722–31. doi: 10.1016/j.jmoldx.2021.08.003
30. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor Aneuploidy Correlates With Markers of Immune Evasion and With Reduced Response to Immunotherapy. *Science* (2017) 355(6322):eaaf8399. doi: 10.1126/science.aaf8399
31. Ben-David U, Amon A. Context Is Everything: Aneuploidy in Cancer. *Nat Rev Genet* (2020) 21(1):44–62. doi: 10.1038/s41576-019-0171-x
32. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* (2017) 376(22):2109–21. doi: 10.1056/NEJMoa1616288
33. Kim HS, Cha H, Kim J, Park WY, Choi YL, Sun JM, et al. Genomic Scoring to Determine Clinical Benefit of Immunotherapy by Targeted Sequencing. *Eur J Cancer* (2019) 120:65–74. doi: 10.1016/j.ejca.2019.08.001
34. Teixeira VH, Pipinikas CP, Pennycuick A, Lee-Six H, Chandrasekharan D, Beane J, et al. Deciphering the Genomic, Epigenomic, and Transcriptomic Landscapes of Pre-Invasive Lung Cancer Lesions. *Nat Med* (2019) 25(3):517–25. doi: 10.1038/s41591-018-0323-0
35. Liu L, Bai X, Wang J, Tang XR, Wu DH, Du SS, et al. Combination of Tmb and Cna Stratifies Prognostic and Predictive Responses to Immunotherapy Across Metastatic Cancer. *Clin Cancer Res* (2019) 25(24):7413–23. doi: 10.1158/1078-0432.Ccr-19-0558
36. Concepcion CP, Ma S, LaFave LM, Bhutkar A, Liu M, DeAngelo LP, et al. Smarca4 Inactivation Promotes Lineage-Specific Transformation and Early Metastatic Features in the Lung. *Cancer Discovery* (2022) 12(2):562–85. doi: 10.1158/2159-8290.Cd-21-0248
37. Farago AF, Taylor MS, Doebele RC, Zhu VW, Kummar S, Spira AI, et al. Clinicopathologic Features of Non-Small-Cell Lung Cancer Harboring an Ntrk Gene Fusion. *JCO Precis Oncol* (2018). doi: 10.1200/po.18.00037
38. Zhang Y, Tseng JT, Lien IC, Li F, Wu W, Li H. Mrnasi Index: Machine Learning in Mining Lung Adenocarcinoma Stem Cell Biomarkers. *Genes (Basel)* (2020) 11(3):257. doi: 10.3390/genes11030257
39. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated With Oncogenic Dedifferentiation. *Cell* (2018) 173(2):338–54.e15. doi: 10.1016/j.cell.2018.03.034
40. Wu J, Li L, Zhang H, Zhao Y, Zhang H, Wu S, et al. A Risk Model Developed Based on Tumor Microenvironment Predicts Overall Survival and Associates With Tumor Immunity of Patients With Lung Adenocarcinoma. *Oncogene* (2021) 40(26):4413–24. doi: 10.1038/s41388-021-01853-y
41. Tan DS, Haaland B, Gan JM, Tham SC, Sinha I, Tan EH, et al. Bosutinib Inhibits Migration and Invasion *Via* Ack1 in Kras Mutant Non-Small Cell Lung Cancer. *Mol Cancer* (2014) 13:13. doi: 10.1186/1476-4598-13-13
42. Zhao H, Yang F, Shen W, Wang Y, Li X, You J, et al. Pazopanib Diminishes Non-Small Cell Lung Cancer (Nsclc) Growth and Metastases in Vivo. *Thorac Cancer* (2015) 6(2):133–40. doi: 10.1111/1759-7714.12138
43. Fatehi Hassanabad A, MacQueen KT. Molecular Mechanisms Underlining the Role of Metformin as a Therapeutic Agent in Lung Cancer. *Cell Oncol (Dordrecht)* (2021) 44(1):1–18. doi: 10.1007/s13402-020-00570-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Yao, Wang, Yang and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Immunology

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fimmu.2022.928742

A Signature Based on Costimulatory Molecules for the Assessment of Prognosis and Immune Characteristics in Patients With Stomach Adenocarcinoma

Bangjie Chen[1†], Yong Yao[2†], Deshen Mao[1†], Conghan Li[1†], Xingyu Wang[3], Shuyan Sheng[1], Lizhi Zhang[1], Xinyi Wang[1], Sanwei Chen[1], Wentao Xu[1], Jianyi Deng[1], Chenyu Sun[4], Qin Zhou[5], Scott Lowe[6], Rachel Bentley[6], Wei Shao[3] and Haiwen Li[7,8*]*

[1] First Clinical Medical College (First Affiliated Hospital), Anhui Medical University, Hefei, China, [2] School of Life Sciences, Anhui Medical University, Hefei, China, [3] School of Basic Medicine, Anhui Medical University, Hefei, China, [4] AMITA Health Saint Joseph Hospital Chicago, University of Illinois Chicago, Chicago, IL, United States, [5] Radiation Oncology, Mayo Clinic, Rochester, MN, United States, [6] Medical College, Kansas City University, Kansas, MO, United States, [7] Department of Gastroenterology, the First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [8] Third Affiliated Hospital (Hefei First People's Hospital), Anhui Medical University, Hefei, China

Although costimulatory molecules have been shown to boost antitumor immune responses, their significance in stomach adenocarcinoma (STAD) remains unknown. The purpose of this study was to examine the gene expression patterns of costimulatory molecule genes in patients with STAD and develop a predictive signature to aid in therapy selection and outcome prediction. We used 60 costimulatory family genes from prior research to conduct the first complete costimulatory molecular analysis in patients with STAD. In the two study groups, consensus clustering analysis based on these 60 genes indicated unique distribution patterns and prognostic differences. Using the least absolute shrinkage and selection operator and Cox regression analysis, we identified nine costimulatory molecular gene pairs (CMGPs) with prognostic value. With these nine CMGPs, we were able to develop a costimulatory molecule-related prognostic signature that performed well in an external dataset. For the patients with STAD, the signature was proven to be a risk factor independent of the clinical characteristics, indicating that this signature may be employed in conjunction with clinical considerations. A further connection between the signature and immunotherapy response was discovered. The patients with high mutation rates, an abundance of infiltrating immune cells, and an immunosuppressive milieu were classified as high-risk patients. It is possible that these high-risk patients have a better prognosis for immunotherapy since they have higher cytolytic activity scores and immunophenoscores of CTLA4 and PD-L1/PD-L2 blockers. Therefore, our signature may help clinicians in assessing patient prognosis and developing treatment plans.

Keywords: costimulatory molecule, STAD, prognostic signature, nomogram, immunotherapy

OPEN ACCESS

Edited by:
Ken Chen,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Holly Hill,
University of Texas MD Anderson Cancer Center, United States
Merve Dede,
University of Texas MD Anderson Cancer Center, United States

****Correspondence:***
Wei Shao
ShaoWei@ahmu.edu.cn
Haiwen Li
lihaiwen1@gmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

Received: *26 April 2022*
Accepted: *22 June 2022*
Published: *22 July 2022*

Citation:
Chen B, Yao Y, Mao D, Li C, Wang X, Sheng S, Zhang L, Wang X, Chen S, Xu W, Deng J, Sun C, Zhou Q, Lowe S, Bentley R, Shao W and Li H (2022) A Signature Based on Costimulatory Molecules for the Assessment of Prognosis and Immune Characteristics in Patients with Stomach Adenocarcinoma. Front. Immunol. 13:928742. doi: 10.3389/fimmu.2022.928742

INTRODUCTION

Gastric cancer (GC) is a global public health burden, affecting more than one million individuals and causing an estimated 769,000 deaths (equating to 1 in every 13 deaths globally) each year (1). It is the fourth leading cause of cancer mortality globally (1) despite breakthroughs in surgical methods, radiation, chemotherapy, and neoadjuvant treatment. GC has high molecular and phenotypic diversity. Endoscopic resection is the most common treatment for early GC and surgery for advanced or intermediate staged GC (2). Owing to the low rate of early detection, surgery as a first-line treatment frequently does not yield the desired outcome (3). The combination of immunotherapy and chemotherapy is considered a powerful treatment for advanced GC (4). Patients with GC and/or gastroesophageal junction cancer may respond to targeted treatment based on four molecular indicators: T-DM1 and PD-L1 expression is necessary for trastuzumab and trastuzumab deruxtecan and MSI and HER2 positivity for pembrolizumab (5). Therefore, finding novel biomarkers that can predict patient survival and responsiveness to targeted medicines or immunotherapies is critical.

Multiple clinical studies are combining immune checkpoint inhibition (ICI) therapy with conventional chemotherapy. Success has been documented in non-small- and small-cell lung carcinomas (6, 7), as well as in esophageal (8), urothelial (9), gastric (10), and head and neck malignancies (11). However, the objective response rate is poor, and some patients develop drug resistance and disease progression after ICI therapy. In addition, immunotherapies, such as vaccine therapy and genome editing, are widely used in patients with GC. Initial attempts of other immunotherapies, such as CAR-T therapy, have prompted the advancement of immunotherapy in GC (12). However, the high heterogeneity of GC makes screening for typical biomarkers difficult. Identification of more biomarkers and mobilization of tumor-reactive lymphocytes from patients in a rapid and accurate manner should be the focus of future studies (13). Through genomic profile analysis, The Cancer Genome Atlas (TCGA) identified four distinct subtypes of stomach adenocarcinoma (STAD) in 2014: microsatellite unstable (MSI), genomically stable, Epstein–Barr virus-positive, and chromosomally unstable cancers (14). Consequently, it may be possible to develop new concepts for more precise molecular subtypes and tailored therapies if representative gene sets are selected for tumor classification and if prediction models are constructed. A number of studies have confirmed that costimulatory molecules are closely related to pathological tumor angiogenesis (15–17). Given the importance of angiogenesis in GC, using costimulatory molecules to enable efficient risk classification and identify possible targets for tailored therapeutic approaches appears to be extremely promising. Previous studies have shown that costimulatory molecules have therapeutic potential in various cancers (18). T-cell activation and proliferation are regulated by costimulatory molecules, making them potential targets for the development of novel ICI therapy. Immunological tumor milieu regulation may also be one of these functions (19, 20). However, it is unclear what specific roles these costimulatory chemicals play in the pathogenesis of GC.

In this study, we examined the expression patterns and prognostic significance of costimulatory molecular gene pairs (CMGPs) in patients with STAD. We then created and verified a predictive signature and nomogram for these patients. In STAD, a risk model based on CMGPs showed promise in predicting survival. Furthermore, a nomogram combining a risk model with clinical parameters effectively predicted the prognosis of patients with STAD. Finally, we assessed the potential response to immunotherapy and chemotherapy among several patient groups classified using the CMGP-based signature. Notably, similar data mining, processing, and model building have been achieved in renal cell carcinoma (21, 22), prostate cancer (23), hepatocellular carcinoma (24), etc.

MATERIALS AND METHODS

Data Collection

The RNA-seq and important clinical features of patients with STAD were downloaded as the modeling cohort from the TCGA database (https://portal.gdc.cancer.gov/), and the dataset was randomly divided into a training cohort and an internal test cohort at a 7:3 ratio. Furthermore, we used data from the Gene Expression Omnibus (GEO) (GEO-GSE15459) database (https://www.ncbi.nlm.nih.gov/geo) as the external validation cohort. In tumor immunotherapy, the tumor immune checkpoint pathways PD-L1/PD-1 and CD86/CTLA4 belong to the B7-CD28 family, and other costimulatory pathways mainly originate from the tumor necrosis factor (TNF) family. At present, 13 molecules are classified as members of the B7-CD28 family, including 8 molecules belonging to the B7 family (CD80, CD86, PD-L1, PD-L2, ICOSLG, B7-H3, B7x, and HHLA2) and 5 molecules belonging to the CD28 family (CD28, CTLA4, ICOS, PD-1, and TMIGD2). The TNF family consists of the TNF ligand superfamily (TNFSF) and the TNF receptor superfamily (TNFRSF) with 47 molecules. Among them, 18 ligands are members of the TNFSF, and the other 29 receptors are members of the TNFRSF. Herein, we identified 60 costimulatory molecule genes (CMGs) from the study by Zhang et al. (25) and downloaded them for further analyses. All the data used in our study are publicly available.

Consensus Clustering Analysis

Consensus clustering was used to further investigate the roles and prognostic importance of the costimulatory molecules in STAD using the "ConsensusClusterPlus" R program (26). The clustering score for the cumulative distribution function curve determined the optimal cluster number. The algorithm first subsampled a proportion of items and features from the data matrix. Thereafter, each subsample was divided into, at most, k groups using an agglomerative hierarchical clustering algorithm. This process was repeated for a specified number of times. The pairwise consensus value, defined as "the running proportion of clusters where two items are together (group)," was computed

for each k and stored in the consistency matrix. Agglomerative hierarchical clustering was performed using a consensus value of 1 for each k, which was then pruned into k groups, called consensus clusters.

Comparison of Immune Cell Infiltration and Tumor Microenvironment Between the STAD Subtypes

We calculated the abundance of eight immune cells and two stromal cells using the "MCPcounter" R package (27). The score indicated the degree of infiltration to the immune microenvironment. The tumor microenvironment (TME) scores (stromal, immune, and estimate scores) for the total STAD cohort were calculated using the "ESTIMATE" package (28).

Functional Analyses

The "Limma" (29) R software was used to identify genes whose expression was different between the two groups. The "GSVA" (30) R package was used to reveal how the signaling pathways differed between the two clusters *via* a gene set variation analysis.

CMGP-Based Prognostic Model Construction and Validation Using the Least Absolute Shrinkage and Selection Operator and Cox Regression Analysis

Sixty costimulatory molecules were pairwise aligned, and 3,540 permutations could be formed according to random permutations; the expression quantity of the gene pairs in each sample in the TCGA database was examined. When the former gene was more highly expressed than the latter, it was labeled as 1; the reverse was marked as 0. When a gene pair was >20% scaled to 1 or 0, it was eliminated. The CMGPs were obtained using pairwise comparisons and gene expression analyses in the same patient, which avoided batch effects associated with multiple platforms and eliminated the need to scale and normalize the data. The CMGPs linked with prognosis were identified using univariate Cox regression analysis ($p < 0.001$). The least absolute shrinkage and selection operator (LASSO)–Cox regression model was built using the prognostic-associated CMGPs derived from the univariate Cox regression analysis. We then used the LASSO method with penalty parameter tweaking, conducted *via* 10-fold cross-validation, to exclude the CMGPs that may be substantially associated with other CMGPs. A subset of CMGPs was identified by decreasing the regression coefficient with a penalty proportional to their size. For future multivariate Cox regression analysis, the CMGPs with nonzero regression coefficients were maintained. We compared the predictive CMGP values with the regression coefficients from the multivariate Cox proportional hazard regression analysis (β) to create a risk score model. For the LASSO regression analysis of the prognostic CMGPs, the "glmnet" (31) R package was employed. The median risk score was used to divide the patients into high- and low-risk categories. The "survminer" (32) and "timeROC" (33) R packages were used to create the Kaplan–Meier survival and receiver operating characteristic (ROC) curves of the risk score, which were used to estimate the model's predictive power. Clinical usefulness was assessed using decision curve analysis (DCA). To compare the two groups in terms of the survival curve, we used the log-rank test and set the statistical significance level at $p < 0.05$. The GSE15459 cohort was used for external validation, whereas the TCGA cohort was split at a 7:3 ratio into a training cohort and an internal validation cohort.

Correlation Analysis Between the Prognostic Model and TME

The TME scores (stromal score, immune score, estimate score, and tumor purity) were calculated using the "ESTIMATE" package (28), and gene expression data were utilized to determine the infiltrating stromal or immune cells in the tumor tissues. Additionally, we used the Wilcoxon test to compare the four types of scores between the two groups, as well as the Pearson correlation test to examine the link between the risk score and the four TME scores. According to the TCGA database, tumor mutation burden (TMB) was defined as the total number of somatic gene coding mistakes, base substitutions, insertions, and deletions detected per million bases (34). We examined the association between the TMB and odds of survival.

Gene Set Enrichment Analysis Based on the GO and KEGG Datasets

We used cp.kegg.v7.1.symbols.gmt and go.v7.4.symbols.gmt in the "cluster profiler" package (35) to analyze the highly expressed genes both in the low- and high-risk groups as a reference gene set and the function gesaplots to plot the results, filtering significantly enriched pathways with $p < 0.05$ as a threshold (FDR < 0.25).

Nomogram Construction and Evaluation

We used the "RMS" (36) R package to integrate variables, such as age, tumor stage, and risk score, and the Cox method to establish a nomogram to evaluate STAD prognosis. The "timeROC" (33) R program was used to assess the prognostic performance of the nomogram model based on the time-dependent ROC curves. The concordance index (C-index) was used to measure the likelihood of the projected result matching the actual result. The 45 dotted line indicated the best prediction. Calibration curves were generated to test the discriminative ability of the nomogram.

Chemical Reaction Prediction

We predicted the treatment response for each sample using the world's largest publicly available pharmacogenomics database, Genomics of Drug Sensitivity in Cancer (https://www.cancerrxgene.org/), and the "pRRophetic" package (37).

Statistical Analyses

To compare two variables, we used the *t*-test or Wilcoxon test. To assess survival differences, we performed the Kaplan–Meier method and log-rank tests(two-stage test was used when curves crossed (38)). The predictive impact of the CMGs was assessed using univariate and multivariate Cox regression models. To

assess variations in the distribution of the clinical variables among the patients with STAD, we performed Pearson's chi-squared test. The R software was used to perform all statistical analyses in this investigation. The statistical significance level was set at $p < 0.05$.

RESULTS

Cluster Analysis Based on the CMG Expression Profiles

The workflow of this study is illustrated in **Figure 1**. We used a consensus clustering technique to stratify the patients with STAD to determine the overall prognostic value of the genes. We discovered that a k value of 2 appeared to be a more stable number between values of 2 and 9 (**Figures 2A, B**). The Kaplan–Meier curves revealed that the patients in cluster 1 showed worse overall survival (OS) (**Figure 2C**) and disease-specific survival (DSS) (**Figure 2D**) than did the patients in cluster 2 in the two molecular subtypes. There is a partial crossover at the end of the survival curve, suggesting that other factors may have an impact on the survival outcome. However, the crossover is located at the end of the curve, and the number of patients is small; thus, it is difficult to analyze hierarchically. Referring to relevant statistical literature, when the survival curves are crossed, the log-rank test is no longer used, but the two-stage test should be used (38). The *p*-value of the two-stage test is still less than 0.05, indicating that costimulatory molecular genes are indeed an important factor affecting survival, and the overall survival of C2 patients is longer than C1.

Immune Status in the Two Clusters

Immunological differences between the two molecular subtypes have been investigated in previous immune studies. According to the estimation algorithm, the patients in cluster 2 had substantially higher immune scores ($p < 0.001$), estimated scores ($p < 0.001$), and stromal scores ($p < 0.001$) than those in cluster 1 (**Figure 3A**). In addition, the abundance of B lineage ($p = 1.9e{-}13$) (**Figure 3B**), $CD8^+$ T cells ($p < 2.22e{-}16$) (**Figure 3C**), cytotoxic lymphocytes ($p < 2.22e{-}16$) (**Figure 3D**), monocyte lineage ($p < 2.22e{-}16$) (**Figure 3E**), myeloid dendritic cells ($p = 3.8e{-}05$) (**Figure 3F**), NK cells ($p < 2.22e{-}16$) (**Figure 3G**), and T cells ($p < 2.22e{-}16$) (**Figure 3H**) was significantly higher in the patients in cluster 2 than in those in cluster 1; meanwhile, no significant difference was detected with respect to the abundance of endothelial cells ($p = 0.1$) (**Figure 3I**), fibroblasts ($p = 0.78$) (**Figure 3J**), and neutrophils ($p = 0.087$) (**Figure 3K**).

Differentially Expressed Genes and Functional Analyses

Differentially expressed genes (DEGs) were identified between the two clusters, and functional investigations were performed to investigate the underlying signaling processes. Cluster 2 had 893 DEGs, 126 of which were downregulated and 767 were upregulated, compared with cluster 1. To evaluate the relationship between the enriched pathways and the prognosis of the patients with STAD, we utilized GSVA analysis to analyze the relative expression differences in the pathways in the two clusters. The heatmap in **Figure 3L** shows a number of differentially expressed pathways enhanced by the GSVA analysis.

Distribution of the Clinical Features in the Two Clusters

We examined the distribution of the clinical characteristics in the two clusters. The analysis showed that there was a significant difference in grade and stage, but none in age, sex, or other clinical features (**Figure 3M**).

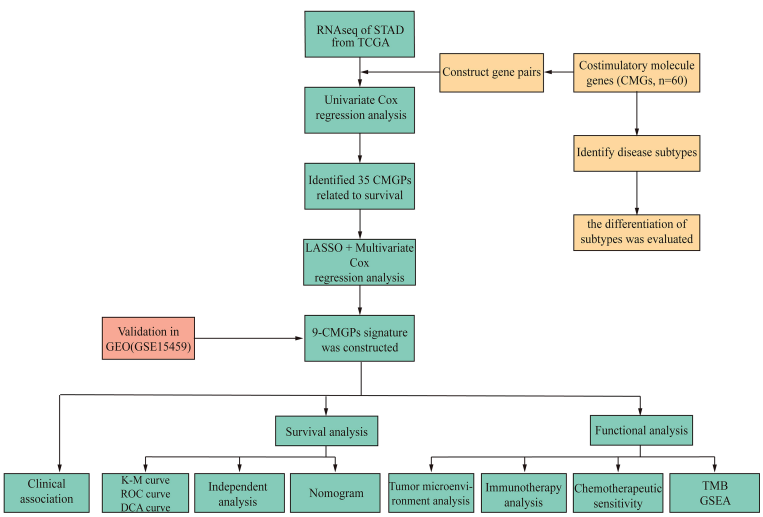


FIGURE 1 | Flowchart of the data analysis.

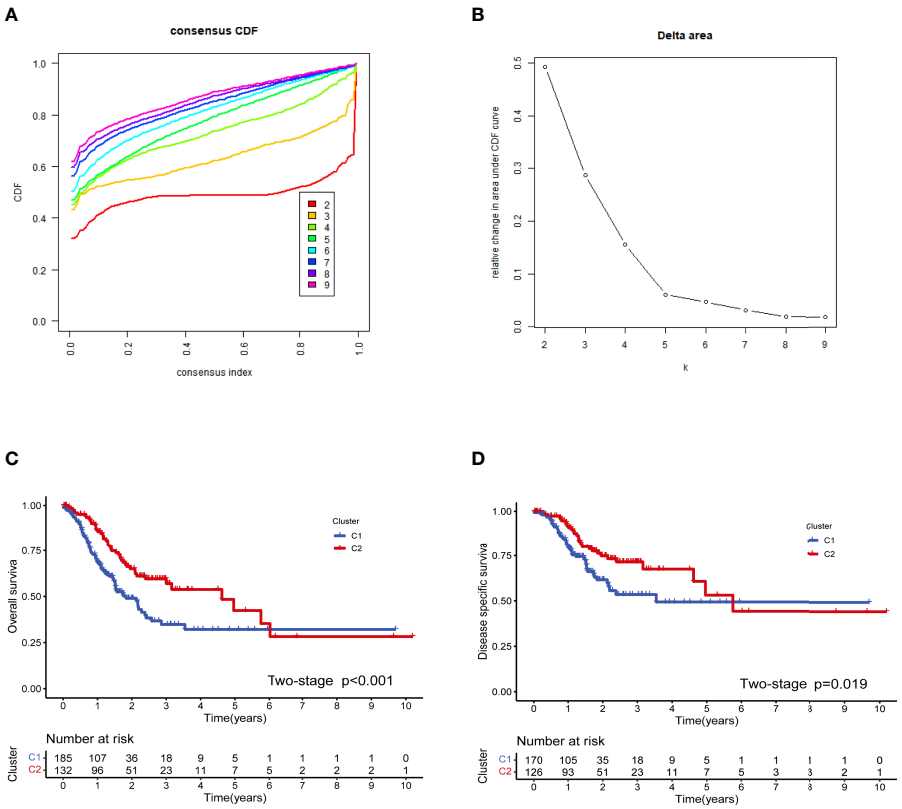


FIGURE 2 | Cluster analysis based on the costimulatory molecule gene (CMG) expression profiles. The optimal value for consensus clustering **(A, B)** was found to be k = 2. Kaplan–Meier curve for the disease-free survival in the stomach adenocarcinoma (STAD) group **(C)**. Kaplan–Meier curve for the total survival in the STAD group **(D)**.

Development of a Risk Model Based on the CMGs in the TCGA Training Cohort

The univariate Cox proportional hazard regression analysis identified 35 CMGPs that were related to survival. After LASSO regression (**Figures 4A, B**) and multivariate Cox regression (**Figure 4C**) analyses, nine CMGPs were selected and utilized to build a prognostic signature as follows: Risk score = (0.54369 *`CD276|LTBR`) + (0.69502 *`CD28|CTLA4`) + (0.58032 *`EDA|VTCN1`) + (−0.53341 *`EDAR|TNFRSF19`) + (−0.66137 *`FASLG|TNFSF8`) + (−0.51890 *`PDCD1|TNFRSF9`) + (−0.43837 *`TNF|TNFSF14`) + (0.47905 *`TNFRSF11B|TNFSF15`) + (−0.49658 *`TNFRSF18|TNFSF9`). We used this method to determine each patient's risk score and divided 223 individuals into high- and low-risk groups based on the median value of the risk score (0.6378) (**Figure 4D**). The OS of the high-risk group was lower than that of the low-risk group ($p < 0.001$), as evidenced by the survival curve (**Figure 4E**). The risk scores that predicted OS at 1, 3, and 5 years had area under the curve of ROC (AUC) values of 0.756, 0.813, and 0.808, respectively (**Figure 4F**). The frequency of fatalities increased as the risk score increased, and this trend was more obvious with the increase in the risk score, especially in the very high-risk population (**Figure 4G**); meanwhile, the DSS rate decreased (**Figure 4H**). These preliminary findings suggest that stratifying prognosis based on the risk score is useful.

Internal Validation of the Prognostic Model in the TCGA Test Cohort

The prognosis of the high-risk group was considerably poorer than that of the low-risk group in the TCGA test cohort ($n = 95$) (**Figure 5A**). The risk scores that predicted OS at 1, 3, and 5 years had AUC values of 0.697, 0.726, and 0.764, respectively (**Figure 5B**). The risk score had good accuracy in predicting STAD prognosis, based on the findings of the internal validation.

External Validation of the Prognostic Model in the GEO Cohort

The GSE15459 cohort from the GEO database was used as an external dataset to validate the prognostic model because of its large sample size ($n = 192$) and complete clinical data. The risk score for each patient in the cohort was determined using a prior method, and the patients were classified into high- or low-risk groups based on the unified cutoff value (0.6378). The high-risk group had significantly lower OS rates than the low-risk group, consistent with earlier research findings (**Figure 5C**). AUC values of 0.595, 0.638, and 0.651 were found in the risk score that predicted OS at 1, 3, and 5 years, respectively (**Figure 5D**).

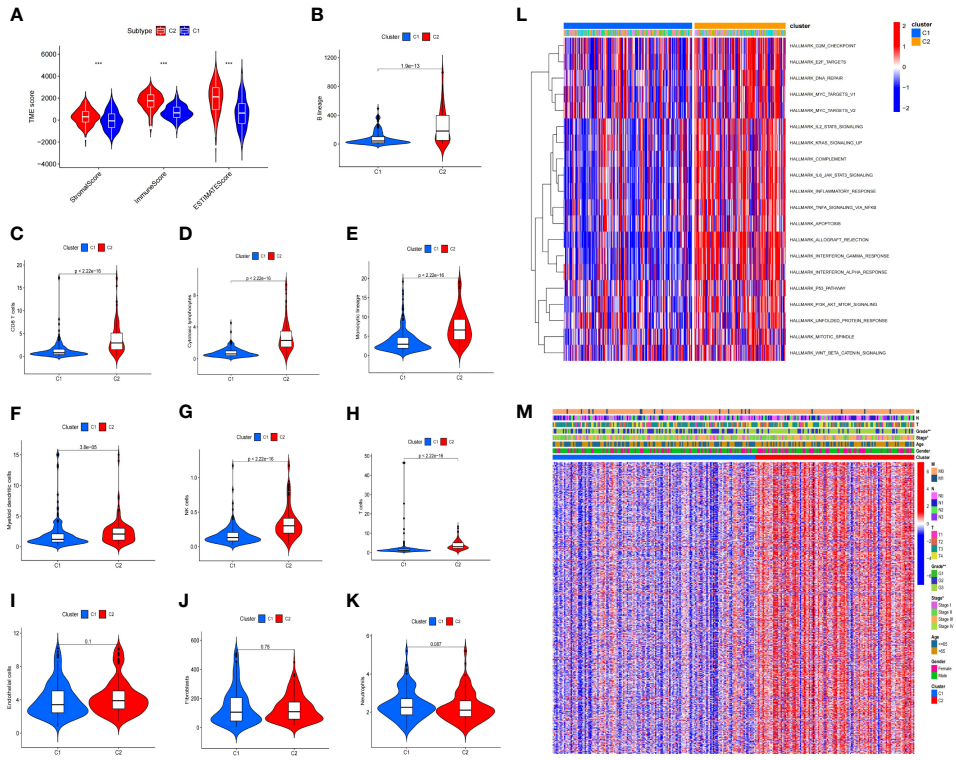


FIGURE 3 | Immune difference between the two clusters. The patients in cluster 2 had substantially higher immunological, estimation, and stromal scores than those in cluster 1 **(A)**. Furthermore, the abundance of B lineage **(B)**, $CD8^+$ T cells **(C)**, cytotoxic lymphocytes **(D)**, monocytic lineage **(E)**, myeloid dendritic cells **(F)**, NK cells **(G)**, and T cells **(H)** was significantly higher in cluster 2 than in cluster 1, while no significant difference was found for the abundance of endothelial cells **(I)**, fibroblasts **(J)**, and neutrophils **(K)**. Gene set variation analysis of the two clusters **(L)**. Differences in the clinicopathological characteristics between the two clusters **(M)** (*p < 0.05, **p < 0.01, ***p < 0.001).

The heatmap in **Figure 5E** shows the expression patterns of 11 CMGs in the patients with varied risk levels. The external validation revealed that the prognostic model created had a wide range of applications and was very stable in predicting STAD prognosis.

Prognostic Assessment of the Prognostic Model in the Entire TCGA Cohort

For analysis, we included all the study items in a single TCGA cohort (n = 318). In the high- and low-risk groups, principal component analysis revealed different distribution patterns (**Figure 5F**). In the TCGA cohort, DCA of the nomogram revealed that the nomogram model had a good net benefit for 1-year (**Figure 5G**), 3-year (**Figure 5H**), and 5-year (**Figure 5I**) OS. Accordingly, the nomogram based on the risk score may be utilized as an effective tool for predicting patient prognosis in clinical practice. The heatmap in **Figure 5J** depicts the expression patterns of 13 CMGs in the patients with varying risk ratings. The risk score may also be used to predict prognosis independently, as revealed in the univariate (**Supplementary Figure 1H**) and multivariate regression analyses (**Supplementary Figure 1I**). These findings demonstrated the dependability and consistency of the predictive signature.

Clinical Correlation Analysis of the Prognostic Model

In the TCGA cohort, we examined sex, age, grade, pathological stage, T stage, M stage, and N stage to determine whether there was a link between immunotyping and common clinical characteristics. The results suggested that our prognostic signature was not significantly associated with the clinical factors in STAD (**Supplementary Figures 1A–G**).

Relationship Between the Prognostic Model and Immune Infiltration

The degree of infiltration of the immune cells varied between the high- and low-risk groups. For example, the high-risk group had significantly more CD4 memory resting T cells (p = 0.003) and eosinophil infiltrates (p = 0.008) than the low-risk group; meanwhile, the low-risk group had significantly more $CD8^+$ T cells (p < 0.001), CD4 memory-activated T cells (p < 0.001), follicular helper T cells (p = 0.002), and M1 macrophages (p < 0.001) than the high-risk group (**Figure 6A**). These findings suggested that the tumor immune microenvironment (TIME) was related to the risk score. The four genes [CD28 (**Figure 6B**), CTLA4 (**Figure 6C**), PDCD1 (**Figure 6D**), and TNF (**Figure 6E**)] in

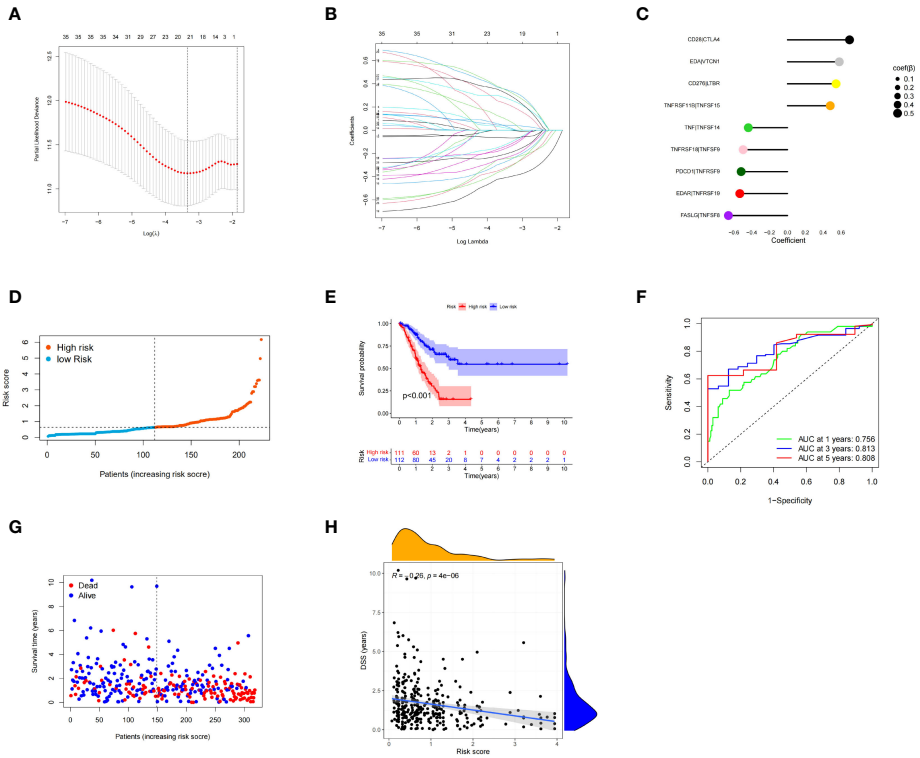


FIGURE 4 | Multivariate least absolute shrinkage and selection operator (LASSO) regression analysis. Nine costimulatory molecular gene pairs (CMGPs) were identified using Cox regression analysis to create the prognostic signature **(A, B)**. Coefficient of the nine CMGPs **(C)**. The risk score of each patient was generated, and the 223 patients were divided into high- and low-risk groups according to the median risk score (0.6378) **(D)**. Survival curves in the high- and low-risk groups **(E)**. Time-dependent receiver operating characteristic (ROC) curve of the risk model **(F)**. The number of deaths increased as the risk score increased; this trend was more obvious with the increase in the risk score, especially in the very high-risk population. There were significant differences in the clinical outcomes at the symmetrical positions of the left and right sides of the median value, especially when the risk score was <100 and >200 **(G)**. Disease-specific survival (DSS) and risk score correlation **(H)**.

the model showed a strong correlation with some immune cells. These findings provide useful information for future research.

Relationship Between the Prognostic Model and TME

The patients with a lower estimate score ($p = 0.047$) (**Figure 7A**) and stromal score (p = 5.9e05) (**Figure 7B**) and greater tumor purity ($p = 0.047$) had a better prognosis (**Figure 7C**) than their counterparts. Patients with a lower immune score ($p = 0.79$) had no statistical significance. Pearson correlation analysis showed that the estimated score ($R = 0.13$, $p = 0.024$) (**Figure 7D**) and stromal score ($R = 0.28$, p = 6e−07) (**Figure 7F**) were positively correlated with the risk score, while tumor purity (R = -0.13, p = 0.024) was negatively correlated with the risk score (**Figure 7E**). In addition, no correlation was found between the immune score ($R = -0.038$, $p = 0.49$) and the risk score. Accordingly, the risk score can be used to analyze the TME.

Differences in the Genomic Alterations Between the High- and Low-Risk Groups

Genomic mutations are closely associated with tumorigenesis. Therefore, the frequency of alterations in patients with STAD was studied. Among the CMGs, TNFRSF11B had the highest genetic alteration rate (**Figure 7G**). Furthermore, there was a significant difference in the TMB ($p = 0.00022$) between high- and low-risk groups (**Figure 7I**). Using Pearson correlation analysis, we also validated the clearly negative relationship between the risk score and TMB (**Figure 7H**). The high TMB group had a somewhat higher OS rate than the low TMB group (**Figure 7J**). The patients were divided into four groups based on their risk score and TMB. We found that the group with the highest TMB and lowest risk score had the best survival rate (**Figure 7K**).

Gene Set Enrichment Analysis Between the Different Risk Groups

We performed gene set enrichment analysis of the various risk groups to identify probable molecular mechanisms for the prognostic model. The analysis revealed that the gene sets in the high-risk group were mostly abundant in receptor- or metastasis-related pathways, such as the KEGG (ECM–receptor interaction, complement and focal adhesion, coagulation cascades, neuroactive ligand–receptor interaction, and PPAR signaling pathway) (**Supplementary Figure 2A**) and

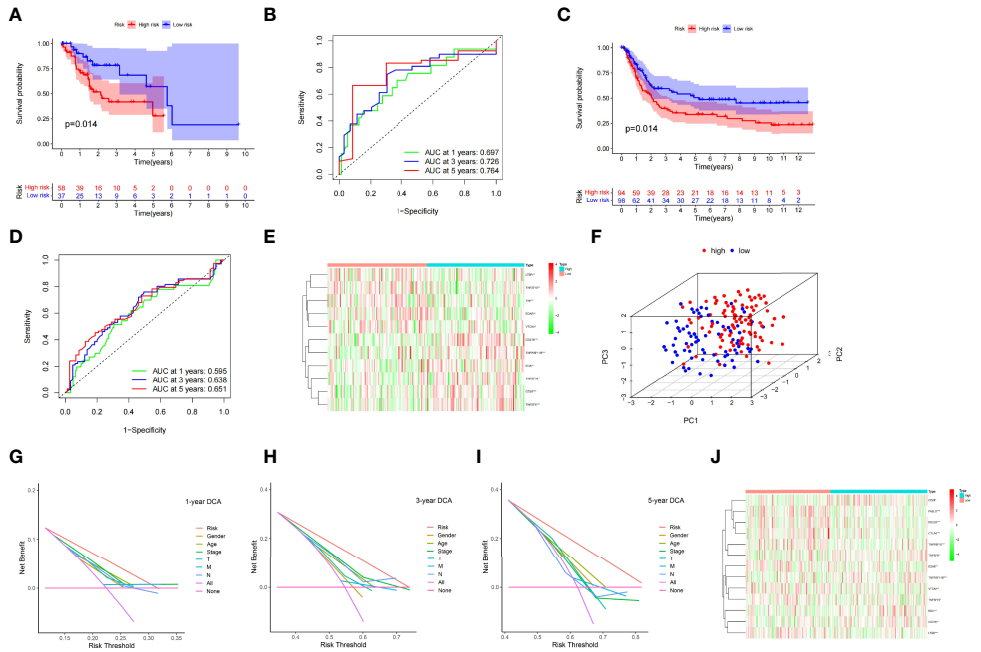


FIGURE 5 | Kaplan–Meier curves in the TCGA test cohort **(A)** and the GEO validation cohort **(C)**. TCGA test cohort **(B)** and GEO validation cohort **(C)** time-dependent receiver operating characteristic curves for the sensitivity and specificity of the prognosis assessment **(D)**. Gene set variation analysis of the GEO validation cohort (the unit of color bars is log2 (actual expression) +1) **(E)**. In the TCGA dataset, principal component analysis was used to evaluate the distribution patterns of the high- and low-risk patients **(F)**. Prognostic signature decision curve analysis at 1 **(G)**, 3 **(H)**, and 5 years **(I)**. Gene set variation analysis of the entire TCGA cohort (the unit of color bars is log2 (actual expression) +1) **(J)** (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

GO pathways (behavior, cell matrix adhesion, cell substrate adhesion, circulatory system process, and external encapsulating structure organization) (**Supplementary Figure 2B**). Most of the pathways that presented a significant enrichment in the low-risk group were related to immunology, including the KEGG (autoimmune thyroid disease, graft versus host disease, antigen processing and presentation, allograft rejection, and proteasome) (**Supplementary Figure 2C**) and GO pathways (activation of immune response, adaptive immunological response based on immunoglobulin superfamily domain-based somatic recombination of immune receptors, B cell-mediated immunity and complement activation, and antigen receptor-mediated signaling pathway) (**Supplementary Figure 2D**). These findings provide important information for future research on the molecular mechanisms underlying STAD.

Nomogram Construction and Validation

A visual nomogram was created to produce a therapeutically useful tool to determine the prognosis of patients with STAD. The nomogram was built using a training set that predicted the OS. It incorporated age, pathological stage, and risk score as integrated clinicopathological variables (**Figure 8A**). The predictive value of the nomogram was assessed using ROC analysis and the C-index. In the TCGA dataset, the AUC values of the predictive value of the nomogram for 1-, 3-, and 5-year OS were 0.712, 0.767, and 0.725, respectively (**Figure 8B**). In terms of the 1-year (**Figure 8C**), 3-year (**Figure 8D**), and 5-year OS (**Figure 8E**) in the TCGA cohort, the calibration plots revealed a sustained concordance between the nomogram-projected probability and actual observation.

Relationship Between the Risk Scores for Immunotherapy and Chemotherapy

The immunophenotype score was used to assess the ICI therapy response. In the comparison between the low- and high-risk groups, we discovered that the proportion of CTLA4 and PD1 was somewhat greater in the low-risk group than in the high-risk group (**Figure 8F**). The low-risk group was more sensitive to the chemotherapy drugs, such as cisplatin, gemcitabine, imatinib, vinblastine, and VX.680, than the high-risk group. In contrast, the high-risk group was more sensitive to bexarotene than the low-risk group (**Figure 8G**).

DISCUSSION

Currently, GC ranks fifth in incidence and fourth in mortality among cancer cases worldwide (1). Owing to the lack of early diagnosis, patients who are detected to have GC are mostly terminal patients who can only benefit slightly from surgical treatment (39). Complete tumor excision and lymph node dissection in combination with preoperative chemotherapy and postoperative adjuvant radiation and chemotherapy have been

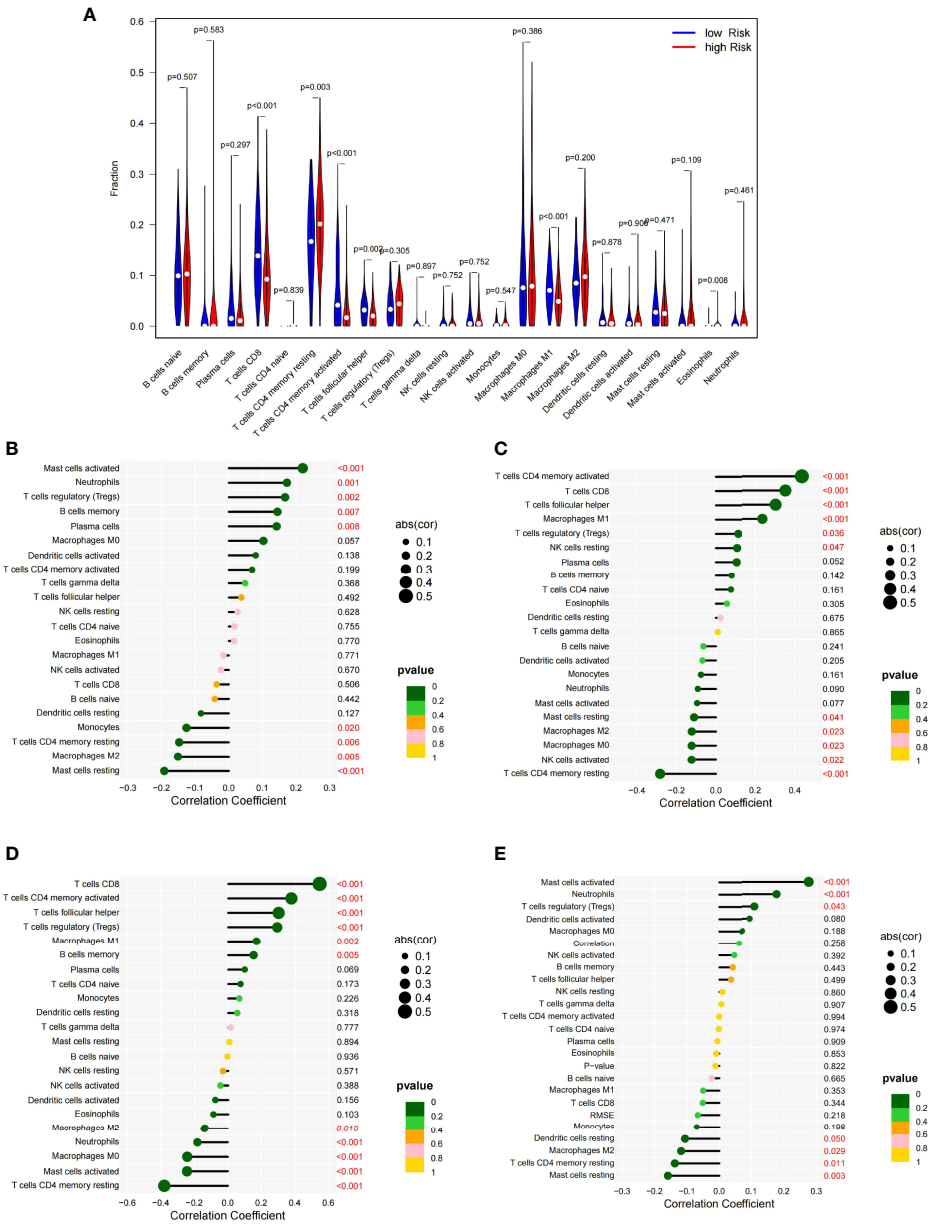


FIGURE 6 | Vioplot of the absolute abundance of 22 immune infiltration cells between the high- and low-risk groups **(A)**. Correlation coefficients between the immune cells and CD28 **(B)**, CTLA4 **(C)**, PDCD1 **(D)**, and TNF **(E)**.

found to considerably enhance the postoperative survival time of patients with GC when the effects of surgery are restricted (40). With a few exceptions for patients with tumors of certain molecular subtypes, chemotherapy remains the mainstay of care (41). For patients with HER2+ tumors, trastuzumab, a HER2 targeting monoclonal antibody, is used for chemotherapy (42). PDL1 immunotherapy has recently emerged as a new treatment option for advanced GC because of advances in the research on the immune microenvironment of gastric tissues. Patients with a high MSI-H phenotype or high TMB (>10 mutations per megabase) may benefit from second-line treatment with pembrolizumab, a monoclonal anti-PD-1 antibody. Furthermore, patients with malignancies that express PD-L1 (combined positive score of 1) may also benefit from third-line treatment with this drug (43). Another anti-PD-1 antibody, nivolumab, improves OS as an advanced treatment for unselected patients with STAD and is combined with chemotherapy as a first-line treatment (44).

Clinical investigations have shown that ICI therapy is effective for STAD. Since the clinical use of this method, the identification of biomarkers for cancer diagnosis, efficacy, and prognosis has become a top priority in oncology immunotherapy research. The

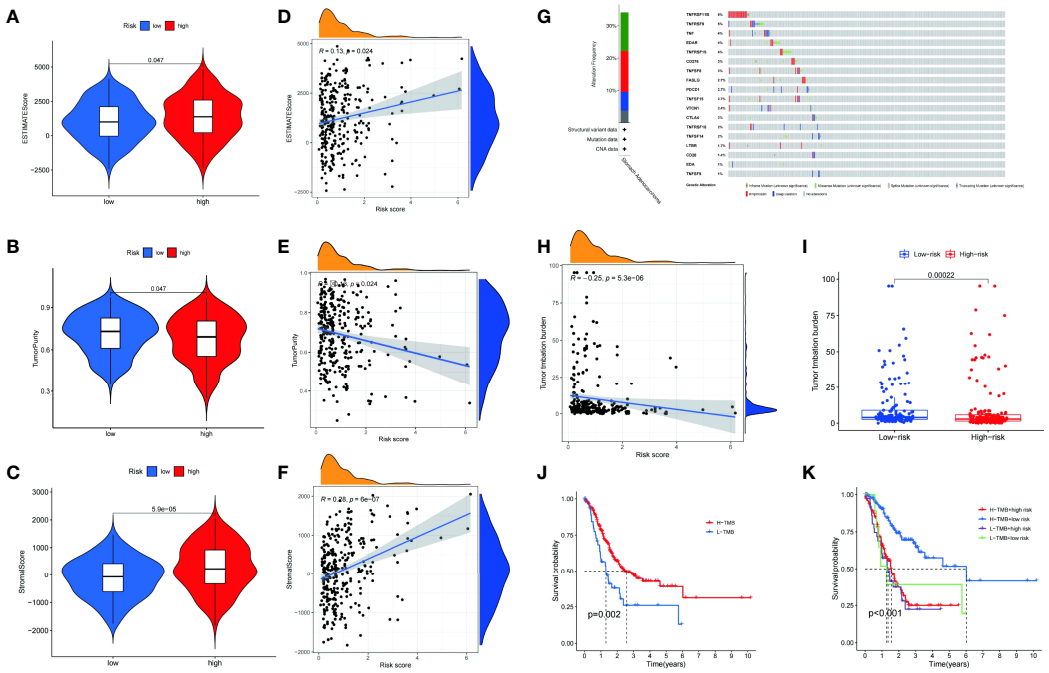


FIGURE 7 | The estimate scores **(A)**, tumor purity **(B)**, and stromal scores **(C)** differed between the high- and low-risk patients. Spearman correlation analysis of the risk scores, estimate scores **(D)**, tumor purity **(E)**, and stromal scores **(F)**. Genetic mutations: types and frequencies **(G)**. Spearman correlation analysis of tumor mutation burden (TMB) and risk score **(H)**. Difference in TMB between the high- and low-risk patients **(I)**. Kaplan–Meier curve for the overall survival of the high- and low-TMB patients **(J)**. Kaplan–Meier curve for the overall survival of the high-TMB/high-risk, high-TMB/low-risk, low-TMB/high-risk, and low-TMB/low-risk patients **(K)**.

regulation of tumor immunity relies heavily on costimulatory molecules (45, 46). Monoclonal antibodies that target the PD-1/PD-L1 (B7-H1) or B7-2/CTLA-4 pathways have been shown to be promising in promoting long-term tumor regression in a range of cancers (47, 48). Costimulatory chemicals are responsible for all the therapeutic targets. However, there are few studies on the role of CMGs in the prognosis of STAD.

We obtained 60 members of the B7-CD28 and TNF families from patients with STAD in our study. To investigate the expression level and prognostic significance of the costimulatory molecules in STAD, we selected nine CMGPs (CD276|LTBR, CD28|CTLA4, EDA|VTCN1, EDAR|TNFRSF19, FASLG|TNFSF8, PDCD1|TNFRSF9, TNF|TNFSF14, TNFRSF11B|TNFSF15, and TNFRSF18|TNFSF9). The B7-CD28 family includes B7-H3 (CD276), an essential immunological checkpoint. B7-H3 is a protein produced by antigen-presenting cells and is involved in the suppression of T-cell activity. More importantly, it is overexpressed in a variety of human solid tumors and is often associated with poor prognosis among patients (49). The importance of the members of the B7-CD28 family and their ligands in immune activity has been demonstrated. However, many parts of CD28 biological activity remain unknown, and its translation into immunomodulatory treatments is uneven (50). TNF superfamily ligands have a wide range of biological activities, including cell death, survival, and proliferation, making them ideal therapeutic targets for cancer immunotherapy (51). Several members of the TNF family investigated in our study play pivotal roles in the immunotherapy of multiple cancers. However, these costimulatory molecules are novel and require further investigation in patients with STAD.

With the advancement of immunotherapy, it is critical to find biomarkers and select the most sensitive individuals to increase immunotherapy response rates. To investigate the overall prognostic value, we used a consensus clustering approach based on the 60 CMGs. According to the Kaplan–Meier curves, the patients in cluster 2 had a worse prognosis. These patients also had a high concentration of immune-related pathways, indicating that CMGPs are closely associated with the TIME. Patients classified under cluster 2 may have a poorer prognosis owing to immune system weakness or a limited immunological response.

Risk profiles based on CMGs might provide fresh insights into the clinical care of patients with STAD. In colorectal cancer (52) and lung adenocarcinoma, risk signatures based on CMGs have been developed (25). All these prognostic markers have been shown to be accurate and perform well. However, we were the first to develop a risk profile for patients with STAD based on CMGs. The performance of our prognostic signature was tested using the TCGA and E-MTAB-3267 datasets, both of which yielded positive results. We also discovered that the predictive signature was strongly linked to the clinical parameters, suggesting that it may be used as a complement to help guide treatment. We also selected genes that were substantially associated with the risk score of our prognostic signature, and

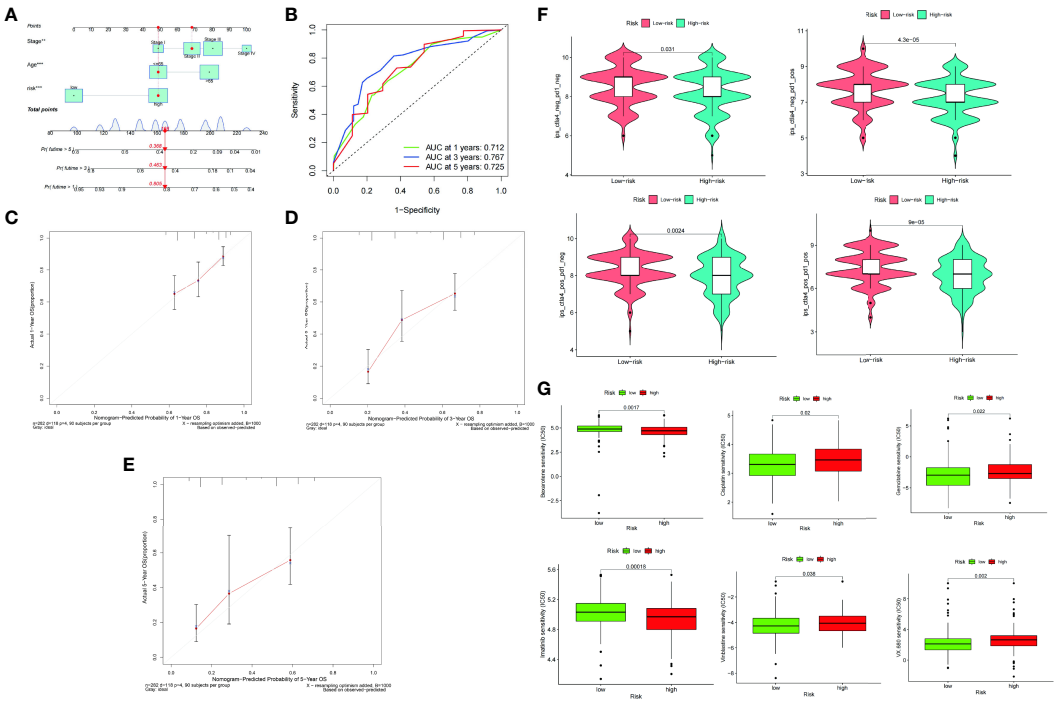


FIGURE 8 | A nomogram was constructed on the basis of age, pathological stage, and risk score **(A)**. ROC curve of the nomogram at 1, 3, and 5 years **(B)**. Calibration plots of overall survival (OS) at 1 **(C)**, 3 **(D)**, and 5 years **(E)**. Differences in PD-1 and CTLA-4 therapy sensitivity between the low- and high-risk populations **(F)**. Differences in chemotherapy sensitivity between the low- and high-risk populations **(G)**.

the functional enrichment analysis revealed that T-cell homeostasis and NF-B signaling were enriched.

To further determine the links between our signature and the TIME, we analyzed immune cell infiltration and tumor mutation patterns in the high- and low-risk groups. In our analysis, the high-risk patients exhibited a much greater immune cell infiltration than did the low-risk patients. In addition, we found that the number of immunosuppressive cells, such as gamma delta T cells, MDSCs, monocytes, immature dendritic cells, macrophages, plasmacytoid dendritic cells, T follicular helper cells, and regulatory T cells, was larger in the high-risk patients than in the low-risk patients, indicating the presence of an immunosuppressive microenvironment. Tumor cells use an immunosuppressive microenvironment to evade immune responses and accelerate disease development. Understanding the immunological microenvironment of each patient will help identify patients who are more likely to respond to immunotherapy and enhance treatment response rates when combined with innovative treatment options.

TMB generally refers to the number of somatic non-synonymous mutations per megabase pair in a specific genomic region. It can indirectly reflect the ability and degree of tumor production of neoantigens and has been proven to predict the efficacy of immunotherapy for a variety of tumors (34, 53). Tumor-specific mutated genes can produce new proteins that are delivered by the major histocompatibility complex as well as their degradation products. They are present on the surface of tumor cells to form tumor neoantigens, which are then recognized by activated $CD8^+$ T cells, thereby triggering tumor-targeted immune responses. Therefore, tumor gene mutations are considered the premise of antitumor immunotherapy (54). In recent years, an increasing number of studies have confirmed that tumors with higher TMB have higher neoantigen loads and are more likely to benefit from ICI therapy (55). To a certain extent, TMB reflects DNA repair damage in tumor cells and is closely related to the ability to generate tumor neoantigens (56). In 2014, TMB was first confirmed to correlate with the efficacy of the CTLA-4 antibody in the treatment of malignant melanoma (57). In 2015, tissue TMB (tTMB) was shown to be associated with the efficacy of PD-1 antibody treatment in patients with non-small-cell lung cancer (58, 59). A meta-analysis conducted in 2017 found that tTMB had a significant predictive effect on the efficacy of immunotherapy for 27 tumor types. There was a significant correlation between tTMB and ORR ($p < 0.001$), suggesting that tTMB is strongly correlated with the efficacy of PD-1/PDL1 antibodies (60). In this study, we found that the high-risk patients had a higher TMB than the low-risk patients, which may boost immunogenicity and result in a higher immunotherapy response rate. However, clinical trials in actual clinical settings are required to corroborate the above-mentioned outcomes.

This study had certain limitations. The data for this study were gathered retrospectively from public sources. The clinical indicators evaluated in this research were incomplete because of the limited number of datasets, including prognostic data for patients with STAD. Calculating the value of a prognostic signature requires

actual prognostic information from patients with STAD. The genes were also limited to costimulatory molecules in this study, although the immunological TME was highly spatially heterogeneous. Consequently, the potency of the predictive signature is restricted. Furthermore, no evidence of CMG expression in the patients with STAD following immunotherapy was observed. Consequently, the risk signature utilized to evaluate immunotherapy response was indirect. Further prospective trials of immunotherapy in patients with STAD are needed to determine the therapeutic applicability of our signature.

In conclusion, we performed the first comprehensive study of costimulatory molecules in patients with STAD and identified nine pairs of genes with prognostic and diagnostic values. We created a costimulatory molecular-based prediction signature for patients with STAD and investigated its molecular underpinnings. With the use of our prognostic signature, the patients with a high mutation frequency, a large quantity of immune cell infiltration, and an immunosuppressive milieu were classified as high-risk patients. Taken together, our signature may help doctors in predicting the prognosis and selecting appropriate therapy for patients with STAD.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analysed during the current study are available in The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/), Gene Expression Omnibus (GEO-GSE15459) database (http://www.ncbi.nlm. nih.gov/geo) and Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/). The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. HL, WS, and BC contributed to the conception of the study. BC, DM, CL, QZ, JD, SL, and RB performed the data analyses. BC, YY, XingW, SS, LZ, XinW, SC, WX, CS, QZ, SL, and RB contributed significantly in writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Scientific Research Cooperation Project Between Anhui Medical University and Hefei First People's Hospital (K202003, chaired by HL), the National-Level Innovation Training Program for Chinese College Students of the Ministry of Education of the People's Republic of China (202110366010, chaired by BC) and the Research Level Improvement Project of Anhui Medical University (2021xkjT001, chaired by WS).

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.cn) for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.928742/full#supplementary-material

Supplementary Figure 1 | Spearman correlation analysis between the risk score and clinical features, including sex **(A)**, age **(B)**, grade **(C)**, pathological stage **(D)**, T stage **(E)**, N stage **(F)**, and M stage **(G)**. Sex, age, stage, and risk score were all found to be independent predictive variables in the univariate regression analysis **(H)**. Age, stage, and risk score were all found to be independent predictive variables in the multivariate regression analysis **(D)**.

Supplementary Figure 2 | Enriched gene sets in the KEGG pathway in the high- **(A)** and low-risk patients **(B)**. Enriched gene sets in the GO pathway in the high- **(C)** and low-risk patients **(D)**.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
2. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric Cancer. *Lancet* (2020) 396(10251):635–48. doi: 10.1016/s0140-67362031288-5
3. Tan Z. Recent Advances in the Surgical Treatment of Advanced Gastric Cancer: A Review. *Med Sci Monit* (2019) 25:3537–41. doi: 10.12659/msm.916475
4. Salas-Benito D, Pérez-Gracia JL, Ponz-Sarvisé M, Rodriguez-Ruiz ME, Martínez-Forero I, Castañón E, et al. Paradigms on Immunotherapy Combinations With Chemotherapy. *Cancer Discov* (2021) 11(6):1353–67. doi: 10.1158/2159-8290.Cd-20-1312
5. Digklia A, Wagner AD. Advanced Gastric Cancer: Current Treatment Landscape and Future Perspectives. *World J Gastroenterol* (2016) 22 (8):2403–14. doi: 10.3748/wjg.v22.i8.2403
6. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Csoszi T, Fulop A, et al. Pembrolizumab Versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med* (2016) 375(19):1823–33. doi: 10.1056/NEJMoa1606774
7. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Csoszi T, Fulop A, et al. Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater. *J Clin Oncol* (2019) 37(7):537–46. doi: 10.1200/JCO.18.00149
8. Sun JM, Shen L, Shah MA, Enzinger P, Adenis A, Doi T, et al. Pembrolizumab Plus Chemotherapy Versus Chemotherapy Alone for First-Line Treatment of Advanced Oesophageal Cancer KEYNOTE-590: A Randomised, Placebo-Controlled, Phase 3 Study. *Lancet* (2021) 398(10302):759–71. doi: 10.1016/S0140-67362101234-4
9. Rosenberg JE, Hoffman-Censits J, Powles T, van der Heijden MS, Balar AV, Necchi A, et al. Atezolizumab in Patients With Locally Advanced and Metastatic Urothelial Carcinoma Who Have Progressed Following Treatment With Platinum-Based Chemotherapy: A Single-Arm, Multicentre, Phase 2 Trial. *Lancet* (2016) 387(10031):1909–20. doi: 10.1016/S0140-67361600561-4

10. Janjigian YY, Shitara K, Moehler M, Garrido M, Salman P, Shen L, et al. First-Line Nivolumab Plus Chemotherapy Versus Chemotherapy Alone for Advanced Gastric, Gastro-Oesophageal Junction, and Oesophageal Adenocarcinoma CheckMate 649: A Randomised, Open-Label, Phase 3 Trial. *Lancet* (2021) 398 (10294):27–40. doi: 10.1016/S0140-67362100797-2
11. Burtness B, Harrington KJ, Greil R, Soulieres D, Tahara M, de Castro G, Jr , et al. Pembrolizumab Alone or With Chemotherapy Versus Cetuximab With Chemotherapy for Recurrent or Metastatic Squamous Cell Carcinoma of the Head and Neck KEYNOTE-048: A Randomised, Open-Label, Phase 3 Study. *Lancet* (2019) 394(10212):1915–28. doi: 10.1016/S0140-67361932591-7
12. Song Y, Tong C, Wang Y, Gao Y, Dai H, Guo Y, et al. Effective and Persistent Antitumor Activity of HER2-Directed CAR-T Cells Against Gastric Cancer Cells. *Vitro Xenotransplanted Tumors Vivo Protein Cell* (2018) 9(10):867–78. doi: 10.1007/s13238-017-0384-8
13. Zhao Q, Cao L, Guan L, Bie L, Wang S, Xie B, et al. Immunotherapy for Gastric Cancer: Dilemmas and Prospect. *Brief Funct Genomics* (2019) 18 (2):107–12. doi: 10.1093/bfgp/ely019
14. Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* (2014) 513(7517):202–9. doi: 10.1038/nature13480
15. Georganaki M, van Hooren L, Dimberg A. Vascular Targeting to Increase the Efficiency of Immune Checkpoint Blockade in Cancer. *Front Immunol* (2018) 9:3081. doi: 10.3389/fimmu.2018.03081
16. Mesri M, Birse C, Heidbrink J, McKinnon K, Brand E, Bermingham CL, et al. Identification and Characterization of Angiogenesis Targets Through Proteomic Profiling of Endothelial Cells in Human Cancer Tissues. *PLoS One* (2013) 8(11):e78885. doi: 10.1371/journal.pone.0078885
17. Seaman S, Stevens J, Yang MY, Logsdon D, Graff-Cherry C, St Croix B, et al. Genes That Distinguish Physiological and Pathological Angiogenesis. *Cancer Cell* (2007) 11(6):539–54. doi: 10.1016/j.ccr.2007.04.017
18. Wei SC, Duffy CR, Allison JP. Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov* (2018) 8(9):1069–86. doi: 10.1158/2159-8290.Cd-18-0367
19. Croft M, Benedict CA, Ware CF. Clinical Targeting of the TNF and TNFR Superfamilies. *Nat Rev Drug Discov* (2013) 12(2):147–68. doi: 10.1038/nrd3930
20. Schildberg FA, Klein SR, Freeman GJ, Sharpe AH. Coinhibitory Pathways in the B7-CD28 Ligand-Receptor Family. *Immunity* (2016) 44(5):955–72. doi: 10.1016/j.immuni.2016.05.002
21. Hua X, Ge S, Zhang J, Xiao H, Tai S, Yang C, et al. A Costimulatory Molecule-Related Signature in Regard to Evaluation of Prognosis and Immune Features for Clear Cell Renal Cell Carcinoma. *Cell Death Discov* (2021) 7(1):252. doi: 10.1038/s41420-021-00646-2
22. Lin G, Yang Y, Feng Q, Zhan F, Sun C, Niu Y, et al. Prognostic Implication and Immunotherapy Response Prediction of a Costimulatory Molecule Signature in Kidney Renal Clear Cell Carcinoma. *Immunogenetics* (2022) 74 (3):285–301. doi: 10.1007/s00251-021-01246-1
23. Ge S, Hua X, Chen J, Xiao H, Zhang L, Zhou J, et al. Identification of a Costimulatory Molecule-Related Signature for Predicting Prognostic Risk in Prostate Cancer. *Front Genet* (2021) 12:666300. doi: 10.3389/fgene.2021.666300
24. Hu Y, Liu J, Yu J, Yang F, Zhang M, Liu Y, et al. Identification and Validation a Costimulatory Molecule Gene Signature to Predict the Prognosis and Immunotherapy Response for Hepatocellular Carcinoma. *Cancer Cell Int* (2022) 22(1):97. doi: 10.1186/s12935-022-02514-0
25. Zhang C, Zhang Z, Sun N, Zhang Z, Zhang G, Wang F, et al. Identification of a Costimulatory Molecule-Based Signature for Predicting Prognosis Risk and Immunotherapy Response in Patients With Lung Adenocarcinoma. *Oncoimmunology* (2020) 9(1):1824641. doi: 10.1080/2162402x.2020.1824641
26. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool With Confidence Assessments and Item Tracking. *Bioinformatics* (2010) 26 (12):1572–3. doi: 10.1093/bioinformatics/btq170
27. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression. *Genome Biol* (2016) 17 (1):218. doi: 10.1186/s13059-016-1070-5
28. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007
30. Hanzelmann S, Castelo R, Guinney J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7
31. Lin H, Zelterman DJT. Modeling Survival Data: Extending the Cox Model. *Technometrics 44.1* (2000) 44:85–6. doi: 10.1198/tech.2002.s656
32. Dong B, Liang J, Li D, Song W, Song J, Zhu M, et al. Identification of a Prognostic Signature Associated With the Homeobox Gene Family for Bladder Cancer. *Front Mol Biosci* (2021) 8:688298. doi: 10.3389/fmolb.2021.688298
33. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and Comparing Time-Dependent Areas Under Receiver Operating Characteristic Curves for Censored Event Times With Competing Risks. *Stat Med* (2013) 32 (30):5381–97. doi: 10.1002/sim.5958
34. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational Burden. *Genome Med* (2017) 9(1):34. doi: 10.1186/s13073-017-0424-2
35. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. Clusterprofiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation Camb.* (2021) 2(3):100141. doi: 10.1016/j.xinn.2021.100141
36. Pan X, Jin X, Wang J, Hu Q, Dai B. Placenta Inflammation is Closely Associated With Gestational Diabetes Mellitus. *Am J Transl Res* (2021) 13 (5):4068–79.
37. Geeleher P, Cox N, Huang RS. Prrophetic: An R Package for Prediction of Clinical Chemotherapeutic Response From Tumor Gene Expression Levels. *PLoS One* (2014) 9(9):e107468. doi: 10.1371/journal.pone.0107468
38. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. *PLoS One* (2015) 10(1): e0116774. doi: 10.1371/journal.pone.0116774
39. Ilson DH. Advances in the Treatment of Gastric Cancer: 2019. *Curr Opin Gastroenterol* (2019) 35(6):551–4. doi: 10.1097/mog.0000000000000577
40. Aoyama T, Nishikawa K, Fujitani K, Tanabe K, Ito S, Matsui T, et al. Early Results of a Randomized Two-By-Two Factorial Phase II Trial Comparing Neoadjuvant Chemotherapy With Two and Four Courses of Cisplatin/S-1 and Docetaxel/Cisplatin/S-1 as Neoadjuvant Chemotherapy for Locally Advanced Gastric Cancer. *Ann Oncol* (2017) 28(8):1876–81. doi: 10.1093/annonc/mdx236
41. Ford HE, Marshall A, Bridgewater JA, Janowitz T, Coxon FY, Wadsley J, et al. Docetaxel Versus Active Symptom Control for Refractory Oesophagogastric Adenocarcinoma COUGAR-02: An Open-Label, Phase 3 Randomised Controlled Trial. *Lancet Oncol* (2014) 15(1):78–86. doi: 10.1016/s1470-20451370549-7
42. Bang YJ, Van Cutsem E, Feyereislova A, Chung HC, Shen L, Sawaki A, et al. Trastuzumab in Combination With Chemotherapy Versus Chemotherapy Alone for Treatment of HER2-Positive Advanced Gastric or Gastro-Oesophageal Junction Cancer ToGA: A Phase 3, Open-Label, Randomised Controlled Trial. *Lancet* (2010) 376(9742):687–97. doi: 10.1016/s0140-67361061121-x
43. Fuchs CS, Doi T, Jang RW, Muro K, Satoh T, Machado M, et al. Safety and Efficacy of Pembrolizumab Monotherapy in Patients With Previously Treated Advanced Gastric and Gastroesophageal Junction Cancer: Phase 2 Clinical KEYNOTE-059 Trial. *JAMA Oncol* (2018) 4(5):e180013. doi: 10.1001/jamaoncol.2018.0013
44. Kang YK, Boku N, Satoh T, Ryu MH, Chao Y, Kato K, et al. Nivolumab in Patients With Advanced Gastric or Gastro-Oesophageal Junction Cancer Refractory to, or Intolerant of, at Least Two Previous Chemotherapy Regimens ONO-4538-12, ATTRACTION-2: A Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial. *Lancet* (2017) 390(10111):2461–71. doi: 10.1016/s0140-67361731827-5
45. Pitt JM, André F, Amigorena S, Soria JC, Eggermont A, Kroemer G, et al. Dendritic Cell-Derived Exosomes for Cancer Therapy. *J Clin Invest* (2016) 126 (4):1224–32. doi: 10.1172/jci81137

46. Topalian SL, Taube JM, Pardoll DM. Neoadjuvant Checkpoint Blockade for Cancer Immunotherapy. *Science* (2020) 367(6477). doi: 10.1126/science.aax0182
47. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, Activity, and Immune Correlates of Anti-PD-1 Antibody in Cancer. *N Engl J Med* (2012) 366(26):2443–54. doi: 10.1056/NEJMoa1200690
48. Topalian SL, Drake CG, Pardoll DM. Targeting the PD-1/B7-H1PD-L1 Pathway to Activate Anti-Tumor Immunity. *Curr Opin Immunol* (2012) 24 (2):207–12. doi: 10.1016/j.coi.2011.12.009
49. Picarda E, Ohaegbulam KC, Zang X. Molecular Pathways: Targeting B7-H3 CD276 for Human Cancer Immunotherapy. *Clin Cancer Res* (2016) 22 (14):3425–31. doi: 10.1158/1078-0432.Ccr-15-2428
50. Esensten JH, Helou YA, Chopra G, Weiss A, Bluestone JA. CD28 Costimulation: From Mechanism to Therapy. *Immunity* (2016) 44(5):973–88. doi: 10.1016/j.immuni.2016.04.020
51. Vanamee É S, Faustman DL. On the TRAIL of Better Therapies: Understanding TNFRSF Structure-Function. *Cells* (2020) 9(3):764. doi: 10.3390/cells9030764
52. Tang J, Jiang W, Liu D, Luo J, Wu X, Pan Z, et al. The Comprehensive Molecular Landscape of the Immunologic Co-Stimulator B7 and TNFR Ligand Receptor Families in Colorectal Cancer: Immunotherapeutic Implications With Microsatellite Instability. *Oncoimmunology* (2018) 7(10): e1488566. doi: 10.1080/2162402x.2018.1488566
53. Schumacher TN, Schreiber RD. Neoantigens in Cancer Immunotherapy. *Science* (2015) 348(6230):69–74. doi: 10.1126/science.aaa4971
54. Fumet JD, Truntzer C, Yarchoan M, Ghiringhelli F. Tumour Mutational Burden as a Biomarker for Immunotherapy: Current Data and Emerging Concepts. *Eur J Cancer* (2020) 131:40–50. doi: 10.1016/j.ejca.2020.02.038
55. Buttner R, Longshore JW, Lopez-Rios F, Merkelbach-Bruse S, Normanno N, Rouleau E, et al. Implementing TMB Measurement in Clinical Practice: Considerations on Assay Requirements. *ESMO Open* (2019) 4(1):e000442. doi: 10.1136/esmoopen-2018-000442
56. Luchini C, Bibeau F, Ligtenberg MJL, Singh N, Nottegar A, Bosse T, et al. ESMO Recommendations on Microsatellite Instability Testing for Immunotherapy in Cancer, and Its Relationship With PD-1/PD-L1 Expression and Tumour Mutational Burden: A Systematic Review-Based Approach. *Ann Oncol* (2019) 30(8):1232–43. doi: 10.1093/annonc/mdz116
57. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N Engl J Med* (2014) 371(23):2189–99. doi: 10.1056/NEJMoa1406498
58. Hellmann MD, Nathanson T, Rizvi H, Creelan BC, Sanchez-Vega F, Ahuja A, et al. Genomic Features of Response to Combination Immunotherapy in Patients With Advanced Non-Small-Cell Lung Cancer. *Cancer Cell* (2018) 33 (5):843–52.e844. doi: 10.1016/j.ccell.2018.03.018
59. Hellmann MD, Ciuleanu TE, Pluzanski A, Lee JS, Otterson GA, Audigier-Valette C, et al. Nivolumab Plus Ipilimumab in Lung Cancer With a High Tumor Mutational Burden. *N Engl J Med* (2018) 378(22):2093–104. doi: 10.1056/NEJMoa1801946
60. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor Mutational Load Predicts Survival After Immunotherapy Across Multiple Cancer Types. *Nat Genet* (2019) 51(2):202–6. doi: 10.1038/s41588-018-0312-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chen, Yao, Mao, Li, Wang, Sheng, Zhang, Wang, Chen, Xu, Deng, Sun, Zhou, Lowe, Bentley, Shao and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author (s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fonc.2022.934711

Oncogenic Role of miR-217 During Clear Cell Renal Carcinoma Progression

Jose María Zamora-Fuentes[1], *Enrique Hernández-Lemus*[1,2] *and Jesús Espinal-Enríquez*[1,2]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

Clear cell renal carcinoma (ccRC) comprises a set of heterogeneous, fast-progressing pathologies with poor prognosis. Analyzing ccRC progression in terms of modifications at the molecular level may provide us with a broader understanding of the disease, paving the way for improved diagnostics and therapeutics. The role of micro-RNAs (miRs) in cancer by targeting both oncogenes and tumor suppressor genes is widely known. Despite this knowledge, the role of specific miRs and their targets in the progression of ccRC is still unknown. To evaluate the action of miRs and their target genes during ccRC progression, here we implemented a three-step method for constructing miR–gene co-expression networks for each progression stage of ccRC as well as for adjacent-normal renal tissue (NT). In the first step, we inferred all miR–gene co-expression interactions for each progression stage of ccRC and for NT. Afterwards, we filtered the whole miR–gene networks by differential gene and miR expression between successive stages: stage I with non-tumor, stage II with stage I, and so on. Finally, all miR–gene interactions whose relationships were inversely proportional (overexpressed miR and underexpressed genes and *vice versa*) were kept and removed otherwise. We found that miR-217 is differentially expressed in all contrasts; however, its targets were different depending on the ccRC stage. Furthermore, the target genes of miR-217 have a known role in cancer progression —for instance, in stage II network, GALNTL6 is overexpressed, and it is related to cell signaling, survival, and proliferation. In the stage III network, WNK2, a widely known tumor suppressor, is underexpressed. For the stage IV network, IGF2BP2, a post-transcriptional regulator of MYC and PTEN, is overexpressed. This data-driven network approach has allowed us to discover miRs that have different targets through ccRC progression, thus providing a method for searching possible stage-dependent therapeutic targets in this and other types of cancer.

Keywords: clear cell renal carcinoma, gene co-expression networks, WKN2, GALNTL6, IGF2BP2, miR-217, cancer progression stages, renal carcinoma progression

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, United States

Reviewed by:
Rebecca Kusko,
Immuneering Corporation, United States
Xuemei Chen,
Chongqing Medical University, China

***Correspondence:**
Jesús Espinal-Enríquez
jespinal@inmegen.gob.mx

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *03 May 2022*
Accepted: *13 June 2022*
Published: *22 July 2022*

Citation:
Zamora-Fuentes JM, Hernández-Lemus E and Espinal-Enríquez J (2022) Oncogenic Role of miR-217 During Clear Cell Renal Carcinoma Progression. Front. Oncol. 12:934711. doi: 10.3389/fonc.2022.934711

INTRODUCTION

The global incidence of renal cell carcinomas (RCC) has notoriously increased since 2008, exerting an important burden in both individuals and health systems (1). A number of basic and clinical endeavors have been implemented to try to alleviate this situation. Important efforts have been made in searching for key regulators in the development of this disease. Oncogenes and tumor suppressors such as VHL (3p26), FH (1q42.1), MET (7q34), or FLCN (17p11.2) genes have been studied in different types of RCC. These have been associated with different syndromes and inheritance patterns (2).

However, up to 70% of RCC cases correspond to the clear cell subtype (ccRC). The progression in this tumor subtype is commonly initiated by mutations in VHL. Some transcription factors are accumulated due to VHL inactivation, which induces the expression of vascular endothelial growth factor (VEGF). Therefore, ccRCCs are often highly vascularized and respondent to anti-angiogenic therapy (3). Subsequent mutations commonly arise in BAP1/PBRM1/SET2/KDM5C, giving rise to DNA repair defects. These genes are then considered as gene drivers for ccRC evolution. Moreover, activation of the PI3K pathway promotes metastases (4). Considering that up to a third of cases will present metastases, the importance of determining with a higher accuracy the molecular factors underlying the progression of ccRC is undeniable. Additionally, there is evidence that VHL inactivation in humans and mice does not directly induce ccRC tumorigenesis (5).

Regarding immune responses, the relationship of the microenvironment with ccRC progression is not clear. However, efforts have been made to find the patterns of macrophages and T cells, characterized by a wide diversity, both in phenotypes and responses (6). To overcome this challenge, efforts have been made to find markers of the different elements of immune response—for instance, in (7), it was found that individuals with inflammatory responses enriched for BAP1 have a worse prognosis. PBRM1 has also been found in both animal models and in human samples with decreased immune infiltration. It has also been observed that PBRM1-knocked-out tumors were more resistant to anti-PD-1 antibody (8).

Tumor microenvironment heterogeneity is indeed just partially responsible for the complexity behind ccRC responses. Regulatory elements and epigenomic modulators are known to be also playing relevant roles—for instance, it has been argued that micro-RNAs (miRs) appear to regulate more than 60% of human genes (9). Furthermore, aberrant expression patterns of miRNA have been reported in many human cancers (10). A number of these genes are considered as key factors in cancer development pathways (11). To mention just a few, miRs such as miR-646, miR-21, and miR-204 have been implicated in the development and progression of renal cell carcinoma (12). miR families, such as miR-200 family, have also been reported to be strongly dysregulated in metastases and met-like primary tumors (13, 14).

The effects of miRNAs over gene regulation are complex and highly context dependent, varying by cell type as well as by the severity and persistence of conditions in cell signaling and other processes, including genomic damage (15). The effects that miRs exert on gene expression are often mostly attributed to the miR–mRNA 3′ untranslated region (UTR) interactions. These interactions lead to target post-translational inhibition or degradation. However, focusing on this mechanism to design miR therapeutics is likely proven to be too simplistic, owing in part to other emerging micro-RNA mechanisms, which include decoy activity and 5′ UTR and direct DNA regulatory activities.

Alternatively, miRNAs can be associated to the development of tumor-suppressive and oncogenic functions, and their ability to modulate different genes may be also context dependent (3). Specifically, miR–gene regulation may repress or promote transcription (non-canonical) or translation (canonical) (15).

It is widely known that a single miR can regulate several mRNAs and that a single mRNA transcript can be targeted by several miRs (16). To broadly understand the intrinsic complexity of a miR–gene interaction, developing integral approaches that combine different sources of information becomes mandatory. Additionally, it is necessary to take into account the complex nature of miR–gene regulation and its many associated mechanisms.

Abnormalities in cell behavior that involve the dysregulation of gene and miR expression have been argued to play relevant roles in triggering carcinogenic processes. Several studies have confirmed, for instance, that overexpression of miRs has the potential to promote cancer development [for a broad review, see (17–19)]—for example, miR-203 has been related to follicular grow factor 2 (FGF2) and CAV1 as a downstream regulator, affecting pathways such as PI3K/AKT/mTOR (20). There is also evidence that miRs such as miR-141, miR-200a, or miR-200b may serve as drivers in the epithelium-to-mesenchymal transition and the complementary process, the mesenchymal-to-epithelium transition, by inhibiting the expression of VIM, ZEB1, or ZEB2 genes (14). Processes such as proliferation, migration, invasion, or apoptosis can be altered by miR-203 by targeting FGF2 (12).

Despite the ever-growing evidence of the role that miRs exert on the oncogenic process, a comprehensive list of oncomiRs or tumor suppressor miRs, particularly for ccRC, is still lacking (11).

Building such a comprehensive catalog is indeed easier said than done. Important steps have been taken, however, in this direction. With the advent of next-generation sequencing, gene expression profiles have been extensively used to discover crucial features based on the expression of certain genes that may drive the reconfiguration of the transcriptional program, often leading to dramatic effects on the phenotype (21–25).

Although the importance of genetic expression in cancer is out of discussion, it is also clear that the gene regulation during the carcinogenic process is strongly altered by several components. Additionally, the gene expression landscape often does not provide information on how those genes are regulated (26).

To overcome the latter challenge, a common approach used for high-throughout-derived datasets is the gene co-expression

network (GCN). These networks are commonly inferred by correlating the expression profile of gene couples with multiple samples. GCNs offer a framework that allows the analysis of global changes in a given phenotype, such as cancer. With this approach, the statistical dependency of a given gene can be quantified by the expression of any other gene (27–31).

In relation to ccRC, we recently demonstrated that differential gene expression profiles are quite similar between progression stages; however, the gene co-expression networks observed in those stages resulted different in terms of structure and also the associated biological processes involved in such networks (32).

The evidence of the role of miRs in the rise and development of cancer, in particular for clear cell renal carcinoma, is increasing. miR alterations may be key factors in the development and progression of ccRC to more advanced stages. However, the specific role of miRs during the progression stages of ccRC is still unknown.

In order to evaluate the role of the miRs–gene relationships in ccRC progression, here we implemented a three-step method for constructing miR–gene co-expression networks for the four progression stages of clear cell renal carcinoma as well as for healthy renal tissue, with data obtained from The Cancer Genome Atlas (TCGA)-GDC consortium.

In the first step, we inferred all miR–gene co-expression interactions of each progression stage of ccRC and for the healthy renal tissue. In the second step, we filtered the whole miR–gene networks by differentially expressed miRs and genes. We assessed the differentially expressed genes between non-tumor adjacent-to-tissue control samples and each progression stage. However, to establish the progression between stages with higher accuracy, we calculated the differentially expressed miRs and genes from contiguous stages: stage I *vs*. control, stage II *vs*. stage I, and so on.

Once each network was constructed, we conserved the miR–gene interaction whose relationships were inversely proportional (overexpressed miR and underexpressed genes and *vice versa*); otherwise, we removed them. Finally, we observed the shared genes and interactions between cancer stages and also those genes and interactions that resulted unique for each stage. With this data-driven network sifting, we were able to discover miRs that have different targets through the clear cell renal carcinoma progression, thus providing a method for searching possible therapeutic targets in ccRC and other types of cancer.

MATERIALS AND METHODS

In order to carry out the research program just outlined, we have implemented a streamlined analytics methodology. A graphical representation of the workflow followed is shown in **Figure 1**. In the following subsections, we will expound on the different aspects of the workflow just presented.

Data Acquisition

ccRC RNA sequencing data was obtained from TCGA collaboration (33–35). To obtain the gene expression profiles for each progression stage, we started by downloading RNA-Seq level 3 gene expression files for ccRC samples. Additionally, we downloaded the miR profiles for the same samples; therefore, all samples were paired RNA/micro-RNA, and the corresponding metadata is indeed harmonized. Hence, we compiled two main datasets (1): miR expression quantification (reads per million) and (2) isoform expression quantification, which contains detailed information about the transcribed species (as coordinates mapped) for each transcript. This can be used to get mature micro-RNA information.

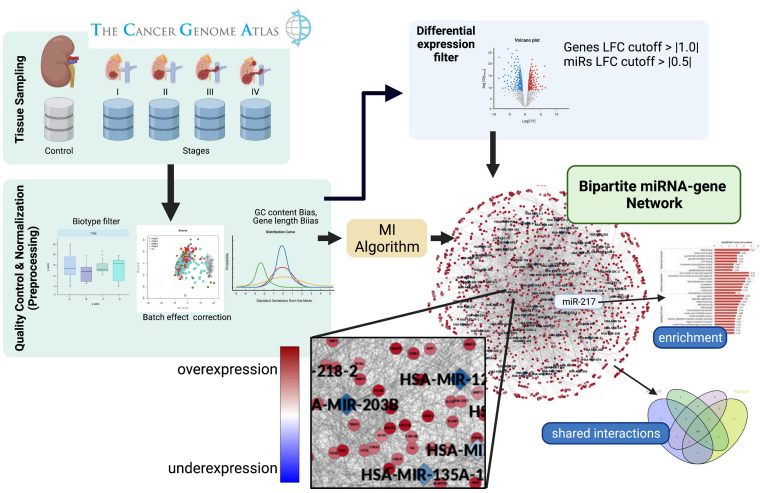


FIGURE 1 | Graphical pipeline. Firstly, sequencing data was obtained from the TCGA-GDC consortium. Secondly, a pre-processing phase was performed, where raw counts were filtered and normalized. Differential expression was calculated to retrieve genes and miRs whose expression became altered between contiguous stages. A bipartite co-expression network (miR–gene) for each progression stage was inferred. After that, we conserved those miR–gene interactions with miRs and genes with opposite differential expression. Finally, an enrichment analysis of the relevant genes obtained by the aforementioned pipeline was performed.

The indexes of both datasets were harmonized to match patient codes as a master key for agglomerating RNAseq and micro-RNA raw counts. A summary of pre-processed data can be seen in **Table 1**.

Clinical Information

We processed the clinical information directly from the TCGA-KIRC project. We categorized all samples by its *tumor_stage* variable. Samples with not-reported stages were removed. The TCGAbiolinks library (V 2.24.1) was used to retrieve data from TCGA.

Pre-processing

We pre-processed gene and miR data as follows: (1) we removed genes without annotation in the BioMart Database, (2) we removed genes with more than 50% of zero counts per sample, and (3) genes with a mean expression of less than 10 counts were also removed. For bias correction, we used the EDASeq R-package (V 2.30.0) (36). In brief, we removed biases in GC content, gene length, and biotype. Finally, in order to correct for possible batch effects, we used the ARSyN method, implemented in R as a function of NOIseq library (V 2.40.0) (37).

After all filters were implemented and the bias removal was performed, the total number of miRs for analysis was 275; meanwhile, the total of genes was 16,224. Those were the entities used to infer miR–gene networks and to perform differential expression analyses. A summary of phenotypes, units of counts, and size of genes and miRs is portrayed in **Table 2**.

Differential miR and Gene Expression

Differential expression analysis was implemented by using the DESeq R package (V 1.8.3) (38). Here we considered differentially expressed genes (DEGs) with the following filters: $|\text{LogF C}| > 1.0$ and FDR-corrected $\text{d p} - \text{value} < 1\text{e} - 5$. Meanwhile for differentially expressed miRs (DEMs), the filters were $|\text{LogF C}| > 0.5$ and p–value < 1e–5. It is worth noticing that the logFC cutoffs depend on the empirical data distributions and the associated dynamic ranges of the measurements of the variables. Even though both RNASeq and miRNASeq were performed with roughly the same technology (Illumina NGS Sequencing), there are indeed differences in the capture rates, the variant calls and annotations, and other issues of the experimental methodologies. Even more important, there are differences in the natural abundance of these two types of transcripts in the samples.

We compared the non-tumor (NT) dataset with all progression stages (st_I , st_{II} , st_{III} , *and* st_{IV}). Additionally, in order to track down the evolution of the tumor progression, we also performed differential expression analysis between contiguous stages (progression contrast) NT- st_I, st_I -st_{II}, and so on. To visualize the DEGs and DEMs, we constructed volcano plots for each contrast with the default specifications of EnhancedVolcano (V1.14.0) package (https://github.com/kevinblighe/EnhancedVolcano).

We observed the number of DEGs and DEMs which appeared for each contrast. We also calculated those unique DEGs and DEMs for each contrast as well as those shared DEGs/DEMs in all contrasts. The code to develop these analyses can be found at the following repository: https://github.com/josemaz/kirc-mirna.

TABLE 1 | Number of harmonized cases for each stage of ccRC and NT.

	NT	st_I	st_{II}	st_{III}	st_{IV}
Number of samples	71	251	55	122	81

TABLE 2 | Summary of genes and miRs.

	Genes	miRNAs
Size	16,224	275
Units	HTSeq - counts	Reads-per-million-miRNA-mapped
Phenotypes	5	5

Network Inference

To analyze the potential role played by miRs in the gene expression program, we inferred five miR–gene networks, one for tumor-adjacent-healthy-tissue (NT) samples, and one for each tumor progression stage. All networks were inferred by using mutual information (MI) as a statistical dependence measure. MI was calculated over the expression values for all miR–gene couples ($275 \times 16,\ 227 \approx 4.5 \times 10^6$ pairwise interactions) for each phenotype. We implemented a multi-thread miR–gene co-expression calculation based on the ARACNe algorithm (27). The code to infer such MI-based networks can be found at https://github.com/josemaz/aracne-multicore.

Network Filtering and Visualization

In order to find dysregulated genes targeted by micro-RNAs, we used both DEGs and DEMs as network filters. Briefly, we conserved the 100,000 highest miR–gene MI interactions to capture the most relevant co-expression relations for any given phenotype. We conserved only those miR–gene interactions in which the micro-RNA and its target have opposite differential expressions: overexpressed miR and underexpressed gene and *vice versa* looking for canonical miR–gene interactions.

Finally, we analyzed the fraction of conserved miR–gene interactions and the fraction of unique interactions for each phenotype. Network visualizations were performed with Cytoscape 3.8.2 (39).

RESULTS AND DISCUSSION

NT and ccRC Contrasts

We performed a multi-group comparison between control and each progression stage. We found a larger number of over-expressed genes and miRs than that of underexpressed ones. **Table 3** shows the comparison between miRs and differentially expressed genes between non-tumor (NT) and each progression stage of ccRC (st_I, st_{II}, st_{III}, *and* st_{IV}). Interestingly, the number of

TABLE 3 | DEGs and DEMs for each progression stage as compared with non-tumor adjacent tissue-derived samples.

	NT–st_I	NT–st_{II}	NT–st_{III}	NT–st_{IV}
Underexpressed genes	1,946	2,012	2,106	2,187
Overexpressed genes	2,187	2,238	2,587	2,630
Overexpressed miRs	88	87	96	91
Underexpressed miRs	87	87	88	88
All genes	4,133	4,250	4,693	4,817
All miRs	175	174	184	179

DEGs increases with progression stages; this may suggest that the whole gene regulatory program becomes more disrupted as the tumor evolves to later stages.

The large difference of DEGs and DEMs between NT and st_I, compared with the rest of contrasts, may be due to the recruitment and accumulation of several different cancer-associated cell types, aside from the intrinsic genomic alterations of cancer cells with respect to normal ones.

Despite the large number of DEGs and DEMs, unique differentially expressed genes or miRs are quite scarce. **Table 4** shows the number of unique DEGs and DEMs per contrast. As observed, the number of unique DEGs/DEMs per contrast is almost 40 times lower than the total amount of DEGs/DEMs.

These results become of particular interest because it appears that most of the DEGs/DEMs are conserved throughout the whole evolution of the disease. However, as we have previously observed in gene–gene co-expression networks (32), differential expression is not sufficient to explain the evolution of the early stages to the more advanced ones.

DEGs and DEMs Are More Abundant Between Control and Stage I Than in Any Other Contrast

Although the differential expression between control and progression stages provides information regarding those genes and miRs that may exert influence on the acquisition of oncogenic traits, a comparison between contiguous stages can be, in some sense, more revealing since it represents the evolution of the gene expression program along tumor progression.

To further investigate on this, we performed a differential expression analysis between sequentially contiguous stages. In **Figures 2**, **3**, we can observe volcano plots showing the DEGs and DEMs between the consecutive stages of ccRC evolution: NT $- st_I$, $st_I - st_{II}$, $st_{II} - st_{III}$, and $st_{III} - st_{IV}$. **Supplementary Material S1** shows the shared and unique genes/miRs for each contrast.

Figure 2 shows the volcano plots for the DEGs. As can be seen, the contrast with the larger number of DEGs is that between NT and st_I , with a total of 2,187 overexpressed genes and 1,946 underexpressed ones. The following contrasts had a number of DEGs more than 100 times lower than the first one. Analogously, in **Figure 3**, we can observe a similar behavior for DEMs. According to these results, in ccRC, the main changes in the gene and miR regulatory programs occur in the initial phase of tumor evolution.

In the latter contrasts, both DEGs and DEMs show specific differences in their expression—for instance, mir-155 (considered as miR regulator of VHL) (40, 41) is overexpressed in the T-stage1 comparison: However, in the following contrasts it is not differentially expressed.

To notice, in the latter contrasts, for both cases of miRs and genes, the list of DEGs and DEMs is different for each contrast. In **Supplementary Material S2**, we provide all contrasts between NT and all ccRC progression stages as well as between sequentially contiguous stages.

The resulting overexpressed and underexpressed miRs and genes were then used to construct the miR–gene networks for each phenotype (NT and the four stages). We conserved only those miR–gene interactions between DEGs and DEMs with an opposite differential expression trend (potentially corresponding to the canonical mechanisms of miR–gene regulation).

miR–Gene Networks Are Mostly Stage Specific

Figure 4 shows an upset plot of the shared miR–gene interactions for each stage. In stark contrast with the high number of shared genes among DEGs and DEMS, in the case of miR–gene networks, there is only a small subset of interactions that are shared between networks. More than 90% of the miR–gene interactions are unique for each network.

This result was seemingly counter-intuitive at first since the number of shared genes and miRs between contrasts was very

TABLE 4 | Unique differentially expressed genes and miRs for each non-tumor stage contrast.

	NT–st_I	NT–st_{II}	NT–st_{III}	NT–st_{IV}
Overexpressed genes	52	54	61	189
Underexpressed genes	56	58	30	141
Overexpressed miRs	2	2	4	4
Underexpressed miRs	1	1	1	3

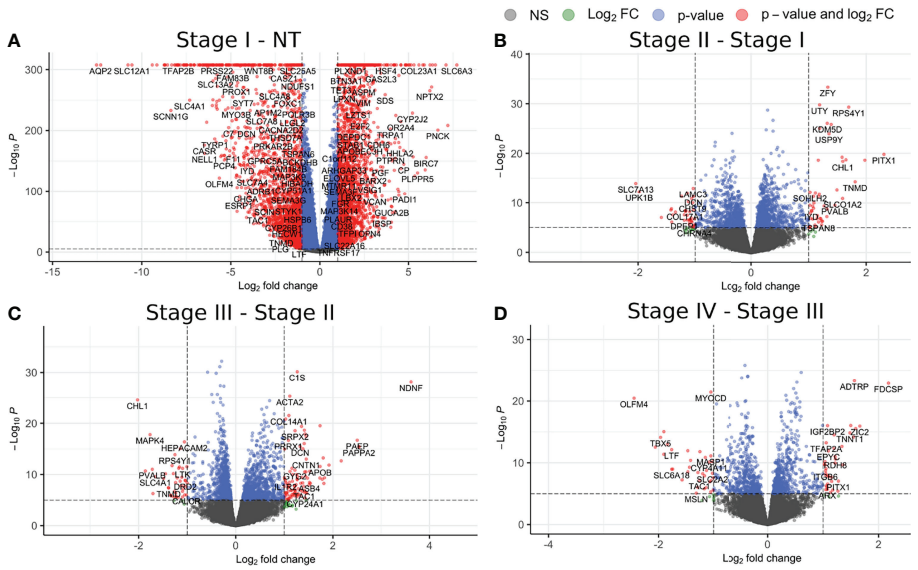


FIGURE 2 | Differentially expressed genes for each contiguous stage of ccRC. **(A)** Contrast between control and stage I; **(B)** stage I V, stage II; **(C)** stage II V, stage III; **(D)** stage III V, stage IV. Red circles represent genes with a $|\log_2 FC| > 1$ and a *p*-value <1e-5; circles depicted in green take account for those genes with a $|\log_2 FC| > 1$ but *p*-value <1e-5; those genes with a $|\log_2 FC| < 1$ but a *p*-value <1e-5 are depicted in blue. Finally, those genes with values lower than those thresholds are depicted in gray. It becomes evident that the contrast with more DEGs is the one between NT and st_I.

high. However, the miR–gene regulatory programs, as represented by high-confidence co-expression networks, are indeed highly specific for each progression stage.

A concomitant result derived from the uniqueness of miR–gene interactions for each progression stage is that there is a small set of shared interactions between cancer stages but not shared with the NT network. Only 33 interactions are common for the four stages and not presented in the non-tumor phenotype.

By looking at those interactions, it can be appreciated that practically all of them correspond to miRs and genes that belong

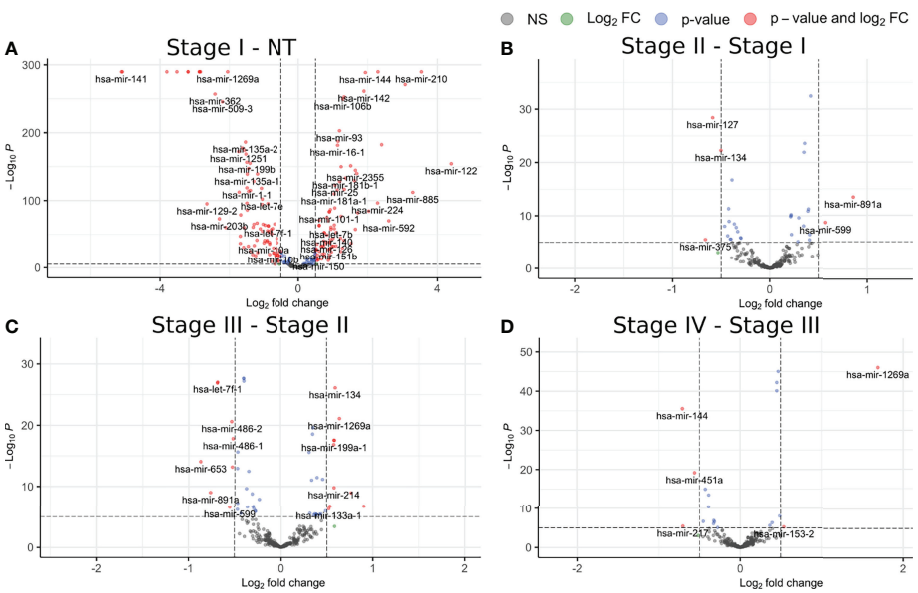


FIGURE 3 | Differentially expressed miRs for each contiguous stage of ccRC. **(A)** Contrast between control and stage I; **(B)** stage I V, stage II; **(C)** stage II V, stage III; **(D)** stage III V, stage IV. The color code is the same as that in **Figure 2**.

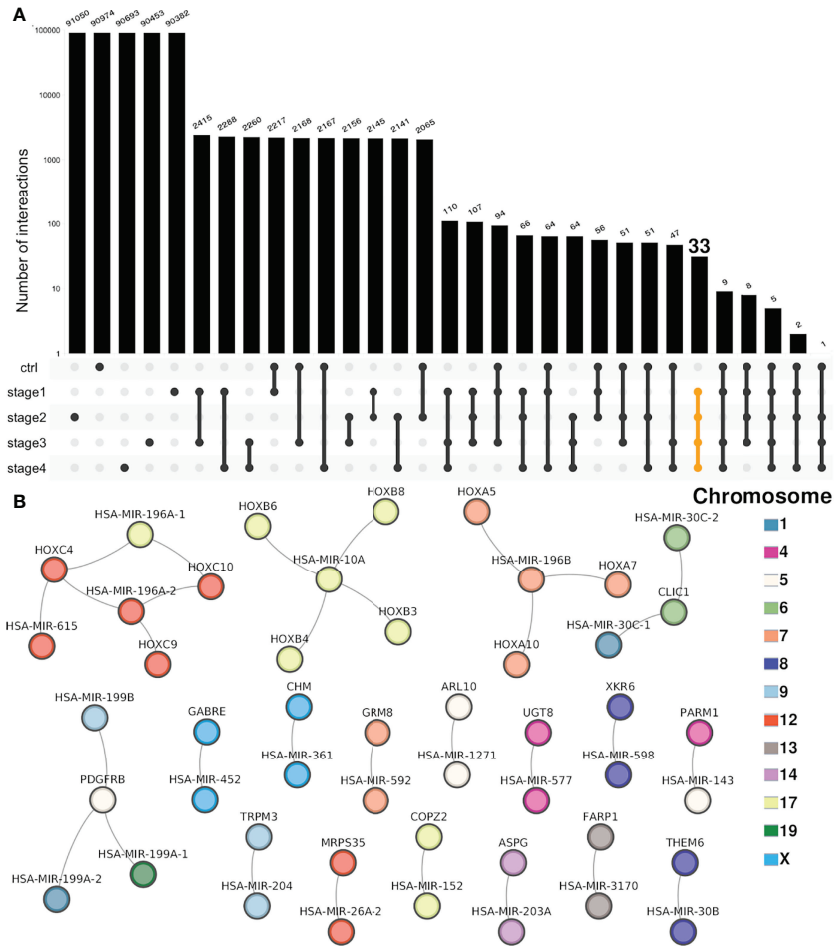


FIGURE 4 | Intersection of miR–gene co-expression networks. **(A)** Each bar in the UpSet plot represents the number of interactions in the selected set, represented by linked points below the bars (log scale). Above each bar, the number of interactions is shown. The first five bars account for unique interactions. From the sixth bar onward, each one of them shows the number of shared interactions between two or more networks. At the right side, the set of shared interactions between the four CCRC progression stages (but not NT) is highlighted in yellow. **(B)** The 33 shared interactions between the four progression stages but not shared with the non-tumor network are depicted. In the figure, the color of the nodes represents the chromosome where miRs and genes are located.

to the same chromosome. Furthermore, they belong to the same cytoband (**Supplementary Material S6**).

Interestingly, among the most connected miRs corresponding to the miR-196 family are the following: miR-10A and miR-196A-1 (Chr17q21.32), miR-196A-2 (Chr12q13.13), and miR-196B (Chr7p15.2). As can be observed in **Figure 4B**, those miRs (upper part of the network) are associated with HOX genes which belong to the same location than the said miRs: HOXC9, HOXC10, and HOXC11 are located at Chr12q13.13. Analog is the case of HOXA5, HOXA7, and HOXA10 (Chr7p15.2) or HOXB3, HOXB4, HOXB6, and HOXB8 (17q21.32).

The role of HOX genes in the rise and development of several types of cancer has been extensively reported (42–44). Additionally, the role of the miR-196 family has been also described in different carcinomas (45–47). The fact that the miR-196-HOX genes complexes are shared between all progression stages but absent in the non-tumor network, may indicate the role of these relationships in ccRC progression.

It is worth to noticing, genes shared by non-tumor (control) and stages were: HOXA9, MEST, TENM4, ARPP21, DIO3. As it was mentioned, HOAX genes have an important role in cancer. The loss of imprinting of MEST gene has been linked to certain types of cancer and may be due to promotor switching. However, all those genes play a critical role in mammalian development as a common feature.

Finally, the neighboring location of miRs and genes observed in **Figure 4B** has been previously described in gene–gene co-expression networks for breast cancer (28, 30, 48), lung cancer (31), and also in clear cell renal carcinoma progression stages (32). In this case, where the inferred networks are obtained by correlating the micro-RNA expression with the gene expression, the effect of loss of long-distance co-expression is not appreciated

in the whole networks. However, in the 33 (out of 100,000 for each stage) cancer-shared interactions, we observe not only miR–gene interactions with molecules from the same chromosome but also the same cytoband and, furthermore, contiguous locations in terms of start positions (**Supplementary Material S6**).

After observing the location of miRs and genes in the shared network, we argue that the appearance of intra-cytoband interactions in cancer-exclusive phenotypes could be related to an anomalous transcriptional event which allows to have similar expression patterns between microRNAs and gene transcripts. Experimental corroboration, however, is needed to fully elucidate the role of those interactions. Additionally, the HOX–miR-196 complex should also be investigated in order to provide a possible explanation of those development-related genes in ccRC progression.

miR–Gene Networks Are Different Between Stages, Both in Size and Composition

As previously stated, we inferred five networks (one for each phenotype), one network for NT, and one network for each ccRC progression stage. To construct all networks, we calculated the mutual information measure between miRs and genes by using the expression matrices for miRs and genes in all stages (see **Table 1**).

We conserved the top 100,000 miR–gene interactions for each one of the five networks (**Supplementary Material S3**). From the top 100,000 edge networks, we conserved only those miR–gene interactions between DEGs and DEMs with opposite differential expression trend (overexpressed miR, underexpressed gene, and *vice versa*). The resulting networks are depicted in **Figure 5**. The

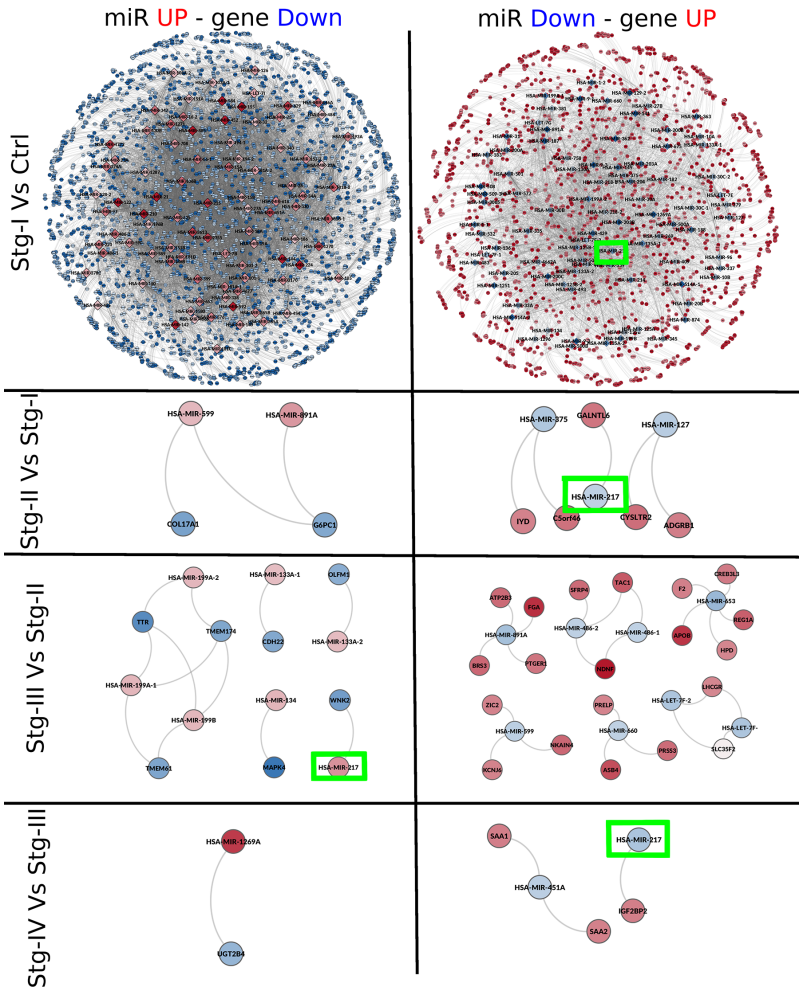


FIGURE 5 | miR–gene networks for each progression stage. In this figure, we can observe networks inferred by mutual information between the expression of miRs and genes for each progression stage. Networks are placed from top to bottom according to the progression stage. The contrast used to depict each network is placed at the left. Red nodes represent overexpressed miRs or genes; meanwhile, underexpressed molecules are depicted in blue. At the left side, networks constructed with overexpressed miRs and underexpressed genes can be found. The right part of the figures contains networks with underexpressed miRs and overexpressed genes. Green squares mark the location of miR-217, the only micro-RNA present in the four networks.

difference in size between the networks of st_I and the rest of networks is evident.

miR-217 Is Differentially Expressed in All Sequentially Contiguous Contrasts Yet Shows Different Target Genes for Each Stage

From **Figure 5**, it can be appreciated that, in each network, miR-217 appears as one of the DEMs and also has a target gene in all cases. In the contrast between stage I and NT, miR-217 is underexpressed (Log_2FC = −1.32). In this stage, this microRNA potentially regulates up to 60 target genes (**Supplementary Material S4**). Among the target genes of miR-217, we can find BIRC7 (Log_2FC = *5.988*), LAMA4 (Log_2FC = *4.024*), or E2F2 (Log_2FC = *2.266*) genes (**Table 5**).

For the contrast between stage I and NT, BIRC7 was the most overexpressed gene. BIRC7 has been reported to be crucial in the development of thyroid cancer by inhibiting apoptosis (49) in several cancer types, such as thyroid (50), leukemia (51), or neuroblastoma (52). In particular, for renal cell carcinoma, the overexpression of BIRC7 has been associated to PTEN-related malignancy and poorer prognosis (53) and metastatic behavior (54). **Supplementary Material S5** is a Cytoscape session file (a.cys network file) containing all top 100,000 networks as well as those for the differentially expressed miR–genes.

LAMA4 is also strongly overexpressed in stage I compared with NT. Its overexpression has been related305 to metastasis in pancreatic Cancer (55). Additionally, it has been observed that miR-200b306 down-regulated LAMA4 and decreases metastasis of renal cell carcinoma (56).

Regarding the stage II network, GALNTL6 (polypeptide N-acetylgalactosaminyltransferase like 6) is the only target of miR-217 present. This gene is related to the metabolism of proteins and O-linked glycosylation (57). The Gene Ontology (GO) annotations related to this gene include carbohydrate binding and polypeptide N-acetylgalactosaminyltransferase activity. GALNTs typically initiate O-glycosylation in the Golgi apparatus, but in cell culture models these enzymes can translocate to the ER *via* a process that involves aberrant Src signaling, leading to an increased density of O-glycosylation of MUC1 repeats (58). GALNTL6 has been reported to be amplified in papillary thyroid carcinomas (59).

For the stage III network, the only target of miR-217 is WNK2 (WNK lysine-deficient protein kinase 2). Diseases associated with WNK2 include hypomagnesemia 4 and renal and angiomatous meningioma. Pathways related to WNK2 are the transport of glucose and other sugars, bile salts and organic acids, and metal ions and amine compounds and ion channel transport. The GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups, and protein tyrosine kinase activity.

We should notice that, in stage III network, WNK2 is underexpressed, and miR-217 is upregulated. WNK2 is considered as a tumor suppressor gene because it inhibits cell proliferation (60), negatively regulating epidermal growth factor receptor signaling *via* the inhibition of MEK1 (61).

Taking these issues into account, the fact that miR-217 resulted overexpressed and its only target in stage III network was WNK2 supports the hypothesis that WNK2 may be a stage-III-specific tumor suppressor gene downregulated by miR-217.

Finally, we found IGF2BP2 as the unique target of miR-217 in the stage IV network. IGF2BP2 is an IGF2 (insulin growth factor 2) post-transcriptional regulator. Other targets of this gene are MYC and PTEN, two crucial participants in pathways associated with tumorigenesis (62). IGF2BP2 was considered as a metabolism regulator. It modulates cellular metabolism in diabetes, obesity, or fatty liver diseases by means of post-transcriptional gene regulation (63). Recently, it has been demonstrated that the overexpression of this gene is a prognostic factor in several types of cancer, such as leukemia (64), breast (65), lung (66), colorectal (67), or hepatocellular carcinoma (68) (**Figure 6**).

In the stage IV network, miR-217 is underexpressed, and its only target is IGF2BP2, which is overexpressed (Log_2FC = 1.0798). The overexpression of this gene may be due to the underexpression of miR-217 in this stage of ccRC.

It is worth noticing that the differential expression of all aforementioned genes occurs between sequentially contiguous stages, *i.e.*, the contrast between those genes is made by the previous phase of ccRC. These results are remarkable since the "control" dataset is an earlier stage of ccRC; that control gene expression dataset is already altered by cancer. Hence, DEGs and DEMs are "more differentiated" than in the control network, which is the traditionally selected contrast.

As shown in **Figures 2**, **3**, the number of statistically significant interactions in the $NT - st_I$ network is much larger than in any other contrast. This implies that the largest alterations occurring between these contiguous progression stages are given by the high number of differentially expressed genes and miRs, allowing the deregulation of several biological processes, which are, in turn, associated with radical changes of the whole phenotype.

On the other hand, the low number of interactions in the subsequent contrasts may also imply that the miR–gene deregulation observed in the advanced stages is a complementary process, which is concomitant to several other phenomena that drive the clear cell renal carcinoma progression.

We want to highlight an apparently counter-intuitive result; this is related to the number of shared differentially expressed

TABLE 5 | Expression statistics for miR-217 and its target genes.

	$NT–st_I$	$st_I–st_{II}$	$st_{II}–st_{III}$	$st_{III}–st_{IV}$
miR-217 Log_2FC	-1.322	-0.5326	0.9033	-0.7065
Number of targets	60	GALNTL6	WNK2	IGF2BP2
Targets Log_2F	1.899 (average)	1.1678	-1.2373	1.0798

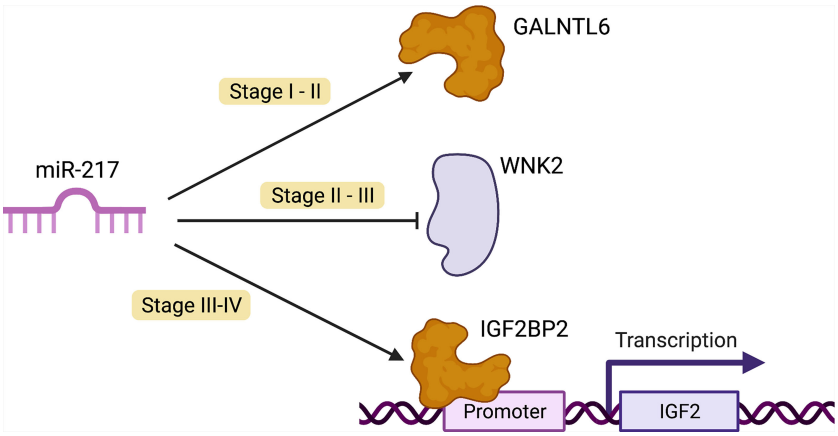


FIGURE 6 | Possible oncogenic role of miR-217. In cancer transition one (stage I–II), miR-217 allows GALNTL6 overexpression. This protein typically initiates post-translational modifications in the Golgi apparatus. Additionally, in cell culture models, these enzymes affect the ER *via* aberrant Src signaling. In stages II–III (transition two), miR-217 represses WNK2 expression, a tumor suppressor which inhibits cell proliferation by negatively modulating the activation of the MEK1 pathway. In the last transition (stages III–IV), miR-217 enables IGF2BP2 overexpression. This gene promotes tumor progression in several types of cancer, such as glioblastoma multiforme and gallbladder cancer. IGF2BP2 also promotes tumor cell proliferation through the PI3K-Akt pathway (69).

genes/miRs in each tumor stage with respect to the shared interactions: the same set of 16,227 genes and 275 miRs was used to construct each network. However, as shown in **Figure 4A**, the number of shared interactions is very low compared with the unique interactions per network (more than 90,000 out of 100,000 for any given phenotype).

This effect of uniqueness in the network interactions most likely obeys the specificities of regulation by micro-RNAs in each context. Despite the fact that the five networks contain the same genes and miRs, the way in which miRs and genes co-express is exclusive. The progression of CCRC apparently modifies the micro-RNA-mediated genetic regulatory processes.

Notwithstanding, the network composed of the shared interactions between the four CCRC progression stages is also informative. From that network, we can observe that almost all interactions occur between genes and miRs from the same chromosome.

The bias to intra-chromosomal interactions has been previously reported by our group in gene co-expression networks for breast cancer (28, 48, 70–72), lung cancer (31), and also CCRC (32). These results show a clear trend to favor close gene correlations in terms of base pair distance. However, for miR–gene co-expression networks in breast cancer (14, 73, 74), we did not observe a trend to present more intra-chromosomal miR–gene interactions over inter-chromosomal ones. To our knowledge, this is the first time that a bias into the intra-chromosomal miR–gene interactions, in the context of breast cancer, was observed.

The finding of those 31 intra-chromosomal miR–gene interactions may be related to the same mechanism behind the bias favoring local correlations over the long-distance ones.

However, the mechanism for which this phenomenon emerges in cancer, but not in control, networks remains elusive. We have investigated the role of other biomolecular processes such as those in transcription factor binding sites, CTCF binding sites (30), or copy number alterations (75). It is worth noticing that none of them has shown to be significantly related to the loss of inter-chromosomal interactions.

Regarding the differences between progression stage networks, the low number of regulated genes by miRs is intriguing since the reports of genes targeted by miRs in the context of renal carcinoma has grown in the recent years [for a systematic review, see (76)]. The latter could be due to the form in which networks were constructed. These networks were obtained by three different filters: (a) the top 100,000 miR–gene co-expression interactions, (b) those miRs and genes that resulted differentially expressed between contiguous stages, and (c) the co-expression relationships between miRs and genes with opposite sign in their differential expression values.

CONCLUDING REMARKS

In this work, we have constructed a set of networks in order to provide a framework for the evolution of the co-expression landscape of micro-RNAs and genes during the progression of clear cell renal carcinoma. As a summary of findings, we may establish the following:

- With this approach we were able to find differentially expressed genes and miRs for each progression stage. At the same time, we were capable of inferring networks filtered to look up for canonical miRs–gene regulatory interactions.
- The largest difference in terms of number of differentially expressed genes as well as in the number of miR–gene interactions occurring between control and stage I.
- Each network behaves differently in terms of miRs and genes involved. Those networks do not share interactions, and the

large majority of miR–gene edges are indeed unique for each progression stage network.
- miR-217 is differentially expressed in all networks. It is the only micro-RNA that is differentially expressed in each stage with oppositely expressed gene targets.
- miR-217 correlates with a completely different set of genes depending on the progression stage. Furthermore, the differential expression of all those genes is in agreement with their role as oncogenes or tumor suppressor genes.
- The finding that LAMA4, BIRC7, GALNTL6, WNK2, and IGFBP2 are potential targets of miR-217 at different times of tumor evolution may help to develop stage-specific strategies, taking into account the differential expression of miR-217 in each stage of clear cell renal carcinoma progression.
- To our knowledge, this is the first time that the evolution of the expression patterns of a micro-RNA is tracked down during all steps of carcinoma progression and, at the same time, its ability to regulate different targets according to the tumor evolution is analyzed.

Possible extensions to the work presented here may include the analysis of other -omic sources, such as the methylation profile, the role of long non-coding RNAs, or the copy number alteration profile. The idea of integrating several sources to provide a more realistic model of the transcriptomic regulation in cancer is important for further steps towards an integrative understanding of gene regulatory programs in cancer.

Additional extensions could be related to the classification of samples based on other clinical features and not only in the progression stage, such as age, gender, or survival status.

A number of the hypotheses that this and other studies have generated must be experimentally tested under different conditions in order to fully capture the potential mechanisms and their implications. However, we believe that approaches such as this one could help the biomedical and clinical research in the search for stage-specific micro-RNA-targeted therapies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JZ-F performed the computational analyses, developed and implemented the programming code, performed pre-processing and low-level data analysis, made the figures, and drafted the manuscript. EH-L developed the theoretical approach, supervised the statistical analysis, designed the figures, and drafted and reviewed the manuscript. JE-E conceived and designed the project, supervised the project, made the figures, and drafted and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by CONACYT (267236 PhD student scholarship to JMZ-F) as well as by federal funding from the National Institute of Genomic Medicine (Mexico). JMZ-F is a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México. This work is part of his PhD thesis.

ACKNOWLEDGMENTS

Figure 1 and **Figure 6** were created with https://Biorender.com.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.934711/full#supplementary-material

Supplementary Material S1 | Venn diagrams of genes and micro-RNAs crossing all elements in contrast non-tumor stages. This zip file contains six Venn diagrams for upregulated, downregulated, and all elements according to 2.4.

Supplementary Material S2 | Quality control and normalization results for gene expression in the five phenotypes. This zip file contains PCA plots before and after normalization for genes and micro-RNAs. All plots were performed after applying filters reported in 2.3.

Supplementary Material S3 | Differential expression values for all combinations between four stages of ccRC and non-tumor. This zip file contains 20 differential expression analyses, including $\log_2$ fold change value, *p*-value, adjusted *p*-value, and *B*-statistic for each gene. BioMart annotation in every gene is included for better use.

Supplementary Material S4 | Targets of miR217. The table contains $\log_2$ fold change value and *p*-value. Additionally, each gene was annotated with BioMart data.

Supplementary Material S5 | Networks processed. The compressed file contains a.*cys* file with all networks described in this work. The networks are merged with differential expression data. All networks were classified by stage, contrast, and $\log_2$ fold change.

Supplementary Material S6 | Molecular features of microRNAs and genes of the 33 cancer-shared interactions.

REFERENCES

1. SEER. *Cancer Stat Facts: Kidney and Renal Pelvis Cancer*. Available at: https://seer.cancer.gov/statfacts/html/kidrp.html (Accessed January, 2022).
2. Haas NB, Nathanson KL. Hereditary Kidney Cancer Syndromes. *Adv Chronic Kidney Dis* (2014) 21:81–90. doi: 10.1053/j.ackd.2013.10.001
3. Weidle UH, Nopora A. Clear Cell Renal Carcinoma: MicroRNAs With Efficacy in Preclinical *In Vivo* Models. *Cancer Genomics - Proteomics* (2021) 18:349–68. doi: 10.21873/cgp.20265

4. Jonasch E, Walker CL, Rathmell WK. Clear Cell Renal Cell Carcinoma Ontogeny and Mechanisms of Lethality. *Nat Rev Nephrol* (2020) 17:245–61. doi: 10.1038/s41581-020-00359-2
5. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal Cell Carcinoma. *Nat Rev Dis Primers* (2017) 3:1–19. doi: 10.1038/nrdp.2017.9
6. Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* (2017) 169:736–749.e18. doi: 10.1016/j.cell.2017.04.016
7. Wang T, Lu R, Kapur P, Jaiswal BS, Hannan R, Zhang Z, et al. An Empirical Approach Leveraging Tumorgrafts to Dissect the Tumor Microenvironment in Renal Cell Carcinoma Identifies Missing Link to Prognostic Inflammatory Factors. *Cancer Discov* (2018) 8:1142–55. doi: 10.1158/2159-8290.cd-17-1246
8. Liu XD, Kong W, Peterson CB, McGrail DJ, Hoang A, Zhang X, et al. PBRM1 Loss Defines a Nonimmunogenic Tumor Phenotype Associated With Checkpoint Inhibitor Resistance in Renal Carcinoma. *Nat Commun* (2020) 11. doi: 10.1038/s41467-020-15959-6
9. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most Mammalian Mrnas are Conserved Targets of Micrornas. *Genome Res* (2009) 19:92–105. doi: 10.1101/gr.082701.108
10. Li M, Marin-Muller C, Bharadwaj U, Chow KH, Yao Q, Chen C. Micrornas: Control and Loss of Control in Human Physiology and Disease. *World J Surg* (2009) 33:667–84. doi: 10.1007/s00268-008-9836-x
11. Abd-Aziz N, Kamaruzman NI, Poh CL. Development of MicroRNAs as Potential Therapeutics Against Cancer. *J Oncol* (2020) 2020:1–14. doi: 10.1155/2020/8029721
12. Xu M, Gu M, Zhang K, Zhou J, Wang Z, Da J. miR-203 Inhibition of Renal Cancer Cell Proliferation, Migration and Invasion by Targeting of FGF2. *Diagn Pathol* (2015) 10:24. doi: 10.1186/s13000-015-0255-7
13. Olson P, Lu J, Zhang H, Shai A, Chun MG, Wang Y, et al. MicroRNA Dynamics in the Stages of Tumorigenesis Correlate With Hallmark Capabilities of Cancer. *Genes Dev* (2009) 23:2152–65. doi: 10.1101/gad.1820109
14. Drago-García D, Espinal-Enríquez J, Hernández-Lemus E. Network Analysis of EMT and MET Micro-RNA Regulation in Breast Cancer. *Sci Rep* (2017) 7:1–17. doi: 10.1038/s41598-017-13903-1
15. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front Endochrinology* (2018) 9:402. doi: 10.3389/fendo.2018.00402
16. Hashimoto Y, Akiyama Y, Yuasa Y. Multiple-To-Multiple Relationships Between Micrornas and Target Genes in Gastric Cancer. *PloS One* (2013) 8: e62589. doi: 10.1371/journal.pone.0062589
17. Di Leva G, Garofalo M, Croce CM. Micrornas in Cancer. *Annu Rev Pathology: Mech Dis* (2014) 9:287–314. doi: 10.1146/annurev-pathol-012513-104715
18. Garzon R, Calin GA, Croce CM. Micrornas in Cancer. *Annu Rev Med* (2009) 60:167–79. doi: 10.1146/annurev.med.59.053006.104707
19. Hayes J, Peruzzi PP, Lawler S. Micrornas in Cancer: Biomarkers, Functions and Therapy. *Trends Mol Med* (2014) 20:460–9. doi: 10.1016/j.molmed.2014.06.005
20. Han N, Li H, Wang H. Microrna-203 Inhibits Epithelial-Mesenchymal Transition, Migration, and Invasion of Renal Cell Carcinoma Cells *via* the Inactivation of the Pi3k/Akt Signaling Pathway by Inhibiting Cav1. *Cell Adhesion Migration* (2020) 14:227–41. doi: 10.1080/19336918.2020.1827665
21. Amar D, Safer H, Shamir R. Dissection of Regulatory Networks That are Altered in Disease *via* Differential Co-Expression. *PloS Comput Biol* (2013) 9: e1002955. doi: 10.1371/journal.pcbi.1002955
22. Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enríquez J, Hernández-Lemus E. Network Modularity in Breast Cancer Molecular Subtypes. *Front Physiol* (2017) 8:915. doi: 10.3389/fphys.2017.00915
23. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene Co-Expression Analysis for Functional Classification and Gene–Disease Predictions. *Briefings Bioinf* (2018) 19:575–92. doi: 10.1093/bib/bbw139
24. Fionda V. Networks in Biology. In: S Ranganathan, M Gribskov, K Nakai, C Schönbach, editors. *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press (2019). p. 915 – 921. Available at: https://doi.org/10.1016/B978-0-12-809633-8.20420-2.
25. Tieri P, Farina L, Petti M, Astolfi L, Paci P, Castiglione F. Network Inference and Reconstruction in Bioinformatics. In: *Encyclopedia of Bioinformatics and Computational Biology*. Amsterdam: Elsevier (2019).
26. Hassler MR, Egger G. Epigenomics of Cancer – Emerging New Concepts. *Biochimie* (2012) 94:2219–30. doi: 10.1016/j.biochi.2012.05.007
27. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. Aracne: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinf* (2006) 7:1–15. doi: 10.1186/1471-2105-7-S1-S7
28. Espinal-Enriquez J, Fresno C, Anda-Jáuregui G, Hernández-Lemus E. Rna-Seq Based Genome-Wide Analysis Reveals Loss of Inter-Chromosomal Regulation in Breast Cancer. *Sci Rep* (2017) 7:1–19. doi: 10.1038/s41598-017-01314-1
29. de Anda-Jáuregui G, Fresno C, García-Cortés D, Espinal-Enríquez J, Hernández-Lemus E. Intrachromosomal Regulation Decay in Breast Cancer. *Appl Mathematics Nonlinear Sci* (2019) 4:223–30. doi: 10.2478/AMNS.2019.1.00020
30. García-Cortés D, de Anda-Jáuregui G, Fresno C, Hernandez-Lemus E, Espinal-Enriquez J. Gene Co-Expression is Distance-Dependent in Breast Cancer. *Front Oncol* (2020) 10:1232. doi: 10.3389/fonc.2020.01232.
31. Andonegui-Elguera SD, Zamora-Fuentes JM, Espinal-Enríquez J, Hernández-Lemus E. Loss of Long Distance Co-Expression in Lung Cancer. *Front Genet* (2021) 12:625741. doi: 10.3389/fgene.2021.625741
32. Zamora-Fuentes JM, Hernández-Lemus E, Espinal-Enríquez J. Gene Expression and Co-Expression Networks are Strongly Altered Through Stages in Clear Cell Renal Carcinoma. *Front Genet* (2020) 11:578679. doi: 10.3389/fgene.2020.578679
33. Creighton CJ, Morgan M, Gunaratne PH, Wheeler DA, Gibbs RA, et al. Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma. *Nature* (2013) 499:43. doi: 10.1038/nature12222
34. Network CGAR. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* (2016) 374:135–45. doi: 10.1056/NEJMoa1505917
35. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep* (2018) 23:313–26. doi: 10.1016/j.celrep.2018.03.075
36. Risso D, Schwartz K, Sherlock G, Dudoit S. Gc-Content Normalization for Rna-Seq Data. *BMC Bioinf* (2011) 12:1–17. doi: 10.1186/1471-2105-12-480
37. Nueda Mj, Ferrer A, Conesa A. Arsyn: A Method for the Identification and Removal of Systematic Noise in Multifactorial Time Course Microarray Experiments. *Biostatistics* (2012) 13:553–66. doi: 10.1093/biostatistics/kxr042
38. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With Deseq2. *Genome Biol* (2014) 15:1–21. doi: 10.1186/s13059-014-0550-8
39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* (2003) 13:2498–504. doi: 10.1101/gr.1239303
40. Kong W, He L, Richards EJ, Challa S, Xu CX, Permuth-Wey J, et al. Upregulation of miRNA-155 Promotes Tumour Angiogenesis by Targeting VHL and is Associated With Poor Prognosis and Triple-Negative Breast Cancer. *Oncogene* (2013) 33:679–89. doi: 10.1038/onc.2012.636
41. Neal CS, Michael MZ, Rawlings LH, der Hoek MBV, Gleadle JM. The VHL-Dependent Regulation of microRNAs in Renal Cancer. *BMC Med* (2010) 8:1–17. doi: 10.1186/1741-7015-8-64
42. Shah N, Sukumar S. The Hox Genes and Their Roles in Oncogenesis. *Nat Rev Cancer* (2010) 10:361–71. doi: 10.1038/nrc2826
43. Bhatlekar S, Fields JZ, Boman BM. Hox Genes and Their Role in the Development of Human Cancers. *J Mol Med* (2014) 92:811–23. doi: 10.1007/s00109-014-1181-y
44. Li B, Huang Q, Wei GH. The Role of Hox Transcription Factors in Cancer Predisposition and Progression. *Cancers* (2019) 11:528. doi: 10.3390/cancers11040528
45. Meyer SE, Muench DE, Rogers AM, Newkold TJ, Orr E, O'Brien E, et al. Mir-196b Target Screen Reveals Mechanisms Maintaining Leukemia Stemness With Therapeutic Potential. *J Exp Med* (2018) 215:2115–36. doi: 10.1084/jem.20171312

46. Rawat VP, Götze M, Rasalkar A, Vegi NM, Ihme S, Thoene S, et al. The Microrna Mir-196b Acts as a Tumor Suppressor in Cdx2-Driven Acute Myeloid Leukemia. *Haematologica* (2020) 105:e285. doi: 10.3324/haematol.2019.223297
47. Xu F, Zhu F, Wang W, Gao W, Chen X, Yu C. Down-Regulation of Mirna-196b Expression Inhibits the Proliferation, Migration and Invasiveness of Hepg2 Cells While Promoting Their Apoptosis *via* the Pi3k/Akt Signaling Pathway. *Cell Mol Biol* (2020) 66:159–64. doi: 10.14715/cmb/2020.66.3.25
48. García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Luminal a Breast Cancer Co-Expression Network: Structural and Functional Alterations. *Front Genet* (2021) 12. doi: 10.3389/fgene.2021.629475
49. Rigato DB, Branco PC, Mateus Reis Silva CS, Machado-Neto JA, Costa-Lotufo LV, Jimenez PC. Birc7 (Baculoviral Iap Repeat Containing 7). *Atlas Genet Cytogenetics Oncol Haematology*.
50. Liu K, Yu Q, Li H, Xie C, Wu Y, Ma D, et al. Birc7 Promotes Epithelial-Mesenchymal Transition and Metastasis in Papillary Thyroid Carcinoma Through Restraining Autophagy. *Am J Cancer Res* (2020) 10:78.
51. Ibrahim L, Aladle D, Mansour A, Hammad A, Al Wakeel AA, Abd El-Hameed SA. Expression and Prognostic Significance of Livin/Birc7 in Childhood Acute Lymphoblastic Leukemia. *Med Oncol* (2014) 31:1–8. doi: 10.1007/s12032-014-0941-4
52. Dasgupta A, Alvarado CS, Xu Z, Findley HW. Expression and Functional Role of Inhibitor-of-Apoptosis Protein Livin (Birc7) in Neuroblastoma. *Biochem Biophys Res Commun* (2010) 400:53–9. doi: 10.1016/j.bbrc.2010.08.001
53. Cheng T, Zhang JG, Cheng YH, Gao ZW, Ren XQ. Relationship Between Pten and Livin Expression and Malignancy of Renal Cell Carcinomas. *Asian Pacific J Cancer Prev* (2012) 13:2681–5. doi: 10.7314/APJCP.2012.13.6.2681
54. Wagener N, Crnković-Mertens I, Vetter C, Macher-Göppinger S, Bedke J, Gröne E, et al. Expression of Inhibitor of Apoptosis Protein Livin in Renal Cell Carcinoma and non-Tumorous Adult Kidney. *Br J Cancer* (2007) 97:1271–6. doi: 10.1038/sj.bjc.6604028
55. Zheng B, Qu J, Ohuchida K, Feng H, Chong SJF, Yan Z, et al. Lama4 Upregulation is Associated With High Liver Metastasis Potential and Poor Survival Outcome of Pancreatic Cancer. *Theranostics* (2020) 10:10274. doi: 10.7150/thno.47001
56. Li Y, Guan B, Liu J, Zhang Z, He S, Zhan Y, et al. Microrna-200b is Downregulated and Suppresses Metastasis by Targeting Lama4 in Renal Cell Carcinoma. *EBioMedicine* (2019) 44:439–51. doi: 10.1016/j.ebiom.2019.05.041
57. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. Uniprot: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res* (2020) 49:D480–ndash;9. doi: 10.1093/nar/gkaa1100
58. Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in Health and Disease. *Nat Rev Nephrol* (2019) 15:346–66. doi: 10.1038/s41581-019-0129-4
59. Passon N, Bregant E, Sponziello M, Dima M, Rosignolo F, Durante C, et al. Somatic Amplifications and Deletions in Genome of Papillary Thyroid Carcinomas. *Endocrine* (2015) 50:453–64. doi: 10.1007/s12020-015-0592-z
60. Moniz S, Verissimo F, Matos P, Brazao R, Silva E, Kotevelets L, et al. Protein Kinase Wnk2 Inhibits Cell Proliferation by Negatively Modulating the Activation of Mek1/Erk1/2. *Oncogene* (2007) 26:6071–81. doi: 10.1038/sj.onc.1210706
61. Jun P, Hong C, Lal A, Wong JM, McDermott MW, Bollen AW, et al. Epigenetic Silencing of the Kinase Tumor Suppressor Wnk2 is Tumor-Type and Tumor-Grade Specific. *Neuro-oncology* (2009) 11:414–22. doi: 10.1215/15228517-2008-096
62. Bell JL, Wächter K, Mühleck B, Pazaitis N, Köhn M, Lederer M, et al. Insulin-Like Growth Factor 2 mRNA-Binding Proteins (IGF2bps): Post-Transcriptional Drivers of Cancer Progression? *Cell Mol Life Sci* (2012) 70:2657–75. doi: 10.1007/s00018-012-1186-z
63. Wang J, Chen L, Qiang P. The Role of Igf2bp2, an M6a Reader Gene, in Human Metabolic Diseases and Cancers. *Cancer Cell Int* (2021) 21:1–11. doi: 10.1186/s12935-021-01799-x
64. He X, Li W, Liang X, Zhu X, Zhang L, Huang Y, et al. Igf2bp2 Overexpression Indicates Poor Survival in Patients With Acute Myelocytic Leukemia. *Cell Physiol Biochem* (2018) 51:1945–56. doi: 10.1159/000495719
65. Li X, Li Y, Lu H. Mir-1193 Suppresses Proliferation and Invasion of Human Breast Cancer Cells Through Directly Targeting Igf2bp2. *Oncol Res Featuring Preclinical Clin Cancer Ther* (2017) 25:579–85. doi: 10.3727/97818823455816X14760504645779
66. Huang Rs, Zheng Yl, Li C, Ding C, Xu C, Zhao J. Microrna-485-5p Suppresses Growth and Metastasis in non-Small Cell Lung Cancer Cells by Targeting Igf2bp2. *Life Sci* (2018) 199:104–11. doi: 10.1016/j.lfs.2018.03.005
67. Ye S, Song W, Xu X, Zhao X, Yang L. Igf2bp2 Promotes Colorectal Cancer Cell Proliferation and Survival Through Interfering With Raf-1 Degradation by Mir-195. *FEBS Lett* (2016) 590:1641–50. doi: 10.1002/1873-3468.12205
68. Wei H, Pu J, Wang J, Chuan W, Xu Z, Wu X, et al. Igf2bp2 Promotes Liver Cancer Growth Through an M6a-Fen1-Dependent Mechanism. *Front Oncol* (2020) 10:2377. doi: 10.3389/fonc.2020.578816
69. Xu X, Yu Y, Zong K, Lv P, Gu Y. Up-Regulation of Igf2bp2 by Multiple Mechanisms in Pancreatic Cancer Promotes Cancer Proliferation by Activating the Pi3k/Akt Signaling Pathway. *J Exp Clin Cancer Res* (2019) 38:497. doi: 10.1186/s13046-019-1470-y
70. Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Multilayer Approach Reveals Organizational Principles Disrupted in Breast Cancer Co-Expression Networks. *Appl Network Sci* (2020) 5:1–23. doi: 10.1007/s41109-020-00291-1
71. González-Espinoza A, Zamora J, Hernandez-Lemus E, Espinal-Enríquez J. Gene Co-Expression in Breast Cancer: A Matter of Distance. *Front Oncol* (2021) 1:4743. doi: 10.3389/fonc.2021.726493
72. Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. K-Core Genes Underpin Structural Features of Breast Cancer. *Sci Rep* (2021) 11:1–17. doi: 10.1038/s41598-021-95313-y
73. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, Highly Connected Micrornas Control Functionality in Breast Cancer Networks. *Int J Genomics* (2018) 2018:1–11. doi: 10.1155/2018/9585383
74. de Anda-Jáuregui G, Espinal-Enríquez J, Hernández-Lemus E. Highly Connected, non-Redundant Microrna Functional Control in Breast Cancer Molecular Subtypes. *Interface Focus* (2021) 11:20200073. doi: 10.1098/rsfs.2020.0073
75. Hernández-Gómez C, Hernández-Lemus E, Espinal-Enríquez J. The Role of Copy Number Variants in Gene Co-Expression Patterns for Luminal B Breast Tumors. *Front Genet* (2022) 13:806607–7. doi: 10.3389/fgene.2022.806607
76. Li M, Wang Y, Song Y, Bu R, Yin B, Fei X, et al. Micrornas in Renal Cell Carcinoma: A Systematic Review of Clinical Implications. *Oncol Rep* (2015) 33:1571–8. doi: 10.3892/or.2015.3799

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zamora-Fuentes, Hernández-Lemus and Espinal-Enríquez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fimmu.2022.895465

Check for updates

OPEN ACCESS

EDITED BY
George Bebis,
University of Nevada, Reno,
United States

REVIEWED BY
Xiaolong Cui,
The University of Chicago,
United States
Runbi Ji,
Jiangsu University Affiliated People's Hospital, China

*Correspondence:
Ren-Xiong Wei
wls0821@126.com

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 13 March 2022
ACCEPTED 01 July 2022
PUBLISHED 28 July 2022

CITATION
Huang Z-D, Fu Y-C, Liu S-Y, Mao Y-J, Zhang Y, Hu C and Wei R-X (2022) m6A-related metabolism molecular classification with distinct prognosis and immunotherapy response in soft tissue sarcoma.
Front. Immunol. 13:895465.
doi: 10.3389/fimmu.2022.895465

COPYRIGHT
© 2022 Huang, Fu, Liu, Mao, Zhang, Hu and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

m6A-related metabolism molecular classification with distinct prognosis and immunotherapy response in soft tissue sarcoma

Zhen-Dong Huang[1,2], Yong-Cheng Fu[3], Shu-Yan Liu[3], Ya-Juan Mao[3], Yan Zhang[3], Chao Hu[1] and Ren-Xiong Wei[1]*

[1]Department of Spine and Orthopedic Oncology, Zhongnan Hospital of Wuhan University, Wuhan, China, [2]School of Stomatology, Southern Medical University, Guangzhou, China, [3]The Third Clinical School, Hubei University of Medicine, Shiyan, China

N6-methyladenosine (m6A) methylation, one of the most crucial RNA modifications, has been proven to play a key role that affect prognosis of soft tissue sarcoma (STS). However, m6A methylation potential role in STS metabolic processes remains unknown. We comprehensively estimated the m6A metabolic molecular subtypes and corresponding survival, immunity, genomic and stemness characteristics based on 568 STS samples and m6A related metabolic pathways. Then, to quantify the m6A metabolic subtypes, machine learning algorithms were used to develop the m6A-metabolic Scores of individual patients. Finally, two distinct m6A metabolic subtypes (Cluster A and Cluster B) among the STS patients were identified. Compared to Cluster B subtype, the Cluster A subtype was mainly characterized by better survival advantages, activated anti-tumor immune microenvironment, lower gene mutation frequency and higher anti-PD-1 immunotherapy response rates. We also found that the m6A-metabolic Scores could accurately predict the molecular subtype of STS, prognosis, the abundance of immune cell infiltration, tumor metastasis status, sensitivity to chemotherapeutics and immunotherapy response. In general, this study revealed that m6A-regulated tumor metabolism processes played a key role in terms of prognosis of STS, tumor progression, and immune microenvironment. The identification of metabolic molecular subtypes and the construction of m6A-metabolic Score will help to more effectively guide immunotherapy, metabolic therapy and chemotherapy in STS.

KEYWORDS

m6A modification, cancer metabolism, tumor microenvironment, molecular classification, machine learning, immunotherapy

Introduction

Soft tissue sarcoma (STS) constitutes more than 50 subtypes of malignant tumors that originate from the interstitial connective tissues (1). The main treatment options for patients with STS include surgical resection, radiotherapy, and chemotherapy. However, metastasis or recurrence is reported in approximately 50% of STS patients after treatment (2). Furthermore, the median survival time of patients with metastatic STS is only 8-12 months (3). Immunotherapy is a promising treatment for several solid tumors. The survival benefits of immune checkpoint inhibitors are higher in several cancers compared to chemotherapy and targeted therapy (4). Moreover, immunotherapies such as IL-15 cytokine therapy in combination with TIGIT blockade therapy (5) as well as PD-1 and CTLA-4 inhibitors (6) have made progress in treating patients with STS. However, the complexity and heterogeneity of tumor microenvironment (TME) affects the efficacy of immunotherapies in STS patients. Currently, the immunotherapy response rate in patients with STS is significantly low (7). Therefore, more precise molecular classification is urgently needed to unravel the TME heterogeneity of STS to select patients more suitable for immunotherapy and improve the response rate of immunotherapy.

N6-methyladenosine (m6A) is the most abundant and common epigenetic modification in the eukaryotic mRNAs. The protein machinery involved in m6A recognition, addition, or removal, including m6A methyltransferases (writers), m6A demethylases (erasers), and m6A readers, has been well characterized. Several reports have shown that alterations in m6A mRNA methylation disrupt gene expression and the downstream cellular processes, and play a significant role in the initiation and progression of tumors (8). Dysregulation of m6A mRNA methylation in the cancer cells alters the expression of metabolic genes and the activities of the related metabolic pathways, thereby significantly affecting proliferation, differentiation, invasion, and metastasis of cancer cells (9, 10). Metabolic pathways play a crucial role in tumorigenesis, cancer cell survival, and regulation of the tumor immune microenvironment (TIME) (11). Therefore, characterization of the m6A-related metabolic signatures in the STS tissues can unravel the status of the tumor microenvironment (TME) and help the clinicians to strategize immunotherapy options for individual patients. Recent study has shown that molecular typing of STS tumors is more accurate than the traditional classification of STS tumors based on pathological staging and TNM staging (10). Therefore, in this study, we integrated 568 STS samples from the TCGA-SARC and GSE21050 cohorts, and comprehensively evaluated the association between m6A-related metabolic pathways and immune characteristics of STS using multi-omics data. We also analyzed the different molecular subtypes of STS based on m6A-related metabolism. Furthermore, we established a scoring system based on m6A-related metabolism and evaluated its accuracy in predicting the prognosis and immunotherapy response of STS patients.

Materials and methods

STS dataset acquisition and processing

RNA-sequencing data were downloaded for the STS samples in the TCGA-SARC cohort (https://gdc.xenahubs.net) in the form of normalized Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Then, the FPKM values were then transformed into transcripts per kilobase million (TPM) values. The microarray data for the STS samples from the GSE21050 cohort were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The TCGA and GEO datasets (n=568) were merged and the batch effect biases were corrected using the "ComBat" algorithm based on R package "sva". All statistical analyses were conducted using R software 4.1.2 (https://www.r-project.org/).

Pre-processing clinicopathological and genomic data

Clinicopathological data, including overall survival rates of the STS patients in the TCGA-SARC and GSE21050 cohorts, were extracted. The baseline clinicopathological characteristics of the STS samples and immune microenvironment tissue classification (12) are shown in Supplementary Table S1. Genomic data, including somatic mutations and copy number alterations (CNA) of the TCGA-SARC cohort, were obtained from the Firehose project (https://gdac.broadinstitute.org/). The mutational landscape of the STS samples was visualized using the "maftools" R package (13). The arm-level and focal-level genome amplifications and deletions were analyzed using the GISTIC 2.0 package based on GenePattern tool (https://cloud.genepattern.org) (14).

Metabolic pathway acquisition and unsupervised clustering

We acquired 114 gene sets related to tumor metabolism from previously published study (15). The single sample Gene Set Enrichment Analysis (ssGSEA) algorithm was used to estimate the enrichment scores of various metabolic pathways based on gene expression profiles. Each sample was assigned a score corresponding to the status of each metabolic pathways. Next, we identified 21 m6A regulators from previous studies (16), including 11 m6A readers (*ELAVL1, FMR1, HNRNPA2B1, HNRNPC, IGF2BP1, LRPPRC, YTHDC1, YTHDC2, YTHDF1, YTHDF2*, and *YTHDF3*), 8 m6A writers (*CBLL1, KIAA1429, METTL14, METTL3, RBM15, RBM15B, WTAP*, and *ZC3H13*), and 2 m6A erasers (*ALKBH5* and *FTO*). Pearson's correlation analysis was used to identify m6A-related metabolic pathways

using |Pearson's r| >0.4 and p <0.001 as the criteria. Univariate Cox regression analysis was used to identify the m6A-related metabolic pathways closely related to the prognosis of STS patients. The m6A-related metabolic pathways were ranked according to their degree of importance to STS patient prognosis (nrep = 100 iterations in the Monte Carlo simulation; nstep = 5) using the random survival forest (RSF) algorithm. Then, a network to demonstrate the interactions between the m6A-related metabolic pathways was constructed based on Spearman and distance correlation analyses.

The unsupervised clustering analysis was performed using the Consensus Clustering algorithm to identify clusters based on distinct statuses of the m6A-related metabolic pathways associated with prognosis. The STS samples were classified by k-means, with k from 2 to 10 using the "ConsensusClusterPlus" R package (17) and 1000 repetitions were performed to ensure clustering stability. The consensus clustering matrix and cumulative distribution function (CDF) curve analysis was used to determine the optimal number of clusters.

Tumor immune microenvironment analysis

The abundance of immune cell types in the TIME was determined using the "immunedeconv" and "gsva" R packages (18, 19), which estimated immune cell infiltration based on seven different algorithms, namely, ssGSEA, TIMER, xCell, MCP-counter, EPIC, quanTIseq, and IPS. "ESTIMATE" R package (20) was then used to determine the immune scores and tumor purity of different molecular subtypes of STS patients.

Gene set variation analysis and functional enrichment analysis

The pathway enrichment profiles of different clusters were evaluated using the "gsva" R package. The enrichment scores of twelve well-defined tumor-related gene sets or pathways (Supplementary Table S2) were calculated for all samples using the GSVA algorithm. Gene ontology (GO) enrichment analysis was performed with the "clusterProfiler" R package (21) using q-value <0.05 as the cutoff.

Calculation of stemness index

The one-class logistic regression (OCLR) machine learning algorithm (22) was used to calculate the mRNA expression-based stemness index (mRNAsi) of the tumor samples. The mRNAsi value indicated the stem cell-like features of the cancer cells in the samples. We then compared the differences in stemness levels (mRNAsi) between the clusters using Wilcoxon test. P values of less than 0.05 were considered statistically significant.

Evaluation of the m6A-metabolic score

Next, we evaluated the m6A-metabolic Score for STS patients. First, we identified the differentially expressed genes (DEGs) using log2 fold change (FC) >1.5 and adjusted P-value <0.05 as the criteria between the between clusters based on the status of the m6A-related metabolic pathways using "limma" R package (23). Then, the unsupervised clustering algorithm consensus clustering was used to assign STS patients into different gene clusters based on the DEGs. The m6A-related metabolism scoring system was developed using the least absolute shrinkage and selection operator (LASSO) algorithm based on the "glmnet" R package with 10-fold cross validation to estimate the penalty parameters (24). The m6A-metabolic Score was calculated using the following formula: m6A-metabolic Score = $\sum_{i=1}^{n}$Coefficient $*$ Exp, where Exp is the expression value of each selected hub gene from LASSO algorithm. Then, we identified hub genes as independent prognostic factors using multivariate Cox regression analysis. We then developed a clinical nomogram to predict the survival of STS patients by integrating the clinicopathological features and m6A-metabolic scores using the "rms" R package.

The cutoff value was determined based on the correlation between the m6A-metabolic score and prognosis using the "surv-cutpoint" function of the "survminer" R package. The STS patients were classified into groups with high- and low-m6A-metabolic Scores based on the cut-off value.

Immunotherapy response prediction

The expression levels of 14 immune checkpoint-related genes were used to predict the immunotherapy response of each STS patient. Next, the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was used to predict responses to immune checkpoint blockade therapy (25). The tumor immune evasion characteristics of the pre-treatment tumor profiles were derived by integrating the expression signatures of T cell dysfunction and T cell exclusion in the tumor tissues. Furthermore, the "tumor inflammation signature" (TIS) score consisted of 18 key gene signatures that were estimated to quantify immune response activation in the TIME (26). TIS score was then used to predict response to anti-PD-1 immunotherapy. In general, a lower TIDE score or a higher TIS score correlated with better response to immunotherapy. The correlation between m6A-related metabolism score, TIDE score, and TIS score was analyzed and the differences in the

TIDE scores and TIS scores between the high and low m6A-metabolic score groups was compared.

The Subclass Mapping (SubMap) algorithm (27) was used to compare the similarities in the expression profiles of the high- and low-m6A-metabolic Score groups among the STS cohorts and a clinical cohort with 47 melanoma patients treated with immunotherapy (28). The lower the p values, the higher the similarity. Recommended default parameters, 1000 random permutations for Fisher's statistics, were used. The "complexHeatmap" R package (29) was used to visualize the SubMap results.

Chemotherapy response prediction

The chemotherapy response was predicted for each STS sample using the predictive model for the cell line data from the publicly available Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/). A lower half-maximal inhibitory concentration (IC50), estimated by ridge regression, indicates a better sensitivity to a given drug. IC50 of a drug was estimated by ridge regression to determine the degree of sensitivity. The chemotherapeutic response prediction was performed using the "pRRophetic" R package (30).

Results

Eleven prognostic m6A-related metabolic pathways are identified in the STS cohort

The study design and strategy are shown in Figure 1A. To reduce the likelihood of batch effects from non-biological technical biases, the "ComBat" algorithm was used between TCGA-SARC and GSE21050 datasets (Supplementary Figure S1). The ssGSEA algorithm was used to analyze the status of 114 metabolic pathways in the STS samples. The metabolic pathways were assigned scores for every STS sample enrolled in this study. Then, 35 m6A-related metabolic pathways were identified by evaluating 21 m6A regulators (8 writers, 2 erasers, and 11 readers) and the 114 metabolic pathways using Pearson correlation analysis (Supplementary Table S3). Univariate Cox regression analysis showed that 11 m6A-related metabolic pathways were significantly associated with the prognosis of STS patients. These 11 metabolic pathways were used for subsequent analysis. Figure 1B shows the prognostic values of the 11 m6A-related metabolic pathways in the STS patients. A total of 8 m6A regulators (*VIRMA, METTL3, LRPPRC, ELAVL1, YTHDC1, ALKBH5, YTHDF2, METTL14*) associated with 11 metabolic pathways that play a key role in the prognosis of STS were screened, and their co-expression networks were showed in Figure 1C. These 11 m6A-related metabolic pathways were ranked using the RSF algorithm. Pyrimidine metabolism, fructose and mannose metabolism, and linoleic acid metabolism were the top 3 m6A-related metabolic pathways (Figure 1D).

Identification and validation of two distinct molecular subtypes of STS based on m6A-related metabolic pathways

Consensus clustering was performed to identify the molecular subtypes of STS patients based on the prognostic m6A-related metabolic pathways. K=2 was selected as the optimal number of clusters based on consensus matrix and the cumulative distribution function (CDF) plots (Figures 2A, Supplementary Figure S2). STS patients were classified into two subtypes, Cluster A (n=262) and Cluster B (n=306). The graph learning-based dimensionality reduction technique was used to distribute individual STS patients into specific branches based on the expression profiles of the prognostic m6A-related metabolic pathways. This analysis also classified STS patients into two groups that were consistent with the defined molecular subtypes (Figure 2B). Furthermore, clustering analysis was performed separately on the TCGA-SARC and GSE21050 cohorts as the validation datasets. The results again confirmed the stability of the K=2 clusters (Supplementary Figure S3A). Kaplan–Meier survival analysis demonstrated that the overall survival (OS) rates of STS patients in Cluster A was significantly higher than those in Cluster B (Figure 2C). The heatmap in Supplementary Figure S3B shows the differences in the scores of the m6A-related metabolic pathways in the STS patients belonging to clusters A and B. The interaction network between the 11 prognostic m6A-related metabolic pathways and their prognostic value were shown in Figure 2D. These results demonstrated significant differences in the survival rates and the metabolic characteristics of the STS patients belonging to these two molecular subtypes.

STS samples in Cluster A subtype demonstrate higher anti-tumor immunity than those in Cluster B subtype

Next, we investigated the differences in the biological characteristics of the two subtypes to determine the mechanisms underlying the differences in prognosis. The evaluation of 11 cancer-related pathways ($CD8^+$ T effector, Antigen processing machinery, Epithelial-Mesenchymal Transition [EMT], Angiogenesis, Cell cycle, DNA replication, Nucleotide excision repair, DNA damage repair, Homologous recombination, Mismatch repair, Hypoxia) using the GSVA

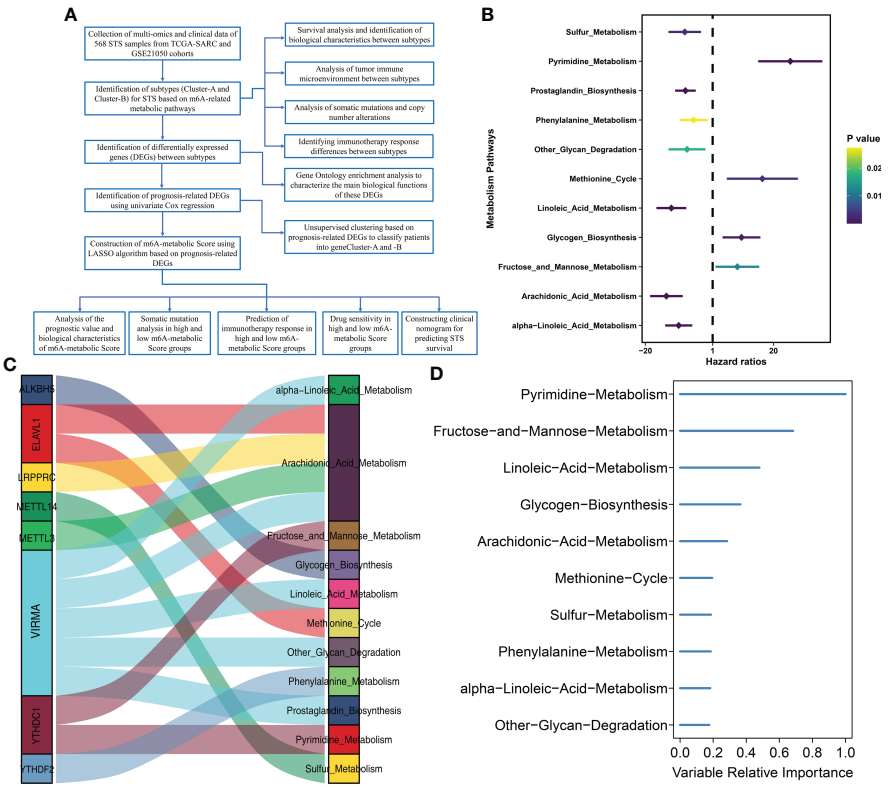


FIGURE 1
Flow diagram and prognostic m6A-related metabolic pathways are identified **(A)** Flow diagram shows this study's systematic analysis process. **(B)** Univariate Cox regression analysis reveals 11 m6A related metabolic pathways significantly correlated with STS prognosis **(C)** The alluvial diagram displays 8 m6A regulators and 11 m6A-related metabolic pathways. **(D)** 11 m6A related metabolic pathways were ranked by Random Survival Forests algorithm.

algorithm showed that pathways related to immune activation were significantly up-regulated in the STS samples from Cluster A, whereas, pathways related to EMT, cell cycle, and DNA replication were significantly up-regulated in the STS samples from Cluster B (Figure 2E). Next, the differences in the TIME and the status of immune cell infiltration were analyzed in the STS tissues from the two clusters using seven algorithms, namely, ssGSEA, MCPcounter, xCell, EPIC, TIMER, quanTIseq, and IPS. STS samples in Cluster A showed significantly higher infiltration of antitumor immune cell types such as dendritic cells, $CD8^+$ T cells, and cytotoxic T cells compared to the STS samples in Cluster B (Supplementary Figure S3C). The immune scores were higher for the STS samples in Cluster A compared to the samples in Cluster B, whereas, tumor purity was higher for the STS samples in Cluster B compared to the STS samples in Cluster A (Figure 2F). Cluster B samples also showed higher degree of tumor metastases (Figure 2G) and stemness levels (mRNAsi) compared to the Cluster A samples (Figure 2H). Furthermore, the m6A regulators were up-regulated in the Cluster B samples compared to the Cluster A samples (Figure 2I). Overall, these results showed that STS samples in Cluster A exhibited higher anti-tumor TIME than the STS samples in Cluster B.

Cluster B subtype shows higher genomic mutations and alterations than Cluster A subtype

Next, we investigated the differences in somatic mutations and CNA between the two molecular subtypes of STS. Among the somatic mutations, missense mutations and single nucleotide polymorphisms (C>T) were the most common mutation types in the STS samples. The five most frequently mutated genes were *TP53, TTN, ATRX, MUC16*, and *MUC4* (Figure 3A). The mutation landscape for the two clusters was displayed using the oncoPrint plot. The mutation frequency was significantly higher for the Cluster B samples (72.52%) compared to the Cluster A samples (59.62%) (Figures 3B, C). Among the commonly mutated genes, the mutation frequency of *TP53* was highest in both groups. STS patients from both clusters exhibited significant differences in the frequency of deletions at the arm level. Overall, the amplification and

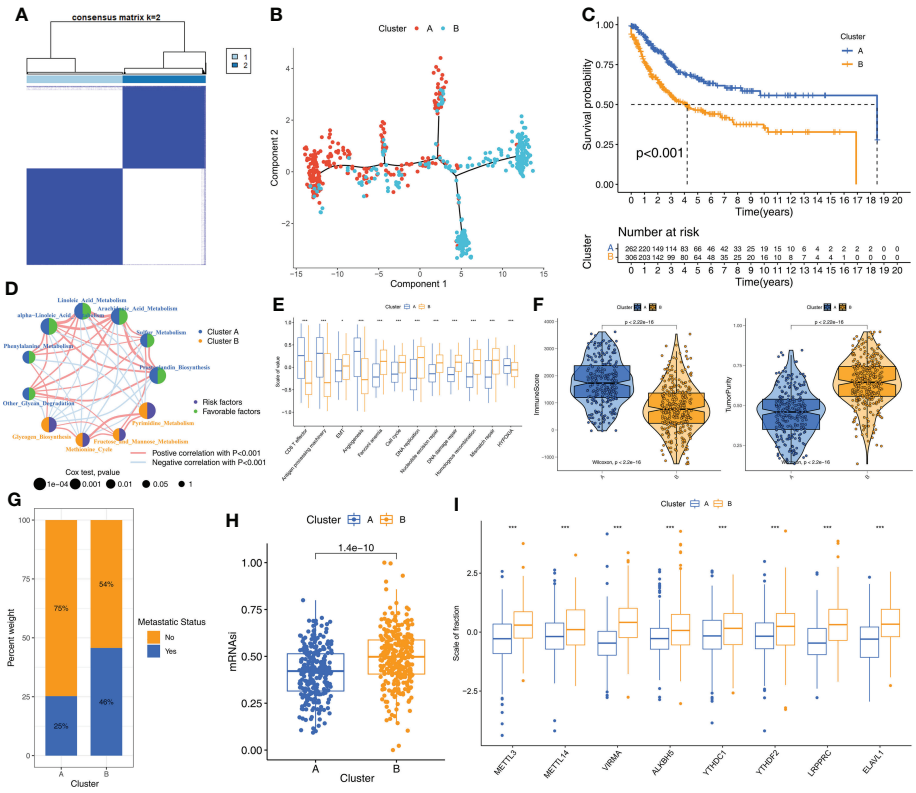


FIGURE 2
The molecular subtypes based on prognostic m6A-related metabolic pathways **(A)** The consensus matrix heatmap corresponding to k=2 obtained by consensus clustering. **(B)** STS patients could be stratified into two subtypes based on m6A-related metabolic pathways; each point represents a patient with colors corresponding to two subtypes. **(C)** Kaplan–Meier curves for overall survival of 568 STS patients from TCGA-SARC and GSE21050 cohorts. Log-rank test revealed P-value <0.001. **(D)** The interactions of 11 m6A-related metabolic pathways in STS. The size of the circles indicated the effect of each m6A-related metabolic pathway on prognosis, respectively. Prognostic risk factors were shown in purple and prognostic favorable factors were shown in green. The lines connecting m6A-related metabolic pathways represented the interactions between m6A-related metabolic pathways, and the thickness of the lines represented the correlation strength estimated by Spearman correlation analysis. Red was a positive correlation, and blue was a negative correlation. **(E)** Differences in immune-related and cancer-related signatures between Cluster A and B. The asterisks represented the statistical p value (*P < 0.05; ***P < 0.001). **(F)** Violin plot shows the difference of ImmuneScore and tumor purity from ESTIMATE algorithms in Cluster A and B. **(G)** The proportion of STS patients with metastatic status in TCGA-SARC and GSE21050 cohorts. **(H)** The box plot shows differences in stemness index (mRNAsi) between Cluster A and B. The statistical difference of clusters was compared through the Wilcoxon test. **(I)** The difference of expression for eight m6A regulators associated with metabolic pathways between Cluster A and B. Wilcoxon test was used to test statistical difference (***P < 0.001).

deletion frequencies were significantly higher in the Cluster B samples at both the arm-level and the focal-level compared to the Cluster A samples (Figure 3D, E).

Cluster A patients show better response to immunotherapy than the Cluster B patients

Since STS samples in clusters A and B show significant differences in TIME, we postulated differential responses to immunotherapy. Therefore, we analyzed the treatment responses of both molecular subtypes to immunotherapy by estimating the gene signatures related to immune checkpoint genes. Cluster A samples showed higher activation levels compared to the Cluster B samples for the immune checkpoint gene set (Figure 4A). This suggested that STS patients in Cluster A group would benefit more from the treatment with immune checkpoint inhibitors compared to the STS patients in Cluster B. Then, we verified these results using the TIS and TIDE scores. TIS score was higher and TIDE score was lower for the STS patients in Cluster A compared to the patients in Cluster B (Figures 4B, C). Finally, we predicted the responses of the two molecular subtypes to PD-1 and CTLA4 immune checkpoint inhibitors by Submap analysis. The results demonstrated that patients in Cluster A were more sensitive to anti-PD-1 therapy compared to the patients in Cluster B (Figure 4D).

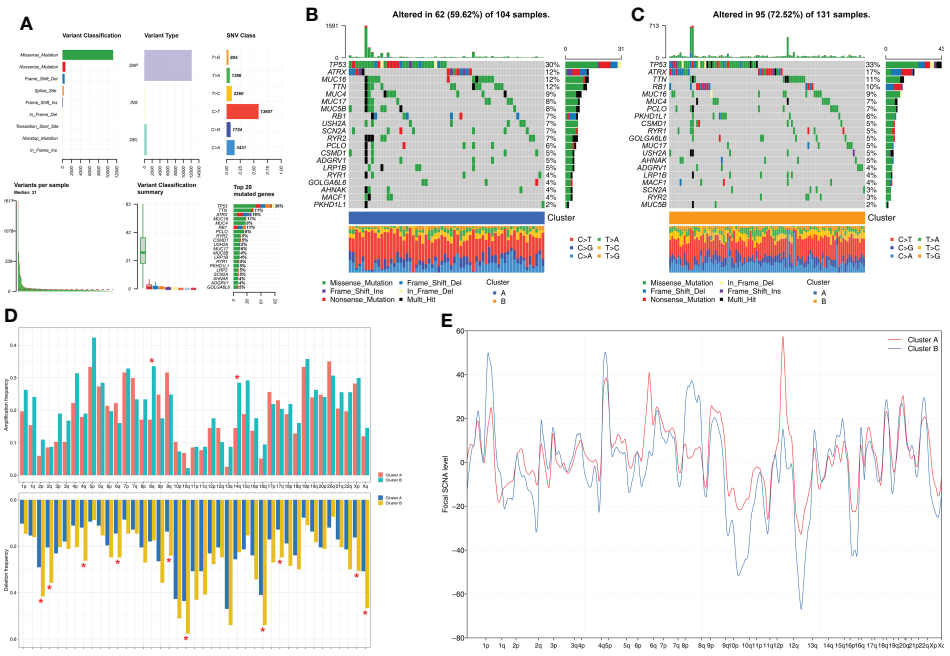


FIGURE 3
Comparison of somatic mutation and CNA analysis between Clusters. **(A)** The summary of the overall distribution of mutation in STS. **(B, C)** The oncoPrint plots show tumor somatic mutation landscape between Cluster A subtype **(B)** and Cluster B subtype **(C)**. **(D, E)** Comparisons of arm-level amplification and deletion frequencies and focal-level amplification and deletion levels between Cluster A and Cluster B *P < 0.05.

Analysis of DEGs shows two geneClusters of STS patients

Next, we analyzed the DEGs between the two molecular subtypes by identifying the key genes and determine the potential biological effects. We identified 150 DEGs between the two molecular subtypes by analyzing gene expression profiles using the "limma" package (Supplementary Table S4). The result of the GO enrichment analysis was observed in the genes related to humoral immune response, complement activation, regulation of immune effect process, and some immune diseases (Figure 5A). Subsequently, unsupervised clustering analysis based on the expression patterns of the DEGs identified two molecular subtypes because K=2 was the optimal setting to distinguish the STS samples accurately (Figures 5B, Supplementary Figure S3D). The 568 STS samples were divided into geneCluster A (n=261) and geneCluster B (n=307). The 11 prognostic m6A-related metabolic pathways were significantly different between the two geneClusters (Figure 5C). Furthermore, 150 DEGs were between geneCluster A and geneCluster B were shown in Figure 5D. Kaplan-Meier survival curve analysis showed that the OS rates were higher for patients in geneCluster A compared to the patients in geneCluster B (Figure 5E). Thus, geneCluster A corresponds to the Cluster A phenotype, and geneCluster B corresponds to the Cluster B phenotype.

The m6A-related metabolic scoring system was established for estimating prognosis of STS patients based on 12 hub genes

We established a scoring system based on m6A-related metabolism (m6A-metabolic Score) to assess the prognosis of all STS patients (Supplementary Table S5). Firstly, we performed univariate Cox regression analysis for the 150 DEGs to identify the prognostic genes. Then, we identified 104 DEGs that were significantly associated with the OS rates of STS patients (Supplementary Table S6). Finally, a machine learning model, LASSO-penalized Cox analysis was used to identify 12 hub genes (*ACTN1, ITGA10, MYLK, CNN1, LYVE1, IGF1, CPVL, C1S, PODN, ALDH1A1, MFAP5*, and *IGHM*) for estimating the m6A-metabolic Score of each STS patients (Figure 6A). Among these hub genes, *ITGA10, MYLK, LYVE1, IGF1, CPVL, C1S, ALDH1A1* and *MFAP5* were identified as independent prognostic factors for the STS patients using multivariate Cox regression analysis (Figure 6B).

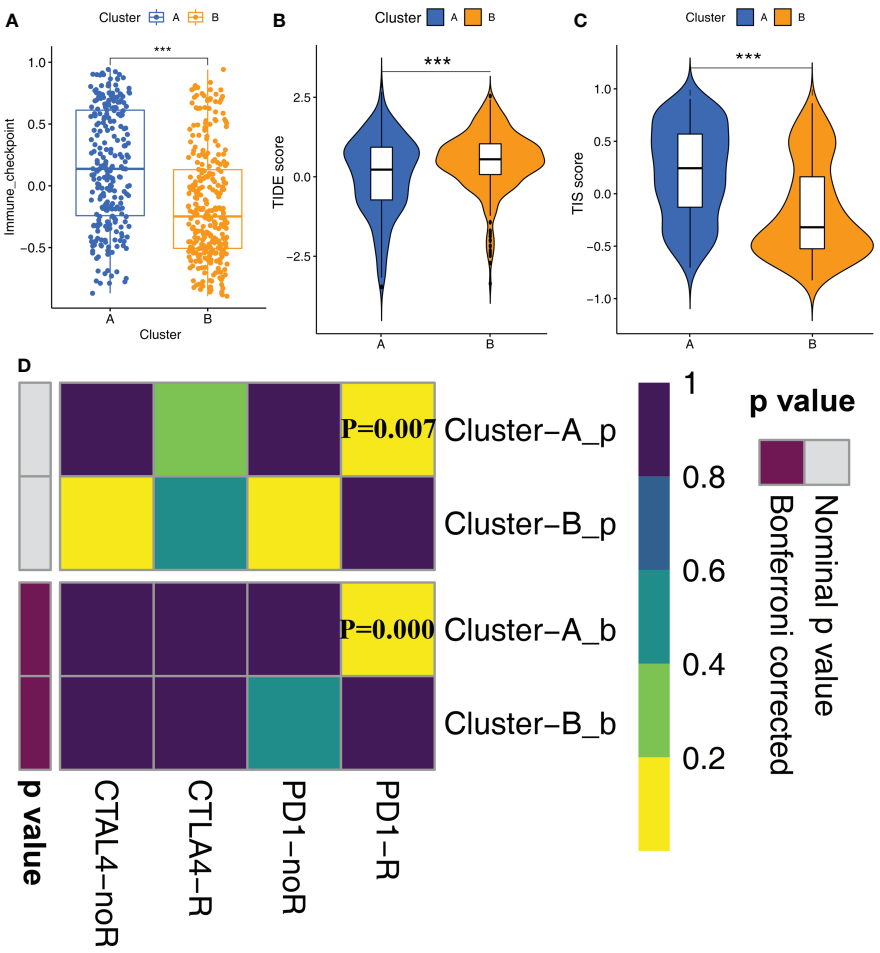


FIGURE 4
Immune checkpoint inhibitor therapy responses between Clusters **(A)** The differences in the scores of immune checkpoint gene set between Cluster A and Cluster B were compared. Statistical differences between the two subtypes were compared by Wilcoxon test (***P < 0.001). **(B, C)** The violin plots demonstrate the difference between TIDE score **(B)** and TIS score **(C)** in the Cluster A and Cluster B groups. Wilcoxon test was used to compare the differences between groups (***P < 0.001). **(D)** Submap analysis shows that Cluster A subtype could be more sensitive to anti-PD-1 treatment (Bonferroni corrected P-value = 0.00).

Low m6A-metabolic scores group show better prognosis than high m6A-metabolic scores

Next, we analyzed the correlation between the m6A-metabolic score and TIME-related biological processes. The m6A-metabolic score showed negative correlation with $CD8^+$ T effectors, antigen processing machinery, immune checkpoints, and other immune-related biological processes (Figure 6C). Then, based on the optimal cutoff m6A-metabolic score of -1.86, the STS samples were divided into two groups with high (n=193) and low (n=375) m6A-metabolic Scores. Kaplan-Meier survival curve analysis showed that the OS rates of STS patients with low m6A-metabolic Scores were significantly higher than those with high m6A-metabolic Scores (Figure 6D). The survival analysis was performed separately with the TCGA-SARC and GSE21050 cohorts to validate the accuracy of the results and the same trend was observed (Supplementary Figures S3E, F). These results demonstrated that the m6A-metabolic Score could be used to predict the survival of STS patients.

The distribution of m6A-metabolic Scores, survival status, survival time, and expression of the 12 hub genes are shown in Figure 6E. We then verified the accuracy in predicting prognosis of STS patients with m6A-metabolic Scores and obtained an AUC value of 0.798 in ROC curve (Figure 6F), thereby demonstrating the accuracy of the m6A-metabolic Score in predicting the survival outcomes of STS patients compared to the other clinicopathological indicators (Age, Histological type, Gender and Race, etc.). We then evaluated the differences in the immune cell proportions in the TME of the two groups using ssGSEA, MCPcounter, xCell, EPIC, TIMER, quanTIseq, and IPS algorithms. The results showed significantly higher infiltration of the anti-tumor immune cells

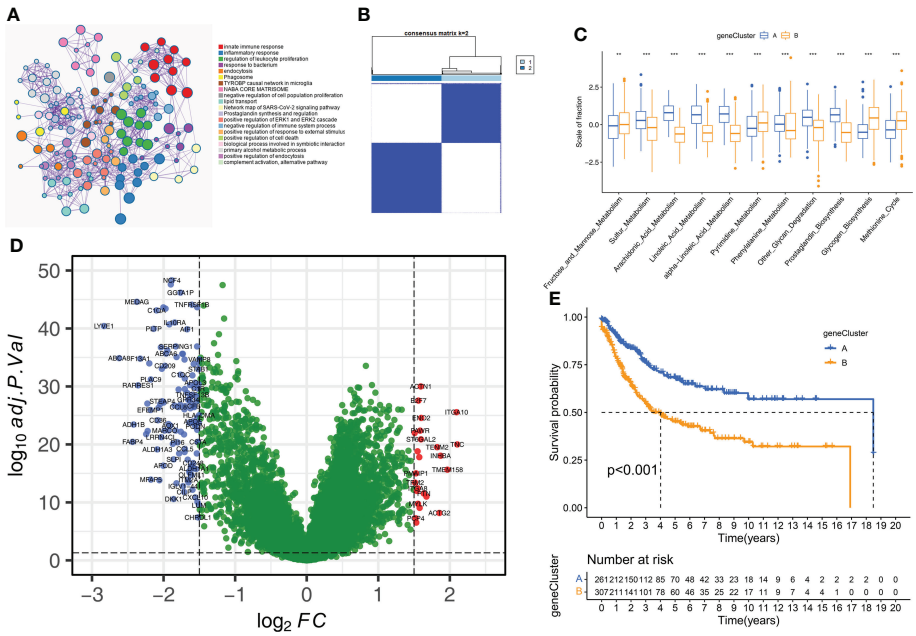


FIGURE 5
Construction of geneClusters. **(A)** Metascape enrichment network displayed the enriched terms for DEGs. Cluster annotations were shown in the color code. **(B)** The consensus clustering matrix for geneCluste (K=2). **(C)** The difference of 11 m6A related metabolic pathways between geneClusters (**P < 0.01; ***P < 0.001). **(D)** The volcano plot shows DEGs of between geneCluster-A and geneCluster-B. **(E)** Kaplan-Meier curves for overall survival of STS patients between geneCluster A and geneCluster B groups.

such as CD8+ T cells, NK cells and dendritic cells in the low m6A-metabolic Score group compared to the high-m6A-metabolic Score group; the infiltration of pro-tumorigenic immune cells such as Th2 cells was higher in the high m6A-metabolic Score group compared to the low m6A-metabolic Score group (Figure 6G). ESTIMATE analysis also showed higher infiltration of the immune cells in the low m6A-metabolic Score group compared to the high m6A-metabolic Score group (Figure 6H).

The Cluster A and geneCluster A groups were associated with low m6A-metabolic Scores and better prognosis compared to the Cluster B and geneCluster B groups (Figures 6I, J and Supplementary Figure S3G). Furthermore, the mutation profiles and landscape were visualized using OncoPrint plot between the high- and low-m6A-metabolic Score groups. The high m6A-metabolic Scores group (Figure 6K) exhibited higher overall gene mutation rates than low m6A-metabolic Scores group (Figure 6L). As shown in Figure 6M, top 13 significant differences were observed in the high- and low-m6A-metabolic Score groups. The number of mutations were significantly higher in the high m6A-metabolic Score group compared to the low m6A-metabolic Score group. Furthermore, STS patients in the high m6A-metabolic Score group showed higher proportion of metastases compared to those in the low m6A-metabolic Score group (Figure 6N). Patients with complete clinical data were used to construct the prognostic nomogram to predict the 1-, 3-, and 5-year survival rates (Figure 6O).

We calculated m6A-metabolic Scores for different immune microenvironment tissue subtypes and found that the Inflammatory subtype had the lowest m6A-metabolic Score (Supplementary Figure S4A). We also found that patients in high ImmuneScore group had lower m6A-metabolic Scores than low ImmuneScore group (Supplementary Figure S4B). Next, according to the median survival time of STS patients, patients greater than the median were classified as longer survival time group, and those less than the median were classified as shorter survival time group. We found that the longer survival time group had lower m6A-metabolic Scores than the shorter survival time group (Supplementary Figure S4C). For patient survival status information, we also found that patients from alive group had lower m6A-metabolic Scores than dead group (Supplementary Figure S4D). Besides, STS patients with metastases had higher m6A-metabolic Scores than primary STS patients (Supplementary Figure S4E).

STS patients with low m6A-metabolic scores are more sensitive to immunotherapy and chemotherapeutic drug sensitivity

The expression levels of 14 immune checkpoint-associated genes including PDCD1 and CD274 were significantly higher in

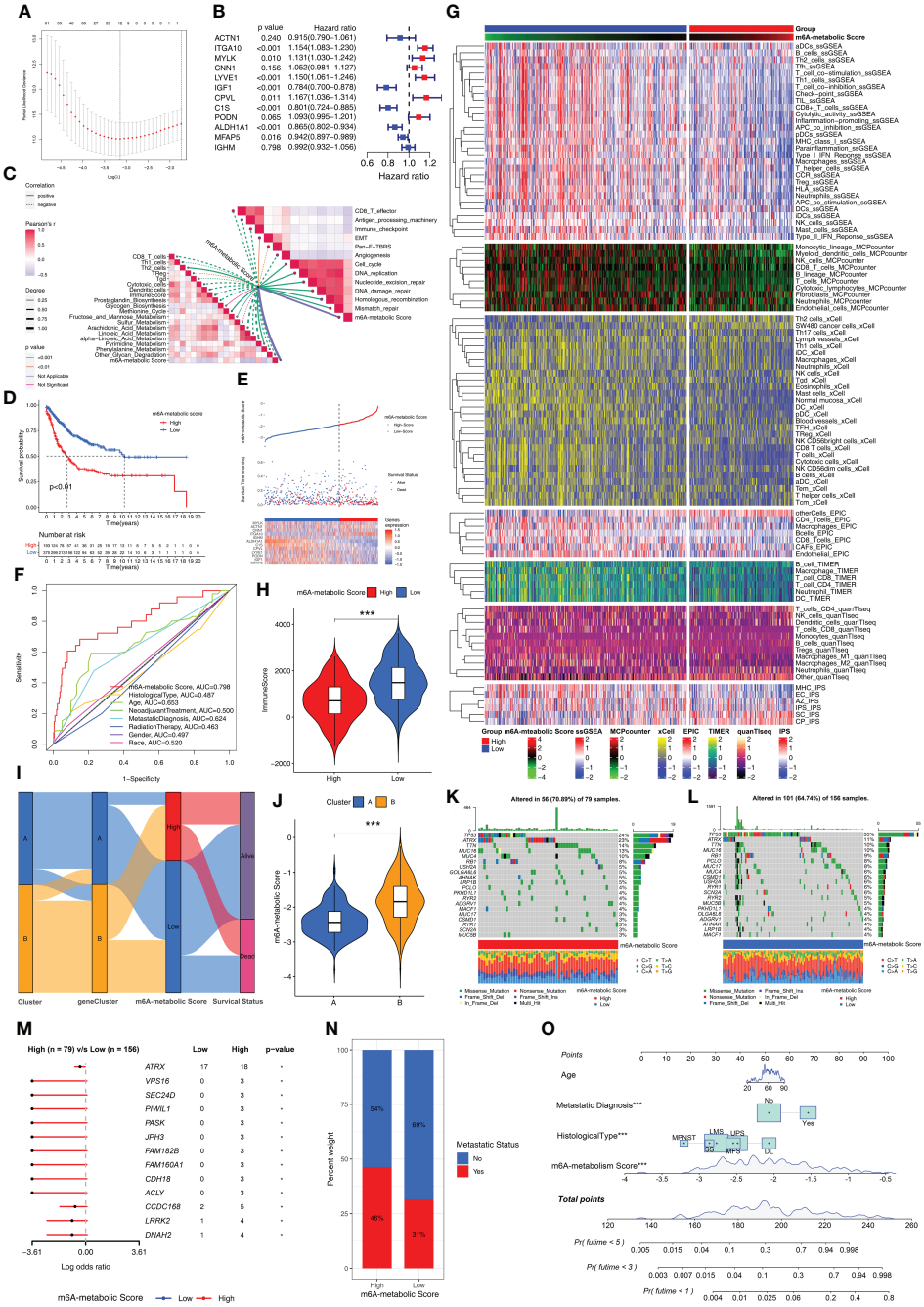


FIGURE 6
Biological characteristics and mutations between high and low m6A-metabolic Scores. **(A)** LASSO Cox regression model construction. λ selection by 10-fold cross-validation. The partial likelihood deviance with changing of log (λ) was plotted. **(B)** Forest plot shows 12 hub genes using multivariate cox regression analysis. **(C)** Correlations between m6A-metabolic Score and immune related biological processes, metabolic pathways, tumor-associated pathways. **(D)** Kaplan-Meier curves of overall survival in high and low m6A-metabolic Score groups for the TCGA-SARC and GSE21050 cohorts. **(E)** Comparison of differences in m6A metabolic Score, overall survival, and gene expression of 12 selected hub genes between high and low m6A-metabolic Score groups. **(F)** The ROC curve was used to calculate the area under the curve (AUC) of m6A-metabolic Score (AUC=0.798). **(G)** The heatmap demonstrates immune cell infiltration of high and low m6A-metabolic Score groups by ssGSEA, MCPcounter, xCell, EPIC, TIMER, quanTIseq and IPS algorithms. **(H)** The violin plot shows differences in immuneScore between the high and low m6A-metabolic Score groups. Statistical differences between the high and low m6A-metabolic Score groups were compared by the Wilcoxon test (***, P < 0.001). **(I)** Alluvial diagram displays changes in Clusters, geneClusters, m6A-metabolic Score and survival outcomes. **(J)** The violin plot reveals the differences of m6A-metabolic Scores between Cluster A and Cluster B The Wilcoxon test was used to compare the statistical difference between two groups. **(K, L)** The OncoPrint plots were drawn to reflect the landscape of mutations in high and low m6A-metabolic Score groups. **(M)** The forest plot revealed differences in 13 mutated genes between high and low m6A-metabolic Score groups (*, P < 0.05). **(N)** The proportion of tumor metastasis status in the high and low m6A-metabolic Score groups. Metastasis, red; No metastasis, blue. **(O)** A prognostic nomogram predicting 1-, 3-, and 5-year overall survival of STS.

the low m6A-metabolic Score group compared to the high m6A-metabolic Score group (Figure 7A). Furthermore, TIDE scores were lower and TIS scores were higher for the low m6A-metabolic Score group (Figures 7B, C). SubMap analysis showed that STS patients in the low m6A-metabolic Score group were more responsive to treatment with the PD-1 inhibitors (Figure 7D). The m6A-metabolic score also showed positive correlation with the TIDE score and negative correlation with the TIS score (Figure 7E). These suggested that STS patients with lower m6A-metabolic Scores were more sensitive to immunotherapy.

Chemotherapy is a standard treatment for the STS patients. Therefore, we analyzed the response of the two groups to chemotherapy. IC50 values indicate the potency of chemotherapeutic drugs to induce tumor cell apoptosis. STS samples in the low m6A-metabolic Score group were associated with lower IC50 values for 44 chemotherapeutic drugs compared with high m6A-metabolic Score group (Figure 7F). This suggested that chemotherapeutic response was significantly higher in the STS patients from the low m6A-metabolic Score group compared to the high m6A-metabolic Score group.

Discussion

RNA methylation has been proven to play a significant role in several key physiological processes. Therefore, alterations in m6A RNA methylation are implicated in human pathology and play a key role in cancers, immune system diseases, neurological diseases, and others (31). Previous studies have demonstrated

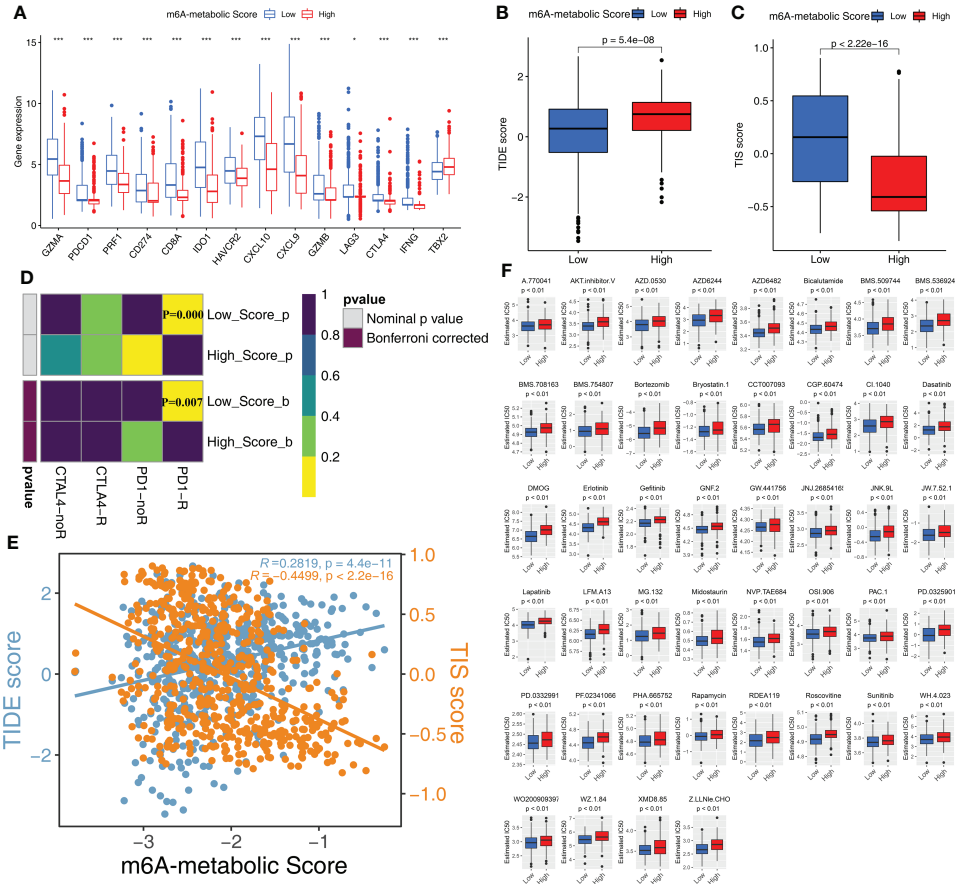


FIGURE 7
Immunotherapy response and drug sensitivity in high and low m6A-metabolic Score. **(A)** Differences in the gene expression of 14 immune checkpoint related genes in high and low m6A-metabolic Score groups. The thick line exhibited the median value. Statistical differences were compared by the Wilcoxon test (*P < 0.05; ***P < 0.001). **(B, C)** TIDE score **(B)** and TIS score **(C)** differences in the high and low m6A-metabolic Score groups. The upper and lower ends of the boxes indicated an interquartile range of values, the lines in the boxes represented median value. The statistical difference of two groups was compared through the Wilcoxon test. **(D)** Submap analysis shows that low m6A-metabolic Score groups could be more sensitive to anti-PD-1/PD-L1 treatment (Bonferroni corrected P-value = 0.007). **(E)** Correlation analysis (Spearman correlation) of m6A-metabolic Score with TIDEscore and TISscore. TIDE score, blue; TIS score, yellow. **(F)** The difference of chemotherapy response for 44 drugs between high and low m6A-metabolic Score groups. The statistical difference of two groups was compared through the Wilcoxon test.

that m6A mRNA methylation regulates tumor immunity, metabolism, and stemness (10). However, the role of m6A mRNA methylation in the regulation of the TME through modulation of metabolic pathways and its effects on the prognosis and immunotherapeutic responses of STS patients is not widely reported. Therefore, in this study, we analyzed the multi-omics data from STS patients to determine the relationship between m6A mRNA methylation and the status of the metabolic pathways in the STS tissues and their impact on tumor immunity, progression, and prognosis. The multi-omics data analysis revealed two distinct and stable subtypes in the STS samples based on m6A-related metabolism. Furthermore, we established a scoring system based on the m6A-related metabolism. We demonstrated that the m6A-metabolic Scores can accurately predict the prognosis of STS patients and their response to immunotherapy.

We identified 11 m6A-related metabolic pathways that were associated with the prognosis of STS, including prostaglandin biosynthesis, glycogen biosynthesis, methionine cycle, fructose and mannose metabolism, sulfur metabolism, arachidonic acid metabolism, linoleic acid metabolism, alpha-linoleic acid metabolism, pyrimidine metabolism, phenylalanine metabolism, other glycan degradation. Among them, pyrimidine metabolic pathway promotes tumor progression by increasing cancer cell proliferation (32, 33). The pyrimidine metabolic pathway is also associated with drug resistance in various cancer patients (34). Linoleic acid metabolic pathway promotes proliferation and migration of breast cancer cells (35). Glycogen biosynthesis is also associated with breast cancer, bladder cancer, and others (36). The hypoxic environment in the tumor tissue promotes proliferation of breast cancer cells by activating the glycogen metabolism pathway. Our results suggested that hypoxia probably altered metabolism in the TME and contributed to the heterogeneity in the STS tissues (37). Besides, the activation of glycogen metabolism has been reported to promote aerobic glycolysis or the "Warburg" effect in the cancer cells, glycogen metabolism is a common metabolic pathway in cancer cells and is a significant marker of malignant tumors (38). Other metabolic pathways have been previously reported to be involved in tumorigenesis and progression. In this study, these important metabolic pathways were observed to be closely related to STS, which will provide new insights and evidence for future metabolic therapies for STS (39–41).

We further analyzed significance of the 11 m6A-related metabolic pathways in altering the tumor microenvironment and their relationship with prognosis in STS patients. Clustering analysis revealed two subtypes among the STS patients (Cluster A and Cluster B) based on the status of the m6A-related metabolic pathways. These two subtypes showed significant differences in OS and biological characteristics. The OS rates of the STS patients in Cluster A were significantly higher than those in Cluster B. These two molecular subtypes of STS also showed distinct profiles of metabolic pathways. STS tissues in Cluster A showed higher expression of genes involved in linoleic acid metabolism, alpha linoleic acid metabolism, prostaglandin biosynthesis, and arachidonic acid metabolism. These metabolic pathways improved the prognosis of STS patients in Cluster A.

The TIME plays a crucial role in the progression and prognosis of multiple cancers. Activation of innate and adaptive immunity increases the survival rates of cancer patients and their sensitivity to immunotherapy. Therefore, we compared the characteristics of the TIME between clusters A and B. We observed enrichment of tumor-suppressing immune cells such as $CD8^+$ T effector cells in the Cluster A subtype, whereas, pro-tumorigenic pathways such as EMT were enriched in the Cluster B subtype. Furthermore, we used GSVA to analyze the differences in the enriched biological processes between Cluster A and Cluster B. The results showed significant enrichment of DNA replication in Cluster B, thereby suggesting increased proliferation of cancer cells. In Cluster A, we observed significant enrichment of the B cell receptor signaling pathway and natural killer (NK) cell mediated cytotoxicity, thereby suggesting enhanced activation of the immune cells. STS patients in Cluster B showed increased rate of metastasis and higher degree of malignancy compared to those in Cluster A. The higher survival rates of STS patients in Cluster A correlated with higher immune scores and increased infiltration to tumor-killing immune cells. In contrast, tumor-promoting immune cells were enriched in the Cluster B patients with STS and were associated with poor prognosis.

Previous studies have reported glucose and lipid metabolism plays a significant role in cancer stem cells originating from various cancers (42). Cancer stem cells rely highly on glucose and lipid metabolism for keeping their stemness features and satisfying their energy requirements, ultimately leading to tumor invasion and metastasis (43). Our results demonstrated that the tumor stemness levels was higher for STS tissues from Cluster B and was associated with increased metastasis. In addition, we explored the effects of alterations in tumor metabolism on EMT, which is often related with acquisition of stemness characteristics. Cluster B showed higher EMT scores, which supported increased incidence of tumor metastasis (44). Furthermore, compared with the Cluster A group, m6A regulators that were associated with metabolism in the STS tissues were up-regulated in Cluster B. Moreover, Cluster B group showed increased glycogen biosynthesis, methionine cycle, and pyrimidine metabolism. These data demonstrated that the m6A regulators modulated immune cell infiltration as well as tumor cell proliferation and progression by altering tumor cell metabolism. Furthermore, Cluster B was associated with higher rates of somatic mutations and higher degree of malignancy.

We also investigated the differences in immunotherapeutic responses between the STS patients in Cluster A and Cluster B. STS patients in Cluster A were more sensitive to immunotherapy. This suggested potential clinical application of the classification system based on m6A-related metabolic

pathways. Therefore, we developed a scoring system (m6A-metabolic Score) for the STS patients based on m6A-related metabolism. This scoring system was based on the expression of 12 hub genes including eight hub genes (*ITGA10, MYLK, LYVE1, IGF1, CPVL, C1S, ALDH1A1* and *MFAP5*) that were identified as independent prognostic biomarkers of STS. The m6A-metabolic Score was significantly lower for STS patients in Cluster A compared to those in Cluster B. Our data suggested that the m6A-metabolic Score was a reliable tool for comprehensive assessment of the two molecular subtypes based on m6A-related metabolism and could be used to determine the status of tumor immune infiltration and patient survival outcomes. We also demonstrated that the m6A-metabolic Score was a robust tool for determining the efficacy of immune checkpoint inhibitors in individual STS patients.

In conclusion, our study performed a comprehensive analysis of the multi-omics data from STS patients and classified them into two molecular subtypes based on m6A-related metabolism. We also established a m6A-related metabolism scoring system and demonstrated its accuracy in predicting the prognosis of the STS patients and predicting their response to immunotherapy. Therefore, the m6A-metabolic Score shows great promise in clinical application for accurately classifying STS patients at a molecular level and may be used as a guide for precision therapy of individual STS patients. However, our study also has several limitations. Firstly, we integrated two large STS cohorts for our analysis. This may have masked the heterogeneity in different cohorts. We corrected potential batch effects to overcome this issue. Secondly, the m6A-metabolic Score cutoff value requires further validation in larger cohorts of STS patients.

Data availability statement

All data used in this work can be acquired from the Gene-Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) and the GDC portal (https://portal.gdc.cancer.gov/). The accession number(s) can be found in the article/supplementary material.

Author contributions

Z-DH, CH, and R-XW conceived and designed this study. Z-DH, R-XW and Y-CF carried out the analysis procedure. Z-DH, S-YL and Y-JM analyzed the results. Z-DH, Y-CF, S-YL and YZ contributed analysis tools. Z-DH, Y-CF, CH and R-XW participated in the manuscript writing and revising. Z-DH and R-XW mainly participated in the manuscript revising. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.895465/full#supplementary-material

SUPPLEMENTARY TABLE 1
Clinical information for STS samples in the TCGA-SARC and GSE21050 cohorts.

SUPPLEMENTARY TABLE 2
The gene sets used in this study including 12 oncogenic pathways.

SUPPLEMENTARY TABLE 3
35 metabolic pathways correlated with m6A regulators (the absolute value of the correlation coefficient was greater than 0.4).

SUPPLEMENTARY FIGURE 1
Sample distributions from TCGA and GEO before batch effect correction. The "ComBat" algorithm was applied to reduce the likelihood of batch effects from non-biological technical biases between TCGA-SARC and GSE21050 cohorts.

SUPPLEMENTARY FIGURE 2
(A) Relative change in area under consensus cumulative distribution functions (CDF) curve (k = 2-9) using consensus clustering based on 11 m6A related metabolic pathways for total patients of TCGA-SARC and GSE21050 cohorts. **(B)** The consensus matrix heatmap corresponding to k=2-9 obtained by consensus clustering.

SUPPLEMENTARY FIGURE 3
(A) Relative change in area under consensus CDF curve (k = 2-9) using consensus clustering based on m6A related metabolic pathways for TCGA-SARC (left) and GSE21050 (right) cohorts, respectively. **(B)** The heatmap demonstrates differences in 11 m6A-related metabolic pathways scores between Cluster A and Cluster B. **(C)** The heatmap demonstrates immune cell infiltration of two molecular subtypes from unsupervised clustering in the TCGA-SARC and GSE21050 cohorts by ssGSEA, MCPcounter, xCell, EPIC, TIMER, quanTIseq and IPS algorithms. **(D)** Relative change in area under consensus CDF curve (k = 2-9) using consensus clustering based on DEGs for total population of TCGA-SARC and GSE21050 cohorts. **(E, F)** Survival analyses for the two Clusters in **(E)** TCGA-SARC and **(F)** GSE21050 cohorts, respectively. **(G)** The violin plot reveals the differences of m6A-metabolic Scores between geneCluster A and geneCluster B.

SUPPLEMENTARY FIGURE 4
(A) The boxplot reveals the differences of m6A-metabolic Scores between between STS patients with different immune microenvironment tissue subtype. **(B-E)** These boxplots reveal the differences of m6A-metabolic Scores between between STS patients with different ImmuneScore level **(B)**, survival time **(C)**, survival status **(D)**, metastatic status **(E)**.

References

1. Dammerer D, VANB A, Schneeweiss V, Schwabegger A. Follow-up strategies for primary extremity soft-tissue sarcoma in adults: A systematic review of the published literature. *In Vivo (Athen Greece)* (2020) 34(6):3057–68. doi: 10.21873/invivo.12140

2. Ayodele O, Razak ARA. Immunotherapy in soft-tissue sarcoma. *Curr Oncol (Toronto Ont)* (2020) 27(Suppl 1):17–23. doi: 10.3747/co.27.5407

3. Comandone A, Petrelli F, Boglione A, Barni S. Salvage therapy in advanced adult soft tissue sarcoma: A systematic review and meta-analysis of randomized trials. *Oncol* (2017) 22(12):1518–27. doi: 10.1634/theoncologist.2016-0474

4. Perrier A, Didelot A, Laurent-Puig P, Blons H, Garinet S. Epigenetic mechanisms of resistance to immune checkpoint inhibitors. *Biomolecules* (2020) 10(7):1061. doi: 10.3390/biom10071061

5. Judge SJ, Darrow MA, Thorpe SW, Gingrich AA, O'Donnell EF, Bellini AR, et al. Analysis of tumor-infiltrating NK and T cells highlights IL-15 stimulation and TIGIT blockade as a combination immunotherapy strategy for soft tissue sarcomas. *J Immunother Cancer* (2020) 8(2):e001355. doi: 10.1136/jitc-2020-001355

6. Martín-Broto J, Moura DS, Van Tine BA. Facts and hopes in immunotherapy of soft-tissue sarcomas. *Clin Cancer Res* (2020) 26(22):5801–8. doi: 10.1158/1078-0432.ccr-19-3335

7. Roulleaux Dugage M, Nassif EF, Italiano A, Bahleda R. Improving immunotherapy efficacy in soft-tissue sarcomas: A biomarker driven and histotype tailored review. *Front Immunol* (2021) 12:775761. doi: 10.3389/fimmu.2021.775761

8. He L, Li H, Wu A, Peng Y, Shu G, Yin G. Functions of N6-methyladenosine and its role in cancer. *Mol Cancer* (2019) 18(1):176. doi: 10.1186/s12943-019-1109-9

9. Han X, Wang L, Han Q. Advances in the role of M(6)A RNA modification in cancer metabolic reprogramming. *Cell Biosci* (2020) 10:117. doi: 10.1186/s13578-020-00479-z

10. Huang ZD, Lin LL, Liu ZZ, Hu C, Gu HY, Wei RX. M6a modification patterns with distinct immunity, metabolism, and stemness characteristics in soft tissue sarcoma. *Front Immunol* (2021) 12:765723. doi: 10.3389/fimmu.2021.765723

11. Xia L, Oyang L, Lin J, Tan S, Han Y, Wu N, et al. The cancer metabolic reprogramming and immune response. *Mol Cancer* (2021) 20(1):28. doi: 10.1186/s12943-021-01316-8

12. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. *Immunity* (2018) 48(4):812–30.e14. doi: 10.1016/j.immuni.2018.03.023

13. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* (2018) 28 (11):1747–56. doi: 10.1101/gr.239244.118

14. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* (2011) 12(4):R41. doi: 10.1186/gb-2011-12-4-r41

15. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using the cancer genome atlas. *Nat Commun* (2018) 9(1):5330. doi: 10.1038/s41467-018-07232-8

16. Wang T, Kong S, Tao M, Ju S. The potential role of RNA N6-methyladenosine in cancer progression. *Mol Cancer* (2020) 19(1):88. doi: 10.1186/s12943-020-01204-7

17. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinf (Oxf Engl)* (2010) 26 (12):1572–3. doi: 10.1093/bioinformatics/btq170

18. Sturm G, Finotello F, List M. Immunedeconv: An r package for unified access to computational methods for estimating immune cell fractions from bulk RNA-sequencing data. *Methods Mol Biol (Clifton NJ)* (2020) 2120:223–32. doi: 10.1007/978-1-0716-0327-7_16

19. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7

20. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612

21. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (New York NY)* (2021) 2(3):100141. doi: 10.1016/j.xinn.2021.100141

22. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* (2018) 173(2):338–54.e15. doi: 10.1016/j.cell.2018.03.034

23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007

24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models *via* coordinate descent. *J Stat Softw* (2010) 33(1):1–22. doi: 10.18637/jss.v033.i01

25. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* (2018) 24(10):1550–8. doi: 10.1038/s41591-018-0136-1

26. Danaher P, Warren S, Lu R, Samayoa J, Sullivan A, Pekker I, et al. Pan-cancer adaptive immune resistance as defined by the tumor inflammation signature (TIS): Results from the cancer genome atlas (TCGA). *J Immunother Cancer* (2018) 6(1):63. doi: 10.1186/s40425-018-0367-1

27. Hoshida Y, Brunet JP, Tamayo P, Golub TR, Mesirov JP. Subclass mapping: Identifying common subtypes in independent disease data sets. *PLoS One* (2007) 2 (11):e1195. doi: 10.1371/journal.pone.0001195

28. Roh W, Chen PL, Reuben A, Spencer CN, Prieto PA, Miller JP, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Trans Med* (2017) 9(379): eaah3560. doi: 10.1126/scitranslmed.aah3560

29. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinf (Oxf Engl)* (2016) 32 (18):2847–9. doi: 10.1093/bioinformatics/btw313

30. Geeleher P, Cox N, Huang RS. Prrophetic: An r package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One* (2014) 9(9):e107468. doi: 10.1371/journal.pone.0107468

31. Jiang X, Liu B, Nie Z, Duan L, Xiong Q, Jin Z, et al. The role of M6a modification in the biological functions and diseases. *Signal Transduct Target Ther* (2021) 6(1):74. doi: 10.1038/s41392-020-00450-x

32. Edwards L, Gupta R, Filipp FV. Hypermutation of DPYD deregulates pyrimidine metabolism and promotes malignant progression. *Mol Cancer Res* (2016) 14(2):196–206. doi: 10.1158/1541-7786.mcr-15-0403

33. Wang X, Yang K, Wu Q, Kim LJY, Morton AR, Gimple RC, et al. Targeting pyrimidine synthesis accentuates molecular therapy response in glioblastoma stem cells. *Sci Trans Med* (2019) 11(504):eaau4972. doi: 10.1126/scitranslmed.aau4972

34. Gu X, Tohme R, Tomlinson B, Sakre N, Hasipek M, Durkin L, et al. Decitabine- and 5-azacytidine resistance emerges from adaptive responses of the pyrimidine metabolism network. *Leukemia* (2021) 35(4):1023–36. doi: 10.1038/s41375-020-1003-x

35. Gonzalez-Reyes C, Marcial-Medina C, Cervantes-Anaya N, Cortes-Reynosa P, Salazar EP. Migration and invasion induced by linoleic acid are mediated through fascin in MDA-MB-231 breast cancer cells. *Mol Cell Biochem* (2018) 443 (1-2):1–10. doi: 10.1007/s11010-017-3205-8

36. Massari F, Ciccarese C, Santoni M, Iacovelli R, Mazzucchelli R, Piva F, et al. Metabolic phenotype of bladder cancer. *Cancer Treat Rev* (2016) 45:46–57. doi: 10.1016/j.ctrv.2016.03.005

37. Tang K, Zhu L, Chen J, Wang D, Zeng L, Chen C, et al. Hypoxia promotes breast cancer cell growth by activating a glycogen metabolic program. *Cancer Res* (2021) 81(19):4949–63. doi: 10.1158/0008-5472.can-21-0753

38. Shulman RG, Rothman DL. The glycogen shunt maintains glycolytic homeostasis and the warburg effect in cancer. *Trends Cancer* (2017) 3(11):761–7. doi: 10.1016/j.trecan.2017.09.007

39. Yarla NS, Bishayee A, Sethi G, Reddanna P, Kalle AM, Dhananjaya BL, et al. Targeting arachidonic acid pathway by natural products for cancer prevention and therapy. *Semin Cancer Biol* (2016), 40–41:48–81. doi: 10.1016/j.semcancer.2016.02.001

40. Locasale JW. Serine, glycine and one-carbon units: Cancer metabolism in full circle. *Nat Rev Cancer* (2013) 13(8):572–83. doi: 10.1038/nrc3557

41. Ward NP, DeNicola GM. Sulfur metabolism and its contribution to malignancy. *Int Rev Cell Mol Biol* (2019) 347:39–103. doi: 10.1016/bs.ircmb.2019.05.001

42. Visweswaran M, Arfuso F, Warrier S, Dharmarajan A. Aberrant lipid metabolism as an emerging therapeutic strategy to target cancer stem cells. *Stem Cells (Dayton Ohio)* (2020) 38(1):6–14. doi: 10.1002/stem.3101

43. Yi M, Li J, Chen S, Cai J, Ban Y, Peng Q, et al. Emerging role of lipid metabolism alterations in cancer stem cells. *J Exp Clin Cancer Res* (2018) 37(1):118. doi: 10.1186/s13046-018-0784-5

44. Kim BN, Ahn DH, Kang N, Yeo CD, Kim YK, Lee KY, et al. TGF-β induced EMT and stemness characteristics are associated with epigenetic regulation in lung cancer. *Sci Rep* (2020) 10(1):10597. doi: 10.1038/s41598-020-67325-7

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fmolb.2022.972146

Check for updates

OPEN ACCESS

EDITED BY
George Bebis,
University of Nevada, Reno,
United States

REVIEWED BY
Mohit Kumar Jolly,
Indian Institute of Science (IISc), India
Aritro Nath,
City of Hope National Medical Center,
United States

*CORRESPONDENCE
Guillermo Lorenzo,
guillermo.lorenzo@utexas.edu,
guillermo.lorenzo@unipv.it

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 17 June 2022
ACCEPTED 17 August 2022
PUBLISHED 12 September 2022

CITATION
Yang EY, Howard GR, Brock A, Yankeelov TE and Lorenzo G (2022), Mathematical characterization of population dynamics in breast cancer cells treated with doxorubicin. *Front. Mol. Biosci.* 9:972146. doi: 10.3389/fmolb.2022.972146

COPYRIGHT
© 2022 Yang, Howard, Brock, Yankeelov and Lorenzo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Mathematical characterization of population dynamics in breast cancer cells treated with doxorubicin

Emily Y. Yang[1], Grant R. Howard[2], Amy Brock[2,3,4], Thomas E. Yankeelov[1,2,3,5,6,7] and Guillermo Lorenzo[1,8]*

[1]Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, United States, [2]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, United States, [3]Livestrong Cancer Institutes, Dell Medical School, The University of Texas at Austin, Austin, TX, United States, [4]Interdisciplinary Life Sciences Program, The University of Texas at Austin, Austin, TX, United States, [5]Department of Diagnostic Medicine, The University of Texas at Austin, Austin, TX, United States, [6]Department of Oncology, The University of Texas at Austin, Austin, TX, United States, [7]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [8]Department of Civil Engineering and Architecture, University of Pavia, Pavia, Italy

The development of chemoresistance remains a significant cause of treatment failure in breast cancer. We posit that a mathematical understanding of chemoresistance could assist in developing successful treatment strategies. Towards that end, we have developed a model that describes the cytotoxic effects of the standard chemotherapeutic drug doxorubicin on the MCF-7 breast cancer cell line. We assume that treatment with doxorubicin induces a compartmentalization of the breast cancer cell population into surviving cells, which continue proliferating after treatment, and irreversibly damaged cells, which gradually transition from proliferating to treatment-induced death. The model is fit to experimental data including variations in drug concentration, inter-treatment interval, and number of doses. Our model recapitulates tumor cell dynamics in all these scenarios (as quantified by the concordance correlation coefficient, CCC > 0.95). In particular, superior tumor control is observed with higher doxorubicin concentrations, shorter inter-treatment intervals, and a higher number of doses ($p < 0.05$). Longer inter-treatment intervals require adapting the model parameterization after each doxorubicin dose, suggesting the promotion of chemoresistance. Additionally, we propose promising empirical formulas to describe the variation of model parameters as functions of doxorubicin concentration (CCC > 0.78). Thus, we conclude that our mathematical model could deepen our understanding of the cytotoxic effects of doxorubicin and could be used to explore practical drug regimens achieving optimal tumor control.

KEYWORDS

mathematical oncology, mathematical modeling, population dynamics, chemotherapy, drug resistance, breast cancer, time-resolved microscopy

Introduction

Breast cancer is the most common cancer among women worldwide and the leading cause of cancer death in over 100 countries (Sung et al., 2021). Chemotherapy is a primary component of cancer treatment and options have both advanced and increased considerably in recent years (Anampa et al., 2015). However, the development of chemoresistance, and resulting tumor recurrence, remains a common cause of treatment failure and a primary cause of cancer death (Holohan et al., 2013; Longley and Johnston, 2005). Indeed, for a standard chemotherapy drug such as doxorubicin, breast cancer patients may develop chemoresistance within just 6–10 months (Rivera and Gomez, 2010; O'Shaughnessy, 2005).

From a biological perspective, the development of chemoresistance is governed by many complex mechanisms, such as treatment-induced genetic and epigenetic alterations, altered metabolic states, and adaptive responses of the tumor microenvironment (Easwaran et al., 2014; Zhao, 2016; Dong et al., 2018; Ji et al., 2019). Tumor cells can also possess an intrinsic phenotypic or genetic resistance that can render the therapy ineffective even before acquired chemoresistance develops (Harris et al., 2007; Bernard et al., 2009; Campbell et al., 2018). Moreover, the existence of intratumoral heterogeneity and its role in tumor regrowth have become increasingly recognized, as the presence of even a minor subpopulation of drug-resistant cells can give rise to tumor relapse (Easwaran et al., 2014; Polyak, 2011; Alizadeh et al., 2015; Sun and Yu, 2015). Furthermore, phenotype switching, in which tumor cells swap between varying degrees of drug-resistant and drug-sensitive phenotypes, can enable the establishment of more permanent chemoresistance mechanisms that hinder complete tumor eradication (Easwaran et al., 2014; Meacham and Morrison, 2013; Echeverria et al., 2019; Kumar et al., 2019). Considering the complex biological processes underlying chemoresistance development, we believe that a robust framework is needed to comprehensively integrate the growing knowledge of this phenomenon and guide future research efforts. To this end, experimentally-validated mathematical models of chemoresistance mechanisms could be a potent tool in understanding the dynamics of overall tumor drug response. The description of cancer growth and therapeutic response by leveraging mechanistic mathematical models is a rich field known as mathematical oncology (Yankeelov et al., 2013; Rockne et al., 2019; Lorenzo et al., 2022). This approach has already shown promise in characterizing breast cancer growth and treatment response in both the preclinical and clinical settings (Atuegwu et al., 2013; Pascal et al., 2013; Weis et al., 2015; Zhang et al., 2015; Geng et al., 2017; Palmer and Sorger, 2017; Jarrett et al., 2018; Jarrett et al., 2020a).

There are several mechanistic approaches to mathematically describe chemoresistance (Sun and Hu, 2018; Craig et al., 2019; Craig et al., 2021; Hori et al., 2021), with the original theoretical models dating back more than two decades (Panetta, 1997; Panetta, 1998). The standard strategy consists of defining a multicompartmental tumor cell population including one or multiple species of both drug-resistant and drug-sensitive cells, which evolve and interact over time following a set of ordinary differential equations, or over both space and time according to a set of partial differential equations (Craig et al., 2019; Craig et al., 2021; Hori et al., 2021; Jackson and Byrne, 2000; Greene et al., 2019; Kim et al., 2021; Strobl et al., 2021; Howard et al., 2022). Alternatively, Sun et al. (2016) utilized a stochastic, multiscale model that incorporated heterogeneous population dynamics with drug pharmacokinetics and microenvironment contributions to drug resistance in melanoma patients. Furthermore, Pisco et al. (2013) and Álvarez-Arenas et al. (2019) applied the evolutionary theories of Darwinian selection and Lamarckian induction to guide their modeling of drug resistance in leukemia cells and non-small cell lung carcinoma, respectively. However, despite these promising studies, there is a still a dearth of experimentally-validated mechanistic models of chemoresistance in breast cancer, with which we could test alternative biological hypotheses to ultimately enhance chemotherapeutic strategies for individual patients. For instance, Chapman et al. (2019) developed a model integrating phenotypic switching of cell differentiation states and tumor heterogeneity to characterize therapeutic escape in the triple-negative subtype, but the empirical validation of their model predictions currently remains limited. Additionally, *in vitro* studies usually label cell lines as homogeneously drug-resistant or drug-sensitive and assume a static drug sensitivity (AbuHammad and Zihlif, 2013; Gottesman, 2002), which overlooks the existence of intratumoral heterogeneity and transient drug resistance. Moreover, preclinical studies often assess tumor cell death at a single time point 24–72 h post-treatment (Martins et al., 2018; Chung et al., 2020; Low et al., 2021). This experimental setting does not enable the characterization of long-term tumor drug responses and, hence, the development of drug-induced chemoresistance.

Here, we present a mechanistic model to describe the dynamics of drug response and chemoresistance development in MCF-7 breast cancer cells treated with doxorubicin, which we fit to time-resolved microscopy measurements of tumor cell number subjected to diverse therapeutic plans over long experimental times (>8 days). Doxorubicin is a cytotoxic anthracycline drug that is extensively used in chemotherapeutic regimens for breast cancer (Jarrett et al., 2020a; Zardavas and Piccart, 2015; Waks and Winer, 2019). As a cytotoxic drug, treatment with doxorubicin primarily induces tumor cell death, but this therapeutic effect may also be preceded by cell cycle arrest (Anampa et al., 2015; Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017). Our work continues the first efforts of Howard et al. in studying

doxorubicin resistance in breast cancer cell populations by leveraging several experimentally-informed mechanistic models (Howard et al., 2022; Howard et al., 2018). While Howard et al. originally proposed multiple models to characterize this phenomenon and selected the best of them for each dataset, we have developed a single model that can be extended for multiple drug doses. To incorporate intratumoral heterogeneity, we assume that the cytotoxic action of doxorubicin treatment induces a compartmentalization of the breast cancer cell population into two subgroups: surviving cells and irreversibly damaged cells, which will ultimately die due to doxorubicin action. We further assume that the surviving cells continue proliferating after exposure to doxorubicin, while the irreversibly damaged cells progressively transition from proliferation to drug-induced death. Hence, the eventual development of chemoresistance will be driven by the surviving cells. Importantly, the model compartmentalization ultimately results from the underlying distribution of diverse drug sensitivity phenotypes in the tumor cell population and its changes after the delivery of each doxorubicin dose (Howard et al., 2022; Pisco et al., 2013; Álvarez-Arenas et al., 2019). To accommodate potentially significant treatment-induced variations in the underlying spectrum of drug resistance phenotypes, we investigate an adaptive parametrization of our model with each doxorubicin dose. Hence, drug sensitivity in the tumor cell population is assumed to be dynamic with time, thereby accounting for tumor cell plasticity (Easwaran et al., 2014; Meacham and Morrison, 2013; Echeverria et al., 2019; Kumar et al., 2019). Additionally, our model is fit to the same time-resolved microscopy experiments used in Howard et al. (2022), in which breast cancer cells were subjected to doxorubicin treatments varying in either drug concentration, inter-treatment interval, or the number of doses. Our results show that our proposed model can fit the data observed in all three scenarios with remarkable accuracy. We have also analyzed the model parameter trends for each experiment and built empirical parameter formulas as functions of doxorubicin concentration, which may provide further insight into the development of chemoresistance.

The remainder of this work is organized as follows. First, given that we utilize the time-resolved microscopy data previously collected by Howard et al. (2022), we briefly outline their acquisition and preprocessing procedures. We also describe the derivation of the model and explain the numerical and statistical methods leveraged in this study. We then present the results from our model fittings for each of the three aforementioned experimental scenarios and analyze the corresponding quality of fit and trends in model parameters. To conclude, we discuss the main implications from our work, its limitations, and future directions.

Methods

Data acquisition and preprocessing

The experimental data leveraged in this study were fully obtained from Howard et al. (2022). In the following, we provide only the salient details of the data acquisition and preprocessing procedures presented in Howard et al. (2022) and directly relevant to our study, to which we added a final outlier assessment.

Cell culture

MCF-7 human breast cancer cells (ATCC HTB-22) were cultured in Minimum Essential Media (Gibco) supplemented with 10% fetal bovine serum (Gibco) and 1% penicillin-streptomycin (Gibco). Cells were maintained at 37°C with 5% CO_2. A stable fluorescent cell line expressing constitutive EGFP with a nuclear localization signal (MCF7-EGFPNLS1) was established to aid in the automated cell quantification of the time resolved microscopy measurements (Howard et al., 2022; Howard et al., 2018). Genomic integration of the EGFP expression cassette was accomplished by leveraging the Sleeping Beauty transposon system. The EGFP-NLS sequence was obtained as a gBlock (IDT) and cloned into the optimized Sleeping Beauty transfer vector pSBbi-Neo (which was a gift from Eric Kowarz, Addgene plasmid #60525) (Kowarz et al., 2015). To mediate genomic integration, this two-plasmid system consisting of the transfer vector containing the EGFP-NLS sequence and the pCMV(CAT)T7-SB100 plasmid containing the Sleeping Beauty transposase was co-transfected into the MCF-7 population utilizing Lipofectamine 2000 (mCMV(CAT)T7-SB100 was a gift from Zsuzsanna Izsvak, Addgene plasmid #34879) (Mátés et al., 2009). After gene integration with Sleeping Beauty transposase, EGFP + cells were collected by fluorescence activated cell sorting and maintained in media supplemented with 200 ng/ml G418 (Caisson Labs) in place of penicillin-streptomycin.

Doxorubicin response experiments

Cells were seeded in a 96-well plate at a target density of 2,000 cells/well and grown for approximately 48 h to allow for cell adhesion and recovery from passaging. An IncuCyte S2 Live Cell Analysis System (Essen/Sartorius, Goettingen, Germany) was used to collect fluorescent and phase contrast images every 2–4 h. Images were collected for periods of 21–56 days to ensure that cultures in which cells recover after exposure to doxorubicin were able to display logistic growth. Doxorubicin treatment was prepared by reconstituting doxorubicin hydrochloride (Cayman Chemical 15,007, Ann Arbor, Michigan) in water and mixing it with 100 µl of growth media at 2× the target concentration, which was then added to each well of the plate. The drug-containing media was then replaced with fresh growth media after 24 h. Three experiment types were run, in which either the

TABLE 1 Experimental conditions. In Experiment 1, one dose of doxorubicin was delivered at concentrations varying from 10 to 300 nM (n = 6). In Experiment 2, two doses of 75 nM doxorubicin were delivered at inter-treatment intervals varying from 0 to 16 days (n = 12). In Experiment 3, one to five doses of 75 nM doxorubicin were delivered at either 2-day or 2-week inter-treatment intervals (n = 12).

Experiment	Concentration [nM]	Inter-treatment Interval [d]	Number of Doses	Number of Replicates
1	0, 10, 20, 35, 50, 75, 100, 125, 150, 300	-	1	6
2	75	0, 2, 4, 6, 8, 10, 12, 14, 16	2	12
3	75	2 or 14	1, 2, 3, 4, 5	12

doxorubicin concentration, the inter-treatment interval, or the number of doses was varied (see Table 1). Each doxorubicin concentration was tested in n = 6 replicates, while each inter-treatment interval and number of doses was tested in n = 12 replicates.

Image analysis

Using IncuCyte's integrated software, the quantification of total tumor cell counts was performed on the fluorescent images using the green fluorescence channel. Individual cells were consistently resolved using standard image analysis techniques of background subtraction, followed by thresholding, edge detection, and minimum area filtering. The phase contrast images were consulted in parallel to aid the validation of image analysis (Howard et al., 2022; Howard et al., 2018).

Data truncation

The tumor cell time courses extracted from some wells did not provide meaningful data throughout the entire time course due to a variety of reasons. These included the cell population growing to confluence and fluctuating with feeding cycles, being disturbed during media replenishment, or growing three-dimensionally resulting in cells overlapping each other and thereby compromising the ability to accurately quantify cell numbers. Thus, for each dataset, the estimated cell number was truncated either just prior to reaching confluence, when the cell number dropped more than 50% due to media handling, or when repeated discontinuities were observed in the time course data.

Data normalization

For smaller discontinuities in which less than 50% of the cells were lost due to media handling, the data was normalized by dividing the cell number at time points before the discontinuity by a constant α (Howard et al., 2022) calculated *via* Eq. 1:

$$\alpha = \frac{\frac{(N_{d-1}-N_{d-2})}{(t_{d-1}-t_{d-2})} + 2\frac{N_{d-1}}{t_d - t_{d-1}}}{2\frac{N_d}{t_d - t_{d-1}} + \frac{N_d - N_{d+1}}{t_{d+1} - t_d}} \tag{1}$$

in which N_d, N_{d-i}, and N_{d+i} are the total tumor cell counts at the discontinuity, i points before the discontinuity, and i points after the discontinuity, respectively. t_d,t_{d-i}, and t_{d+i} are the times of the discontinuity, i points before the discontinuity, and i points after the discontinuity, respectively. The objective of this normalization was to smooth the first and second derivatives of the total tumor cell counts across the discontinuity, as proposed in (Howard et al., 2022). Supplementary Appendix A in the Supplementary Information provides further details about the purpose and derivation of Eq. 1.

Outlier removal

For datasets that possessed outliers, the *rmoutliers* function from MATLAB R2020b (The Mathworks, Natick, MA) was used to remove data points using median filtering. A visual inspection of the resulting data confirmed that this method removed evident outliers from the original series, while maintaining the natural fluctuations in tumor cell counts (see Supplementary Figures S1–S5).

Mathematical model

We present a mathematical model to describe the response of MCF-7 breast cancer cells to the cytotoxic action of doxorubicin in the three experimental scenarios listed in Table 1. We begin by describing the biological mechanisms captured by the model assuming a single dose of doxorubicin (Experiment 1, Table 1). Then, we show how the model can be generalized to multiple doses (Experiments 2 and 3, Table 1), and can also be modified to vary specific parameters with each dose. Figure 1 illustrates the main tumor cell dynamics described by our model after each dose of doxorubicin, which are further detailed in the following paragraphs. The reader can refer to Supplementary Table S1 for a consolidated list of model parameter definitions and their units.

Single-dose model

We start with a population of tumor cells (N) that grow untreated for a specified period of time prior to doxorubicin treatment (approximately 48 h). We assume that these untreated cells follow logistic growth:

$$\frac{dN}{dt} = g_0 N\left(1 - \frac{N}{\theta_u}\right) \tag{2}$$

$$N(0) = N_0 \tag{3}$$

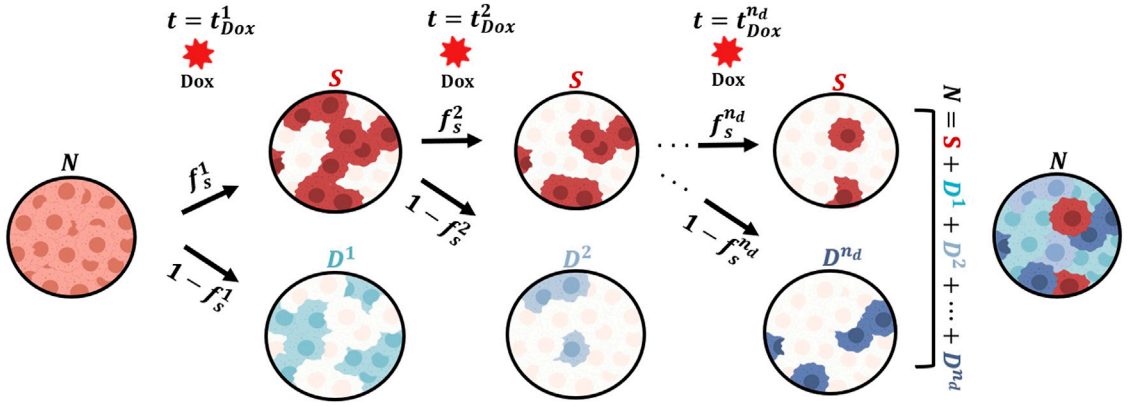


FIGURE 1
Generalized model of tumor cell response to multiple doses of doxorubicin treatment. We start with a population of untreated tumor cells and let them grow for approximately 48 h. At time t^1_{Dox}, we add a dose of doxorubicin (Dox) to each well. We assume that after the treatment, the tumor cells either survive (S) or are irreversibly damaged (D^1) and ultimately die due to the cytotoxic action of doxorubicin. The fraction of cells in either subpopulation is determined by f^1_s, the fraction of surviving cells after the first dose. After the subsequent n_d doses of doxorubicin ($i = 2, 3, \ldots, n_d$), we assume that a fraction f^i_s of the surviving cells survive the treatment, while a fraction $(1 - f^i_s)$ induces a new subpopulation of irreversibly damaged cells (D^i), such that the total number of tumor cells is $N = S + D^1 + D^2 + \ldots + D^{n_d}$ for times $t > t^{n_d}_{Dox}$. In this study, we further assess whether the collection of f^i_s can be assumed to take on the same value or whether they require an independent parameterization with each doxorubicin dose. This Figure was created using BioRender.com.

where g_0 is the untreated proliferation rate, θ_u is the untreated tumor cell carrying capacity, and N_0 is the initial number of tumor cells. We set θ_u = 53,873 cells, which corresponds to the mean value resulting from the fitting of Eqs 2, 3 to the untreated datasets in Experiment 1 (i.e., 0 nM doxorubicin; further details can be found in Supplementary Tables S2–S5 and Supplementary Figure S1).

Let t^1_{Dox} denote the time at which a single dose of doxorubicin is delivered, as described in Experiment 1 (Table 1). At this time point, we assume that a fraction f_s of the tumor cells survives the treatment (S), whereas the complementary fraction $(1 - f_s)$ is irreversibly damaged by the cytotoxic action of doxorubicin and will ultimately die (D) (Anampa et al., 2015; Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017). Hence, the fraction f_s ultimately depends on the underlying spectrum of drug sensitivities in the tumor cell population as well as on the amount of drug delivered with each dose (Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017; Zoli et al., 1995). We denote the initial number of surviving and irreversibly damaged tumor cells immediately after treatment with doxorubicin as $S(t^{1+}_{Dox})$ and $D(t^{1+}_{Dox})$, respectively, which are defined based on the number of untreated cells immediately before the delivery of doxorubicin, $N(t^{1-}_{Dox})$, as

$$S(t^{1+}_{Dox}) = f_s N(t^{1-}_{Dox}) \quad (4)$$

$$D(t^{1+}_{Dox}) = (1 - f_s) N(t^{1-}_{Dox}) \quad (5)$$

such that the total tumor cell number N for time $t > t^1_{Dox}$ is calculated as

$$N(t) = S(t) + D(t) \quad (6)$$

Note that Eqs 4–6 ensure the continuity in the tumor cell count before and after the treatment with doxorubicin, as observed in the corresponding experimental data (see Supplementary Figure S2).

We assume that the surviving cells also follow logistic growth with a different rate and carrying capacity:

$$\frac{dS}{dt} = g_s S\left(1 - \frac{N}{\theta_{Dox}}\right) \quad (7)$$

where g_s is the proliferation rate of surviving tumor cells and θ_{Dox} is the treated tumor cell carrying capacity. For the irreversibly damaged cells, we assume that their logistic growth dynamics gradually transition from proliferation to treatment-induced death at an exponentially-decaying rate:

$$\frac{dD}{dt} = (k_d + (g_d - k_d)\exp(-\gamma_d(t - t^1_{Dox})))D\left(1 - \frac{N}{\theta_{Dox}}\right), \quad (8)$$

where k_d and g_d denote the drug-induced death rate and the proliferation rate of the irreversibly damaged tumor cells, respectively, while γ_d represents the drug-induced death delay rate. The latter mechanism represents the varying duration of the cascade of biological events that takes place between drug exposure and the ultimate doxorubicin-induced tumor cell death (e.g., uptake by tumor cells, damage to DNA, cell cycle arrest, induction of tumor cell death) (Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017). In Eq. 8, the use of a common logistic model formulation to describe

the initial growth after treatment and the ensuing drug-induced death in the irreversibly-damaged tumor cell compartment facilitates the modeling of this transition in the dynamics of this subpopulation within the growth rate of the logistic model (i.e., the first factor in parenthesis in the right-hand side of Eq. 8). Additionally, the cytotoxic action of doxorubicin targets proliferating cells (Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017). The logistic model formulation in Eq. 8 further enables to account for the limitation to tumor cell proliferation depending on the total tumor cell density, which may hence limit drug-induced tumor cell death.

In Eqs 7, 8, we introduce a treated tumor cell carrying capacity (θ_{Dox}), which may be different from θ_u defined for untreated cells in Eq. 2. The rationale for this modeling choice is inspired by the experimental data from Howard et al. (2022), which shows that the maximum tumor cell counts in the replicates treated with doxorubicin could reach either larger or smaller values at confluence with respect to their untreated counterparts (i.e., the 0 nM replicates in Experiment 1; see Supplementary Figures S1–S5). To estimate θ_{Dox} in Eqs 7, 8, we used either one of two approaches. If the last tumor cell count in a dataset was greater than 30% of θ_u, then θ_{Dox} was fit along with the other model parameters. Conversely, if the last tumor cell count was less than 30% of θ_u,then we fixed θ_{Dox} to the mean of the values obtained from the replicates of the same experiment in which this parameter was directly fit. The rationale for this approach is that we observed that final tumor cell counts below 30% of θ_u did not provide enough identifiability for θ_{Dox}, which ultimately induced significant model fitting errors.

Multiple-dose model

Let us now consider a treatment schedule consisting of n_d doses of doxorubicin delivered at times t^i_{Dox}, $(i = 1, 2, \ldots, n_d)$, as described in Experiments 2 and 3 (Table 1). For the first dose, we assume that a fraction f^1_s of the tumor cells survives treatment with doxorubicin and, thus, the multiple-dose model remains identical to the single-dose model described in the previous section. For each of the subsequent drug doses, we assume that a fraction f^i_s $(i = 2, \ldots, n_d)$ of the tumor cells that escaped the cytotoxic action of the previous doxorubicin doses, $S(t^{i-}_{Dox})$, survives the new dose, while a corresponding fraction $1 - f^i_s$ gives rise to a new irreversibly damaged population D^i. We further assume that each new subpopulation D^i is characterized by a distinct value of the death delay rate γ^i_d. The rationale for considering an adaptive parameterization of parameters f_s and γ_d with each doxorubicin dose is that the underlying spectrum of tumor cell sensitivities may significantly change with each doxorubicin dose and inter-treatment interval (Howard et al., 2022; Pisco et al., 2013; Álvarez-Arenas et al., 2019; Lyman, 2009; De Souza et al., 2011; Ponnusamy et al., 2017), such that the surviving fraction may exhibit an increased resistance to the drug. We hypothesize that this phenomenon results in a higher potential to survive a new drug dose (i.e., larger values of f_s) or to partially hinder the cytotoxic action of doxorubicin before ultimately succumbing (i.e., lower values of γ_d).

Thus, after the delivery of the i^{th} dose ($i \geq 2$), the number of surviving tumor cells $S(t^{i+}_{Dox})$ and the initial number of the new subpopulation of irreversibly damaged cells $D^i(t^{i+}_{Dox})$ are calculated as

$$S(t^{i+}_{Dox}) = f^i_s S(t^{i-}_{Dox}) \tag{9}$$

$$D^i(t^{i+}_{Dox}) = \left(1 - f^i_s\right) S(t^{i-}_{Dox}) \tag{10}$$

such that the total number of tumor cells during and after treatment with multiple doses of doxorubicin is given by

$$N(t) = S(t) + \sum_{i=1}^{n_d} D^i(t) \mathcal{H}(t - t^i_{Dox}), \tag{11}$$

where $\mathcal{H}(t - t^i_{Dox})$ is the Heaviside step function, which equals 0 for t < t^i_{Dox} and 1 for t ≥ t^i_{Dox}. Note that Eqs 8–11 ensure the continuity in the total tumor cell number before and after each doxorubicin dose, as observed in the data from Experiments 2 and 3 (Table 1; see Supplementary Figures S3–S5).

In this multiple-dose model, we assume that the surviving cells continue to follow logistic growth after each of the consecutive doxorubicin doses, as described by Eq. 7. Additionally, each of the i^{th} irreversibly damaged subpopulations are assumed to follow the growth dynamics defined in Eq. 8. Thus, for $i = 1, \ldots, n_d$, the dynamics of each irreversibly damaged subpopulation D^i is given by

$$\frac{dD^i}{dt} = \left(k_d + \left(g_d - k_d\right) \exp\left(-\gamma_d \left(t - t^i_{Dox}\right)\right)\right) D^i \left(1 - \frac{N}{\theta_{Dox}}\right) \tag{12}$$

Finally, we further investigate two versions of the multiple-dose model: 1) the general formulation outlined above in which we vary f_s and γ_d with the delivery of each dose, and 2) a simplified version in which we assume a constant parameterization for all doses (i.e., $f^1_s = f^2_s = \ldots = f^{n_d}_s$ and $\gamma^1_d = \gamma^2_d = \ldots = \gamma^{n_d}_d$). Our underlying hypothesis is that longer inter-treatment intervals require an adaptive parameterization because they contribute to the development of chemoresistance (Howard et al., 2022; Lyman, 2009; De Souza et al., 2011; Ponnusamy et al., 2017), which would be represented in our model by higher fractions of surviving cells (f^i_s) along with irreversibly damaged subpopulations (D^i) exhibiting longer transition times from proliferation to treatment-induced death (i.e., lower values of γ^i_d). Conversely, we hypothesize that shorter inter-treatment intervals may not introduce significant changes to the survival fractions (f^i_s) and death delay rates (γ^i_d) associated with each drug dose, such that a constant parameterization would suffice to capture the tumor cell population response to the cytotoxic action of the prescribed doxorubicin treatment. In the Results section, we show that these hypotheses are

significantly supported by the fitting of these two model versions to the data from Experiments 2 and 3 (Table 1).

Numerical methods

Model fitting

We fit the single-dose model to each individual time course of total tumor cell counts from each replicate from Experiment 1 (n = 60; Table 1), and we fit the multiple-dose model to each individual time course of total tumor cell counts from each replicate from Experiments 2 and 3 (n = 108 and 120, respectively; Table 1). Model fitting was carried out with a nonlinear least-squares method, *via* the MATLAB (R2020b) function *lsqnonlin*. We leveraged a trust-region reflective algorithm with function, step, and optimality tolerances of 10^{-6}, while the maximum number of function evaluations and iterations was set to 20,000. The parameter bounds and initial guesses were guided by the results from Howard et al. (2022), and are summarized in Supplementary Tables S2, S8, S12, S16, S17. The ordinary differential equations in our models were solved using a Runge-Kutta method as provided by *ode45* in MATLAB (R2020b). Supplementary Appendix B in the Supplementary Information provides further details about the model fitting approach used in this study.

Empirical parameter formulas

We constructed empirical formulas for the single-dose model parameters as a function of doxorubicin concentration based on the model fittings to the datasets from Experiment 1 (Table 1). To this end, we also applied a nonlinear least-squares method using a trust-region reflective algorithm provided by *lsqnonlin* in MATLAB (R2020b), as described in the previous section. The initial guess and bounds for the empirical parameters in these formulas were chosen according to the range of the single-dose model parameter values obtained from the fittings to the datasets from Experiment 1 (Table 1). The medians of the distributions of these fitted model parameters at each doxorubicin concentration were used as the observed values for the empirical parameter formula fits. The choice of the of the empirical formula for each parameter was based on the observed trend of the fitted parameter values obtained with the single-dose model as a function of increasing doxorubicin concentration (e.g., an exponentially decaying trend was represented with an exponential function; see the Results section and Supplementary Tables S6, S7 for further details).

Statistical analysis

To assess our model's quality of fit to the time course data, we calculated the coefficient of determination (R^2), the normalized root mean squared error (NRMSE), the Pearson correlation coefficient (PCC), and the concordance correlation coefficient (CCC) (Lin, 1989). In the Results section, we report the median and range of these metrics across all the replicates of each experiment (i.e., n = 60, 108, 60, and 60 for Experiment 1, Experiment 2, Experiment 3 with 2-day inter-treatment interval, and Experiment 3 with 2-week inter-treatment interval, respectively). More detailed values can be found in Supplementary Tables S4, S10, S14, S20. We further assessed our model parameterizations and fits to experimental data through 95% nonlinear regression parameter confidence intervals and 95% nonlinear regression prediction confidence intervals calculated using *nlparci* and *nlpredci* in MATLAB (R2020b), respectively (see Supplementary Appendix C). To test for significant differences between two values of a model parameter or quality-of-fit metric within each experimental scenario, we performed two-sided Wilcoxon rank sum tests with 5% significance using *ranksum* in MATLAB (R2020b).

To assess the validity of the proposed empirical formulas using the single-dose model fittings, we ran a simulation test in which we qualitatively compared the model outcomes based on these formulas with the corresponding experimental observations at each drug concentration. To this end, Latin hypercube sampling based on *lhsdesign* in MATLAB (R2020b) was used to define 200 parameter combinations assuming uniform distributions over the 95% confidence intervals of the fitted empirical parameter formulas at each doxorubicin concentration, as calculated by *nlpredci* in MATLAB (R2020b).

Results

Fitting the single-dose model to experiment 1 data: Varying doxorubicin concentrations

Figure 2 shows representative model fits for the observed growth of MCF-7 cell populations treated with only one dose of doxorubicin at concentrations ranging from 10 to 300 nM (Experiment 1, Table 1). Model fits for all replicates at each drug concentration (n = 6) can be found in Supplementary Figure S2. We report the median and range of all the fitted model parameters for each doxorubicin concentration in Supplementary Table S3, while Figure 3 shows the boxplots of the fitted parameter distributions for each doxorubicin concentration. The median and range of the quality of fit metrics for the single-dose model fits to Experiment 1 data (n = 60) were: NRMSE (3.33 [0.80, 12.20]), R^2 (>0.99 [0.96, >0.99]), PCC (>0.99 [0.98, >0.99]), and CCC (0.99 [0.97, >0.99]). Supplementary Table S4 further provides detailed quality of fit metrics for each doxorubicin concentration. Figure 2, Supplementary Figure S2 and Supplementary Table S3 show that, as doxorubicin concentration is increased, the surviving cells exhibit a decrease in growth rate and number,

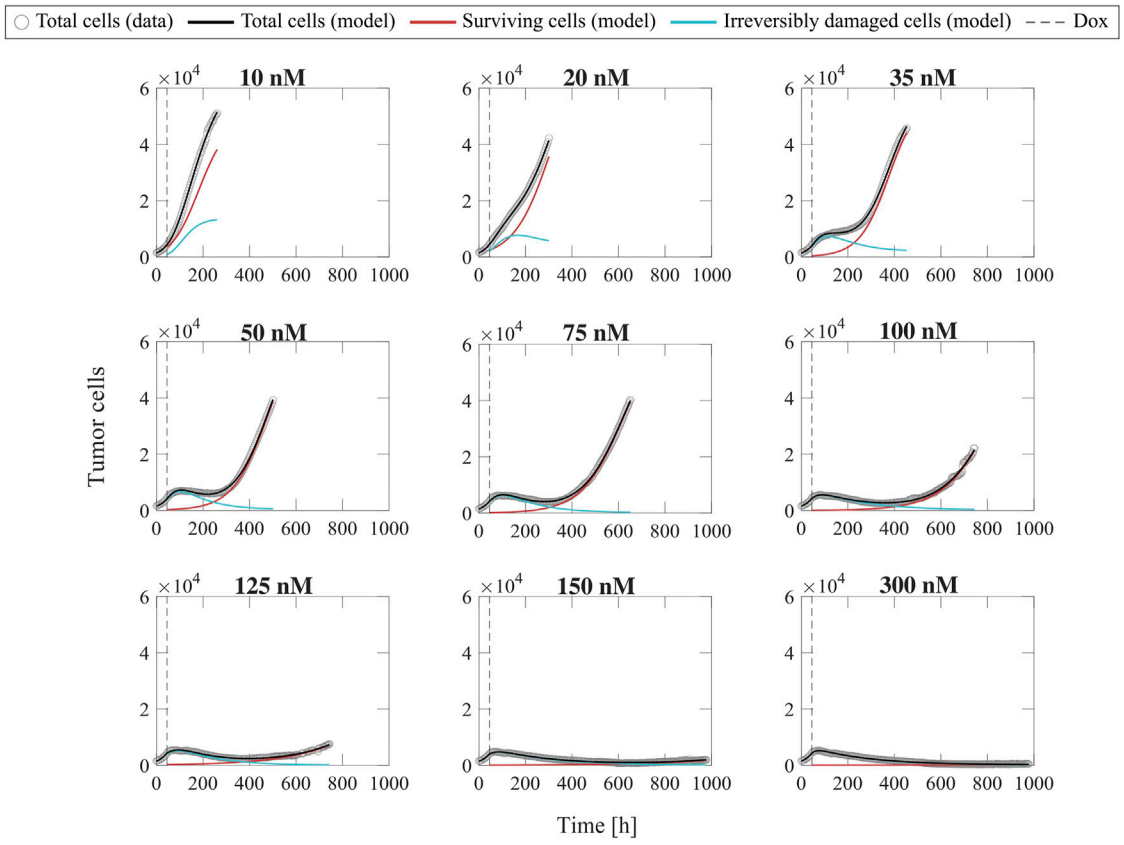


FIGURE 2
Representative fits of the single-dose model for varying concentrations of doxorubicin. Data and model fittings are shown for a representative replicate treated with 10–300 nM doxorubicin concentrations (Experiment 1, Table 1). Experimental data are shown in gray circles. The number of total cells, surviving cells, and irreversibly damaged cells obtained with the fitted single-dose model are shown in black, red, and blue solid lines, respectively. The time of doxorubicin delivery (Dox) is represented with a vertical grey dashed line. As doxorubicin concentration is increased, we observe a decrease in the growth rate of surviving cell subpopulation and a faster transition from growth to treatment-induced death in irreversibly damaged cell subpopulation. These drug-induced effects ultimately translate into a longer delay (or even suppression) of total tumor cell population growth post-treatment and lower total tumor cell count for higher doxorubicin concentrations, indicating superior tumor control overall. The median and range of the quality of fit metrics across all replicates in Experiment 1 (n = 60, Table 1) are NRMSE: 3.33 [0.80, 12.20], R^2: >0.99 [0.96, >0.99], PCC: >0.99 [0.98, >0.99], and CCC: 0.99 [0.97, >0.99].

while the irreversibly damaged cells undergo a faster transition from proliferation to treatment-induced death. These trends ultimately lead to significantly lower final total tumor cell counts ($p < 0.05$, see Supplementary Table S5) and larger delay or even suppression of tumor regrowth in the cells exposed to higher doxorubicin concentrations (see Supplementary Figure S2), suggesting that tumor control improves as the doxorubicin dose is increased.

Figure 3 shows the fitted empirical formulas for the fraction of surviving cells (f_s), the proliferation rate of the surviving cells (g_s), the proliferation rate of the irreversibly damaged cells (g_d), the doxorubicin-induced death rate of the irreversibly damaged cells (k_d), and the doxorubicin-induced death delay rate of irreversibly damaged cells (γ_d). These empirical formulas are functions of doxorubicin concentration, which is denoted with C. The fitted empirical parameter values and their confidence intervals can be found in Supplementary Table S6, while the corresponding quality of fit metrics can be found in Supplementary Table S7. For f_s, g_s, and g_d, we observe a clear exponentially decaying trend as drug concentration is increased (Figures 3A–C). In the case of f_s, we added an additional constant empirical parameter to the decaying exponential to ensure that the empirical formula captures the low nonzero values of this parameter for the higher doxorubicin concentrations (otherwise, the exponential decay would reach the horizontal asymptote at $f_s = 0$ for low doxorubicin concentrations). The parameter k_d exhibits a complex trend, consisting of a steep decreasing branch for doxorubicin concentrations under 50 nM, followed by an increasing branch that plateaus for doxorubicin concentrations over 150 nM. We found that an empirical formula based on a Morse-potential relationship (Girifalco and Weizer, 1959)

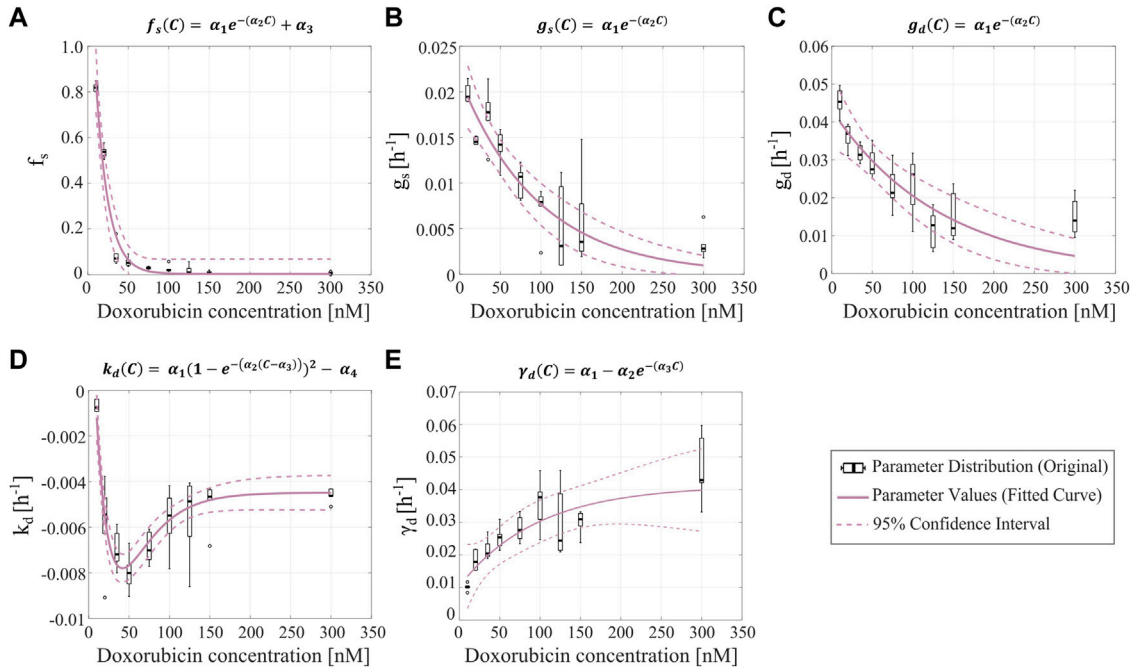


FIGURE 3
Empirical parameter formulas for varying doxorubicin concentrations. The proposed empirical formulas indicated at the top of each panel **(A–E)** were fit to the median of the corresponding parameter distributions obtained from fitting the single-dose model to the varying concentration datasets from Experiment 1 (Table 1). *C* denotes doxorubicin concentration in nM, while α_i ($i = 1, 2, \ldots$) are empirical parameters. The distributions of the single-dose model parameters are represented with black boxplots, in which outliers are represented as black circles. The resulting curves from fitting the empirical parameter formulas are shown as purple solid lines, and their corresponding 95% confidence intervals are plotted as purple dashed lines. Panel **(A)** shows the parameter formula for the fraction of surviving cells (f_s). Panel **(B)** shows the parameter formula for the proliferation rate of the surviving tumor cells (g_s). Panel **(C)** shows the parameter formula for the proliferation rate of the irreversibly damaged tumor cells (g_d). In panels **(A–C)**, we observe that as the drug concentration increases, the corresponding single-dose model parameter values decrease exponentially. Panel **(D)** shows the parameter formula for the doxorubicin-induced death rate of irreversibly damaged cells (k_d), which we approximated with an equation based on a Morse-potential relationship. Panel **(E)** shows the parameter formula for the doxorubicin-induced death delay rate of irreversibly damaged cells (γ_d), which increases and then plateaus as the drug concentration increases. Median and range of quality of fit metrics for the empirical parameter formulas ($n = 5$) are NRMSE: 17.65 [5.45, 153.6], R^2: 0.91 [0.78, 0.98], PCC: 0.95 [0.88, 0.99], and CCC: 0.85 [0.78, 0.88].

captured this trend (Figure 3D). For γ_d, we chose a decaying exponential flipped with respect to the horizontal axis to capture the increasing trend that ultimately plateaus at a nonzero value (Figure 3E). The median and range of the quality of fit metrics for the five proposed empirical formulas were: NRMSE (17.65 [5.45, 153.6]), R^2 (0.91 [0.78, 0.98]), PCC (0.95 [0.88, 0.99]), and CCC (0.85 [0.78, 0.88]). We note that the NRMSE of the fitted empirical formula for f_s reached values beyond 100%. This is due to the small values of f_s at high concentrations of doxorubicin, where the NRMSE is not relevant to modeling outcomes; for reference, the RMSE for the fitted empirical formula for f_s is 0.0476.

Once the parameter formulas had been established, we wanted to qualitatively assess the range of tumor cell population dynamics that our formulas could reproduce. For each doxorubicin concentration, we sampled the 95% confidence intervals of the fitted empirical parameter formulas (dashed purple lines in Figure 3) using Latin hypercube sampling to obtain 200 parameter combinations, with which we ran corresponding model simulations. Figure 4 presents the median and range of the model simulations plotted against the median and range of the experimental data measured at each time point for each doxorubicin concentration tested in Experiment 1 (Table 1). We observe that the proposed empirical parameter formulas (Figure 3) are able to predict a wide range of model solutions and that the simulations are able to capture the overall tumor cell population dynamics observed in the datasets from Experiment 1 (Table 1).

Fitting the multiple-dose model to experiment 2 data: Varying inter-treatment intervals

To fit the experimental data for varying inter-treatment intervals (Experiment 2, Table 1), we initially used the two versions of the multiple-dose model; i.e., with all parameters held constant or varying f_s and γ_d with each drug dose. Figure 5

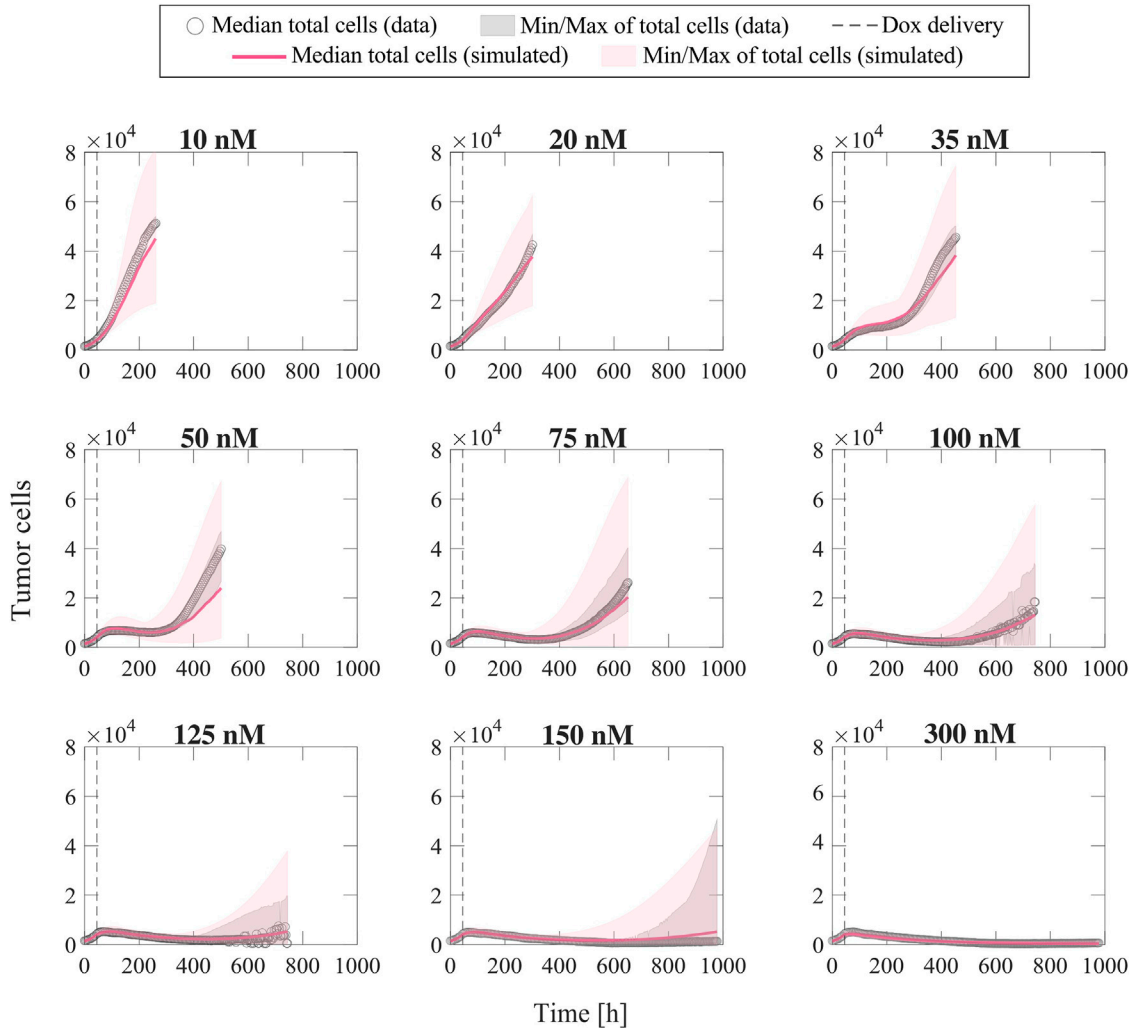


FIGURE 4
Comparison of simulated tumor cell population growth based on empirical parameter formulas with respect to experimental data for varying doxorubicin concentrations. We sampled the 95% confidence intervals for the fitted empirical parameter formulas in Figure 3 using Latin hypercube sampling to obtain 200 parameter combinations for each doxorubicin concentration, with which we carried out corresponding simulations with the single-dose model. The median and range of the model simulations are plotted with the median and range of the experimental data from Experiment 1 (Table 1) at each time point for comparison. The median of the experimental data is shown with gray circles, and the range of the experimental data is represented with gray shaded regions. The median of the model simulations is plotted as a pink solid line, and the range of the simulations is shown as pink shaded regions. The time of doxorubicin delivery is represented with a vertical grey dashed line. We observe that our fitted parameter formulas from Figure 3 can reproduce a wide range of tumor cell population dynamics, including those observed in the varying concentration datasets (Experiment 1, Table 1).

shows the distribution of the NRMSE in fitting the datasets at each inter-treatment interval (n = 12) for both models. We observe a significant difference between the NRMSEs obtained with either version of the multiple-dose model, such that the varying f_s and γ_d model provides a significantly lower NRMSE at 8-, 10-, 12-, 14-, and 16-day inter-treatment intervals (p: 0.0061, 0.0017, 1.56×10^{-4}, 5.92×10^{-4}, 3.66×10^{-5}, respectively). Thus, the results shown in Figure 5 justify the use of the model with constant parameters for inter-treatment intervals shorter than 8 days and the model with varying f_s and γ_d for inter-treatment intervals ≥8 days. We followed this model selection criterion for fitting the datasets from Experiments 2 and 3 (Table 1) for the remainder of this work.

Figure 6 shows representative model fits for the observed growth of MCF-7 cell populations treated with two doses of 75 nM doxorubicin delivered at inter-treatment intervals ranging from 0 to 16 days (Experiment 2, Table 1). Model fits for all the replicates at each inter-treatment interval (n = 12) can be found in Supplementary Figure S3. Additionally, Supplementary Table S9 summarizes the median and range of the fitted model

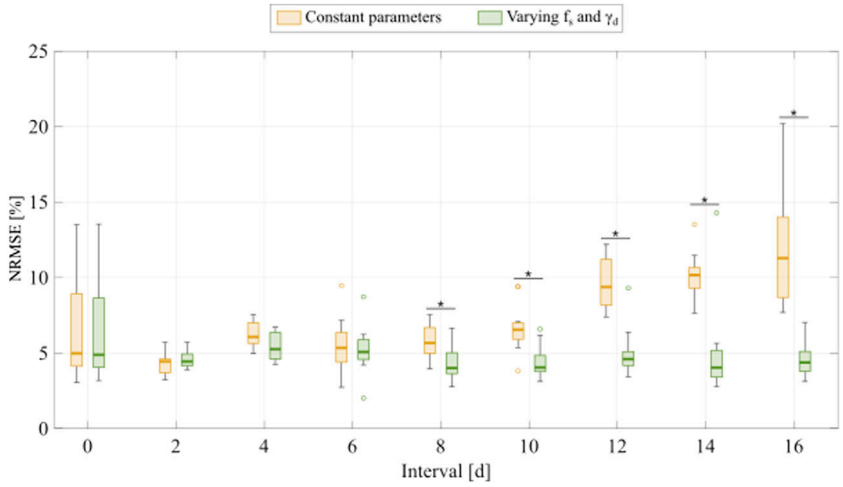


FIGURE 5
Comparison of fitting the experimental data for varying inter-treatment intervals with the multiple-dose model with constant versus varying parameters. For each inter-treatment interval tested in Experiment 2 (Table 1), we compared the normalized root mean squared error (NRMSE) calculated from the fittings using the multiple-dose model with constant parameters (yellow boxplots) with the NRMSE calculated from the fittings using the multiple-dose model with varying f_s and γ_d (green boxplots). Outliers are represented with circles. At inter-treatment intervals of 8, 10, 12, 14, and 16 days, there is a significantly lower NRMSE when the model with varying f_s and γ_d is used (p: 0.0061, 0.0017, 1.56×10^{-4}, 5.92×10^{-4}, 3.66×10^{-5}, respectively). An asterisk (*) indicates $p < 0.05$ (two-sided Wilcoxon rank sum test).

parameters for each inter-treatment interval. The median and range of the quality of fit metrics across all replicates in Experiment 2 (n = 108) were: NRMSE (4.64 [2.74, 14.3]), R^2 (0.99 [0.80, >0.99]), PCC (>0.99 [0.91, >0.99]), and CCC (0.99 [0.90, >0.99]). More detailed quality of fit metrics for each inter-treatment interval are reported in Supplementary Table S10. As the inter-treatment interval lengthens, the surviving cells tend to adopt an increasingly larger proliferation rate and the irreversibly damaged cells transition more slowly from proliferation to drug-induced death after two doses of doxorubicin treatment (see Figure 6, Supplementary Figure S3 and Supplementary Table S9). These effects appear to promote the tumor cell population regrowth after the second dose in most replicates for inter-treatment intervals of 6 days or longer and after the first dose for inter-treatment intervals of 12 days or longer. Overall, this ultimately leads to significantly higher final total tumor cell counts as the inter-treatment interval is lengthened ($p < 0.05$, see Supplementary Table S11), suggesting that increased time spans between consecutive doses of doxorubicin is conducive to poorer tumor control.

Additionally, Figure 7 shows the distributions of the fitted f_s and γ_d values from fitting the multiple-dose model to the data with varying inter-treatment intervals (Experiment 2, Table 1). When the model with constant parameters is used (inter-treatment intervals from 0 to 6 days), we observe a trend towards higher surviving fractions and delayed transitions to treatment-induced death in irreversibly damaged cells as the two doses are further spaced in time. This observation is further supported by the distributions of varying f_s and γ_d obtained from fitting the multiple-dose model to the data for inter-treatment intervals from 8 to 16 days. After the second dose in each of these longer intervals, the surviving fraction significantly increases and the transition from proliferation to treatment-induced death in irreversibly damaged cells significantly slows ($p < 0.05$, see Figure 7), thereby suggesting an enhanced chemoresistance in both tumor cell subpopulations for longer inter-treatment intervals. Additionally, comparing the distributions of parameter f_s^2 obtained for the different inter-treatment intervals considered in Experiment 2, the values obtained in the 0-day and 2-day cases are significantly lower than those obtained in any larger intervals, the f_s^2 values obtained in the 4-day and 6-day scenarios are significantly lower than those obtained for any inter-treatment interval larger or equal to 8 days, and the f_s^2 values obtained in the 8-day case are significantly lower than those obtained for the 12-day inter-treatment interval ($p < 0.05$; see Supplementary Appendix D for further detail). Likewise, for parameter γ_d^2, the values obtained for the 0-day, 2-day, 4-day, 6-day, and 8-day cases are significantly larger than those obtained for any inter-treatment interval larger or equal to 6, 4, 6, 10, and 12 days, respectively ($p < 0.05$; see Supplementary Appendix D). Furthermore, the γ_d^2 values obtained for the 10-day inter-treatment interval are significantly larger than those obtained for the 12-day and 14-day intervals, and the γ_d^2 values obtained for the 14-day inter-treatment interval are significantly lower than those obtained for the 16-day interval ($p < 0.05$; see Supplementary Appendix D).

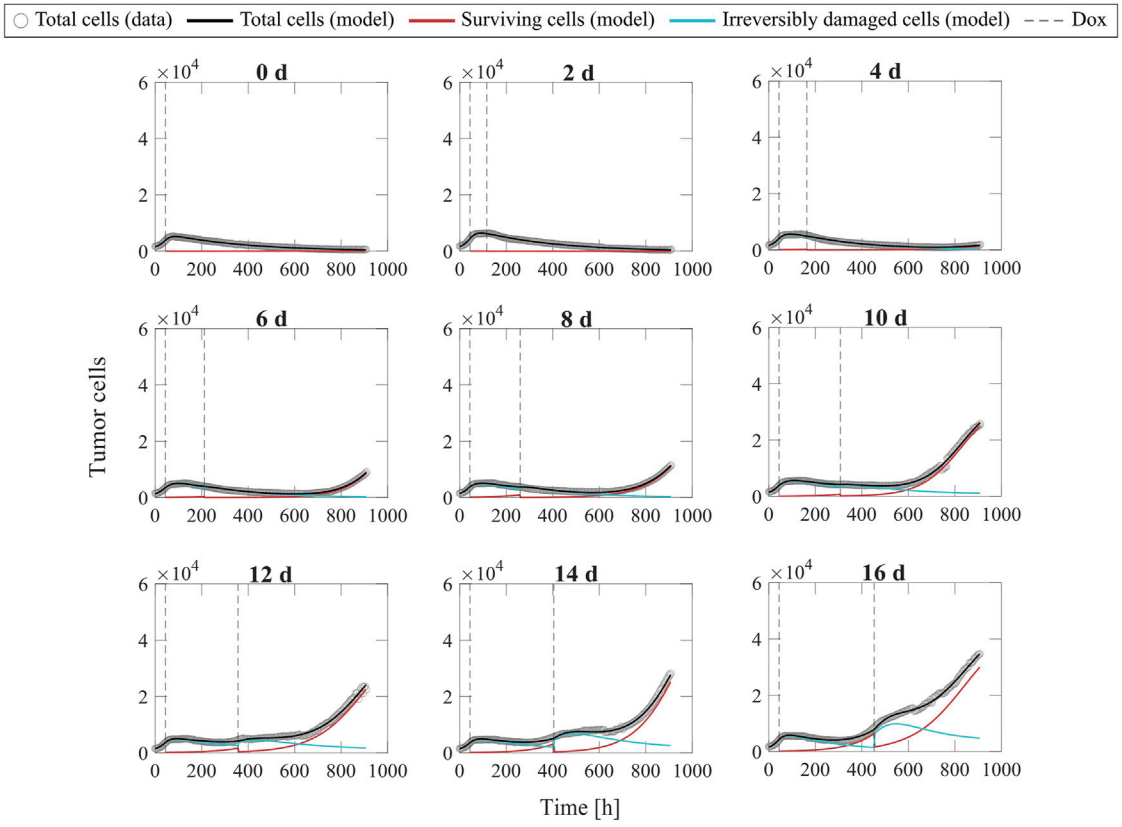


FIGURE 6
Representative fits of the multiple-dose model for varying inter-treatment intervals. Data and model fittings are shown for a representative replicate exposed to two doses of 75 nM doxorubicin delivered at inter-treatment intervals ranging from 0 to 16 days (Experiment 2, Table 1). Experimental data are shown in gray circles. The number of total cells, surviving cells, and irreversibly damaged cells obtained with the fitted multiple-dose model are shown in black, red, and blue solid lines, respectively. The times of doxorubicin (Dox) delivery are represented with vertical grey dashed lines. For the 0-day case, a single line represents a continuous treatment with no interval between the doses. For inter-treatment intervals of 0–6 days, the multiple-dose model with constant parameters was used for data fitting. For inter-treatment intervals of 8–16 days, we used the multiple-dose model with varying f_s and γ_d. As the inter-treatment interval is lengthened, we observe an increase in the proliferation rate of the surviving cells and a slower transition from proliferation to treatment-induced death in the irreversibly damaged cells. These drug-induced effects ultimately lead to a tumor cell population relapse after the second dose for inter-treatment intervals of 6 days or longer in most replicates, as well as tumor cell population regrowth after the first dose for inter-treatment intervals of 12 days or longer. These observations suggest increasingly poor tumor control as the two doses of 75 nM of doxorubicin are spaced further out in time. The median and range of the quality of fit metrics across all datasets in Experiment 2 (n = 108, Table 1) are NRMSE: 4.64 [2.74, 14.3], R^2: 0.99 [0.80, >0.99], PCC: >0.99 [0.91, >0.99], and CCC: 0.99 [0.90, >0.99].

Thus, these results along with the distributions plotted in Figure 7 show that there is a tendency towards a larger f_s^2 value for larger inter-treatment intervals, which is suggestive of increased chemoresistance in the surviving cell compartment. However, this trend becomes less clear among the longest intervals considered in Experiment 2 (i.e., >8 days), for which f_s^2 appears to plateau at a value between 0.10 and 0.15. Additionally, our results also show a decreasing trend in the values of γ_d^2 between the 0-day and the 14-day inter-treatment interval scenario, which suggests an increase in the chemoresistance of the irreversibly-damaged cells (i.e., they transition more slowly from proliferation to drug-induced cell death); although this tendency is reverted for the 16-day inter-treatment interval. Moreover, the distributions shown in Figure 7 further support the use of the multiple-dose model with varying f_s and γ_d for longer inter-treatment intervals.

Fitting the multiple-dose model to experiment 3 data: Varying number of doses

Figure 8 shows representative model fits for the observed growth of MCF-7 cell populations treated with 1–5 doses of 75 nM doxorubicin delivered at either 2-day or 2-week inter-treatment intervals (Experiment 3, Table 1). The datasets from the cells treated with a 2-day inter-treatment interval were fitted with the multiple-dose model with constant parameters, while

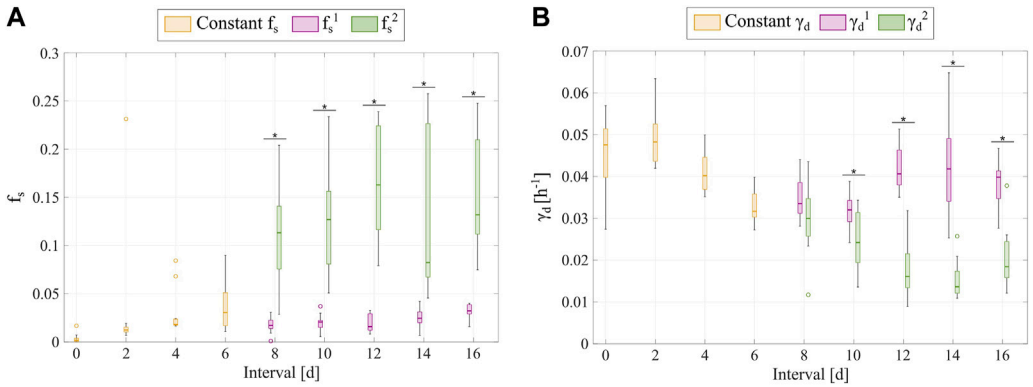


FIGURE 7
Comparison of the f_s and γ_d distributions obtained from fitting the multiple-dose model to the experimental data for varying inter-treatment intervals. The parameter distributions are represented as boxplots and were obtained from fitting the multiple-dose model to the varying inter-treatment interval datasets from Experiment 2 (Table 1). Outliers are represented with circles. Panel **(A)** shows the distributions for the fraction of surviving cells (f_s). Panel **(B)** shows the distributions for the doxorubicin-induced death delay rate of the irreversibly damaged tumor cells (γ_d). For 0–6 days inter-treatment intervals, f_s and γ_d are kept constant in the model (yellow boxplots); whereas, for 8–16 days inter-treatment intervals, we vary f_s and $\gamma_d$with each doxorubicin dose (f_s^1, γ_d^1: purple boxplots, f_s^2, γ_d^2: green boxplots). As the inter-treatment interval is lengthened from 0 to 6 days, the constant f_s and γ_d show a trend towards higher surviving fractions and slower transitions to doxorubicin-induced death, suggesting increasingly poorer tumor control. When f_s and γ_d are varied with each dose, we observe that the second f_s values correspond to significantly higher surviving fractions for 8-, 10-, 12-, 14-, and 16- day inter-treatment intervals (p: 4.7×10^{-5}, 3.7×10^{-5}, 3.7×10^{-5}, 3.7×10^{-5}, and 3.7×10^{-5}, respectively) and that the second γ_d values represent significantly slower transitions to treatment-induced death for 10-, 12-, 14-, and 16- day inter-treatment intervals (p: 0.0141, 3.7×10^{-5}, 6.0×10^{-5}, and 9.7×10^{-5}, respectively). These changes in f_s and γ_d after the second dose also suggest an increasingly poorer tumor control after the second dose with a longer inter-treatment interval. An asterisk (*) indicates $p < 0.05$ in two-sided Wilcoxon rank sum tests comparing the distributions of the two f_s and γ_d values obtained for each inter-treatment interval where the multiple-dose model with varying parameters was used.

the datasets from the cells treated with a 2-week inter-treatment interval were fitted with the multiple-dose model with varying f_r and γ_d. Model fits for all the replicates for each number of doses and both inter-treatment intervals (n = 12) can be found in Supplementary Figures S4, S5. The median and range of the fitted model parameters for each dose number are summarized in Supplementary Tables S13, S18, S19. For the replicates treated every 2 days, the median and range of the quality of fit metrics (n = 60) were: NRMSE (12.2 [2.72, 19.1]), R^2 (0.99 [0.87, >0.99]), PCC (0.99 [0.93, >0.99]), and CCC (0.99 [0.93, >0.99]). Likewise, for the replicates treated every 2 weeks, the median and range of the quality of fit metrics (n = 60) were: NRMSE (3.21 [1.91, 8.57]), R^2 (>0.99 [0.93, >0.99]), PCC (>0.99 [0.97, >0.99]), CCC (0.99 [0.96, >0.99]). More detailed quality of fit metrics for each number of doses and both inter-treatment intervals can be found in Supplementary Tables S14, S20.

The model fittings plotted in Figure 8 and Supplementary Figures S4, S5 show that increasing the number of doses contributed to improved tumor control for the two inter-treatment intervals investigated in this work. In general, for the cells treated every 2 days, we observed significantly lower final total tumor cell counts as the number of doses was increased ($p < 0.05$, see Supplementary Table S15). Furthermore, delivering two or more doses effectively suppressed tumor growth at the end of the experiment, typically showing a decreasing branch in the total tumor cell count right after the first dose. When the inter-treatment interval was extended to 2 weeks, delivering more than one dose of doxorubicin also contributed to limited tumor cell growth ($p < 0.05$, see Supplementary Table S21); however, most of the replicates showed an increasing trend in total tumor cell count over the experiment duration. Thus, with a 2-week inter-treatment interval, an increasing number of doses can decelerate tumor cell growth, but it cannot suppress it as observed with a 2-day inter-treatment interval. Furthermore, the model fitting results reported in Figure 8, Supplementary Figures S4, S5 and Supplementary Tables S3, S18, S19 show that, as the inter-treatment interval is lengthened from 2 days to 2 weeks, the surviving cells exhibit a larger proliferation rate, while the irreversibly damaged cells emerging after the second and subsequent doses undergo a slower transition to treatment-induced death. These effects, induced by the lengthened inter-treatment interval, contribute to explaining the superior tumor control in the 2-day experiments and align with the corresponding results shown in Figure 6, Supplementary Figure S3 and Supplementary Table S9.

We further investigated tumor cell dynamics for the Experiment 3 data with 2-week inter-treatment intervals by analyzing the evolving distributions of parameters f_s and γ_d, which are shown in Figure 9. We observe that the surviving fraction corresponding to the first to fourth doses (f_s^1, f_s^2, f_s^3, f_s^4) shows an increasing trend, which is indicative of progressive chemoresistance during treatment and aligns with the

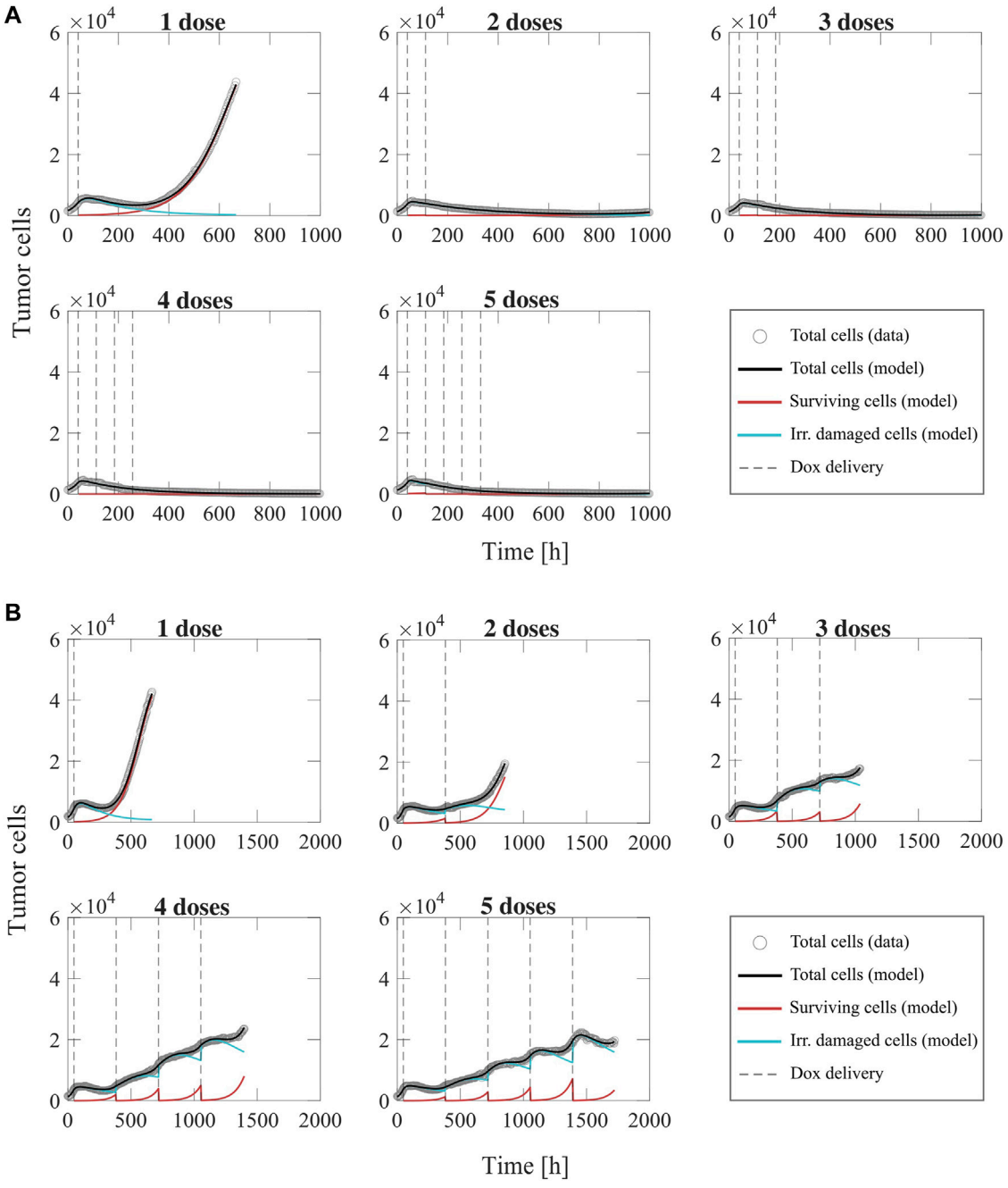


FIGURE 8
Representative fits of the multiple-dose model for a varying number of doxorubicin doses. Data and model results are shown for a representative replicate treated with 1–5 doses of 75 nM doxorubicin delivered at either 2-day or 2-week inter-treatment intervals (Experiment 3, Table 1). Experimental data are shown in gray circles. The number of total cells, surviving cells, and irreversibly damaged cells obtained with the fitted multiple-dose model are shown in black, red, and blue solid lines, respectively. The times at which doxorubicin (Dox) is delivered are represented with vertical grey dashed lines. Panel **(A)** shows fittings for 1 to 5 doxorubicin doses delivered at 2-day inter-treatment intervals obtained with the model with constant parameters. The median and range of the quality of fit metrics across all replicates for this Experiment 3 subgroup (n = 60, Table 1) are NRMSE: 12.2 [2.72, 19.1], R^2: 0.99 [0.87, >0.99], PCC: 0.99 [0.93, >0.99], CCC: 0.99 [0.93, >0.99]. Panel **(B)** shows fittings for 1 to 5 doxorubicin doses delivered at 2-week inter-treatment intervals obtained with the model with varying f_s and γ_d. The median and range of the quality of fit metrics across all replicates for this Experiment 3 subgroup (n = 60, Table 1) are NRMSE: 3.21 [1.91, 8.57], R^2: >0.99 [0.93, >0.99], PCC: >0.99 [0.97, >0.99], CCC: 0.99 [0.96, >0.99]. Overall, we observe that there is superior tumor control with an increased number of doses, which is further improved when the doses are delivered at shorter inter-treatment intervals. As the inter-treatment interval is lengthened from 2 days to 2 weeks, we observe that the population growth rate and number of the surviving cells increase, while the irreversibly damaged cells exhibit a slower transition from proliferation to treatment-induced death.

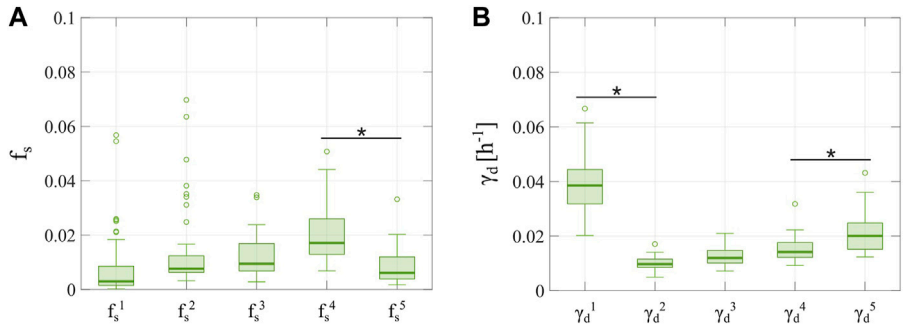


FIGURE 9
Distributions of f_s and γ_d obtained from fitting the multiple-dose model to the experimental data for a varying number of doses with a 2-week inter-treatment interval. The parameter distributions are represented as boxplots and were obtained from fitting the multiple-dose model to the 2-week inter-treatment interval datasets from Experiment 3 (Table 1), in which the model with varying f_s and γ_d was used. Outliers are represented as circles. Panel **(A)** shows the distributions of the fraction of surviving cells, such that a new value for $f_s$is defined for each drug dose ($f_s^1, f_s^2, \ldots, f_s^5$). We observe an increasing trend in the first four f_s parameters, which suggests an increasing chemoresistance with each dose. However, f_s^5 takes on significantly lower values than f_s^4 ($p = 0.0015$), which suggests that adding more doses may limit the trend towards chemoresistance. Panel **(B)** shows the distributions of the doxorubicin-induced death delay rate of irreversibly damaged cells, such that a new value of γ_d is defined for each drug dose ($\gamma_d^1, \gamma_d^2, \ldots, \gamma_d^5$). The values for γ_d^2 are significantly lower than those of γ_d^1 ($p = 5.6 \times 10^{-19}$). Hence, the second irreversibly damaged subpopulation shows a slower transition to treatment-induced death. However, the subsequent doxorubicin doses induce irreversibly damaged subpopulations exhibiting an increasing γ_d, with γ_d^5 being significantly larger than γ_d^4 ($p = 0.0093$). This observation further suggests that past a certain number of doses, initial chemoresistance appears to be reverted. An asterisk (*) indicates $p < 0.05$ (two-sided Wilcoxon rank sum test).

corresponding results shown in Figure 7. However, the fitted values for f_s^5 are significantly lower than the value obtained for f_s^4 ($p = 0.0015$). Additionally, we observe that the values for γ_d^2 are significantly lower than that of γ_d^1 ($p = 5.6 \times 10^{-19}$), following the trend observed in Figure 7 for the data from Experiment 2. However, the values for γ_d^2, γ_d^3, γ_d^4, and γ_d^5 exhibit an increasing trend, with γ_d^5 being significantly larger than γ_d^4 ($p = 0.0093$). These changes in f_s and γ_d suggest that delivering multiple doses of doxorubicin may progressively limit or even revert the chemoresistance observed in the initial surviving and irreversibly damaged subpopulations.

Discussion

We have presented a mathematical framework to describe the therapeutic response of MCF-7 breast cancer cells to treatment with doxorubicin and the development of chemoresistance in the *in vitro* setting. Our mathematical models rely on a compartmentalization of the tumor cell population after the delivery of each drug dose into either surviving cells or irreversibly damaged cells, the latter of which ultimately die due to the cytotoxic action of doxorubicin (Anampa et al., 2015; Howard et al., 2022; Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017). With this dichotomy, we aim to capture the underlying diverse spectrum of drug sensitivities in the tumor cell population, as well as its changes after the delivery of subsequent doxorubicin doses considering different inter-treatment interval lengths (Howard et al., 2022; Pisco et al., 2013; Álvarez-Arenas et al., 2019). We presented a single-dose model that can be extended to a multiple-dose model, in which parameterization can vary with each dose. We fitted our models to various time-resolved microscopy datasets, which enabled us to evaluate tumor cell population dynamics with our models in three experimental scenarios that varied either the doxorubicin concentration, the inter-treatment interval, or the number of doses (see Table 1). In all three cases, our models recapitulated the experimental observations, achieving a remarkable quality of fit.

In Experiment 1 (Table 1), we evaluated the effect of a single dose of doxorubicin on MCF-7 breast cancer cell population growth and we found that tumor control was significantly improved with increased drug concentration ($p < 0.05$, see Supplementary Table S5). Our single-dose model showed that, at a subpopulation level, these dynamics emerged from a lower proliferation rate of surviving cells and a faster transition from proliferation to treatment-induced death in irreversibly damaged cells. The dynamics observed in our varying concentration experiment have also been reported in other studies of doxorubicin effects on breast cancer cell lines, both as monotherapy and in combination with other therapeutic agents (McKenna et al., 2017; Zoli et al., 1995; Czeczuga-Semeniuk et al., 2004).

We used the parameter distributions obtained from our single-dose model fits to the varying drug concentration datasets to empirically fit various parameter formulas as functions of doxorubicin concentration, as shown in Figure 3. The model simulations generated from our proposed empirical parameter formulas were able to capture a spectrum of model solutions that encompass the dynamics observed in our data

from Experiment 1 (see Figure 4). We observed clear exponentially decaying trends for the fraction of surviving cells (f_s) and the proliferation rates of surviving and irreversibly damaged tumor cells (g_s and g_d, respectively) as doxorubicin concentration increases (see Figures 3A–C). These trends seem to capture the growth-inhibition effect of doxorubicin as well as the dose-response curve for this drug within our mechanistic modeling framework, in which doxorubicin efficacy has been observed to plateau at high concentrations (El-Kareh and Secomb, 2005; Neale et al., 2000). The distributions of the doxorubicin-induced death rate in the irreversibly damaged cells (k_d) exhibited a non-monotonic trend as doxorubicin concentration was varied, which we approximated with a Morse-potential relationship (Girifalco and Weizer, 1959). This result was counterintuitive, as we had initially anticipated a strictly decreasing trend in k_d for higher doxorubicin doses, which would indicate an increasingly more intense effect of treatment-induced death. However, the cytotoxic action of doxorubicin (Gewirtz, 1999; Carvalho et al., 2009; McKenna et al., 2017) also induces cell cycle arrest. The interplay between these two drug-induced effects may ultimately lead to nonlinear tumor cell responses, such as the one captured by the empirical formula for k_d proposed in this work. The relative participation of cell death and cell cycle arrest in the overall doxorubicin action on breast cancer cells may follow more complex dynamics that are not fully captured by our models, and thus requires further investigation. Indeed, to refine the description of these doxorubicin effects, our models could also be extended to account for the dynamics of doxorubicin uptake and binding (McKenna et al., 2017); although this would require additional data reporting on those phenomena. From a modeling point of view, previous studies have leveraged other formulations to represent cytotoxic drug action (e.g., an exponential decay, alternative transition terms from proliferation to drug-induced death) (Lorenzo et al., 2022; Howard et al., 2022; Colli et al., 2021) which could be explored with our modeling framework in a model selection study (Lorenzo et al., 2022) to assess the optimal approach to capture the cytotoxic action of doxorubicin.

Experiment 2 involved the delivery of two doses of doxorubicin to each replicate of MCF-7 breast cancer cells at varying-inter-treatment intervals ranging from 0 to 16 days (Table 1). We fit two versions of our multiple-dose model to these datasets: either with constant parameters or with f_s and γ_d varied at each drug dose. The model with constant parameters sufficed to describe the observed tumor cell population dynamics for inter-treatment intervals from 0 to 6 days, while the model with varying f_s and γ_d was superior for inter-treatment intervals from 8 to 16 days ($p < 0.05$, see Figure 5). For two consecutive doses of doxorubicin delivered at varying inter-treatment intervals, our results showed significantly poorer tumor control with longer inter-treatment intervals ($p < 0.05$, see Supplementary Table S11). In the fittings from the model with varying f_s and γ_d, we observed that the second dose induced a significantly larger f_s^2 and a lower γ_d^2 compared to the corresponding values of f_s^1 and γ_d^1 ($p < 0.05$, see Figure 7), further supporting the adoption of an adaptive model parameterization for inter-treatment intervals from 8 to 16 days. Additionally, comparing the distributions of f_s^2 and γ_d^2 obtained for the different inter-treatment intervals, we observed a tendency towards higher survival fraction and a slower death delay rate for higher inter-treatment intervals. Nevertheless, f_s^2 appears to plateau in long inter-treatment intervals and the general trend observed for γ_d^2 is reverted in the 16-day interval scenario. Thus, the changes observed in f_s and γ_d in Experiment 2 suggest that longer inter-treatment intervals contribute to the development of chemoresistance in both tumor cell subpopulations in our model; although our results also suggest to further investigate whether the trends in f_s^2 and γ_d^2 are reverted for inter-treatment intervals larger than 16 days. From a biological perspective, long inter-treatment intervals may allow cancer cells to acquire chemoresistance through processes like treatment-induced mutations, altered epigenetics, and phenotype switching, which ultimately limit the efficacy of the second dose and may lead to tumor regrowth (Easwaran et al., 2014; Zhao, 2016; Dong et al., 2018; Ji et al., 2019; Meacham and Morrison, 2013; Echeverria et al., 2019; Kumar et al., 2019). This phenomenon has been observed in preclinical studies (Lyman, 2009; De Souza et al., 2011; Ponnusamy et al., 2017), but the trends are less clear in the clinical setting (Lyman, 2009; Richards et al., 1992; Citron et al., 2003; Untch et al., 2009; Foukakis et al., 2016).

In Experiment 3, we treated MCF-7 breast cancer cells with multiple doses of doxorubicin at either 2-day or 2-week inter-treatment intervals (Table 1). We observed significantly improved tumor control with an increased number of doses delivered at a 2-day inter-treatment interval ($p < 0.05$, see Supplementary Table S15), with tumor cell population growth effectively suppressed after two or more doxorubicin doses. When the treatment interval was extended to 2 weeks, tumor cell population growth was significantly decelerated ($p < 0.05$, see Supplementary Table S21) but not suppressed, aligning with our previous conclusions that longer inter-treatment intervals may promote chemoresistance. Moreover, these results underscore that, in comparison to the total number of doses, it is the treatment interval that holds a critical impact on determining overall tumor control. Indeed, as most patients receive chemotherapy treatments delivered every 1–3 weeks, our results point to the clinical importance of optimizing treatment interval in designing effective drug regimens (Jarrett et al., 2020a; Lyman, 2009; Richards et al., 1992; Citron et al., 2003; Untch et al., 2009; Foukakis et al., 2016). Additionally, the evolving distributions for the varying f_s and γ_d from the model fits to the 2-week inter-treatment interval datasets (see Figure 9) exhibit trends that potentially explain the relationship between the number of doses and the resulting chemoresistance

dynamics. For f_s, the initially increasing trend for the first four doses ($f_s^1, f_s^2, f_s^3, f_s^4$), suggests a progressive increase in chemoresistance with each dose. However, the values of f_s^5 were significantly lower than those of f_s^4 ($p = 0.0015$), potentially indicating that increasing the number of doses may ultimately hinder chemoresistance. This is further corroborated by the trends for γ_d, in which γ_d^2 drops significantly with respect to γ_d^1 ($p = 5.6 \times 10^{-19}$), but γ_d^2, γ_d^3, γ_d^4, and γ_d^5 exhibit an increasing trend. This result suggests that further doses of doxorubicin can promote increasingly faster transitions to treatment-induced death, thus reverting the initial chemoresistance observed in the irreversibly damaged subpopulation. We do note that because we have only tested up to five doses of doxorubicin, further studies with a larger number of doses would be needed to further probe these trends.

Although our work presents promising insights into the mechanisms of chemoresistance, this study does have its limitations. First, we used a limited number of replicates within the scenarios explored in each experiment (n = 6 or 12, see Table 1). Since we do not observe uniform tumor cell populations dynamics across all replicates within each scenario, we would like to re-assess the observations in this study over a larger experimental setup, for example involving a higher number of replicates exposed to more diverse combinations of drug concentration, inter-treatment interval, and number of drug doses. This would enable us to investigate whether these observations are from doxorubicin effects altering tumor cell dynamics or whether the experimental conditions influence the development of a representative distribution of drug-resistant and drug-sensitive cells (e.g., ~2,000 seeded cells/well might potentially limit the emergence of a resistant subpopulation, which may skew the observed response to treatment). Second, we also acknowledge the general limitations of extrapolating from *in vitro* systems to tumors in patients (Katt et al., 2016), as cell lines do not capture the unique, heterogeneous nature of each patient's tumor. To address this limitation, we plan to evaluate our models on clinically-relevant breast cancer cells other than MCF-7 cells (ER + breast cancer), such as the BT-474 (ER + HER2+ breast cancer) and MD-MBA-231 (triple-negative breast cancer) considered by Howard et al. (2022). Third, our cells were grown in monolayers, which are not representative of the three-dimensional tumor geometry *in vivo*. However, our mathematical models could be made readily applicable to tumor cell spheroid data. In particular, our models could be extended to a set of partial differential equations, accounting for tumor cell mobility and spatially-resolved parameters and variables, which would allow for a spatiotemporal description of spheroid growth in both *in vitro* and *in vivo* settings (Jarrett et al., 2020a; Kazerouni et al., 2020). Indeed, these extended models could incorporate other spatially-varying mechanisms beyond tumor cell dynamics, such as drug diffusion, mechanics, and angiogenesis, which have also been recognized as key components of chemoresistance and drug action (Lankelma et al., 1999; Mascheroni et al., 2017; Yonucu et al., 2017; Jarrett et al., 2020b; Kazerouni et al., 2020). Fourth, given that our model requires a moderate number of parameters that may increase with the number of delivered doses of doxorubicin, their estimation from specific experimental data may exhibit a certain degree of uncertainty (see Supplementary Appendix C). Thus, future studies should investigate whether and how the levels of uncertainty obtained for the parameters of our models affect the description of the therapeutic action of doxorubicin on tumor cells, for example, by leveraging a robust Bayesian framework (Lorenzo et al., 2022). Fifth, we only analyzed a constant versus an adaptive parameterization for the surviving fraction (f_s) and the death delay rate (γ_d) because we hypothesized that these would suffice to account for the development of chemoresistance. While this choice was supported by the results presented herein, an uncertainty quantification approach could also be exploited to conduct a model selection study aiming to investigate the optimal combination of constant and adaptive parameters in Experiments 2 and 3 (Lorenzo et al., 2022), which may provide new insights in the development of chemoresistance to doxorubicin. Furthermore, while experimental observations and modeling results in our study support the adoption of a fixed value of θ_{Dox} after treatment, the aforementioned modeling selection analysis could also be extended to investigate whether the change in the carrying capacity after treatment (i.e., from θ_u to θ_{Dox}) is permanent or temporal; although this analysis most likely requires additional experiments and data types to investigate the biological mechanisms underlying either of these two modeling alternatives (e.g., doxorubicin-induced changes in cell size or genetic alterations). Finally, we acknowledge the limitations in modeling tumor cell subpopulation dynamics with total tumor cell data, and that our study thus lacks methods for specifically validating the proposed mathematical description of surviving and irreversibly damaged tumor cell dynamics. This issue could potentially be addressed by incorporating methods to trace cell lineage, which would enable the collection of time-resolved measurements of the therapeutic response of diverse drug-sensitive and drug-resistant phenotypes in the tumor cell population. For instance, Al'Khafaji et al. (2018) have developed a functionalized lineage tracing tool to track both cell lineages and direct lineage-specific gene expression using barcoded gRNAs. Then, fitting these data to an extension of our model to a multicompartment formulation describing the dynamics of the various detected drug sensitivity phenotypes could provide a more precise insight into the dose-dependent response (including refined parameter empirical formulas) and how timing and the number of doses mediate the global response of the tumor cell population.

In future studies, we intend to explore a refinement of our model to account for the mechanisms underlying the trends observed in the empirical parameter formulas from this work, which will help us further understand doxorubicin effects.

Additionally, we plan to extend the construction of these empirical formulas over the three-dimensional space spanned by the dosage, the inter-treatment interval, and the number of doses by leveraging a larger collection of replicate datasets exhibiting variations across those three treatment regimen variables. This experimental campaign would expand Experiments 2 and 3 in our study beyond a fixed drug concentration of 75 nM per dose, and an inter-treatment interval of 2 days or 2 weeks in Experiment 3 (see Table 1). We hypothesize that nonlinear dependencies will govern the relationship between the three regimen variables and the model parameters, thereby enabling the capture of a diverse spectrum of therapeutic responses to doxorubicin that we already observed in Experiments 2 and 3 (see Supplementary Figures S3–S5). Hence, the resulting three-dimensional empirical formulas could allow for predicting the outcome of any doxorubicin treatment regimen *a priori* (i.e., before running the corresponding experiment) by just selecting the treatment schedule (i.e., dosage, inter-treatment interval, and number of doses), but this would require previous validation by leveraging a different collection of experimental datasets than the one used to construct the three-dimensional empirical formulas. Further experimentally informed studies with our mechanistic models could also contribute to identifying the optimal timing and frequency for doxorubicin delivery in preclinical scenarios. Indeed, we would like to explore optimal control theory (Jarrett et al., 2020b; Colli et al., 2021) *in vitro* through heterogeneous multiclonal cultures to identify optimal treatment combinations of doxorubicin concentration, treatment interval, and number of doses.

High-throughput, time-resolved microscopy *in vitro* systems enable the collection of vast amounts of time-resolved data on the dynamics of tumor cell populations across multiple, diverse scenarios. We (and others) (Zhang et al., 2015; Strobl et al., 2021; Howard et al., 2022; Pisco et al., 2013; Álvarez-Arenas et al., 2019; McKenna et al., 2017; Kazerouni et al., 2020) posit that these time-resolved datasets can be integrated in mathematical models of tumor cell population dynamics to systematically investigate the effects of drugs on tumor cells across a much broader variety of regimens than are possible to test *in vivo*. Then, our ultimate goal is to exploit the knowledge gained from a model constructed and validated in a data-rich *in vitro* preclinical environment (e.g., where hundreds of data points are available for each replicate) to refine mathematical models and their predictions of therapeutic response in a data-poor *in vivo* clinical environment (i.e., where less than five data points may be available for each patient) (Lorenzo et al., 2022; Weis et al., 2015; Jarrett et al., 2018; Jarrett et al., 2020a). In particular, we think that the mechanistic insights provided by the models and empirical formulas proposed in this study could be leveraged to identify the minimal dose range required to effectively inhibit breast cancer growth *in vivo* and achieve optimal tumor control, both of which are of great clinical interest (Carvalho et al., 2009; Jarrett et al., 2020b; Harahap et al., 2020; Chan et al., 1999). Thus, we believe that the complex dynamics underlying the dose-dependent effect of doxorubicin deserve further research coupling extensive experiments with mechanistic modeling.

Conclusion

We have developed a biologically-based, mathematical model of MCF-7 breast cancer cell response to the cytotoxic action of doxorubicin accounting for the development of chemoresistance, which significantly extends the experimentally-informed mechanistic models by Howard et al. (2022). To this end, we proposed a modeling framework that can accommodate multiple doxorubicin doses as well as an adaptive parameterization with each drug dose. We show that model fittings to longitudinal, time-resolved microscopy data of MCF-7 breast cancer cells could remarkably recapitulate the observed tumor cell population dynamics for all experimental scenarios varying in either drug concentration, inter-treatment interval, or number of doses. We also propose empirical formulas that describe model parameters as functions of doxorubicin concentration, which could contribute to refining our mechanistic model and further our understanding of doxorubicin action. We report significantly improved tumor control with higher doxorubicin concentrations, shorter inter-treatment intervals, and a higher number of doses. We also observe that longer inter-treatment intervals potentially promote chemoresistance, which manifests as higher surviving fractions and delayed transitions to treatment-induced death in irreversibly damaged subpopulations. Our findings show promise in furthering our understanding of doxorubicin action and chemoresistance progression, while also representing a step towards systematically exploring optimal treatment regimens *in vitro*.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://doi.org/10.5281/zenodo.6973776.

Author contributions

Conceptualization: EYY, GL, and TEY. Methodology: EYY, GL, GRH, AB, and TEY. Software: EYY and GL. Validation: EYY and GL. Formal analysis: EYY, GL, and TEY. Investigation: EYY, GL, GRH, AB, and TEY. Resources: GRH, AB, and TEY. Data curation: EYY, GRH, and AB. Writing-original draft: EYY, GL, and TEY. Writing-review and editing: EYY, GL, GRH, AB, and TEY. Visualization: EYY, and GL. Supervision: GL and TEY. Project Administration: GL. Funding acquisition: TEY.

Funding

GL was partially supported by a Peter O'Donnell Jr. Postdoctoral Fellowship from the Oden Institute for Computational Engineering and Sciences at The University of Texas at Austin and acknowledges funding from the

European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 838786. We thank the National Institutes of Health for funding through NCI R01CA240589, U01CA174706, R01CA186193, U24CA226110, and U01CA253540. We thank the Cancer Prevention and Research Institute of Texas for support through CPRIT RR160005; TY is a CPRIT Scholar in Cancer Research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.972146/full#supplementary-material

References

AbuHammad, S., and Zihlif, M. (2013). Gene expression alterations in doxorubicin resistant MCF7 breast cancer cell line. *Genomics* 101, 213–220. doi:10.1016/j.ygeno.2012.11.009

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., et al. (2015). Toward understanding and exploiting tumor heterogeneity. *Nat. Med.* 21, 846–853. doi:10.1038/nm.3915

Al'Khafaji, A. M., Deatherage, D., and Brock, A. (2018). Control of lineage-specific gene expression by functionalized gRNA barcodes. *ACS Synth. Biol.* 7, 2468–2474. doi:10.1021/acssynbio.8b00105

Álvarez-Arenas, A., Podolski-Renic, A., Belmonte-Beitia, J., Pesic, M., and Calvo, G. F. (2019). Interplay of darwinian selection, lamarckian induction and microvesicle transfer on drug resistance in cancer. *Sci. Rep.* 9, 9332. doi:10.1038/s41598-019-45863-z

Anampa, J., Makower, D., and Sparano, J. A. (2015). Progress in adjuvant chemotherapy for breast cancer: An overview. *BMC Med.*, 195. doi:10.1186/s12916-015-0439-8

Atuegwu, N. C., Arlinghaus, L. R., Li, X., Chakravarthy, A. B., Abramson, V. G., Sanders, M. E., et al. (2013). Parameterizing the logistic model of tumor growth by DW-MRI and DCE-MRI data to predict treatment response and changes in breast cancer cellularity during neoadjuvant chemotherapy. *Transl. Oncol.* 6, 256–264. doi:10.1593/tlo.13130

Bernard, P. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi:10.1200/JCO.2008.18.1370

Campbell, K. J., Dhayade, S., Ferrari, N., Sims, A. H., Johnson, E., Mason, S. M., et al. (2018). MCL-1 is a prognostic indicator and drug target in breast cancer. *Cell Death Dis.* 9, 19–14. doi:10.1038/s41419-017-0035-2

Carvalho, C., Santos, R. X., Cardoso, S., Correia, S., Oliveira, P. J., Santos, M. S., et al. (2009). Doxorubicin: The good, the bad and the ugly effect. *Curr. Med. Chem.* 16, 3267–3285. doi:10.2174/092986709788803312

Chan, S., Friedrichs, K., Noel, D., PinTer, T., Van Belle, S., Vorobiof, D., et al. (1999). Prospective randomized trial of docetaxel versus doxorubicin in patients with metastatic breast cancer. *J. Clin. Oncol.* 17, 2341–2354. doi:10.1200/JCO.1999.17.8.2341

Chapman, M. P., Risom, T., Aswani, A. J., Langer, E. M., Sears, R. C., and Tomlin, C. J. (2019). Modeling differentiation-state transitions linked to therapeutic escape in triple-negative breast cancer. *PLoS Comput. Biol.* 15, e1006840. doi:10.1371/journal.pcbi.1006840

Chung, L. Y., Tang, S. J., Wu, Y. C., Yang, K. C., Huang, H. J., Sun, G. H., et al. (2020). Platinum-based combination chemotherapy triggers cancer cell death through induction of BNIP3 and ROS, but not autophagy. *J. Cell. Mol. Med.* 24, 1993–2003. doi:10.1111/jcmm.14898

Citron, M. L., Berry, D. A., Cirrincione, C., Hudis, C., Winer, E. P., Gradishar, W. J., et al. (2003). Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: First report of intergroup trial C9741/cancer and leukemia group B trial 9741. *J. Clin. Oncol.* 21, 1431–1439. doi:10.1200/JCO.2003.09.081

Colli, P., Gomez, H., Lorenzo, G., Marinoschi, G., Reali, A., and Rocca, E. (2021). Optimal control of cytotoxic and antiangiogenic therapies on prostate cancer growth. *Math. Models Methods Appl. Sci.* 31, 1419–1468. doi:10.1142/S0218202521500299

Craig, M., Jenner, A. L., Namgung, B., Lee, L. P., and Goldman, A. (2021). Engineering in medicine to address the challenge of cancer drug resistance: From micro-and nanotechnologies to computational and mathematical modeling. *Chem. Rev.* 121, 3352–3389. doi:10.1021/acs.chemrev.0c00356

Craig, M., Kaveh, K., Woosley, A., Brown, A. S., Goldman, D., Eton, E., et al. (2019). Cooperative adaptation to therapy (CAT) confers resistance in heterogeneous non-small cell lung cancer. *PLoS Comput. Biol.* 15, e1007278. doi:10.1371/journal.pcbi.1007278

Czeczuga-Semeniuk, E., Wołczyński, S., Dabrowska, M., Dziecioł, J., and Anchim, T. (2004). The effect of doxorubicin and retinoids on proliferation, necrosis and apoptosis in MCF-7 breast cancer cells. *Folia histochem. Cytobiol.* 42, 221–227.

De Souza, R., Zahedi, P., Badame, R. M., Allen, C., and Piquette-Miller, M. (2011). Chemotherapy dosing schedule influences drug resistance development in ovarian cancer. *Mol. Cancer Ther.* 10, 1289–1299. doi:10.1158/1535-7163.MCT-11-0058

Dong, H., Wang, W., Mo, S., Chen, R., Zou, K., Han, J., et al. (2018). SP1-induced lncRNA AGAP2-AS1 expression promotes chemoresistance of breast cancer by epigenetic regulation of MyD88. *J. Exp. Clin. Cancer Res.* 37, 202–215. doi:10.1186/s13046-018-0875-3

Easwaran, H., Tsai, H. C., and Baylin, S. B. (2014). Cancer epigenetics: Tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Mol. Cell* 54, 716–727. doi:10.1016/j.molcel.2014.05.015

Echeverria, G. V., Ge, Z., Seth, S., Zhang, X., Jeter-Jones, S., Zhou, X., et al. (2019). Resistance to neoadjuvant chemotherapy in triple-negative breast cancer mediated by a reversible drug-tolerant state. *Sci. Transl. Med.* 11, eaav0936. doi:10.1126/scitranslmed.aav0936

El-Kareh, A. W., and Secomb, T. W. (2005). Two-mechanism peak concentration model for cellular pharmacodynamics of doxorubicin. *Neoplasia* 7, 705–713. doi:10.1593/neo.05118

Foukakis, T., von Minckwitz, G., Bengtsson, N. O., Brandberg, Y., Wallberg, B., Fornander, T., et al. (2016). Effect of tailored dose-dense chemotherapy vs standard 3-weekly adjuvant chemotherapy on recurrence-free survival amongwomen with high-risk early breast cancer: A randomized clinical trial. *JAMA* 316, 1888–1896. doi:10.1001/jama.2016.15865

Geng, C., Paganetti, H., and Grassberger, C. (2017). Prediction of treatment response for combined chemo- and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Sci. Rep.* 7, 13542. doi:10.1038/s41598-017-13646-z

Gewirtz, D. A. (1999). A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. *Biochem. Pharmacol.* 57, 727–741. doi:10.1016/s0006-2952(98)00307-4

Girifalco, L. A., and Weizer, V. G. (1959). Application of the Morse potential function to cubic metals. *Phys. Rev.* 114, 687–690. doi:10.1103/physrev.114.687

Gottesman, M. M. (2002). Mechanisms of cancer drug resistance. *Annu. Rev. Med.*, 53 615–627. doi:10.1146/annurev.med.53.082901.103929

Greene, J. M., Gevertz, J. L., and Sontag, E. D. (2019). Mathematical approach to differentiate spontaneous and induced evolution to drug resistance during cancer treatment. *JCO Clin. Cancer Inf.* 3, 1–20. doi:10.1200/cci.18.00087

Harahap, Y., Ardiningsih, P., Winarti, A. C., and Purwanto, D. J. (2020). Analysis of the doxorubicin and doxorubicinol in the plasma of breast cancer patients for monitoring the toxicity of doxorubicin. *Drug Des. devel. Ther.* 14, 3469–3475. doi:10.2147/DDDT.S251144

Harris, L. N., You, F., Schnitt, S. J., Witkiewicz, A., Lu, X., Sgroi, D., et al. (2007). Predictors of resistance to preoperative trastuzumab and vinorelbine for HER2-positive early breast cancer. *Clin. Cancer Res.* 13, 1198–1207. doi:10.1158/1078-0432.CCR-06-1304

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: An evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi:10.1038/nrc3599

Hori, S. S., Tong, L., Swaminathan, S., Liebersbach, M., Wang, J., Gambhir, S. S., et al. (2021). A mathematical model of tumor regression and recurrence after therapeutic oncogene inactivation. *Sci. Rep.* 11, 1341. doi:10.1038/s41598-020-78947-2

Howard, G. R., Johnson, K. E., Rodriguez Ayala, A., Yankeelov, T. E., and Brock, A. (2018). A multi-state model of chemoresistance to characterize phenotypic dynamics in breast cancer. *Sci. Rep.* 8, 12058. doi:10.1038/s41598-018-30467-w

Howard, G. R., Jost, T. A., Yankeelov, T. E., and Brock, A. (2022). Quantification of long-term doxorubicin response dynamics in breast cancer cell lines to direct treatment schedules. *PLoS Comput. Biol.* 18, e1009104. doi:10.1371/journal.pcbi.1009104

Jackson, T. L., and Byrne, H. M. (2000). A mathematical model to study the effects of drug resistance and vasculature on the response of solid tumors to chemotherapy. *Math. Biosci.* 164, 17–38. doi:10.1016/s0025-5564(99)00062-0

Jarrett, A. M., Faghihi, D., Ii, D. A. H., Lima, E. A. B. F., Virostko, J., Biros, G., et al. (2020). Optimal control theory for personalized therapeutic regimens in oncology: Background, history, challenges, and opportunities. *J. Clin. Med.* 9, 1314. doi:10.3390/jcm9051314

Jarrett, A. M., Hormuth, D. A., Barnes, S. L., Feng, X., Huang, W., and Yankeelov, T. E. (2018). Incorporating drug delivery into an imaging-driven, mechanics-coupled reaction diffusion model for predicting the response of breast cancer to neoadjuvant chemotherapy: Theory and preliminary clinical results. *Phys. Med. Biol.* 63, 105015. doi:10.1088/1361-6560/aac040

Jarrett, A. M., Hormuth, D. A., Wu, C., Kazerouni, A. S., Ekrut, D. A., Virostko, J., et al. (2020). Evaluating patient-specific neoadjuvant regimens for breast cancer via a mathematical model constrained by quantitative magnetic resonance imaging data. *Neoplasia* 22, 820–830. doi:10.1016/j.neo.2020.10.011

Ji, X., Lu, Y., Tian, H., Meng, X., Wei, M., and Cho, W. C. (2019). Chemoresistance mechanisms of breast cancer and their countermeasures. *Biomed. Pharmacother.* 114, 108800. doi:10.1016/j.biopha.2019.108800

Katt, M. E., Placone, A. L., Wong, A. D., Xu, Z. S., and Searson, P. C. (2016). *In vitro* tumor models: Advantages, disadvantages, variables, and selecting the right platform. *Front. Bioeng. Biotechnol.* 4, 12. doi:10.3389/fbioe.2016.00012

Kazerouni, A. S., Gadde, M., Gardner, A., Hormuth, D. A., Jarrett, A. M., Johnson, K. E., et al. (2020). Integrating quantitative assays with biologically based mathematical modeling for predictive oncology. *iScience* 23, 101807. doi:10.1016/j.isci.2020.101807

Kim, E., Brown, J. S., Eroglu, Z., and Anderson, A. R. A. (2021). Adaptive therapy for metastatic melanoma: Predictions from patient calibrated mathematical models. *Cancers (Basel)* 13, 823. doi:10.3390/cancers13040823

Kowarz, E., Löscher, D., and Marschalek, R. (2015). Optimized Sleeping Beauty transposons rapidly generate stable transgenic cell lines. *Biotechnol. J.* 10, 647–653. doi:10.1002/biot.201400821

Kumar, N., Cramer, G. M., Dahaj, S. A. Z., Sundaram, B., Celli, J. P., and Kulkarni, R. V. (2019). Stochastic modeling of phenotypic switching and chemoresistance in cancer cell populations. *Sci. Rep.* 9, 10845. doi:10.1038/s41598-019-46926-x

Lankelma, J., Dekker, H., Luque, F. R., Luykx, S., HoeKman, K., van der Valk, P., et al. (1999). Doxorubicin gradients in human breast cancer. *Clin. Cancer Res.* 5, 1703–1707.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255. doi:10.2307/2532051

Longley, D. B., and Johnston, P. G. (2005). Molecular mechanisms of drug resistance. *J. Pathology*, 205 275–292.

Lorenzo, G., Hormuth II, D. A, Jarrett, A. M., Lima, E. A. B. F., Subramanian, S., Biros, G., et al. (2022). "Quantitative *in vivo* imaging to enable tumor forecasting and treatment optimization" in *Cancer, Complexity, Computation* Editors I. Balaz and A. Adamatzky (Springer), 55–97.

Low, H. B., Wong, Z. L., Wu, B., Kong, L. R., Png, C. W., Cho, Y. L., et al. (2021). DUSP16 promotes cancer chemoresistance through regulation of mitochondria-mediated cell death. *Nat. Commun.* 12, 2284. doi:10.1038/s41467-021-22638-7

Lyman, G. H. (2009). Impact of chemotherapy dose intensity on cancer patient outcomes. *J. Natl. Compr. Canc. Netw.* 7, 99–108. doi:10.6004/jnccn.2009.0009

Martins, I., Raza, S. Q., Voisin, L., Dakhli, H., Allouch, A., Law, F., et al. (2018). Anticancer chemotherapy and radiotherapy trigger both non-cell-autonomous and cell-autonomous death. *Cell Death Dis.* 9, 716–718. doi:10.1038/s41419-018-0747-y

Mascheroni, P., Boso, D., Preziosi, L., and Schrefler, B. A. (2017). Evaluating the influence of mechanical stress on anticancer treatments through a multiphase porous media model. *J. Theor. Biol.* 421, 179–188. doi:10.1016/j.jtbi.2017.03.027

Mátés, L., Chuah, M. K. L., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., et al. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* 41, 753–761. doi:10.1038/ng.343

McKenna, M. T., Weis, J. A., Barnes, S. L., Tyson, D. R., Miga, M. I., Quaranta, V., et al. (2017). A predictive mathematical modeling approach for the study of doxorubicin treatment in triple negative breast cancer. *Sci. Rep.* 7, 5725. doi:10.1038/s41598-017-05902-z

Meacham, C. E., and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, 501 328–337. doi:10.1038/nature12624

Neale, M. H., Lamont, A., Hindley, A., Kurbacher, C. M., and Cree, I. A. (2000). The *ex vivo* effect of high concentrations of doxorubicin on recurrent ovarian carcinoma. *Anticancer. Drugs* 11, 865–871. doi:10.1097/00001813-200011000-00011

O'Shaughnessy, J. (2005). Extending survival with chemotherapy in metastatic breast cancer. *Oncol.* 10, 20–29. doi:10.1634/theoncologist.10-90003-20

Palmer, A. C., and Sorger, P. K. (2017). Combination cancer therapy can confer benefit via patient-to-patient variability without drug additivity or synergy. *Cell* 171, 1678–1691. doi:10.1016/j.cell.2017.11.009.e13

Panetta, J. C. A. (1997). A logistic model of periodic chemotherapy with drug resistance. *Appl. Math. Lett.* 10, 123–127. doi:10.1016/s0893-9659(96)00123-1

Panetta, J. C. (1998). A mathematical model of drug resistance: Heterogeneous tumors. *Math. Biosci.* 147, 41–61. doi:10.1016/s0025-5564(97)00080-1

Pascal, J., Bearer, E. L., Wang, Z., Koay, E. J., Curley, S. A., and Cristini, V. (2013). Mechanistic patient-specific predictive correlation of tumor drug response with microenvironment and perfusion measurements. *Proc. Natl. Acad. Sci. U. S. A.* 110, 14266–14271. doi:10.1073/pnas.1300619110

Pisco, A. O., Brock, A., Zhou, J., Moor, A., Mojtahedi, M., Jackson, D., et al. (2013). Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nat. Commun.* 4, 2467. doi:10.1038/ncomms3467

Polyak, K. (2011). Heterogeneity in breast cancer. *J. Clin. Invest.* 121, 3786–3788. doi:10.1172/JCI60534

Ponnusamy, L., Mahalingaiah, P. K. S., and Singh, K. P. (2017). Treatment schedule and estrogen receptor-status influence acquisition of doxorubicin resistance in breast cancer cells. *Eur. J. Pharm. Sci.* 104, 424–433. doi:10.1016/j.ejps.2017.04.020

Richards, M. A., HoPwood, P., Ramirez, A. J., Twelves, C. J., Ferguson, J., Gregory, W. M., et al. (1992). Doxorubicin in advanced breast cancer: Influence of schedule on response, survival and quality of life. *Eur. J. Cancer* 28, 1023–1028. doi:10.1016/0959-8049(92)90447-a

Rivera, E., and Gomez, H. (2010). Chemotherapy resistance in metastatic breast cancer: The evolving role of ixabepilone. *Breast Cancer Res.* 12 S2, S2. doi:10.1186/bcr2573

Rockne, R. C., Hawkins-Daarud, A., Swanson, K. R., Sluka, J. P., Glazier, J. A., Macklin, P., et al. (2019). The 2019 mathematical oncology roadmap. *Phys. Biol.* 16, 041005. doi:10.1088/1478-3975/ab1a09

Strobl, M. A. R., West, J., Viossat, Y., Damaghi, M., Robertson-Tessi, M., Brown, J. S., et al. (2021). Turnover modulates the need for a cost of resistance in adaptive therapy. *Cancer Res.* 81, 1135–1147. doi:10.1158/0008-5472.CAN-20-0806

Sun, X., Bao, J., and Shao, Y. (2016). Mathematical modeling of therapy-induced cancer drug resistance: Connecting cancer mechanisms to population survival rates. *Sci. Rep.* 6, 22498. doi:10.1038/srep22498

Sun, X., and Hu, B. (2018). Mathematical modeling and computational prediction of cancer drug resistance. *Brief. Bioinform.* 19, 1382–1399. doi:10.1093/bib/bbx065

Sun, X. X., and Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol. Sin.* 36, 1219–1227. doi:10.1038/aps.2015.92

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Untch, M., Mobus, V., Kuhn, W., Muck, B. R., Thomssen, C., Bauerfeind, I., et al. (2009). Intensive dose-dense compared with conventionally scheduled preoperative chemotherapy for high-risk primary breast cancer. *J. Clin. Oncol.* 27, 2938–2945. doi:10.1200/JCO.2008.20.3133

Waks, A. G., and Winer, E. P. (2019). Breast cancer treatment: A review. *JAMA*, 321 288–300. doi:10.1001/jama.2018.19323

Weis, J. A., Miga, M. I., Arlinghaus, L. R., Li, X., Abramson, V., Chakravarthy, A. B., et al. (2015). Predicting the response of breast cancer to neoadjuvant therapy using a mechanically coupled reaction-diffusion model. *Cancer Res.* 75, 4697–4707. doi:10.1158/0008-5472.CAN-14-2945

Yankeelov, T. E., Atuegwu, N., Hormuth, D., Weis, J. A., Barnes, S. L., Miga, M. I., et al. (2013). Clinically relevant modeling of tumor growth and treatment response. *Sci. Transl. Med.* 5, 187ps9. doi:10.1126/scitranslmed.3005686

Yonucu, S., Yılmaz, D., Phipps, C., Unlu, M. B., and Kohandel, M. (2017). Quantifying the effects of antiangiogenic and chemotherapy drug combinations on drug delivery and treatment efficacy. *PLoS Comput. Biol.* 13, e1005724. doi:10.1371/journal.pcbi.1005724

Zardavas, D., and Piccart, M. (2015). Neoadjuvant therapy for breast cancer. *Annu. Rev. Med.* 66, 31–48. doi:10.1146/annurev-med-051413-024741

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput. Biol.* 11, e1004498. doi:10.1371/journal.pcbi.1004498

Zhao, J. (2016). Cancer stem cells and chemoresistance: The smartest survives the raid. *Pharmacol. Ther.*, 160 145–158.

Zoli, W., FlAmigni, A., Frassineti, G. L., Bajorko, P., De PaolaF.Milandri, C., et al. (1995). *In vitro* activity of taxol and taxotere in comparison with doxorubicin and cisplatin on primary cell cultures of human breast cancers. *Breast Cancer Res. Treat.* 34, 63–69. doi:10.1007/BF00666492

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fimmu.2022.987881

Check for updates

OPEN ACCESS

EDITED BY
Kathleen Wilkie,
Ryerson University, Canada

REVIEWED BY
Albino Eccher,
Integrated University Hospital
Verona, Italy
Wenle Li,
Xiamen University, China
Jie Tian,
Shanghai Jiao Tong University, China
Haosheng Wang,
Nanjing Drum Tower Hospital, China

*CORRESPONDENCE
Meiyun Liu
liumeiyun0@126.com
Yue Yu
15021567170@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 06 July 2022
ACCEPTED 11 August 2022
PUBLISHED 23 September 2022

CITATION
Wang S, Wei J, Guo Y, Xu Q, Lv X, Yu Y and Liu M (2022) Construction and validation of nomograms based on the log odds of positive lymph nodes to predict the prognosis of lung neuroendocrine tumors. *Front. Immunol.* 13:987881. doi: 10.3389/fimmu.2022.987881

COPYRIGHT
© 2022 Wang, Wei, Guo, Xu, Lv, Yu and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Construction and validation of nomograms based on the log odds of positive lymph nodes to predict the prognosis of lung neuroendocrine tumors

Suyu Wang[1,2,3†], Juan Wei[1†], Yibin Guo[4†], Qiumeng Xu[5†], Xin Lv[1], Yue Yu[3*] and Meiyun Liu[1*]

[1]Department of Anesthesiology, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai, China, [2]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai, China, [3]Department of Cardiothoracic Surgery, Changzheng Hospital, Naval Medical University, Shanghai, China, [4]Department of Health Statistics, Naval Medical University, Shanghai, China, [5]Department of Orthopaedics, Changzheng Hospital, Naval Medical University, Shanghai, China

Background: This research aimed to investigate the predictive performance of log odds of positive lymph nodes (LODDS) for the long-term prognosis of patients with node-positive lung neuroendocrine tumors (LNETs).

Methods: We collected 506 eligible patients with resected N1/N2 classification LNETs from the Surveillance, Epidemiology, and End Results (SEER) database between 2004 and 2015. The study cohort was split into derivation cohort (n=300) and external validation cohort (n=206) based on different geographic regions. Nomograms were constructed based on the derivation cohort and validated using the external validation cohort to predict the 1-, 3-, and 5-year cancer-specific survival (CSS) and overall survival (OS) of patients with LNETs. The accuracy and clinical practicability of nomograms were tested by Harrell's concordance index (C-index), integrated discrimination improvement (IDI), net reclassification improvement (NRI), calibration plots, and decision curve analyses.

Results: The Cox proportional-hazards model showed the high LODDS group (-0.79≤LODDS) had significantly higher mortality compared to those in the low LODDS group (LODDS<-0.79) for both CSS and OS. In addition, age at diagnosis, sex, histotype, type of surgery, radiotherapy, and chemotherapy were also chosen as predictors in Cox regression analyses using stepwise Akaike information criterion method and included in the nomograms. The values of C-index, NRI, and IDI proved that the established nomograms were better than the conventional eighth edition of the TNM staging system. The calibration plots for predictions of the 1-, 3-, and 5-year CSS/OS were in excellent agreement. Decision curve analyses showed that the nomograms had value in terms of clinical application.

10.3389/fimmu.2022.987881

Conclusions: We created visualized nomograms for CSS and OS of LNET patients, facilitating clinicians to bring individually tailored risk assessment and therapy.

KEYWORDS

lung neuroendocrine tumor, log odds of positive lymph nodes, predictor, survival, nomogram

Introduction

Lung neuroendocrine tumors (LNETs) originate from pulmonary neuroendocrine cells, accounting for approximately 25% of primary lung neoplasms (1). The incidence of LNET is less than 0.002% as reported in some countries (2–4). Owing to the increased lung cancer screening, the annual incidence of LNET has substantially increased, rising from 0.0003% in 1973 to 0.0014% in 2004 in the United States (5–7). Currently, the 2015 World Health Organization (WHO) classification has grouped LNETs into four histologic variants based on their histopathologic features: typical carcinoid (TC), atypical carcinoid (AC), large cell neuroendocrine carcinoma (LCNEC), and small cell lung carcinoma (SCLC) (8). The 5-year survival rate of non-small cell lung cancer (NSCLC) was approximately 16%, while the corresponding rate of SCLC, LCNEC, AC, and TC patients was 5%, 17%, 64%, and 84%, respectively (9, 10). Due to the rarity and morphological heterogeneity of these tumors, there have been limited clinical data available regarding LNETs, thus making their diagnosis, staging, risk assessment, and treatment challenging (1). Although specific to NSCLC, the international American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) tumor-node-metastasis (TNM) staging system has been applied to LNETs (11, 12). However, several studies have shown an overlapping survival of patients with LNETs, particularly in stages II and III (11–15). Therefore, further investigation is warranted to optimize the staging system for LNETs.

Lymph node (LN) involvement is a significant prognostic factor for staging and risk stratification. A combination of the number of positive lymph nodes (NPLN), the number of dissected lymph nodes (NDLN), and the anatomic location of LN metastasis have been applied in the staging of many malignancies (16–20). However, the latest version of AJCC/UICC TNM classification of lung cancer did not take account of any number- or ratio-based LN staging system, which could affect the precision of prognosis evaluation (21, 22). A novel prognostic indicator known as log odds of positive lymph nodes (LODDS) is being utilized to identify patients with a homogeneous prognosis in many malignant tumors, including NSCLC (23–27). However, no specific study has focused on its prognostic significance in LNETs till now.

The present study aimed to determine whether LODDS could be utilized to predict the cancer-specific survival (CSS) or overall survival (OS) of patients with node-positive (N1/N2 classification) LNET using the Surveillance, Epidemiology, and End Results (SEER) database. Based on LODDS, the research intended to construct two visualized and online nomograms which are practical tools for clinical prediction used in many diseases (28–32). To facilitate clinical use, we also constructed two visualized and online nomograms for LNETs.

Abbreviations: AC, atypical carcinoid; AJCC, American Joint Committee on Cancer; CI, confidence interval; C-index, concordance index; CSS, cancer-specific survival; DCA, decision curve analysis; HR, hazard ratio; IASLC, International Association for the Study of Lung Cancer; IDI, integrated discrimination improvement; IQR, interquartile range; LCNEC, large cell neuroendocrine carcinoma; LN, lymph node; LNET, lung neuroendocrine tumor; LNR, lymph node ratio; LODDS, log odds of positive lymph nodes; NDLN, number of dissected lymph nodes; NPLN, number of positive lymph nodes; NRI, net reclassification improvement; NSCLC, non-small cell lung cancer; OS, overall survival; PRCDA, purchased/referred care delivery area; SCLC, small cell lung carcinoma; SD, standard deviation; SEER, Surveillance, Epidemiology, and End Results; TC, typical carcinoid; TNM, tumor-node-metastasis; UICC, Union for International Cancer Control; WHO, World Health Organization.

Methods

Study design, data screening, and ethical statement

This is a multi-center retrospective cohort study according to the parts of the methods described in our previous studies (25, 26, 33). We use the Transparent Reporting of a multivariate prediction model for Individual Prediction or Diagnosis (TRIPOD) for reporting (34). The data of this study were downloaded from the SEER 18 registries research database, covering approximately 28% of the population of USA (35). Data were extracted using the SEER*Stat version 8.3.9 software.

The requirement for approval by the institutional review board and individual patient consent was waived since the study made use of the database's anonymous data. In summary, this study complied with the Declaration of Helsinki (36).

Population selection

Data on the patients with lung cancer was obtained from the SEER database. Inclusion criteria were as followed (1): diagnosed from 2004 to 2015; (2) site recode "ICD-O-3/WHO 2008" restricted to "Lung and Bronchus"; and (3) pathologically confirmed as TC (ICD-O-3 code: 8240/3), AC (ICD-O-3 code: 8249/3) or LCNEC (ICD-O-3 code: 8013/3). The study period was set from 2004 to 2015, as the sixth or seventh edition of TMN classification and Collaborative Stage information was available in the database since 2004. Besides, we reclassified the TNM staging system according to the 8th version of TNM classification because the TNM staging system had multiple versions in the SEER database and did not apply to all patients (37). Furthermore, considering its strong invasion ability and unique pathological characteristics limiting the surgical options, SCLC was not included in the present study (38, 39). Although LCNEC was reported to contain subgroups of tumors showing SCLC characteristics and others with NSCLC-like features, surgery could be considered for early and locally advanced LCNEC (40–42). Therefore, we enrolled patients with LCNEC in this study. Patients were excluded who (1) aged<18 years; (2) had a diagnosis of any other cancer; (3) did not undergo radical surgery with systematic LN dissection; (4) had the diagnosis lacking pathological evidence; (5) were at pN0/pN3 disease; (6) had distant metastasis (M1); (7) received preoperative radiotherapy; (8) died within one month after surgery; (9) had unknown information of race, laterality, tumor location, radiotherapy, TNM staging system, and CSS/OS.

Variable extraction, preparation, grouping, and calculation

The baseline demographics data including age at diagnosis (<65 and ≥65), sex (male and female), and race (white, black, and other) were extracted from the SEER database. Baseline tumor-related characteristics included primary site (upper lobe, middle lobe, lower lobe, and other), laterality (right and left), histotype (TC, AC, and LCNEC), tumor differentiation (well/moderately differentiated, poorly differentiated/undifferentiated, and unknown), T classification (T1, T2, T3, and T4), and N classification (N1 and N2). In addition, treatment information including surgical intervention (sublobectomy, lobectomy, and pneumonectomy), radiotherapy (yes and no/unknown), chemotherapy (yes and no/unknown), NDLN, and NPLN was also extracted from the database. LODDS was calculated as: $LODDS = \ln \frac{NPLN+0.50}{NDLN-NPLN+0.5}$. To avoid an infinite number, 0.50 was added to both the numerator and denominator. CSS and OS were two of the study endpoints. The period from diagnosis to all-cause death was referred to as OS, while the time from diagnosis to LNET-related death was referred to as CSS. For censored data, the follow-up duration was computed as the number of months between diagnosis and death or the last follow-up (December 31, 2016).

Construction and validation of nomograms

Baseline features of the study groups stratified by LODDS were compared using Pearson's $\chi 2$ test, Fisher's exact test, Student t test, or Mann-Whitney test as appropriate. Categorical variables were presented as counts and percentages, while continuous variables were reported as the mean (standard deviation [SD]) or the median (interquartile range [IQR]).

Patients from purchased/referred care delivery areas (PRCDA) of Northern plains, East, and Alaska were considered to be the derivation cohort, whereas the external validation cohort includes patients from PRCDA of Southwest and Pacific coast. LODDS was dichotomized *via* the X-tile software to achieve the largest difference in survival outcome by selecting the highest χ^2 value in survival analysis indicating the largest survival difference (43). First, age at diagnosis, sex, race, laterality, primary site, histotype, tumor differentiation, T classification, N classification, surgery, radiotherapy, chemotherapy, and LODDS were analyzed by univariable Cox-proportional-hazards regression analysis. Second, the potential predictors with $P<0.1$ were put into the multivariable Cox-proportional-hazards regression analysis using the stepwise Akaike information criterion method (stepAIC) to select the optimal predictors for the final models (44). The results are presented as hazard ratios (HRs) with 95% confidence intervals (CIs). Using these identified prognostic factors, we constructed two nomograms for predicting 1-, 3-, and 5-year CSS and OS in LNET patients.

Several indexes and methods were used to assess the precision of our nomograms. First, Harrell's concordance index (C-index) was used to evaluate the discrimination power of the nomograms. Second, a calibration plot, a curve presenting the conformity between predicted and actual survival rate at 1, 3, and 5 years, along with bootstrapped sampling of the population, was used to assess the calibration. Third, a comparison between our nomograms and the conventional 8th version of the TNM staging system was conducted by calculating the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (45). Z test was used to examine the difference. Fourth, the receiver operating characteristic (ROC) curve and decision curve analysis (DCA) were performed to test the clinical usefulness of the nomograms

and TNM classification. Kaplan-Maier curves were applied for comparing the discriminative power of nomograms and TNM staging system in the entire study cohort. Finally, to facilitate patients and doctors in using the models, two user-friendly webservers for our nomograms were established.

All statistical analyses were performed using R software (version.3.6.14.1.0; The R Project for Statistical Computing, TX, USA; http://www.r-project.org) and EmpowerStats (version 2.0; http://www.empowerstats.com). Two-tailed P<0.05 was deemed as statistical significance.

Results

Patients characteristics and cutoff value for LOODS

The SEER database collected 11,870 patients diagnosed with LNETs from January 2004 to December 2015. After employing the inclusion and exclusion criteria, 506 patients remained in the final study cohort. The selection process was summarized in Figure 1. According to the X-tile software's calculations, the optimal cutoff value of LODDS was set as -0.79. Baseline demographic and clinicopathological variables of participants in the derivation dataset and external validation dataset were summarized in Table 1. The median number of NDLN was 9.00 (IQR: 6.00-14.00) in the derivation dataset and 9.00 (IQR: 6.00-15.00) in the validation dataset, whereas the median number of NPLN was 1.00 (IQR: 1.00-2.00) in the derivation dataset and 2.00 (IQR: 1.00-3.00) in the validation dataset. Compared with the derivation dataset, the validation dataset had more patients with other race (P-value<0.001). No difference was observed in other variables (all P-values>0.05).

Survival analysis

The median follow-up time for the entire cohort was 42 months (IQR: 19-77 months). Among 506 participants, 180 (35.57%) participants died from any cause, and 135 (26.68%) participants died from LNET-related death in the end of the follow-up. Multivariable Cox regression analysis showed that participants in the high LODDS group (-0.79≤LODDS) was associated with reducing CSS compared to those in the low LODDS group (LODDS<-0.79) (HR=2.21, 95% CI: 1.38-3.52, P<0.001) (Table 2). The multivariable Cox regression analysis for OS yielded similar results (HR=1.76, 95% CI: 1.20-2.58, P=0.004, Table 3). To diminish the potential bias caused by the LN fragments, we performed a sensitivity analysis by restricting the resected LN count to fewer than or equal to 20, and found that LODDS remained statistically significant (CSS: HR=2.06, 95% CI: 1.28-3.31, P=0.003; OS: HR=1.68, 95% CI: 1.12-2.53, P=0.012) (Tables S1, S2). Furthermore, the extent of LN management should be in accordance with the IASLC recommendations, which recommended examination of at least 6 nodes/stations, therefore we excluded patients with the examined LN count to fewer than 6. The multivariable Cox regression analysis yielded similar results (CSS: HR=3.64, 95% CI: 1.99-6.68, P<0.001; OS: HR=2.00, 95% CI: 1.23-3.23, P=0.005) (Tables S3, S4).

Construction and validation of the nomograms

Prognostic nomograms for CSS and OS were established including optimal indicators selected by stepAIC method in multivariable Cox regression analysis. For nomogram construction and validation, among the final study cohort

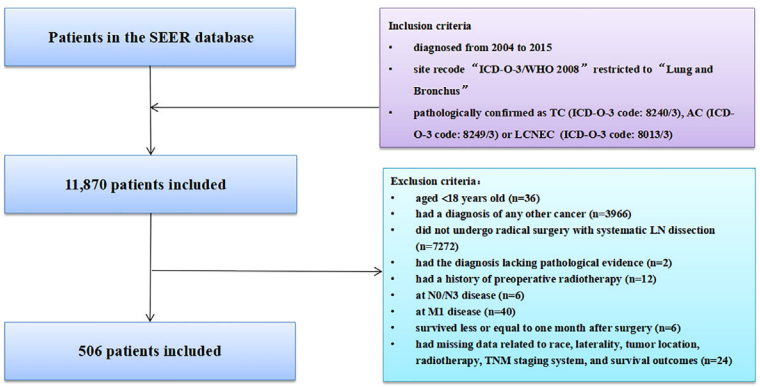


FIGURE 1
Selection of study cohort from the SEER database. AC, atypical carcinoid; LCNEC, large cell neuroendocrine carcinoma; SCLC, small cell lung carcinoma; SEER, Surveillance, Epidemiology, and End Results; LN, lymph node; TNM, tumor-node-metastasis.

TABLE 1 Baseline characteristics of derivation dataset and external validation dataset.

Characteristic	Total (n = 506)	Derivation dataset (n = 300)	External validation dataset (n = 206)	*P* value*
Age at diagnosis (year)				0.533
<65	321 (63.44)	187 (62.3)	134 (65.0)	
≥65	185 (36.56)	113 (37.7)	72 (35.0)	
Sex				0.314
Male	232 (45.85)	132 (44.00)	100 (48.54)	
Female	274 (54.15)	168 (56.00)	106 (51.46)	
Race				<0.001
White	453 (89.53)	281 (93.67)	172 (83.50)	
Black	29 (5.73)	15 (5.00)	14 (6.80)	
Other	24 (4.74)	4 (1.33)	20 (9.71)	
Laterality				0.308
Right	281 (55.53)	161 (53.67)	120 (58.25)	
Left	225 (44.47)	139 (46.33)	86 (41.75)	
Site				0.360
Upper lobe	194 (38.34)	124 (41.33)	70 (33.98)	
Middle lobe	64 (12.65)	34 (11.33)	30 (14.56)	
Lower lobe	196 (38.74)	113 (37.67)	83 (40.29)	
Other	52 (10.28)	29 (9.67)	23 (11.17)	
Histotype				0.100
Large cell neuroendocrine carcinoma	129 (25.49)	86 (28.67)	43 (20.87)	
Carcinoid tumor	274 (54.15)	152 (50.67)	122 (59.22)	
Atypical carcinoid tumor	103 (20.36)	62 (20.67)	41 (19.90)	
Differentiation				0.237
Well/moderately differentiated	201 (39.72)	120 (40.00)	81 (39.32)	
Poorly differentiated/undifferentiated	97 (19.17)	64 (21.33)	33 (16.02)	
Unknown	208 (41.11)	116 (38.67)	92 (44.66)	
T classification				0.387
T1	209 (41.30)	128 (42.67)	81 (39.32)	
T2	190 (37.55)	114 (38.00)	76 (36.89)	
T3	74 (14.62)	43 (14.33)	31 (15.05)	
T4	33 (6.52)	15 (5.00)	18 (8.74)	
N classification				0.468
N1	319 (63.04)	193 (64.33)	126 (61.17)	
N2	187 (36.96)	107 (35.67)	80 (38.83)	
Surgery				0.483
Sublobectomy	44 (8.70)	24 (8.00)	20 (9.71)	
Lobectomy	406 (80.24)	246 (82.00)	160 (77.67)	
Pneumonectomy	56 (11.07)	30 (10.00)	26 (12.62)	
Radiotherapy				0.733
No/unknown	439 (86.76)	259 (86.33)	180 (87.38)	
Yes	67 (13.24)	41 (13.67)	26 (12.62)	
Chemotherapy				0.111
No/unknown	374 (73.91)	214 (71.33)	160 (77.67)	
Yes	132 (26.09)	86 (28.67)	46 (22.33)	
Number of dissected lymph nodes	9.00 (6.00-14.75)	9.00 (6.00-14.00)	9.00 (6.00-15.00)	0.080
Number of positive lymph nodes	2.00 (1.00-3.00)	1.00 (1.00-2.00)	2.00(1.00-3.00)	0.139

(Continued)

TABLE 1 Continued

Characteristic	Total (n = 506)	Derivation dataset (n = 300)	External validation dataset (n = 206)	*P* value*
LODDS as a categorical variable				0.864
LODDS<-0.79	334 (66.01)	190 (63.3)	132 (64.1)	
LODDS≥-0.79	172 (33.99)	110 (36.7)	74 (35.9)	

Data are listed as median (IQR) or n (%). *P values for Mann-Whitney test or Chi square test by comparing the basic characteristics in the derivation and external validation datasets. LODDS, log odds of positive lymph node.

including 506 patients, 300 of them were assigned to the derivation cohort (PRCDA=Northern plains, East, and Alaska) and 206 of them were assigned to the validation cohort (PRCDA=Southwest and Pacific coast). The nomogram of CSS showed that histotype made the largest contribution to prediction, followed by LODDS, surgery, age at diagnosis, and chemotherapy (Figure 2A and Table 2). Similarly, the nomogram of OS showed that histotype made the largest contribution to prediction, followed by chemotherapy, age at diagnosis, LODDS, radiotherapy and sex (Figure 2B and Table 3). The top point reference scale of the nomograms assigned a score for each category of these predictive variables. After adding up the total score and locating the sum on the total points reference scale, a straight line was drawn to the bottom survival probability scale to find the estimated 1-/3-/5- survival rate. For example, for a LNET patient aged 65 (21 points) who had a diagnosis of large cell neuroendocrine carcinoma (100 points) and underwent lobectomy (0 points) with LODDS<-0.79 (0 points) and received chemotherapy (0 points), the total points were 121 points, corresponding to a 1-year CSS of 83%, in addition, the online dynamic nomogram could give a 95%CI of 74-92% (Figure 2A).

The nomograms were validated internally in the derivation cohort and externally in the validation cohort. C-indexes of nomogram for CSS in derivation cohort and validation cohort were 0.843 (0.801-0.885), and 0.809 (0.755-0.863). C-indexes of OS for OS in derivation cohort and validation cohort were 0.813 (0.774-0.852), and 0.801 (0.753-0.848). Additionally, a good similarity between the nomogram-predicted and actual survival rates was demonstrated by the calibration plots (Figure 3), which showed that the spots were near to the 45-degree line.

Comparison of the nomograms and the eighth edition TNM staging system

The comparisons between the nomograms and the TNM staging system were also performed (Table S5). Analysis of accuracy showed that the IDI or NRI for the 1-, 3-, and 5-year CSS or OS in derivation or validation dataset were all larger than 0 with all $P<0.001$, indicating a better prediction power of the nomograms compared with TNM staging system. The area under ROC curve of the nomograms was larger than TNM staging system for the 1-, 3-, and 5-year CSS or OS prediction in derivation or validation dataset (Figure 4). Furthermore, the DCA was applied to determine the clinical applicability. The DCA showed that our nomograms were better than the TNM staging system, as it added more net benefits than the TNM classification for nearly all threshold probabilities based on both the derivation cohort and validation cohort (Figure 5). The nomogram-calculated total points of the patients were divided into low- (CSS: <80; OS: <83), medium- (CSS: 80 to 133; OS: 83 to 135), and high-risk (CSS: >133; OS: >135) subgroups using X-tile software, and this classification method exhibited better discriminative power than the eighth edition of TNM staging system as shown in the Kaplan-Maier curve analysis (Figure 6).

Establishment of online models for convenient clinical use

Two online nomograms with interactive interface based on the multivariable Cox regression models were built (CSS: https://drboidedwater.shinyapps.io/DynNom-CSS-lungneuroendocrinecarcinoma/; OS: https://drboidedwater.shinyapps.io/DynNom-OS-lungneuroendocrinecarcinoma/). To facilitate the access of these two links, the quick response codes were demonstrated in Figure 2. The webservers may generate estimated survival rate and Kaplan-Meier curves by entering the covariates.

Discussion

LNETs constitute a unique clinical subgroup of primary pulmonary tumors. Due to their relatively low incidence, no specific staging system exists for LNETs. An exact and reasonable classification of the lymph node status is critical in the staging and prognosis evaluation of patients with LNCTs. In this study, LODDS was proved to be independently related to long-term clinical prognosis among patients with resectable LNETs. These results were robust to a series of sensitivity analyses. Second, we constructed two visualized and publicly accessible online nomograms, incorporating LODDS and routinely available demographic, staging and treatment

10.3389/fimmu.2022.987881

TABLE 2 Univariable and stepwise multivariable Cox proportional regression analysis for CSS of the derivation dataset.

Characteristic	N (%)	Univariable analysis		Stepwise multivariable analysis	
		HR (95%CI)	*P* value	HR (95%CI)	*P* value
Age at diagnosis (year)					
<65	187 (62.3)	1		1	
≥65	113 (37.7)	2.45 (1.57-3.82)	<0.001	1.82 (1.13-2.93)	0.014
Sex					
Male	132 (44.00)	1			
Female	168 (56.00)	0.68 (0.44-1.06)	0.087		
Race					
White	281 (93.67)	1			
Black	15 (5.00)	1.03 (0.38-2.81)	0.958		
Other	4 (1.33)	0.00 (0.00-Inf)	0.995		
Laterality					
Right	161 (53.67)	1			
Left	139 (46.33)	0.85 (0.54-1.32)	0.467		
Site					
Upper lobe	124 (41.33)	1			
Middle lobe	34 (11.33)	0.25 (0.09-0.71)	0.009		
Lower lobe	113 (37.67)	0.59 (0.36-0.96)	0.033		
Other	29 (9.67)	0.41 (0.16-1.04)	0.061		
Histotype					
Large cell neuroendocrine carcinoma	86 (28.67)	1		1	
Carcinoid tumor	152 (50.67)	0.08 (0.04-0.14)	<0.001	0.06 (0.03-0.12)	<0.001
Atypical carcinoid tumor	62 (20.67)	0.19 (0.10-0.35)	<0.001	0.16 (0.08-0.31)	<0.001
Differentiation					
Well/moderately differentiated	120 (40.00)	1			
Poorly differentiated/undifferentiated	64 (21.33)	6.97 (3.87-12.52)	<0.001		
Unknown	116 (38.67)	1.52 (0.81-2.85)	0.195		
T classification					
T1	128 (42.67)	1			
T2	114 (38.00)	0.93 (0.56-1.54)	0.777		
T3	43 (14.33)	1.54 (0.82-2.88)	0.178		
T4	15 (5.00)	0.95 (0.34-2.68)	0.923		
N classification					
N1	193 (64.33)	1			
N2	107 (35.67)	1.74 (1.12-2.71)	0.014		
Surgery					
Sublobectomy	24 (8.00)	1		1	
Lobectomy	246 (82.00)	0.28 (0.15-0.52)	<0.001	0.69 (0.35-1.34)	0.270
Pneumonectomy	30 (10.00)	0.45 (0.20-1.00)	0.049	1.33 (0.56-3.14)	0.516
Radiotherapy					
No/unknown	259 (86.33)	1			
Yes	41 (13.67)	2.34 (1.40-3.89)	0.001		
Chemotherapy					
No/unknown	214 (71.33)	1		1	
Yes	86 (28.67)	2.63 (1.69-4.10)	<0.001	0.63 (0.37-1.07)	0.084
LODDS as a categorical variable					
LODDS<-0.79	190 (63.3)	1		1	
LODDS≥-0.79	110 (36.7)	2.23 (1.43-3.48)	<0.001	2.21 (1.38-3.52)	<0.001

CSS, cancer-specific survival; LODDS, log odds of positive lymph node; HR, hazard ratio; CI, confidence interval.

10.3389/fimmu.2022.987881

TABLE 3 Univariable and stepwise multivariable Cox proportional regression analysis for OS of the derivation dataset.

Characteristic	N (%)	Univariable analysis		Stepwise multivariable analysis	
		HR (95%CI)	*P* value	HR (95%CI)	*P* value
Age at diagnosis (year)					
<65	187 (62.3)	1		1	
≥65	113 (37.7)	3.10 (2.12-4.55)	<0.001	2.33 (1.57-3.47)	<0.001
Sex					
Male	132 (44.00)	1		1	
Female	168 (56.00)	0.62 (0.43-0.91)	0.014	0.74 (0.51-1.09)	0.131
Race					
White	281 (93.67)	1			
Black	15 (5.00)	1.11 (0.49-2.53)	0.801		
Other	4 (1.33)	0.00 (0.00-Inf)	0.994		
Laterality					
Right	161 (53.67)	1			
Left	139 (46.33)	0.87 (0.59-1.26)	0.453		
Site					
Upper lobe	124 (41.33)	1			
Middle lobe	34 (11.33)	0.41 (0.20-0.87)	0.02		
Lower lobe	113 (37.67)	0.74 (0.49-1.12)	0.149		
Other	29 (9.67)	0.69 (0.35-1.36)	0.284		
Histotype					
Large cell neuroendocrine carcinoma	86 (28.67)	1		1	
Carcinoid tumor	152 (50.67)	0.13 (0.08-0.20)	<0.001	0.09 (0.05-0.15)	<0.001
Atypical carcinoid tumor	62 (20.67)	0.23 (0.13-0.39)	<0.001	0.18 (0.10-0.33)	<0.001
Differentiation					
Well/moderately differentiated	120 (40.00)	1			
Poorly differentiated/undifferentiated	64 (21.33)	6.22 (3.70-10.46)	<0.001		
Unknown	116 (38.67)	1.98 (1.18-3.34)	0.01		
T classification					
T1	128 (42.67)	1			
T2	114 (38.00)	1.12 (0.74-1.70)	0.586		
T3	43 (14.33)	1.35 (0.76-2.40)	0.301		
T4	15 (5.00)	0.91 (0.36-2.31)	0.849		
N classification					
N1	193 (64.33)	1			
N2	107 (35.67)	1.76 (1.21-2.57)	0.003		
Surgery					
Sublobectomy	24 (8.00)	1			
Lobectomy	246 (82.00)	0.38 (0.21-0.67)	<0.001		
Pneumonectomy	30 (10.00)	0.49 (0.23-1.03)	0.058		
Radiotherapy					
No/unknown	259 (86.33)	1		1	
Yes	41 (13.67)	1.98 (1.26-3.12)	0.003	1.51 (0.86-2.66)	0.151
Chemotherapy					
No/unknown	214 (71.33)	1		1	
Yes	86 (28.67)	1.86 (1.26-2.73)	0.002	0.43 (0.25-0.74)	0.002
LODDS as a categorical variable					
LODDS<-0.79	190 (63.3)	1		1	
LODDS≥-0.79	110 (36.7)	1.71 (1.17-2.48)	0.005	1.76 (1.20-2.58)	0.004

LODDS, log odds of positive lymph node; OS, overall survival; HR, hazard ratio; CI, confidence interval.

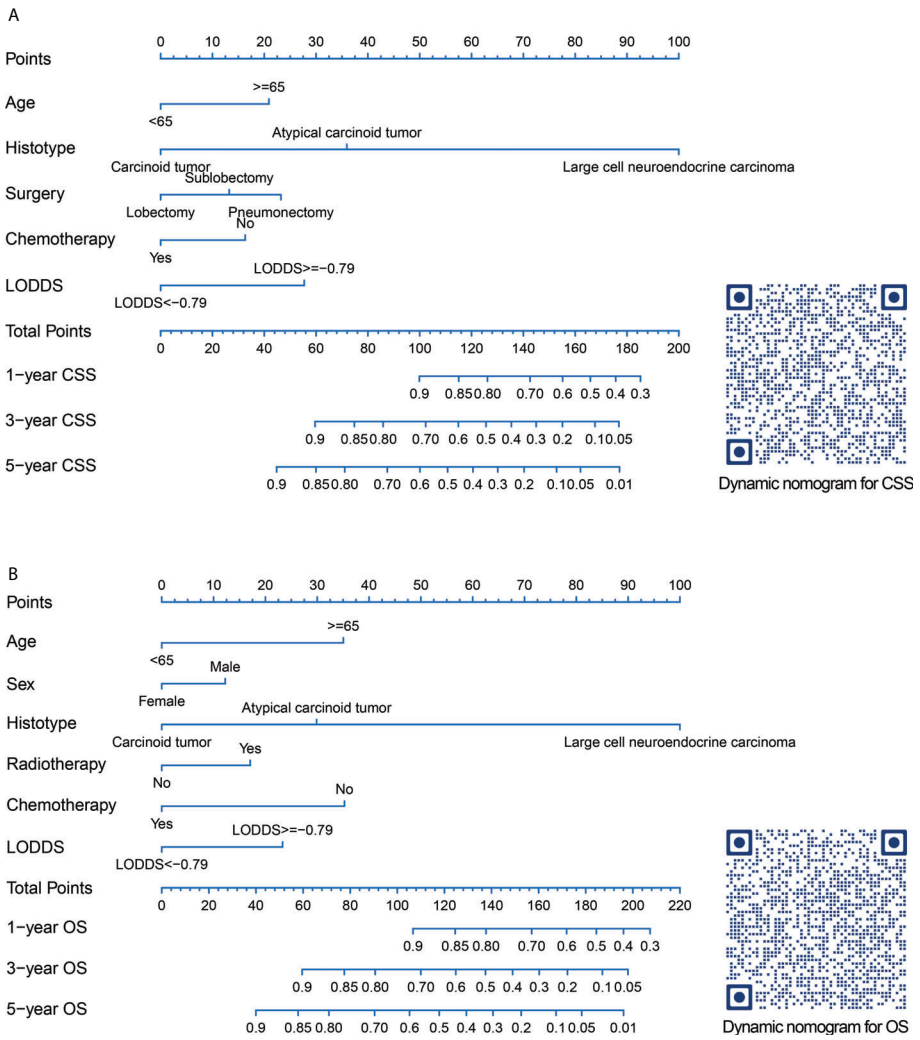


FIGURE 2
Nomograms and quick response codes of the webservers of the nomograms to predict 1-, 3- and 5-year CSS **(A)** and OS **(B)** for patients with node-positive lung neuroendocrine carcinoma after surgery. CSS, lung cancer-specific survival; OS, overall survival; LODDS, log odds of positive lymph nodes.

information, to predict the survival probability for individual LNET patients. To our knowledge, the present study is the first to explore the prognostic value of LODDS for LNET based on a multi-center cohort with a relatively large sample size.

The involvement of regional lymph nodes in malignancies has been considered as one of the most important prognostic factors. The 8th version of the TNM staging system of NSCLC divided metastasis to intrapulmonary LNs and ipsilateral peribronchial and/or hilar LNs as N1 classification, and metastasis into ipsilateral mediastinal and/or subcarinal LNs into N2 classification without taking into account numbers of examined and metastatic lymph nodes (46). LODDS is a new LN ratio-based index and has been reported to be as an independent predictor in many malignancies such as rectal cancer (47), pancreatic cancer (48), gallbladder cancer (49), gastric cancer (50), and colon cancer (51). Recently, several studies were attempting to explore the prognostic value of LODDS for NSCLC. In 2020, we did a research and found that the high value of LODDS>-0.37 was independently related to worse CSS/OS in patients with node-positive lung squamous cell cancer (25). Dziedzic et al. (24) found that it is possible to discriminate NSCLC patients more effectively by using LODDS compared to conventional N classification. Deng et al. (23) found LODDS and lymph node ratio (LNR) staging schemes outperformed NPLN for predicting CSS/OS among patients with node-positive NSCLC. However, most previous studies only focused on NSCLC, and few reports have detected the predictive value of LODDS in LNETs.

In the present study, the high value of LODDS was associated with worse survival for N1/N2 stage patients with LNETs. However, LODDS must be used and calculated with caution, because the value of LODDS is influenced by the number of dissected LN. Therefore, sensitivity analyses were performed. For LNETs, the extent of LN management should conform to the International Association for the Study of Lung Cancer (IASLC) recommendations, which suggested examination of at least 6 nodes/stations, 3 of which should be mediastinal including the subcarinal station (52). Considering that the accurate value of LODDS was dependent on the adequate NDLN, we excluded patients with examined LN count less than 6 and found that LODDS could still serve as an independent predictor for LNETs. What's more, it is quite easy to break the integrity of the LN during surgery, especially when the LNs are adherent to one another or challenging to be removed from the dissected tissues (53). To avoid the potential bias led by fragmented LNs, we excluded patients with examined LN count of more than 20, and the results did not change. In summary, LODDS is the ratio-based LN staging system that combines NPLN and NDLN, which might be superior to some number-based LN assessment methods. Furthermore, in LODDS, the numerator and denominator are both added with a value of 0.5, eliminating the singularities caused by null data, therefore LODDS might be used to estimate survival of node-negative patients, as opposed to LNR.

The LODDS was not the only prognostic factor included in our nomograms. Similar to previous studies (54–56), age at diagnosis, sex, histotype, surgery, and radiotherapy were chosen as prognostic factors of CSS or OS. In this study, the nomograms

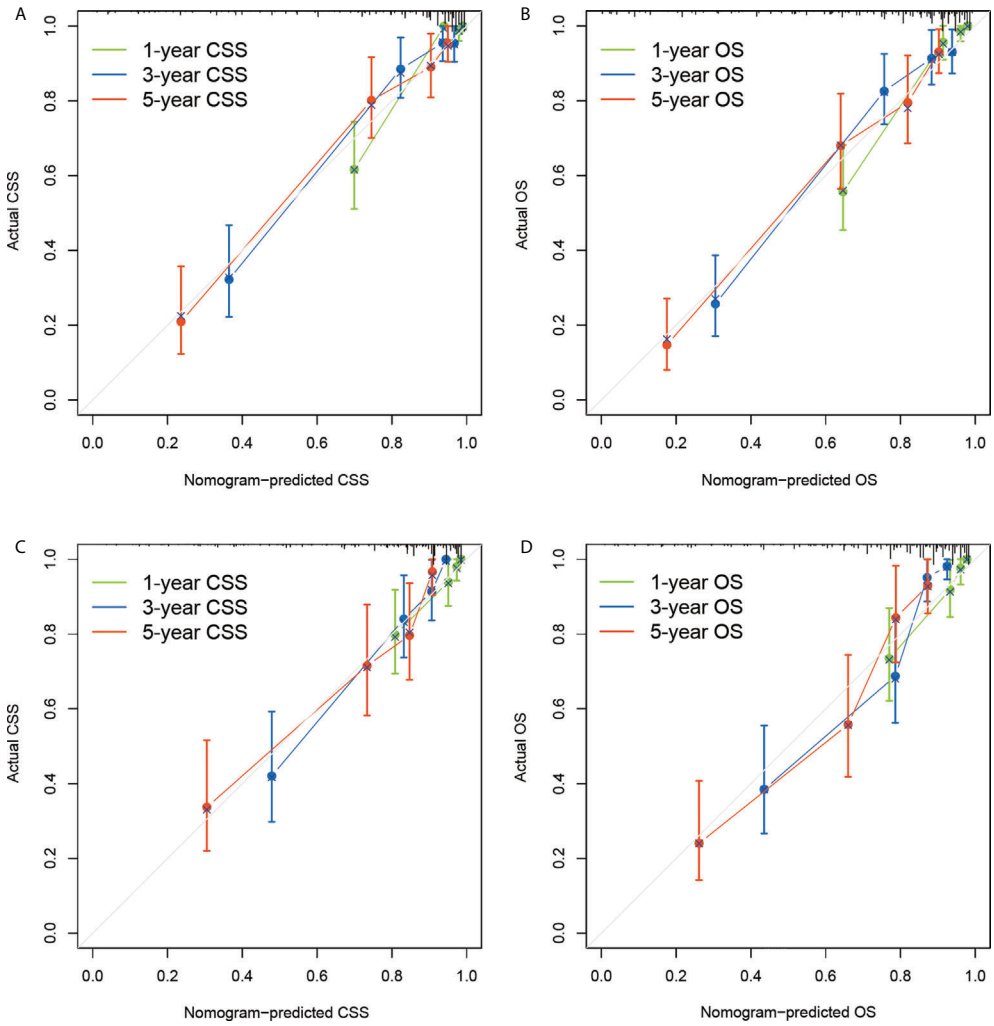


FIGURE 3
Calibration plots of the nomograms to predict CSS and OS of the derivation dataset **(A, B)** and external validation dataset **(C, D)**. CSS, lung cancer-specific survival; OS, overall survival.

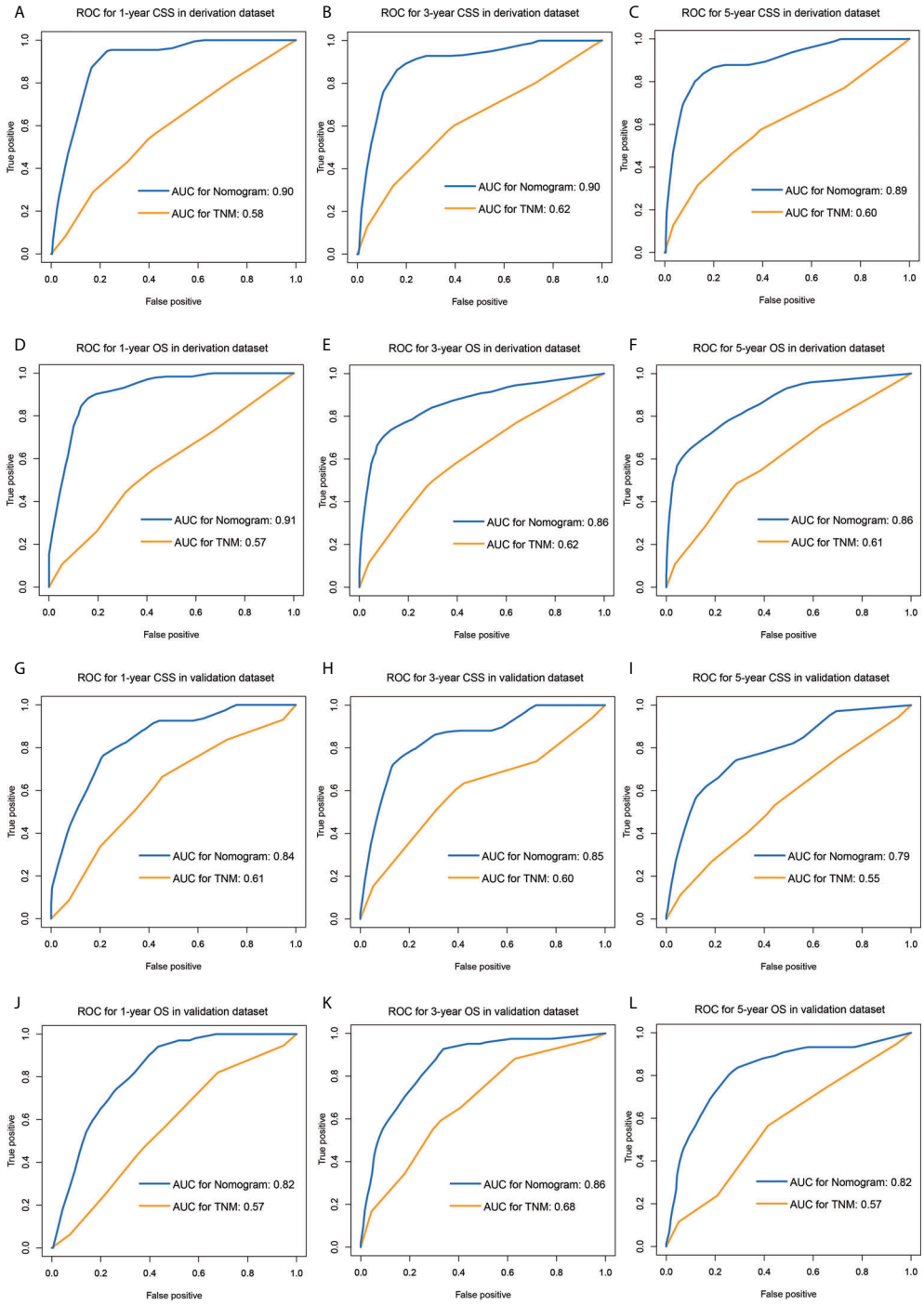


FIGURE 4
ROC curves of the nomograms and TNM staging system for predicting 1-, 3-, 5-year CSS in the derivation dataset **(A–C)** and external validation dataset **(D–F)**, and 1-, 3-, 5-year OS in the derivation dataset **(G–I)** and external validation dataset **(J–L)**. ROC, receiver operating characteristic; CSS, lung cancer-specific survival; OS, overall survival.

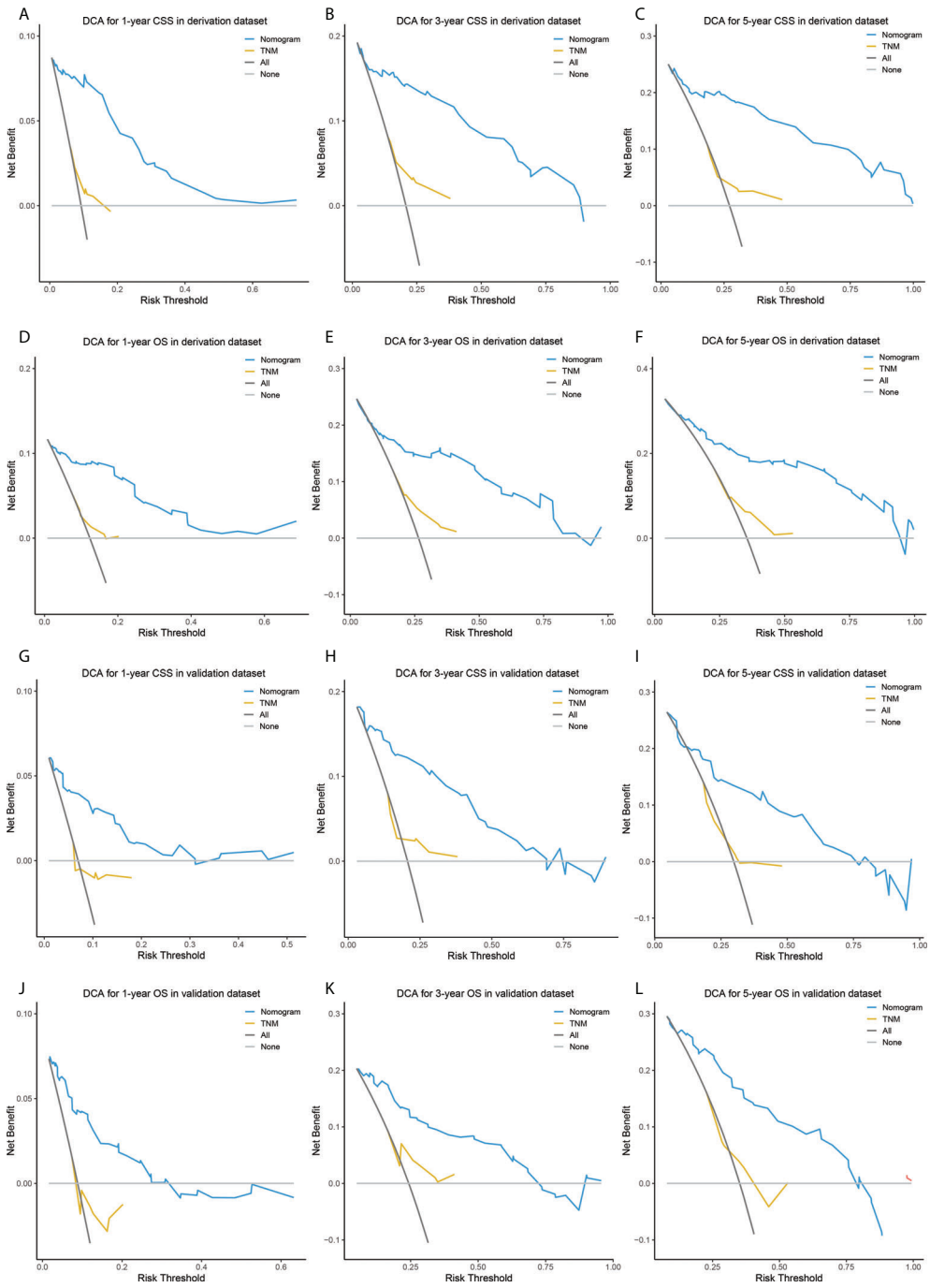


FIGURE 5
DCA of the nomograms and TNM staging system for predicting 1-, 3-, 5-year CSS in the derivation dataset **(A–C)** and external validation dataset **(D–F)**, and 1-, 3-, 5-year OS in the derivation dataset **(G–I)** and external validation dataset **(J–L)**. DCA, decision curve analysis; CSS, lung cancer-specific survival; OS, overall survival.

showed that histotype contributed the most to the prognosis, which indicated tumor histotype is a crucial determinant of the clinical behavior of LNETs. Our nomograms indicated that LCNEC showed the worst prognosis followed by AC and TC. Growing evidence also suggests that high-grade LCNEC is biologically distinct from low-grade TC and intermediate-grade AC in view of clinical behavior, pathologic features, molecular alterations as well as possible precursor lesions (57).

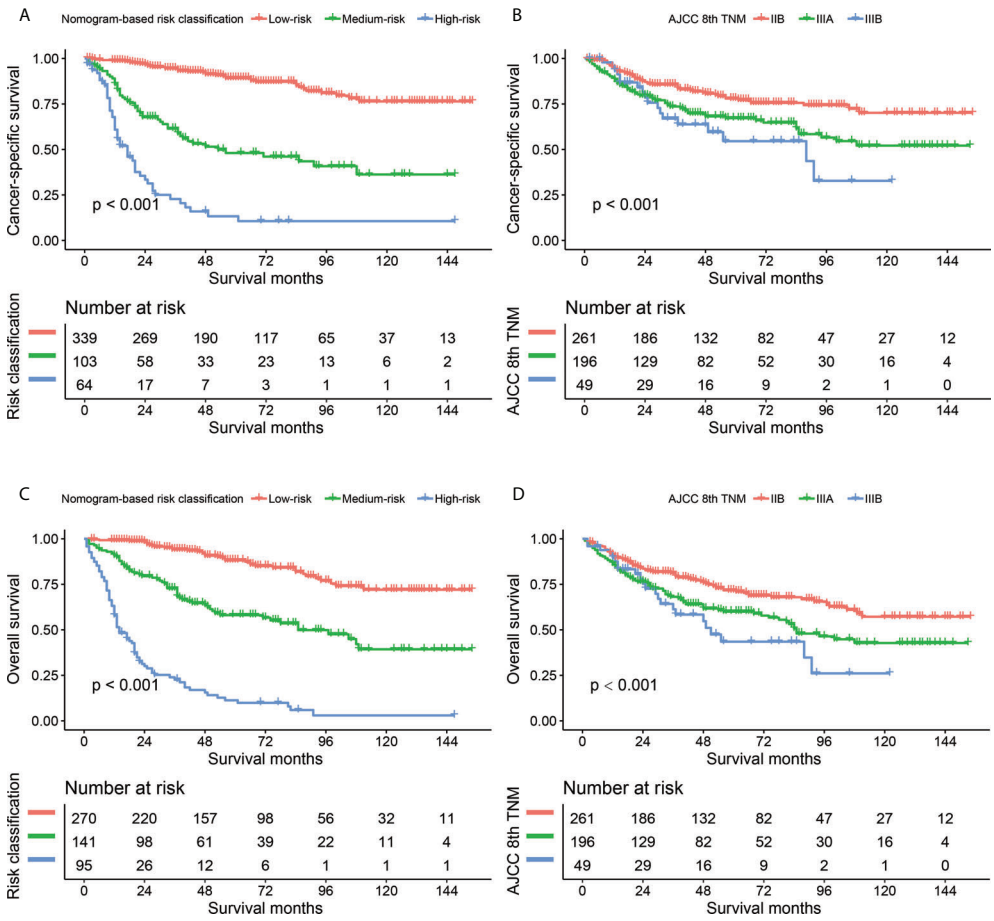


FIGURE 6
Kaplan-Maier curves comparing nomogram-based classification with 8th AJCC TNM staging system in CSS **(A, B)** and OS **(C, D)** prediction. CSS, cancer-specific survival; OS, overall survival.

All-stage 5-year OS for LCNEC fluctuated between 13% and 57% (58). Different from LCNEC, TC and AC are more commonly found in younger patients without smoking histories. AC is significantly more aggressive than TC, with a higher frequency of nodal and distant metastases, and 5-year survival of 60%. In this study, we did not include SCLC patients, because SCLC is usually deemed as nonsurgical (59).

The optimal surgical choice for LNETs is controversial. The surgical type is associated with tumor site, tumor size, and precise assessment of preoperative biopsy specimen. Several studies reported that wedge resection might be correlated with the increased tumor recurrence rate especially in node-positive TC or AC (60, 61). Lobectomy is reported as superior to segmentectomy in terms of OS in some, but not in all pulmonary carcinoids (62–64). Similar to these studies, our study showed that lobectomy was superior to sublobectomy and pneumonectomy in the nomogram for CSS. Furthermore, there is an absence of high-quality evidence to show whether or not chemotherapy could provide clinical benefits for LNETs. Although our nomogram showed that chemotherapy might be associated with more favorable prognoses. However, the results need to be interpreted with caution. In the SEER database, patients without receiving chemotherapy and those with unknown information about adjuvant therapy were classified into one category, which might lead to potential bias. Until now, for pulmonary carcinoids, routine adjuvant therapy may only be considered in selected fit patients (AC, N2 stage) with a particularly high risk of relapse (65). Besides, Iyoda et al. (66) suggested that platinum-based adjuvant chemotherapy after surgery could help patients with LCNEC prevent recurrence.

Because LNET is a heterogeneous disease, each LNET patient requires an individualized and timely risk assessment, which allows for more precise therapeutic strategies and medical resource allocation decisions. In this study, we developed and validated two nomograms to predict prognosis in patients with LNET. Our nomograms based on LODDS were more accurate

and obtained more clinical net benefit than the conventional AJCC/UICC TNM staging system. In summary, the online nomograms, composed of several easily obtained predictors, could be a simpler way to engage clinicians in death risks, patient counseling, and decision-making. To put it another way, LNET patients with poorer clinical results estimated by nomograms may require more aggressive therapy (39).

Several limitations of this study should be noted. First, the SEER database lacks some detailed data, such as smoking history, some promising molecular markers (e.g. Ki-67), imaging techniques used before surgery, histological and morphological data (e.g. mitotic rate), type of resection (R0, R1, or R2), and use of systemic therapies. Therefore, they could not be included as covariables in the multivariable Cox models. Second, information about comorbidities, and tumor recurrence is also not available in the database. Third, although we reclassified the TNM classification according to the eighth edition of AJCC/UICC TNM classification, the TNM staging system, which was derived from the SEER database's collaborative stage system, is a combination of clinical and pathologic stages. Because of the distinction between clinical and pathologic stages, more subgroup analysis is required using a single clinical or pathologic staging system.

Conclusions

LODDS was found to be useful in predicting CSS/OS in LNET patients who underwent surgery. Webservers of nomograms including LODDS to assess CSS and OS were established. The well-executed nomograms may aid clinicians in providing reasonable, customized therapeutic strategies for LNET patients.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

Conception/Design: SW, YY, and ML. Collection and/or assembly of data: JW, YG and ML. Data analysis and interpretation: SW and QX. Manuscript writing: YY, XL, and QX. Final approval of manuscript: All authors. Funding support: XL. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by Program of Shanghai Academic Research Leader (No. 21XD1402800) and Shanghai "Rising Stars of Medical Talent" Youth Development Program: Outstanding Youth Medical Talents.

Acknowledgments

The authors would like to thank all patients and staff who have participated in the SEER program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.987881/full#supplementary-material

References

1. Hendifar AE, Marchevsky AM, Tuli R. Neuroendocrine tumors of the lung: Current challenges and advances in the diagnosis and management of well-differentiated disease. *J Thorac Oncol* (2017) 12(3):425–36. doi: 10.1016/j.jtho.2016.11.2222

2. Hemminki K, Li X. Incidence trends and risk factors of carcinoid tumors: a nationwide epidemiologic study from Sweden. *Cancer* (2001) 92(8):2204–10. doi: 10.1002/1097-0142(20011015)92:8<2204::AID-CNCR1564>3.0.CO;2-R

3. Quaedvlieg PF, Visser O, Lamers CB, Janssen-Heijen ML, Taal BG. Epidemiology and survival in patients with carcinoid disease in the netherlands. an epidemiological study with 2391 patients. *Ann Oncol* (2001) 12(9):1295–300. doi: 10.1023/A:1012272314550

4. Hauso O, Gustafsson BI, Kidd M, Waldum HL, Drozdov I, Chan AK, et al. Neuroendocrine tumor epidemiology: contrasting Norway and north America. *Cancer* (2008) 113(10):2655–64. doi: 10.1002/cncr.23883

5. Dasari A, Shen C, Halperin D, Zhao B, Zhou S, Xu Y, et al. Trends in the incidence, prevalence, and survival outcomes in patients with neuroendocrine tumors in the united states. *JAMA Oncol* (2017) 3(10):1335–42. doi: 10.1001/jamaoncol.2017.0589

6. Hallet J, Law CH, Cukier M, Saskin R, Liu N, Singh S. Exploring the rising incidence of neuroendocrine tumors: a population-based analysis of epidemiology, metastatic presentation, and outcomes. *Cancer* (2015) 121(4):589–97. doi: 10.1002/cncr.29099

7. Naalsund A, Rostad H, Strøm EH, Lund MB, Strand TE. Carcinoid lung tumors–incidence, treatment and outcomes: a population-based study. *Eur J Cardio-thoracic Surg* (2011) 39(4):565–9. doi: 10.1016/j.ejcts.2010.08.036

8. Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 world health organization classification of lung tumors: Impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol* (2015) 10(9):1243–60. doi: 10.1097/JTO.0000000000000630

9. Kumar V, Yadavilli S, Kannan R. A review on RNAi therapy for NSCLC: Opportunities and challenges. *Wiley Interdiscip Rev Nanomed Nanobiotechnol* (2021) 13(2):e1677. doi: 10.1002/wnan.1677

10. Shah S, Gosain R, Groman A, Gosain R, Dasari A, Halfdanarson TR, et al. Incidence and survival outcomes in patients with lung neuroendocrine neoplasms in the united states. *Cancers (Basel)* (2021) 13(8):1753. doi: 10.3390/cancers13081753

11. Yoon JY, Sigel K, Martin J, Jordan R, Beasley MB, Smith C, et al. Evaluation of the prognostic significance of TNM staging guidelines in lung carcinoid tumors. *J Thorac Oncol* (2019) 14(2):184–92. doi: 10.1016/j.jtho.2018.10.166

12. Yeh YC, Chou TY. Pulmonary neuroendocrine tumors: study of 90 cases focusing on clinicopathological characteristics, immunophenotype, preoperative biopsy, and frozen section diagnoses. *J Surg Oncol* (2014) 109(3):280–6. doi: 10.1002/jso.23497

13. Robelin P, Hadoux J, Forestier J, Planchard D, Hervieu V, Berdelou A, et al. Characterization, prognosis, and treatment of patients with metastatic lung carcinoid tumors. *J Thorac Oncol* (2019) 14(6):993–1002. doi: 10.1016/j.jtho.2019.02.002

14. Filosso PL, Oliaro A, Ruffini E, Bora G, Lyberis P, Asioli S, et al. Outcome and prognostic factors in bronchial carcinoids: a single-center experience. *J Thorac Oncol* (2013) 8(10):1282–8. doi: 10.1097/JTO.0b013e31829f097a

15. Aydin E, Yazici U, Gulgosteren M, Agackiran Y, Kaya S, Gulhan E, et al. Long-term outcomes and prognostic factors of patients with surgically treated pulmonary carcinoid: our institutional experience with 104 patients. *Eur J Cardio-thoracic Surg: Off J Eur Assoc Cardio-thoracic Surg* (2011) 39(4):549–54. doi: 10.1016/j.ejcts.2010.08.010

16. Kodera Y, Yamamura Y, Shimizu Y, Torii A, Hirai T, Yasui K, et al. The number of metastatic lymph nodes: a promising prognostic determinant for gastric carcinoma in the latest edition of the TNM classification. *J Am Coll Surgeon* (1998) 187(6):597–603. doi: 10.1016/S1072-7515(98)00229-4

17. Sinn HP, Helmchen B, Wittekind CH. TNM classification of breast cancer: changes and comments on the 7th edition. *Der Pathologe* (2010) 31(5):361–6. doi: 10.1007/s00292-010-1307-0

18. Gallegos-Hernandez JF, Hernandez-Hernandez DM, Flores-Diaz R, Sierra-Santiesteban I, Pichardo-Romero P, Arias-Ceballos H, et al. The number of sentinel nodes identified as prognostic factor in oral epidermoid cancer. *Oral Oncol* (2005) 41(9):947–52. doi: 10.1016/j.oraloncology.2005.05.010

19. Yang L, Xiong Z, Xie Q, He W, Liu S, Kong P, et al. Prognostic value of total number of lymph nodes retrieved differs between left-sided colon cancer and right-sided colon cancer in stage III patients with colon cancer. *BMC Canc* (2018) 18 (1):558. doi: 10.1186/s12885-018-4431-5

20. Chu X, Yang ZF. Impact on survival of the number of lymph nodes resected in patients with lymph node-negative gastric cancer. *World J Surg Oncol* (2015) 13:192. doi: 10.1186/s12957-015-0602-x

21. Tsim S, O'Dowd CA, Milroy R, Davidson S. Staging of non-small cell lung cancer (NSCLC): A review. *Respir Med* (2010) 104(12):1767–74. doi: 10.1016/j.rmed.2010.08.005

22. Lv P, Chen G, Zhang P. Log odds of positive lymph nodes are superior to other measures for evaluating the prognosis of non-small cell lung cancer. *Thorac Canc* (2014) 5(6):570–5. doi: 10.1111/1759-7714.12145

23. Deng W, Xu T, Wang Y, Xu Y, Yang P, Gomez D, et al. Log odds of positive lymph nodes may predict survival benefit in patients with node-positive non-small cell lung cancer. *Lung Canc* (2018) 122:60–6. doi: 10.1016/j.lungcan.2018.05.016

24. Dziedzic D, Piotr R, Langfort R, Orlowski T, Polish Lung Cancer Study G. Log odds of positive lymph nodes as a novel prognostic indicator in NSCLC staging. *Surg Oncol* (2017) 26(1):80–5. doi: 10.1016/j.suronc.2017.01.004

25. Yu Y, Zhang P, Yao R, Wang J, Wang P, Xue X, et al. Prognostic value of log odds of positive lymph nodes in node-positive lung squamous cell carcinoma patients after surgery: a SEER population-based study. *Trans Lung Cancer Res* (2020) 9(4):1285–301. doi: 10.21037/tlcr-20-193

26. Wang S, Yu Y, Xu W, Lv X, Zhang Y, Liu M. Dynamic nomograms combining n classification with ratio-based nodal classifications to predict long-term survival for patients with lung adenocarcinoma after surgery: a SEER population-based study. *BMC Canc* (2021) 21(1):653. doi: 10.1186/s12885-021-08410-6

27. Wang Q, Wang S, Sun Z, Cao M, Zhao X. Evaluation of log odds of positive lymph nodes in predicting the survival of patients with non-small cell lung cancer treated with neoadjuvant therapy and surgery: a SEER cohort-based study. *BMC Canc* (2022) 22(1):801. doi: 10.1186/s12885-022-09908-3

28. Huang Y, Liang C, Liu Z. Nomogram for predicting pulmonary hypertension in patients without pulmonary embolism. *Radiology* (2016) 280 (1):327–8. doi: 10.1148/radiol.2016152584

29. Eisner BH, Goldfarb DS. A nomogram for the prediction of kidney stone recurrence. *J Am Soc Nephrol* (2014) 25(12):2685–7. doi: 10.1681/ASN.2014060631

30. Li W, Dong S, Wang H, Wu R, Wu H, Tang ZR, et al. Risk analysis of pulmonary metastasis of chondrosarcoma by establishing and validating a new clinical prediction model: a clinical study based on SEER database. *BMC Musculoskelet Disord* (2021) 22(1):529. doi: 10.1186/s12891-021-04414-2

31. Li W, Wang H, Dong S, Tang ZR, Chen L, Cai X, et al. Establishment and validation of a nomogram and web calculator for the risk of new vertebral compression fractures and cement leakage after percutaneous vertebroplasty in patients with osteoporotic vertebral compression fractures. *Eur Spine J* (2022) 31 (5):1108–21. doi: 10.1007/s00586-021-07064-z

32. Wang S, Chen A, Zhu W, Feng D, Wei J, Li Q, et al. Characterization of fatty acid metabolism in lung adenocarcinoma. *Front Genet* (2022) 13:905508. doi: 10.3389/fgene.2022.905508

33. Yu Y, Huang R, Wang P, Wang S, Ling X, Zhang P, et al. Sublobectomy versus lobectomy for long-term survival outcomes of early-stage non-small cell lung cancer with a tumor size ≤2 cm accompanied by visceral pleural invasion: a SEER population-based study. *J Thorac Dis* (2020) 12(3):592–604. doi: 10.21037/jtd.2019.12.121

34. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ (Clinical Res ed)* (2015) 350:g7594. doi: 10.1161/CIRCULATIONAHA.114.014508

35. Cronin KA, Ries LA, Edwards BK. The surveillance, epidemiology, and end results (SEER) program of the national cancer institute. *Cancer* (2014) 120 Suppl 23:3755–7. doi: 10.1002/cncr.29049

36. Goodyear MD, Krleza-Jeric K, Lemmens T. The declaration of Helsinki. *BMJ (Clinical Res ed)* (2007) 335(7621):624–5. doi: 10.1136/bmj.39339.610000.BE

37. Detterbeck FC. The eighth edition TNM stage classification for lung cancer: What does it mean on main street? *J Thorac Cardiovasc Surg* (2018) 155(1):356–9. doi: 10.1016/j.jtcvs.2017.08.138

38. Jackson AS, Rosenthal A, Cattoni M, Bograd AJ, Farivar AS, Aye RW, et al. Staging system for neuroendocrine tumors of the lung needs to incorporate histologic grade. *Ann Thorac Surg* (2020) 109(4):1009–18. doi: 10.1016/j.athoracsur.2019.09.053

39. Yi C, Dai J, Song N, Wu C, Zhang L, Zhu Y, et al. Improvement of pathological staging system for neuroendocrine tumors of the lung. *Ann Transl Med* (2021) 9(6):447. doi: 10.21037/atm-20-5910

40. Iyoda A, Makino T, Koezuka S, Otsuka H, Hata Y. Treatment options for patients with large cell neuroendocrine carcinoma of the lung. *Gen Thorac Cardiovasc Surg* (2014) 62(6):351–6. doi: 10.1007/s11748-014-0379-9

41. Iyoda A, Azuma Y, Sano A. Neuroendocrine tumors of the lung: clinicopathological and molecular features. *Surg Today* (2020) 50(12):1578–84. doi: 10.1007/s00595-020-01988-7

42. Raman V, Jawitz OK, Yang CJ, Voigt SL, Tong BC, D'Amico TA, et al. Outcomes for surgery in Large cell lung neuroendocrine cancer. *J Thorac Oncol* (2019) 14(12):2143–51. doi: 10.1016/j.jtho.2019.09.005

43. Camp RL, Dolled-Filhart M, Rimm DL. X-Tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res* (2004) 10(21):7252–9. doi: 10.1158/1078-0432.CCR-04-0713

44. Jungk C, Reinhardt A, Warta R, Capper D, Deimling AV, Herold-Mende C, et al. Extent of resection, MGMT promoter methylation status and tumor location independently predict progression-free survival in adult sporadic pilocytic astrocytoma. *Cancers (Basel)* (2019) 11(8):1072. doi: 10.3390/cancers11081072

45. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Dec Mak* (2006) 26(6):565–74. doi: 10.1177/0272989X06295361

46. Taylor MD, LaPar DJ, Thomas CJ, Persinger M, Stelow EB, Kozower BD, et al. Lymph node ratio predicts recurrence and survival after R0 resection for non-small cell lung cancer. *Ann Thorac Surg* (2013) 96(4):1163–70. doi: 10.1016/j.athoracsur.2013.04.031

47. Huang B, Ni M, Chen C, Cai G, Cai S. LODDS is superior to lymph node ratio for the prognosis of node-positive rectal cancer patients treated with preoperative radiotherapy. *Tumori* (2017) 103(1):87–92. doi: 10.5301/tj.5000560

48. Ramacciato G, Nigri G, Petrucciani N, Pinna AD, Ravaioli M, Jovine E, et al. Prognostic role of nodal ratio, LODDS, pN in patients with pancreatic cancer with venous involvement. *BMC Surg* (2017) 17(1):109. doi: 10.1186/s12893-017-0311-1

49. Xiao Z, Shi Z, Hu L, Gao Y, Zhao J, Liu Y, et al. A new nomogram from the SEER database for predicting the prognosis of gallbladder cancer patients after surgery. *Ann Trans Med* (2019) 7(23):738. doi: 10.21037/atm.2019.11.112

50. Liu H, Deng J, Zhang R, Hao X, Jiao X, Liang H. The RML of lymph node metastasis was superior to the LODDS for evaluating the prognosis of gastric cancer. *Int J Surg (London England)* (2013) 11(5):419–24. doi: 10.1016/j.ijsu.2013.03.009

51. Fortea-Sanchis C, Martínez-Ramos D, Escrig-Sos J. The lymph node status as a prognostic factor in colon cancer: comparative population study of classifications using the logarithm of the ratio between metastatic and nonmetastatic nodes (LODDS) versus the pN-TNM classification and ganglion ratio systems. *BMC Canc* (2018) 18(1):1208. doi: 10.1186/s12885-018-5048-4

52. Caplin ME, Baudin E, Ferolla P, Filosso P, Garcia-Yuste M, Lim E, et al. Pulmonary neuroendocrine (carcinoid) tumors: European neuroendocrine tumor society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann Oncol* (2015) 26(8):1604–20. doi: 10.1093/annonc/mdv041

53. Dai J, Liu M, Yang Y, Li Q, Song N, Rocco G, et al. Optimal lymph node examination and adjuvant chemotherapy for stage I lung cancer. *J Thorac Oncol* (2019) 14(7):1277–85. doi: 10.1016/j.jtho.2019.03.027

54. He Y, Zhao F, Han Q, Zhou Y, Zhao S. Prognostic nomogram for predicting long-term cancer-specific survival in patients with lung carcinoid tumors. *BMC Canc* (2021) 21(1):141. doi: 10.1186/s12885-021-07832-6

55. Dong S, Liang J, Zhai W, Yu Z. Development and validation of an individualized nomogram for predicting overall survival in patients with typical lung carcinoid tumors. *Am J Clin Oncol* (2020) 43(9):607–14. doi: 10.1097/COC.0000000000000715

56. Chen X, Pang Z, Wang Y, Bie F, Zeng Y, Wang G, et al. The role of surgery for atypical bronchopulmonary carcinoid tumor: Development and validation of a model based on surveillance, epidemiology, and end results (SEER) database. *Lung Canc* (2020) 139:94–102. doi: 10.1016/j.lungcan.2019.11.006

57. Hiroshima K, Mino-Kenudson M. Update on large cell neuroendocrine carcinoma. *Trans Lung Cancer Res* (2017) 6(5):530–9. doi: 10.21037/tlcr.2017.06.12

58. Rossi G, Cavazza A, Marchioni A, Longo L, Migaldi M, Sartori G, et al. Role of chemotherapy and the receptor tyrosine kinases KIT, PDGFRalpha, PDGFRbeta, and met in large-cell neuroendocrine carcinoma of the lung. *J Clin Oncol* (2005) 23(34):8774–85. doi: 10.1200/JCO.2005.02.8233

59. Lim E, Goldstraw P, Nicholson AG, Travis WD, Jett JR, Ferolla P, et al. Proceedings of the IASLC international workshop on advances in pulmonary neuroendocrine tumors 2007. *J Thorac Oncol* (2008) 3(10):1194–201. doi: 10.1097/JTO.0b013e3181861d7b

60. García-Yuste M, Matilla JM, Cañizares MA, Molins L, Guijarro R. Surgical treatment of low and intermediate grade lung net. *J Thorac Dis* (2017) 9(Suppl 15):S1435–s41. doi: 10.21037/jtd.2017.09.83

61. García-Yuste M, Matilla JM, Cueto A, Paniagua JM, Ramos G, Cañizares MA, et al. Typical and atypical carcinoid tumours: analysis of the experience of the Spanish multi-centric study of neuroendocrine tumours of the lung. *Eur J Cardio-thoracic Surg* (2007) 31(2):192–7. doi: 10.1016/j.ejcts.2006.11.031

62. Filosso PL, Rena O, Guerrera F, Moreno Casado P, Sagan D, Raveglia F, et al. Clinical management of atypical carcinoid and large-cell neuroendocrine carcinoma: a multicentre study on behalf of the European association of thoracic surgeons (ESTS) neuroendocrine tumours of the lung working group†. *Eur J Cardio-thoracic Surg* (2015) 48(1):55–64. doi: 10.1093/ejcts/ezu404

63. Steuer CE, Behera M, Kim S, Chen Z, Saba NF, Pillai RN, et al. Atypical carcinoid tumor of the lung: a surveillance, epidemiology, and end results database analysis. *J Thorac Oncol* (2015) 10(3):479–85. doi: 10.1097/JTO.0000000000000419

64. Marciello F, Mercier O, Ferolla P, Scoazec JY, Filosso PL, Chapelier A, et al. Natural history of localized and locally advanced atypical lung carcinoids after complete resection: A joined French-Italian retrospective multicenter study. *Neuroendocrinology* (2018) 106(3):264–73. doi: 10.1159/000480015

65. Baudin E, Caplin M, Garcia-Carbonero R, Fazio N, Ferolla P, Filosso PL, et al. Lung and thymic carcinoids: ESMO clinical practice guidelines for diagnosis, treatment and follow-up(☆). *Ann Oncol* (2021) 32(4):439–51. doi: 10.1016/j.annonc.2021.01.003

66. Iyoda A, Hiroshima K, Moriya Y, Takiguchi Y, Sekine Y, Shibuya K, et al. Prospective study of adjuvant chemotherapy for pulmonary large cell neuroendocrine carcinoma. *Ann Thorac Surg* (2006) 82(5):1802–7. doi: 10.1016/j.athoracsur.2006.05.109

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 30 September 2022
DOI 10.3389/fimmu.2022.940774

Check for updates

OPEN ACCESS

EDITED BY
Jinzhuang Dou,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Liang Zhang,
University of Texas MD Anderson Cancer Center, United States
Samantha Sharma,
Indiana University Bloomington, United States
Guang Hu,
Soochow University, China

*CORRESPONDENCE
Ye Xu
yexu@shmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 10 May 2022
ACCEPTED 14 September 2022
PUBLISHED 30 September 2022

CITATION
Zhu Z, Zhao Q, Song W, Weng J, Li S, Guo T, Zhu C and Xu Y (2022) A novel cuproptosis-related molecular pattern and its tumor microenvironment characterization in colorectal cancer. *Front. Immunol.* 13:940774. doi: 10.3389/fimmu.2022.940774

COPYRIGHT
© 2022 Zhu, Zhao, Song, Weng, Li, Guo, Zhu and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A novel cuproptosis-related molecular pattern and its tumor microenvironment characterization in colorectal cancer

Zhonglin Zhu[1,2†], Qiuyan Zhao[3,4†], Wang Song[1,2†], Junyong Weng[1,2†], Shanbao Li[5], Tianan Guo[1,2], Congcong Zhu[1,2] and Ye Xu[1,2*]

[1]Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, Shanghai, China, [2]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China, [3]Department of Gastroenterology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [4]Shanghai Key Laboratory of Pancreatic Diseases, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [5]Department of General Surgery, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Cuproptosis, or copper-induced cell death, has been reported as a novel noncanonical form of cell death in recent times. However, the potential roles of cuproptosis in the alteration of tumor clinicopathological features and the formation of a tumor microenvironment (TME) remain unclear. In this study, we comprehensively analyzed the cuproptosis-related molecular patterns of 1,274 colorectal cancer samples based on 16 cuproptosis regulators. The consensus clustering algorithm was conducted to identify cuproptosis-related molecular patterns and gene signatures. The ssGSEA and ESTIMATE algorithms were used to evaluate the enrichment levels of the infiltrated immune cells and tumor immune scores, respectively. The cuproptosis score was established to assess the cuproptosis patterns of individuals with principal component analysis algorithms based on the expression of cuproptosis-related genes. Three distinct cuproptosis patterns were confirmed and demonstrated to be associated with distinguishable biological processes and clinical prognosis. Interestingly, the three cuproptosis patterns were revealed to be consistent with three immune infiltration characterizations: immune-desert, immune-inflamed, and immune-excluded. Enhanced survival, activation of immune cells, and high tumor purity were presented in patients with low cuproptosisScore, implicating the immune-inflamed phenotype. In addition, low scores were linked to high tumor mutation burden, MSI-H and high CTLA4 expression, showing a higher immune cell proportion score (IPS). Taken together, our study revealed a novel cuproptosis-related molecular pattern

associated with the TME phenotype. The formation of cuproptosisScore will further strengthen our understanding of the TME feature and instruct a more personalized immunotherapy schedule in colorectal cancer.

KEYWORDS

cuproptosis, tumor microenvironment, immunotherapy, colorectal cancer, molecular subtype

Introduction

Colorectal cancer is the third most common cancer and the second most deadly cancer worldwide (1, 2). With the spread of multiple-disciplinary treatment, the death rate gradually declined from declining 3% per year during the 2000s to declining 1.8% per year from 2012 to 2017 (3). The overall 5-year survival of colorectal cancer has been more than 60% in recent years. However, it decreases to 14% for patients with distant metastases (3). Therefore, in-depth understanding of the multiple tumor features and identifying effective prognostic indicators contribute to constructing a more significant therapy schedule for the individual person.

Copper is an essential trace element for eukaryotes. It is involved in numerous fundamental biological processes, such as iron transport, oxygen radical detoxification, and mitochondrial respiration (4). Intracellular copper concentration is in a state of dynamic balance based on the across homeostatic gradients, in which the dynamic signaling influences a diverse number of cellular processes including lipolysis, proliferation, and autophagy (5–9). Dancis et al. (10) and Knight et al. (11) discovered the copper transport proteins 1–5 (Ctr1–5), especially Ctr1 (SLC31A1), with the strongest copper transport capacity. In addition, the excretion of copper irons is mediated by P-type ATPases (ATP7A, ATP7B), whose N-terminal possesses metal binding sites (12). Due to the dysregulation of transmembrane transport of copper, the accumulation of intracellular copper results in cytotoxicity and cell death (13). Nevertheless, the detailed mechanism of copper-induced cell death remains uncertain. It has been reported that excessive copper elevated the level of intracellular reactive oxygen species (ROS), induced endoplasmic reticulum (ER) stress, enhanced damage-associated molecular patterns (DAMPs), and promoted phagocytosis by macrophages (14). In recent, a novel mechanism of copper-induced cell death has been uncovered on *Science*: copper directly binds to lipoylated proteins of mitochondrial tricarboxylic acid (TCA) cycle, then lipoylated protein aggregation, loss of Fe–S cluster and induction of HSP70 lead to proteotoxic stress and cell death (15). The ancient protein lipoylation mechanism of copper homeostasis presents a novel pathway of cell death, termed cuproptosis, which is distinguishable from other known mechanisms including necrosis (16), apoptosis (17), autophagy (18), necroptosis (19), pyrotosis (20), and ferroptosis (21). Also, cuproptosis provides a new sight in disease treatment (22). For Wilson's disease (23) and Menke's disease (24) derived from copper homeostasis dysregulation, less import or more export is an ideal therapy. In several cancer types, copper has been found at higher levels in both serum and tissues compared with those of normal people (25–28). It is evident that intracellular copper accumulation contributes to tumor cell proliferation, angiogenesis, and metastasis (29, 30). However, how to dissect the correlation of cuproptosis with tumor promotion of copper accumulation and how to employ copper toxicity in clinical tumor therapy needs to be further exploited.

Increasing evidence has shown that the tumor microenvironment (TME) is widely involved in tumor development and progression, chemoradiotherapy and immune therapy (31–33). TME is mainly composed of tumor cells and stromal components, of which stromal components consist of residential fibroblasts, endothelial cells, infiltrating immune cells, secreted cytokines and chemokines, and nascent blood and lymphatic vessels. TME is diverse, complex, and plastic toward both phenotypes of tumor promotion and immune escape or tumor suppression and immune enhancement (34–36). For example, macrophages, accounting for the majority of tumor stromal cells, can be molded into classically (M1) or (M2) activated cells in different TMEs. M1 macrophages exert pro-inflammatory and antitumor roles, while M2 macrophages exhibit high levels of anti-inflammatory cytokines and promote cancer cell growth and metastasis (37). Caner-associated fibroblasts (CAFs) are also a main component of TME and play dual roles through matrix deposition and remodeling and extensive crosstalk with cancer cells and infiltrating leukocytes (38). According to the diverse components, TME is widely accepted to be classified into three subtypes: immune-desert, immune-inflamed, and immune-excluded (39). With more and more focus on TME, researchers have found that TME is closely associated with the prognosis of multiple tumors and response to

chemoradiotherapy and immunotherapy, such as colorectal cancer, melanoma, gastric cancer, and intrahepatic cholangiocarcinoma (40–44). In colorectal cancer, the high level of infiltrated cytotoxic $CD8^+$ T cells at the center or margin of the tumor predicts a low risk of recurrence at 5 years (45). Considering the different TME components in an individual patient, there is an urgent need to thoroughly analyze the TME infiltration patterns to administer a better therapy regime.

In this study, we systemically integrated the expression profiles of 1,274 colorectal cancer samples to evaluate the cuproptosis-related molecular patterns. Further analysis confirmed the close relationships between the three cuproptosis patterns and TME infiltration characteristics. Based on the differentially expressed genes, patients were classified into three gene clusters. Finally, a novel cuproptosis score system was constructed to characterize the TME phenotype, which may serve as a biomarker of prognosis evaluation and a target for immunotherapy in colorectal cancer.

Materials and methods

Colorectal cancer data sources

The process of this work is exhibited in Figure S1A. RNA-sequencing data, clinical annotation, and survival time of colon cancer and rectal cancer were downloaded from The Cancer Genome Atlas (TCGA database, https://portal.gdc.cancer.gov/) and the Gene Expression Omnibus (GEO database, https://www.ncbi.nlm.nih.gov/geo/). Transcriptome profiles of 689 samples in the TCGA-colon adenocarcinoma/rectum adenocarcinoma (COAD/READ) were obtained in the format of fragments per kilobase million (FPKM), including 51 normal samples and 638 cancer samples. Then, FPKM values were transformed into transcripts per kilobase million (TPM) for identical analysis with the GEO data (GSE39582, 19 normal samples, and 566 cancer samples) (46). The three datasets were merged with a combat algorithm to correct batch effects with the 'sva' R package. All the data were analyzed with the R program (version 4.1.2).

Unsupervised consensus cluster for cuproptosis regulators

A total of 16 cuproptosis regulators were retrieved, including 13 regulators of the lipoylated TCA cycle pathway (FDX1, LIPT1, LIAS, DLD, MTF1, GLS, CDKN2A, DLAT, PDHA1, PDHB, DBT, GCSH, and DLST) (15) and three copper transport proteins (SLC31A1, ATP7A, and ATP7B) (10–12). Based on their roles in the lipoylated TCA cycle pathway, these regulators were classified into four groups: seven upregulators, three downregulators, three enzymes, and three carriers. According to the expression of these genes, the unsupervised consensus clustering analysis was employed to classify these samples into three distinct molecular patterns with the R package 'ConsensusClusterPlus'.

Gene set variation analysis and functional annotation

GSVA was performed to detect the different biological functions between the distinct cuproptosis clusters with the R package 'GSVA'. The gene sets of hallmark gene sets (v7.5.1) derived from the MSigDB database were used for GSVA analysis (47). Function annotation of a gene list was analyzed by KEGG_Pathway with the Bioconductor package "clusterProfiler" (48) and by the Metascape database (https://metascape.org/).

Three ways of TME cell infiltration analysis

The single-sample gene set enrichment analysis (ssGSEA) was performed to evaluate the relative immune cell infiltration and immune functions in each sample (49). The stromal score and tumor purity of each sample were quantified with the ESTIMATE algorithm. The consensus molecular subtype (CMS) of colorectal cancer was computed with the 'CMScaller' package. CMS1 is classified as a microsatellite unstable and immune activated phenotype, which is characterized by mismatch repair gene mutation and microsatellite instability. CMS4 is a stromal type, consistent with an immune-exclude phenotype. CMS2 is a classic type, with abnormal activation of Wnt and myc signaling pathways and significant variation in somatic copy number. CMS3 is a metabolic type with a high mutation rate of KRAS. Relatively, CMS2 and CMS3 are classified as immune-desert phenotypes.

Differentially expressed genes between the three cuproptosis clusters and establishment of cuproptosis gene signature

To identify cuproptosis-related genes, DEGs between the three cuproptosis clusters were compared in pairs three times with the R package 'limma' (significant criteria, adjusted P-value <0.01). The intersect genes of the three DEGs were next employed to generate cuproptosis gene signatures and to construct a cuproptosis score system to assess the cuproptosis pattern of individuals. First, 965 intersect genes were subjected to GO and KEGG enrichment analysis to explore potential

functions and pathways. Next, univariate Cox regression analysis for each gene was employed to filter genes with significant prognostic correlation. Then, the unsupervised consensus clustering analysis was employed again to classify these samples into distinct cuproptosis gene signature patterns with the R package 'ConsensusClusterPlus'. Finally, principal component analysis (PCA) was carried out to separate the cuproptosis gene signature patterns. This algorithm makes full use of the score on the set with the largest block of well-correlated (or anticorrelated) genes. Also, the algorithm downweights attributions that do not track other set members. Followed by obtaining the prognostic value of each gene, the cuproptosisScore was defined with principal components 1 and 2, similar to the gene expression grade index (GGI) (50, 51): cuproptosisScore = $\Sigma(PC1_i + PC2_i)$. $PC1_i$ and $PC2_i$ represent the expression score of each intersecting gene in two dimensions, respectively.

Somatic mutation, copy number variation, microsatellite instability, and immune cell proportion score analysis

The somatic mutation data of TCGA-COAD/READ was accessed from the TCGA database in varscan file format. The CNV data were downloaded from the UCSC Xena (https://xenabrowser.net/datapages/). The significant mutated genes and tumor mutation burden (TMB) were calculated using the R package 'maftool'. The percentages of microsatellite stability (MSS), high microsatellite instability (MSI-H) and low microsatellite instability (MSI-L) were computed in the different cuporptosis scoring groups. The IPS data were accessed from The Cancer Immunome Altas (https://tcia.at/home). The IPS scores of anti-CTLA4 drugs were compared between the cuporptosis score groups.

Kaplan–Meier survival analysis

The Kaplan–Meier survival curves were plotted and analyzed by the R package 'Survminer'. The samples were stratified into different subgroups with different gene expression, cuproptosisCluster subgroups, geneCluster subgroups, cuproptosisScore, and TMB.

Statistical analyses

All the data analysis was performed using R software (version 4.1.2) and GraphPad Prism (version 9.2). The measurement data were expressed as ($\bar{x} \pm s$), and the t-test was used for the comparison of two groups, while analysis of variance was used for the comparison of more than two groups. The count data were compared between groups by χ^2 test. The statistical significance level was set at a P-value of <0.05.

Results

Landscape of genetic variation of cuproptosis regulators in colorectal cancer

A total of 16 cuproptosis regulators, including seven upregulators, three downregulators, three enzymes, and three carriers, were analyzed in this study. We first summarized the incidences of somatic mutation of 16 cuproptosis regulators in 536 colorectal cancer samples with somatic mutation data. Among these, 79 samples experienced mutations with a frequency of 14.74% (Figure 1A). It was found that the mutation frequency of each regulator is relatively low, of which ATP7A showed the highest mutation rate, followed by ATP7B and LIPT1, while FDX1, CDKN2A, SLC31A1, and GCSH showed no mutation in all colorectal cancer tissues. The CNV analysis showed a copy number loss in most genes, while ATP7B, GLS, PDHA1, and LIPT1 presented a widespread frequency of CNV gain (Figure 1B). The location of CNV of all cuproptosis regulators on chromosomes was exhibited in Figure 1C. To further explore the influence of somatic mutation and CNV on the genomic expression of cuproptosis regulators, mRNA levels of these regulators were investigated in colorectal cancer tissues and normal tissues. The results indicated that the expression level of FDX1, LIAS, DLD, DLAT, PDHA1, MTF1, GLS, DBT, and DLST was lower in colorectal cancer tissues compared with normal tissues (Figure 1D), consistent with the CNV alteration, which implies that CNV may be the dominant effector on the expression alteration of cuproptosis regulators. In addition, we observed that PDHB and CDKN2A with a high frequency of CNV loss showed no significant differences in expression between cancer and normal tissues, and ATP7A with the highest frequency of mutation also showed no difference. Thus, CNV and somatic mutation cannot explain all the mRNA expression alterations, while epigenetic regulation, pre-mRNA splicing, and transcription factors also participate in the regulation of mRNA expression. The above results showed the high heterogeneity of genetic alteration of cuproptosis regulators in colorectal cancer patients.

Identification of cuproptosis patterns in colorectal cancer

To explore the biological functions of these 16 cuproptosis regulators, we performed KEGG enrichment analysis. The

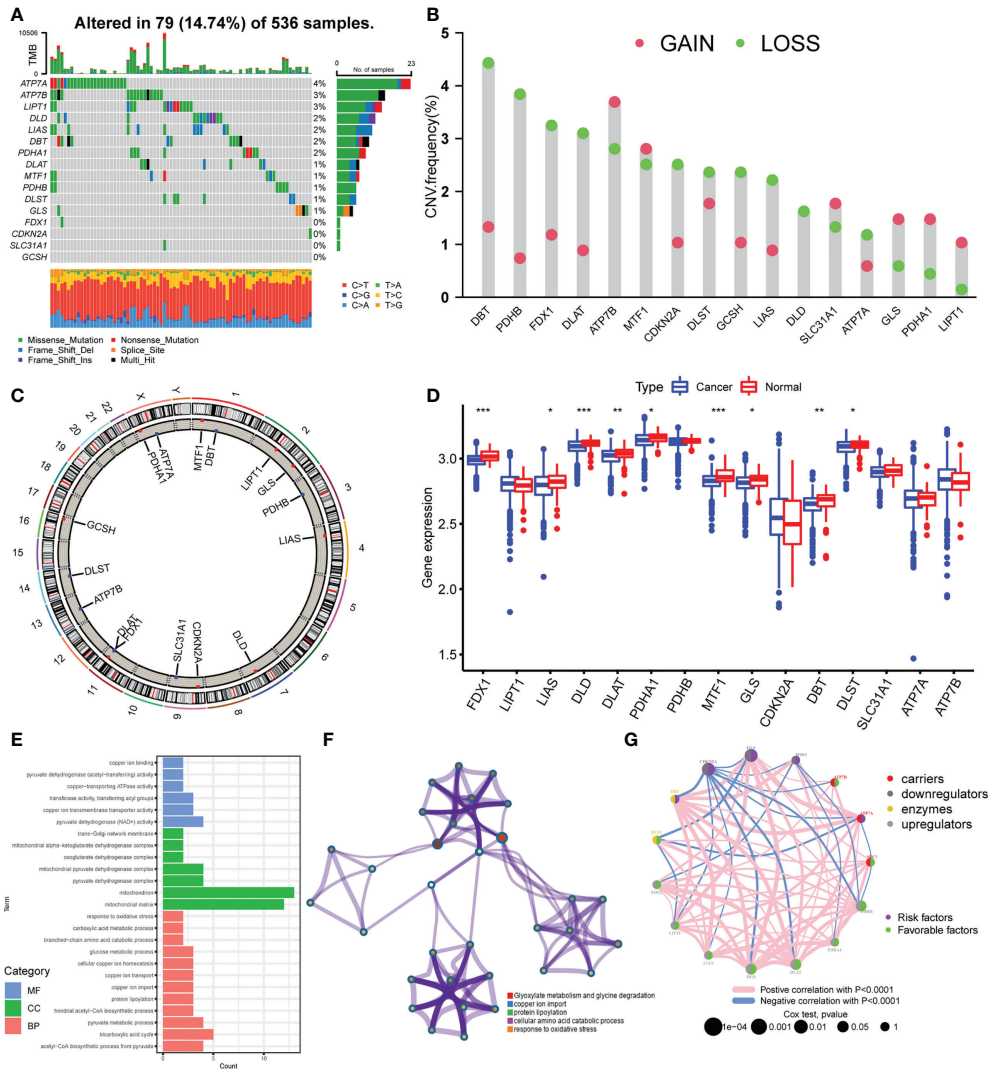


FIGURE 1
Landscape of genetic variation of cuproptosis regulators in colorectal cancer. **(A)** The mutation frequency of 16 cuproptosis regulators in 536 samples from TCGA-COAD/READ. **(B)** The CNV variation frequency of 16 cuproptosis regulators in TCGA-COAD/READ. **(C)** The location diagram of CNV of 16 cuproptosis regulators on 23 chromosomes. **(D)** The expression of 16 cuproptosis regulators in cancer and normal samples of TCGA-COAD/READ. **(E)** Go analysis results of 16 cuproptosis regulators **(F)**. Metascape analysis of 16 cuproptosis regulators **(G)**. The overall landscape of the interaction between cuproptosis regulators and the prognostic significance of the regulators in colorectal cancer patients. *P <0.05; **P <0.01; ***P <0.001.

function annotation results indicated that the cuproptosis regulators participated in cellular copper ion homeostasis, copper transport, protein lipoylation, TCA cycle, and response to oxidative stress (Figure 1E; Supplementary Figure S1B, Supplementary Table S1), consistent with the mechanism of cuproptosis. The metascape analysis also showed the same results (Figure 1F; Supplementary Table S2). Then, we analyzed the interaction correlations of the expression of each regulator (Supplementary Figure S1C). It was found that the expression of ATP7A was positively correlated with the expression of GLS, DBT, and MTF1, while the expression of CDKN2A was negatively correlated with the expression of DBT, DLD, DLAT, and PDHB. Furthermore, a univariate Cox regression model was constructed to analyze the prognostic roles of 16 cuproptosis regulators in colorectal cancer patients (Supplementary Figure S2A, Supplementary Table S3). The Kaplan–Meier curves indicated that patients with high expression of ATP7A, DLAT, DLD, FDX1, LIAS, PDHA1, and PDHB showed a longer overall survival compared with patients with low expression of these genes, respectively. Nonetheless, patients with high expression of CDKN2A and GLS showed an opposite overall survival (Supplementary Figures S2B–J). The

cuproptosis regulators network describes the overall landscape of the interactions between cuproptosis regulators and the prognostic significance of these regulators for colorectal cancer patients (Figure 1G), indicating that the expression of cuproptosis regulators may serve a critical role in the progression and prognosis of colorectal cancer.

We merged the colorectal cancer samples of the TCGA data and GEO (GSE39582) data into one meta-cohort and performed PCA to explore whether the expression of 16 cuproptosis regulators could separate the cancer and normal samples. The scatter diagram showed that normal and cancer samples overlapped (Supplementary Figure S3A). The less normal samples may have led to this result. Then, based on the expression of 16 cuproptosis regulators, we performed an unsupervised consensus clustering algorithm to classify samples with distinct cuproptosis patterns. K from 2 to 9 was conducted and $K = 3$ showed the best results in terms of clustering (Supplementary Figures S3B–D). Three different cuproptosis patterns were finally confirmed. The three patterns were termed CuproptosisClusters A–C, of which 516 cases were in CuproptosisCluster A, 468 cases in CuproptosisCluster B, and 204 cases in CuproptosisCluster C (Supplementary Table S4). As shown in Figure 2A, CuproptosisCluster C has a high expression of CDKN2A, MTF1, DLST, and SLC31A1. CuproptosisCluster A shows a high expression of GLS, ATP7A, LIPT1, LIAS, and DBT, while CuproptosisCluster B exhibits a low expression of ATP7B, PDHB, FDX1, DLAT, DLD, and PDHA1. In addition, we observed that CuproptosisCluster B was preferentially related to the lower stage, and CuproptosisCluster C to the more dead status. The Kaplan–Meier curves showed no significant differences in overall survival between the three CuproptosisClusters, but a rising trend from CuproptosisCluster A to CuproptosisCluster B to CuproptosisCluster C (Supplementary Figure S3E).

TME characteristics in distinct cuproptosis patterns

To appraise the biological functions in the three CuproptosisClusters, GSVA was performed with hallmark gene sets. CuproptosisCluster C was markedly enriched in immune and carcinogenic activation pathways, such as allograft rejection, interferon gamma response, and interferon alpha response (Figure 2B). CuproptosisCluster B was prominently related to stromal activation biological processes: epithelial-to-mesenchymal transition (EMT), myogenesis, and apical junction (Supplementary Figure S4A). However, there was no immune-related pathway enriched in CuproptosisCluster A. The ssGSEA results showed that CuproptosisCluster B was predominately enriched in immunosuppressive cells and innate immune cells including myeloid-derived suppressor cells (MDSCs), macrophage, monocyte, regulatory T cell, Th1 cell, Th17 cell, and dendritic cell, eosinophil, and mast cells compared with CuproptosisClusters A and C, while CuproptosisCluster C was enriched in immunocompetent cells such as activated CD4 T cell and CD8 T cell, natural killer cell, and Gamma delta T cell (Figure 2C). Amounts of research demonstrated that tumors with immune-excluded phenotype always exhibited with stromal activation and significant enrichment of immunosuppressive cells, which could not infiltrate in the tumor parenchyma but surround tumor cell focus. Therefore, we classified CuproptosisCluster B as an immune-excluded phenotype. Next, we examined the stromal activity related pathways, of which genes in EMT and transforming growth factor beta (TGFβ) pathways-ACTA2, CLDN3, COL4A1, SMAD9, TGFBR2, TWIST1, and VIM, were all highly expressed in CuproptosisCluster B (Supplementary Figure S4B), which confirmed the above conclusion. CuproptosisCluster C was remarkably rich in adaptive immune cells, including activated CD8 T cells, natural killer cells, and gamma delta T cells (Figure 2C). To be as it is, CuproptosisCluster C was termed as immune-inflamed phenotype. Similarly, compared with CuproptosisClusters A and B, immune activity-related genes such as CD8A, CXCL10, CXCL9, GZMA, GZMB, IFNG, PRF1, TBX2A, and TNF were more enriched in CuproptosisCluster C (Figure 2D). Then, we adapted the ESTIMATE algorithm to evaluate the stromal score and tumor purity (Supplementary Table S5). To our surprise, the results showed that CuproptosisCluster B was presented with less tumor purity and a higher stromal score compared with CuproptosisClusters A and C (Figure 2E), implying that samples in CuproptosisClusters A and C possessed more tumor parenchyma and CuproptosisCluster B more stromal cell components. In addition, we observed that CuproptosisCluster A presented with high tumor purity, combined with less enrichment of all the immune cells and lower expression levels of immune activity-related genes, indicating the immune-desert phenotype.

To further verify the TME characteristics of the three CuproptosisClusters, a consensus molecular subtype (CMS) of colorectal cancer was computed with the CMScaller package (Supplementary Table S6). At present, the recognized classification of CMS by the academic community is five types as follows: CMS1 is classified as an immune-inflamed phenotype, which is characterized by mismatch repair gene mutation and microsatellite instability. CMS4 is a stromal cell type with abnormal activation of the TGFβ signaling pathway, consistent with an immune-exclude phenotype. Relatively, CMS2 and CMS3 are classified as immune-desert phenotypes. The remaining samples, called mixed phenotypes, cannot be grouped into any single type. The results indicated that CuproptosisCluster A was predominated in CMS2, while CuproptosisCluster B was CMS4 and CuproptosisCluster C was CMS1 (Figure 2F). In addition, the samples of the GSE39582 dataset are divided into six molecular subtypes (Supplementary Table S7) (46). Samples with dMMR are preferably related to immune activation. CINnormal and CSC are closely correlated with EMT, stem feature, and stromal activation. KRASm is prone to being an immune-desert phenotype. Our results indicated that dMMR

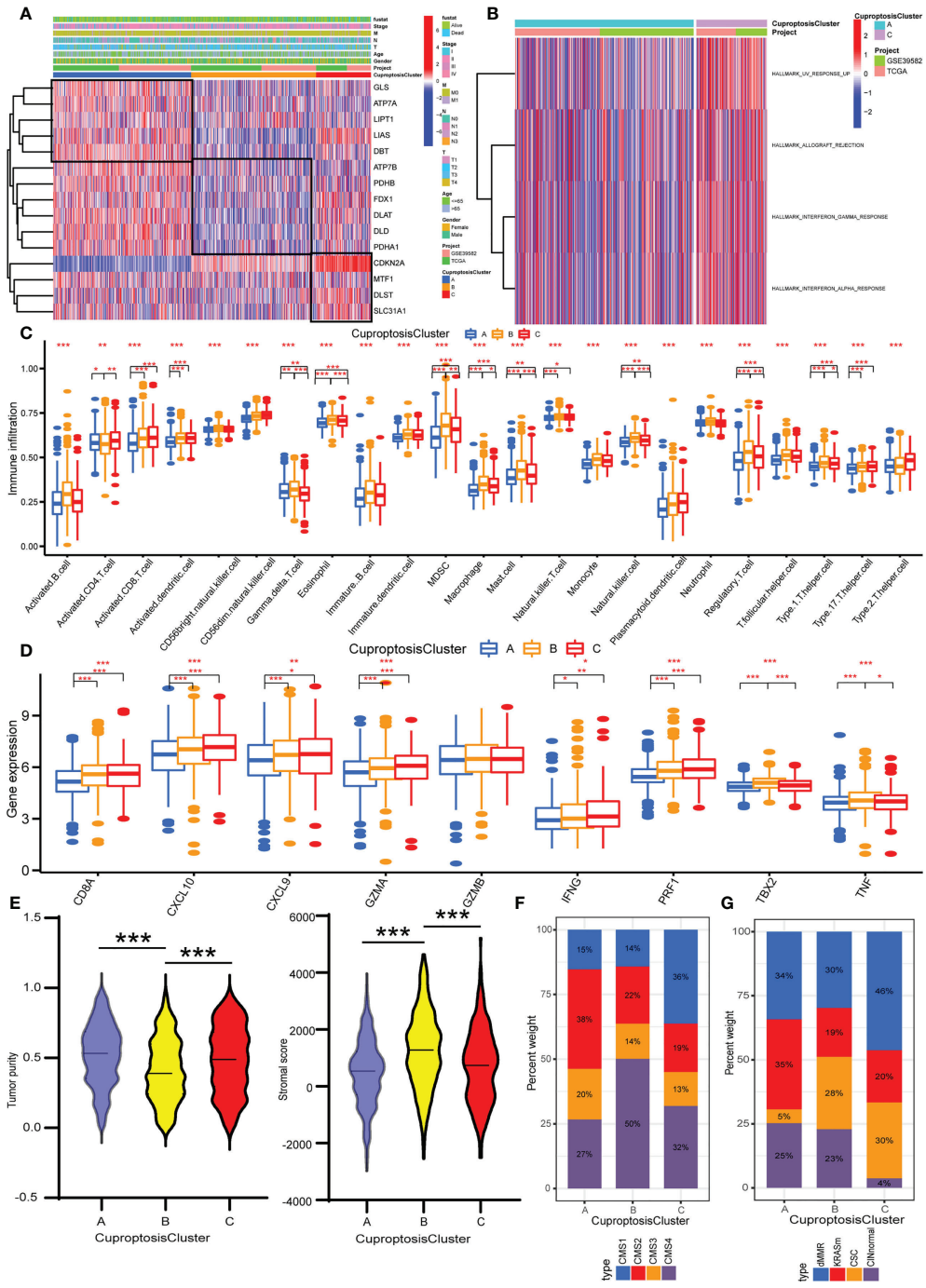


FIGURE 2
Identification of cuproptosis patterns and their TME characteristics in colorectal cancer. **(A)** Unsupervised consensus clustering of 16 cuproptosis regulators in the TCGA and GEO data. The CuproptosisCluster, gender, age, T stage, N stage, M stage, TNM stage, and survival status were used as annotations. **(B)** GSVA enrichment analysis showing the biological processes in CuproptosisClusters A and C. **(C)** The abundance analysis of each immune cell in the TME of three CuproptosisClusters. **(D)** The expression levels of immune activity related genes in three CuproptosisClusters. **(E)** The stromal score and tumor purity in three CuproptosisClusters with ESTIMATE algorithm. **(F)** The CMS analysis of the three CuproptosisClusters. **(G)** The molecular subtypes analysis of GSE39582 dataset in three CuproptosisClusters. *P <0.05; **P <0.01; ***P <0.001.

accounted for the most part of the CuproptosisCluster C, while CINnormal and CSC were the main part of CuproptosisCluster B, and KRASm is the major part of CuproptosisCluster A (Figure 2G), consistent with our above results. In addition, almost all the cuproptosis regulators showed a significant difference between the three CuproptosisClusters (Supplementary Figure S4C). Taken together, we identified the TME infiltration characteristics in three distinct CuproptosisClusters.

Generation of cuproptosis gene signatures

To further explore the underlying biological functions of the three cuproptosis patterns, DEGs between the three CuproptosisClusters were compared in pairs three times (significant criteria, adjusted P-value <0.01). Then, the intersecting genes of the three DEGs were accessed (Figure 3A). GO and KEGG enrichment analyses were performed using the 965 intersected genes. GO results showed that these genes participated in the biological processes of electron carrier activity, mitochondrial inner membrane, adherens junction, T-cell activation, B-cell differentiation, apoptosis process, and lymphatic endothelial cell differentiation (Supplementary Figure S5A). KEGG results revealed the pathways of colorectal cancer, reactive oxygen species, apoptosis, mismatch repair, platinum drug resistance, B-cell receptor signaling pathway, and Fc gamma R-mediated phagocytosis (Figure 3B). To our surprise, these biological processes and pathways are extraordinarily correlated with cuproptosis and immunity, which further verified the vital roles of cuproptosis patterns on the immune variation of TME.

In order to investigate the vital roles forward, univariate Cox regression was constructed to analyze the prognostic roles of the 965 intersect genes in colorectal cancer patients (Supplementary Table S8). A total of 260 genes were identified with significant prognostic values by P <0.05. Then, consensus clustering was performed again based on the expression of the 260 cuproptosis-related genes in order to group samples into different gene signature subtypes. Eventually, three cuproptosis-related gene phenotypes were generated, termed geneClusters A–C (Supplementary Figures S5B–D, Supplementary Table S9). The PCA results indicated an obvious distinction between the three geneClusters (Figure 3C), demonstrating the authentic

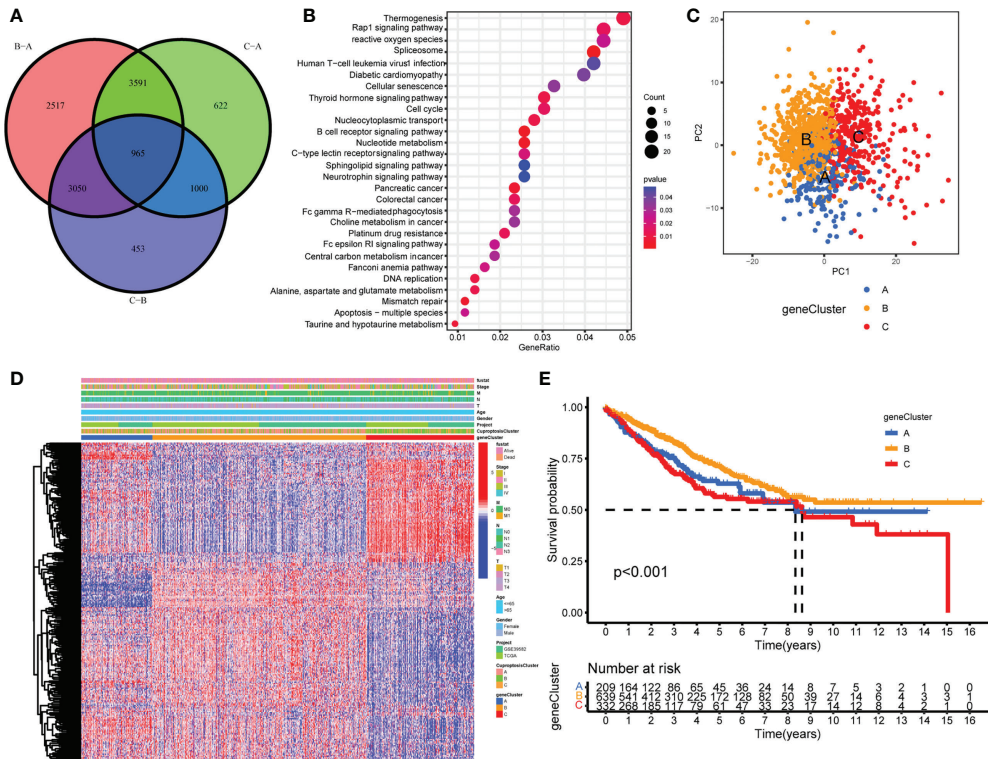


FIGURE 3
Generation of cuproptosis-related geneclusters. **(A)** The Venn diagram of DEGs between three CuproptosisClusters. **(B)** KEGG results of 965 DEGs. **(C)** PCA for three geneClusters to distinguish samples in TCGA COAD/READ and GSE39582. **(D)** Unsupervised clustering of 260 cuproptosis-related genes. **(E)** The Kaplan–Meier curves of overall survival between different geneClusters.

existence of the cuproptosis molecular patterns in colorectal cancer. The correlations between geneClusters with clinicalpathological features were shown in Figure 3D. We observed significantdifferences in overall survival between the three geneClusters, with geneCluster B presenting the longest overall survival (Figure 3E).

TME characteristics in the three cuproptosis gene feature patterns

GSVA was performed with hallmark gene sets to evaluate the biological functions of the three geneClusters. In comparison with geneClusters B and C, geneCluster A was markedly enriched in immune and carcinogenic activation pathways, such as hypoxia, p53 pathway, glycolysis, TNFA signaling *via* NFKB, allograft rejection, inflammatory response, interferon gamma response, interferon alpha response, IL2/STAT5 pathway, IL6/JAK/STAT3 pathway, and DNA repair (Figure 4A). GeneCluster C was prominently related to stromal activation biological process: EMT, angiogenesis, and Wnt/β-catenin signaling (Figure 4B). Then, the ssGSEA was performed to quantify the enrichment level of 22 immune cells in the three geneClusters. As shown in Figure 4C, geneCluster A was predominately enriched in adaptive immune cells including activated CD4 T cells and CD8 T cells, activated dendritic cells, and natural killer cells, and geneCluster C was remarkably rich in immunosuppressive cells and innate immune cells including MDSC, immature dendritic cells, macrophages, monocytes, regulatory T cells, Th1 cells, T follicular helper cells, and dendritic cells, eosinophils, and mast cells, while geneCluster B presented a lower enrichment level of immune cells. Furthermore, the stromal score and tumor purity of all samples were analyzed. GeneCluster B showed the highest tumor purity compared with geneCluster A and geneCluster C (Figure 4D), indicating the lowest infiltration of immune cells and stromal components. Therefore, geneCluster B was in accord with the immune-desert phenotype. Similarly, geneCluster C presented with the highest stromal score compared with geneCluster A and geneCluster B (Figure 4D). To further verify our speculation, two other known molecular subtypes were analyzed in the three gene clusters. GeneCluster A predominated in CMS1, while geneCluster C was CMS4, and geneCluster B was CMS2 and CMS3 (Figure 4E). The subtype of dMMR accounted for the majority of geneCluster A, while CINnormal and CSC were the main parts of geneCluster C, and KRASm is the major part of geneCluster B (Figure 4F). In addition, the gene levels of EMT and TGFβ pathways were higher in geneCluster C compared with geneClusters A and B (Supplementary Figure S5E), whereas immune activity related-genes were higher in geneCluster A compared with geneClusters C and B (Figure 4G). Combined with the features of immune cell infiltration, geneCluster C was consistent with the immune-excluded phenotype, while geneCluster A was the immune-inflamed phenotype. We also observed that all the cuproptosis regulators presented differentiated expression levels in the three geneClusters (Supplementary Figure S5F).

Construction and characteristics of cuproptosisScore

To assess the cuproptosis patterns of individual patients, we constructed a cuproptosis score system with these cuproptosis-related genes, termed CuproptosisScore (Supplementary Table S10). The alluvial diagram showed the attribute changes of individual samples (Figure 5A). The Kaplan–Meier curves indicated that patients with a high CuproptosisScore had shorter overall survival compared with patients with a low CuproptosisScore (Figure 5B). Also, more patients in the high CuproptosisScore group are in the status of death (Figure 5C). This Kruskal–Wallis test indicated significant differences in CuproptosisScore between cuproptosis-related gene clusters (Figure 5D). Genecluster C showed the highest median score, which indicated that a high CuproptosisScore may be correlated with stromal activation-related signatures. whereas geneCluster A had the lowest median score, showing that low CuproptosisScore may be related to an immune activation phenotype. Next, the analysis of stromal score and tumor purity found that samples with a high CuproptosisScore exhibited a high stromal score and low tumor purity (Figure 5E), implicating a stromal activation phenotype. While low CuproptosisScore presented low stromal score and high tumor purity, it also showed high expression of immune activated genes (Figure 5F), implying an immune activation phenotype. These results give us some hints that CuproptosisScore may contribute to predicting the immune cell filtrating in the cancer samples and further evaluating the immune response to the targeted drugs. The analysis of CMS subtypes showed that CMS4 comprises most of the high CuproptosisScore group, while the low CuproptosisScore group consisted of more CMS1 and less CMS4 (Figure 5G). In the GSE39582 dataset, the low CuproptosisScore group mainly belongs to the dMMR type, and the high CuproptosisScore group is mostly divided into the CINnormal and CSC types (Figure 5H). The above results strongly demonstrated that a high CuproptosisScore was remarkably related to an immune-exclude phenotype, while a low CuproptosisScore was correlated with an immune-inflamed phenotype. Therefore, the distinct cuproptosis-related genes play indispensable regulatory roles in shaping different TME characteristics. The CuproptosisScore could better assess the cuproptosis-related molecular patterns of individual samples and TME infiltration characteristics, which is conducive to selecting appropriate targeted drugs in the clinical treatment of colorectal cancer.

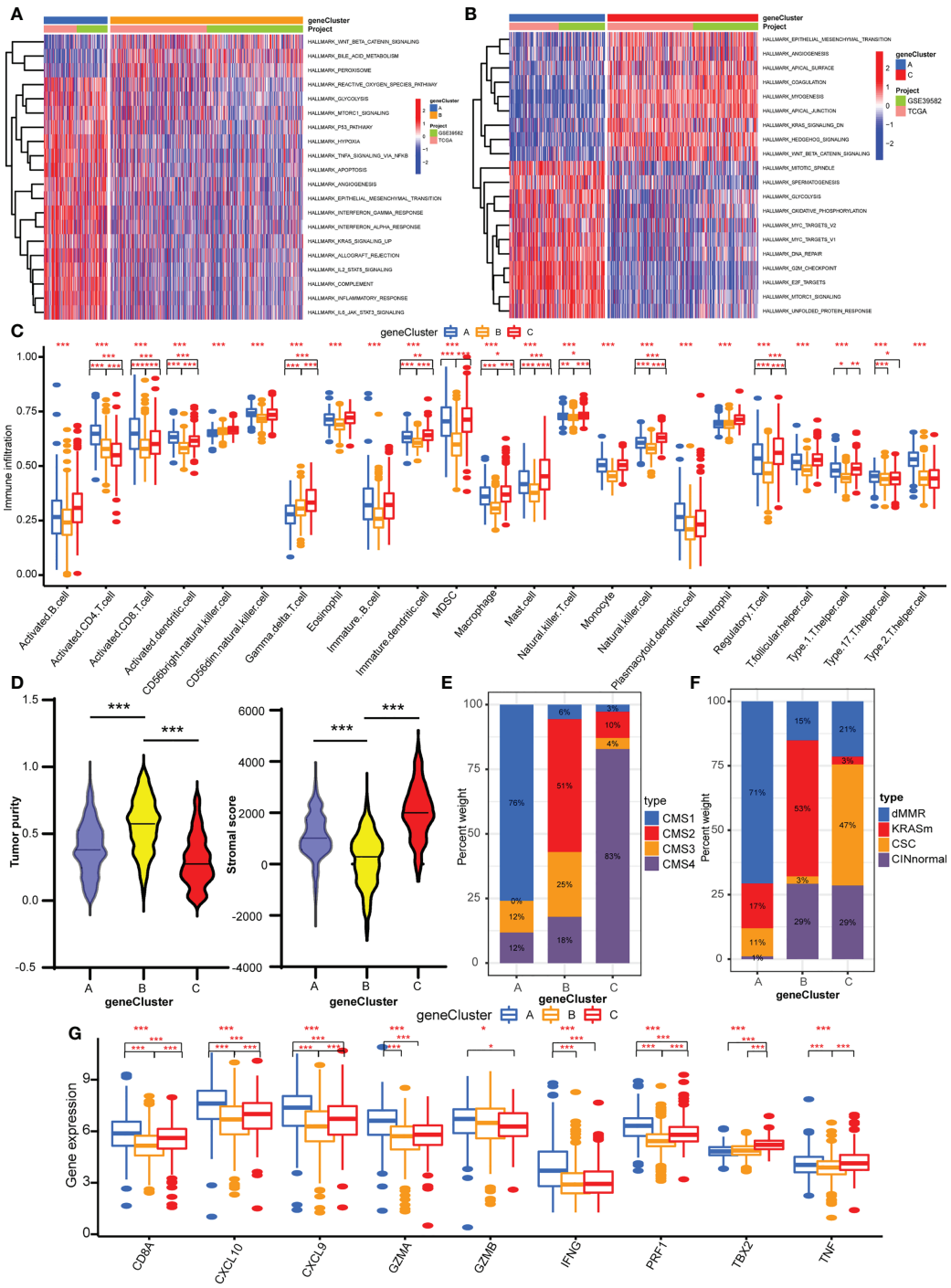


FIGURE 4
TME characteristics of three geneClusters. **(A)** The heatmap of GSVA results between geneCluster A and geneCluster B. **(B)** The heatmap of GSVA results between geneCluster A and geneCluster C. **(C)** The abundances of TME infiltrating immune cells in three geneClusters by ssGSVA algorithm. **(D)** The stromal score and tumor purity in three geneClusters by ESTIMATE algorithm. **(E)** The CMS analysis of the three geneClusters. **(F)** The molecular subtypes analysis of GSE39582 dataset in three geneClusters. **(G)** The expression of immune activated genes between three geneClusters. *P <0.05; **P <0.01; ***P <0.001.

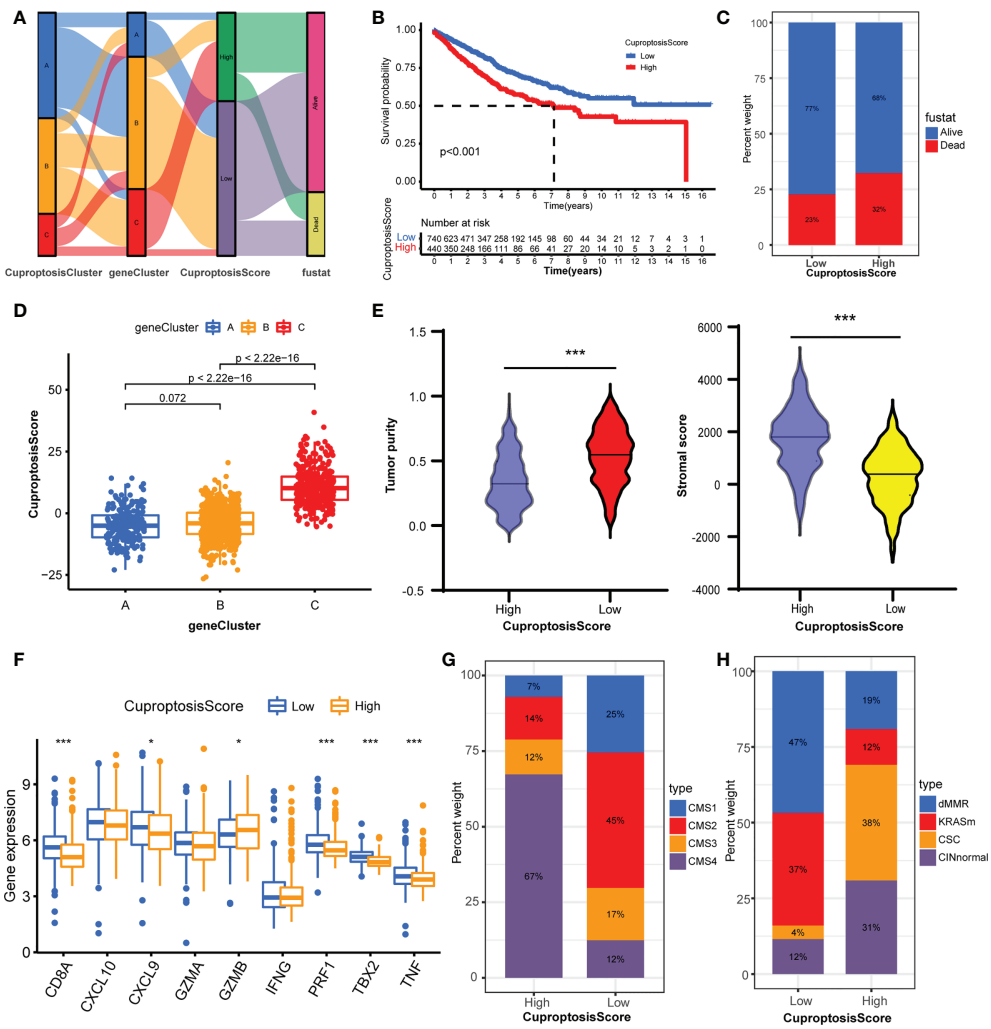


FIGURE 5
Construction of CuproptosisScore and their TME characteristics in colorectal cancer. **(A)** Alluvial diagram showing the attribute changes from CuproptosisClusters to gene Clusters to CuproptosisScore to survival status. **(B)** The Kaplan–Meier curves of overall survival between high and low CuproptosisScore group. **(C)** The frequencies of alive and dead status in high and low CuproptosisScore group. **(D)** The CuproptosisScores of the three geneClusters. **(E)** The stromal score and tumor purity in high and low CuproptosisScore group. **(F)** The expression levels of immune activity-related genes in high and low CuproptosisScore groups. **(G)** The CMS analysis of the two CuproptosisScore groups. **(H)** The molecular subtypes analysis of GSE39582 dataset in the two CuproptosisScore groups. *P <0.05; ***P <0.001.

Relationship of cuproptosis patterns with tumor somatic mutation and immunotherapy

In this part, we analyzed the differences in tumor somatic mutation, MSI, and immunotherapy response between different CuproptosisScore groups. First, we depicted the landscape of gene mutation in the high and low CuproptosisScore groups (Figures 6A, B), and then computed the TMB. The TMB of the low CuproptosisScore group was higher than that of the high CuproptosisScore group (Figure 6C), and CuproptosisScore was negatively correlated with TMB (Supplementary Figure S6A). Second, the Kaplan–Meier curves showed no significant difference in overall survival between patients with high and low TMB (Supplementary Figure S6B), but a trend of shorter 5-year survival in the low TMB group. However, when the combination of TMB and CuproptosisScore served as a prognostic indicator, the differences in overall survival between different groups were obvious (Figure 6D). The patients with low CuproptosisScore had longer overall survival compared with patients with high CuproptosisScore, both in the high and low TMB groups. Subsequently, we evaluated the correlation between CuproptosisScore and MSI. The results showed that MSI-H group scored low CuproptosisScore, while MSS and MSI-L group

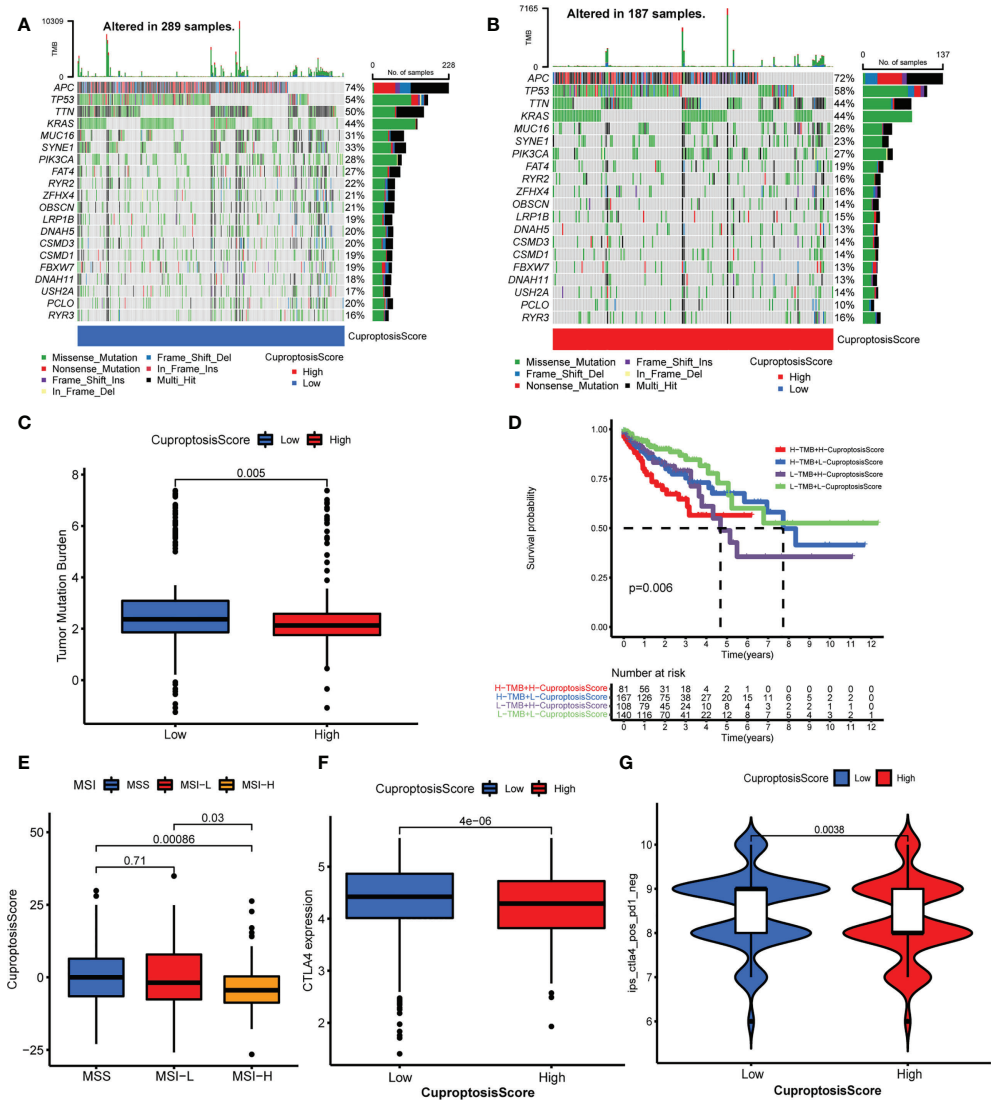


FIGURE 6
Relationship of cuproptosis patterns with tumor somatic mutation and immunotherapy. **(A, B)** The waterfall plot of tumor somatic mutation of low CuproptosisScore group **(A)** and high CuproptosisScore group **(B)**. **(C)** The TMB of high and low CuproptosisScore group. **(D)** The Kaplan–Meier curves of overall survival in different groups of TMB combination with CuproptosisScore. **(E)** The CuproptosisScores of samples with MSS, MSI-L, and MSI-H. **(F)** CTLA4 expression of the two CuproptosisScore groups. **(G)** IPS of anti-CTLA4 drug in the two CuproptosisScore groups.

scored high CuproptosisScore (Figure 6E). Considering that CuproptosisScore was closely associated with TME infiltration features, we evaluated the influence of CuproptosisScore on immunotherapy. High CTLA4 expression was observed in the low CuproptosisScore group compared with the high CuproptosisScore group (Figure 6F). The IPS of the TCGA samples was downloaded online (https://tcia.at/home, Supplementary Table S11). When treated with an anti-CTLA4 drug, the low CuproptosisScore group presented a high IPS, meaning better immunotherapy response (Figure 6G). When treated with an anti-PD1 drug, no significant difference in IPS was found between the two CuproptosisScore groups (Supplementary Figure S6C). In addition, CuproptosisScore can serve as a prognostic marker in different stages of colorectal cancer patients. Longer overall survival was observed in low CuproptosisScore group compared with high CuproptosisScore group, in patients with T3–T4 stage, N1–N3 stage, M0 stage and TNM III–IV stage (Supplementary Figures S7A–H).

Discussion

Cuproptosis is a novel cell death format regulated by an ancient mechanism, distinct from all other known mechanisms

(15). Both copper chelators and copper ionophores have been exploited as antitumor drugs and tested in clinical trials (9, 52, 53). However, the correlations between cuproptosis regulators with molecular patterns, clinicalpathological subtypes, prognostic values, TME infiltration features, and immunotherapy response have not been investigated. Therefore, identifying the roles of cuproptosis-related molecular patterns will promote our cognition of cuproptosis and its features in colorectal cancer, which contributes to exploring the potential of cuproptosis-related gene signatures to serve as a marker for evaluating prognosis and antitumor immune response.

In this study, we accessed 16 cuproptosis regulators and analyzed their mutation, CNV, expression level, and biological functions in colorectal cancer. CNV was the main interfering factor in the expression level of 16 cuproptosis regulators. These cuproptosis regulators participate in copper ion homeostasis, protein lipoylation, TCA cycle, and response to oxidative stress. Clinically, most of these cuproptosis regulators are the prognostic effectors of colorectal cancer patients. Then, we constructed three distinct cuproptosis clusters based on the expression of these cuproptosis regulators. CuproptosisCluster B was preferentially related to lower stage and CuproptosisCluster C more dead status, consistent with the trend of longer overall survival in CuproptosisCluster B than CuproptosisCluster C. Through multiple evidences of different research methods, we demonstrated that the three CuproptosisClusters were characterized by distinct TME characteristics. CuproptosisCluster B was characterized by stromal activation and immunosuppressive phenotype, corresponding to immune-exclude subtype; CuproptosisCluster C was enriched in immune and carcinogenic activation pathway, related to immune-inflamed subtype; while no immune-related pathway and little immune cells were enriched in CuproptosisCluster A, grouped into immune-desert subtype. ESTIMATE algorithm showed high tumor purity and little stromal component in CuproptosisCluster C, whereas CuproptosisCluster B had more stromal component. Surprisingly, the ssGSEA research found unique high level of adaptive immune cells including activated CD8 T cell, natural killer cell, and gamma delta T cell in CuproptosisCluster C, while marked enrichment of immunosuppressive cells and innate immune cells including macrophage, monocyte, regulatory T cell, Th1 cell, Th17 cell, and dendritic cell, eosinophil, and mast cell in CuproptosisCluster B. Activated CD8 T cell by neoantigen is the predominate effector cell exerting tumor killing and at the hotspot of antitumor immunotherapy (54). Nonetheless, the immune system consists of a complicated process of antitumor response in which multiple immune cells, cytokines, and chemical factors participate. These mechanisms must coordinate with each other and play a mutually reinforcing role. Natural killer cells are also a key antitumor effector cell working through perforin, granular enzyme or antibody dependent cell-mediated cytotoxicity (ADCC) (55). γδ T cells can be elicited by butyrophilin or butyrophilin-like molecules in a major histocompatibility complex (MHC)-independent manner and bridge innate and adaptive immunity, therefore responding to multiple types of cancers (56, 57). In addition, regulatory T cells act as immunosuppressive cells through many cytokines and cells. However, macrophages are known to be classified into two main phenotypes: proinflammatory and antitumor M1 macrophages and immunosuppressive M2 macrophages (58, 59). M0 macrophages can be selectively polarized into M1 macrophages or M2 macrophages according to the TME characteristics. M1 macrophages produce type I pro-inflammatory cytokines to promote immune response, while M2 macrophages enhance matrix remodeling, EMT, and Wnt/β-catenin pathway (60, 61). Therefore, a broad range of immune cells were activated and played nonnegligible roles in the antitumor immune response, which provided great plasticity for TME characteristics. Although CuproptosisCluster B was rich in immune cells, these cells played immunosuppressive roles or surrounded the tumor focus but penetrated into the tumor parenchyma, rendering them unable to exert antitumor effects (62, 63). Combined with the above results, CuproptosisCluster B was classified as an immune-excluded phenotype. Furthermore, two molecular subtypes of colorectal cancer (the recognized classification of CMS and molecular subtypes of the GSE39582 dataset) were adapted to verify the above speculation. The results indicated that CMS1 or dMMR accounted for the most part of the CuproptosisCluster C, while CMS4 or CINnormal and CSC were the main part of CuproptosisCluster B, and CMS2 and CMS3 or KRASm consisted of the major part of CuproptosisCluster A. The similar TME characteristics of different molecular subtypes proved the availability and effectiveness of CuproptosisClusters. Finally, we examined the level of immune activity related genes and stromal activity related pathways. The immune activity related genes were significantly highly expressed in CuproptosisCluster C, and the stromal activity related pathways were remarkably rich in CuproptosisCluster C. In general, after comprehensively exploring the TME characteristics of distinct cuproptosis clusters, the novel molecular subtype was proved to be a reliable and effective classification of colorectal cancer patients and immune phenotype.

The differentially expressed genes between the three cuproptosis clusters were proved to be correlated with cuproptosis and immune activation pathways, and considered as cuproptosis-related signature genes. Based on the expression of these genes, three gene clusters were identified to group samples with distinct clinical–pathological features and a TME phenotype. GeneCluster A was markedly enriched in immune and carcinogenic activation pathways, CMS1 subtype, dMMR subtype and high levels of immune activation genes with high tumor purity and low stromal component, classified as the immune-inflamed phenotype. GeneCluster C was remarkably rich in immunosuppressive cells, CMS4 subtype, CINnormal,and CSC subtypes, and high levels of stromal activation genes with high stromal component and low tumor purity, classified as an immune-excluded phenotype. While GeneCluster B was short of immune cells, it was classified as an

immune-desert phenotype. This indicates that the cuproptosis-related genes play crucial roles in shaping different TME characteristics. Considering the crucial roles of cuproptosis patterns in the TME formation of colorectal cancer, we constructed a score system, termed CuproptosisScore, to quantify the cuproptosis patterns of individual samples. The low CuproptosisScore indicated the immune-inflamed phenotype with longer overall survival, while the high CuproptosisScore indicated the immune-excluded phenotype with shorter overall survival. Integrated analysis demonstrated that CuproptosisScore could serve as an effective prognostic marker and indicator of immune subtype (64). For clinical patients, the cuproptosis-related genes can be detected with transcriptome sequencing to calculate the CuproptosisScore for individuals, therefore evaluating their prognosis and immune subtype.

With the development of tumor behaviors and immunological molecular mechanisms, immunotherapy provides a novel site for tumor targeting therapy, especially immune checkpoint inhibitors (ICIs) (65–67), including CTLA4, PD-1, and PD-L1. MSI-H or dMMR are indicators of immunotherapy response for colorectal cancer. We found low CuproptosisScore presented with high TMB, MSI-H, and high expression of CTLA4, and better response to anti-CTLA4 immunotherapy. Previous research reported that TGFβ and EMT related pathways impaired the penetrating of T cells into tumor focus and weakened antitumor effects (21). This is consistent with our results that stromal activation phenotypes with activated TGFβ and EMT pathways exhibited the immune-exclude phenotype and were resistant to ICI response. Combined with tumor stage, TMB, CTLA4, and PD-L1 expression, MSI status, and TME phenotype, CuproptosisScore can serve as an effective predictive schedule for prognosis and contribute to performing patient stratification for the determination of immunotherapy regimen in colorectal cancer patients.

Although we first reveal the clinical pathology and TME phenotype of cuproptosis-related gene patterns in colorectal cancer, there are several shortcomings in the study. First, cuproptosis was established in recent years, so more cuproptosis-related genes remain to be discovered, which will provide more profound insights for cuproptosis. Second, there is no report about cuproptosis with tumor progression and therapy up to now. Therefore, more mechanism studies need to be performed to enhance our cognition of the correlation between cuproptosis and cancer. Third, despite that we provided a novel direction for cuproptosis and TME phenotype, no report about the mechanisms of cuproptosis on shaping TME infiltration has been retrieved. Fourth, the CuproptosisScore is short of verification from other data. More data needs to be collected to confirm the authenticity and reliability of CuproptosisScore. Finally, all the data were accessed from public dataset, and selection bias of samples may exist. Large-scale prospective studies need to be performed to verify the findings.

Conclusion

In conclusion, this study revealed a novel cuproptosis-related gene pattern with different clinical–pathological and TME phenotypes. The integrated analysis of different cuproptosis patterns contributes to our understanding of TME and provides an effective marker for prognosis and immunotherapy of colorectal cancer.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YX and ZZ designed the idea of the article. ZZ and QZ carried out the experiments and analyzed the data. WS and JW wrote the manuscript. SL, TG, and CZ supervised and verified the data analysis. YX validated the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Science and Technology Commission of Shanghai Municipality (20DZ1100101), the Shanghai Hospital Development Center (SKXZ2028), and the National Natural Science Foundation of China (82003060).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.940774/full#supplementary-material

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin* (2021) 71(1):7–33. doi: 10.3322/caac.21654

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660

3. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin* (2020) 70(3):145–64. doi: 10.3322/caac.21601

4. Ruiz LM, Libedinsky A, Elorza AA. Role of copper on mitochondrial function and metabolism. *Front Mol Biosci* (2021) 8:711227. doi: 10.3389/fmolb.2021.711227

5. Krishnamoorthy L, Cotruvo JAJr., Chan J, Kaluarachchi H, Muchenditsi A, Pendyala VS, et al. Copper regulates cyclic-AMP-dependent lipolysis. *Nat Chem Biol* (2016) 12(8):586–92. doi: 10.1038/nchembio.2098

6. Brady DC, Crowe MS, Turski ML, Hobbs GA, Yao X, Chaikuad A, et al. Copper is required for oncogenic BRAF signalling and tumorigenesis. *Nature* (2014) 509(7501):492–6. doi: 10.1038/nature13180

7. Dodani SC, Firl A, Chan J, Nam CI, Aron AT, Onak CS, et al. Copper is an endogenous modulator of neural circuit spontaneous activity. *Proc Natl Acad Sci U.S.A.* (2014) 111(46):16280–5. doi: 10.1073/pnas.1409796111

8. Turski ML, Brady DC, Kim HJ, Kim BE, Nose Y, Counter CM, et al. A novel role for copper in ras/mitogen-activated protein kinase signaling. *Mol Cell Biol* (2012) 32(7):1284–95. doi: 10.1128/mcb.05722-11

9. Tsang T, Posimo JM, Gudiel AA, Cicchini M, Feldser DM, Brady DC. Copper is an essential regulator of the autophagic kinases ULK1/2 to drive lung adenocarcinoma. *Nat Cell Biol* (2020) 22(4):412–24. doi: 10.1038/s41556-020-0481-4

10. Dancis A, Haile D, Yuan DS, Klausner RD. The saccharomyces cerevisiae copper transport protein (Ctr1p). biochemical characterization, regulation by copper, and physiologic role in copper uptake. *J Biol Chem* (1994) 269 (41):25660–7. PMID:7929270.

11. Graden JA, Winge DR. Copper-mediated repression of the activation domain in the yeast Mac1p transcription factor. *Proc Natl Acad Sci U.S.A.* (1997) 94(11):5550–5. doi: 10.1073/pnas.94.11.5550

12. Lukanović D, Herzog M, Kobal B, Černe K. The contribution of copper efflux transporters ATP7A and ATP7B to chemoresistance and personalized medicine in ovarian cancer. *BioMed Pharmacother* (2020) 129:110401. doi: 10.1016/j.biopha.2020.110401

13. Gul NS, Khan TM, Chen M, Huang KB, Hou C, Choudhary MI, et al. New copper complexes inducing bimodal death through apoptosis and autophagy in A549 cancer cells. *J Inorg Biochem* (2020) 213:111260. doi: 10.1016/j.jinorgbio.2020.111260

14. Kaur P, Johnson A, Northcote-Smith J, Lu C, Suntharalingam K. Immunogenic cell death of breast cancer stem cells induced by an endoplasmic reticulum-targeting Copper(II) complex. *Chembiochem* (2020) 21(24):3618–24. doi: 10.1002/cbic.202000553

15. Tsvetkov P, Coy S, Petrova B, Dreishpoon M, Verma A, Abdusamad M, et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* (2022) 375(6586):1254–61. doi: 10.1126/science.abf0529

16. Ahn SJ, Woo TY, Lee DW, Lew DH, Song SY. Nipple-areolar complex ischemia and necrosis in nipple-sparing mastectomy. *Eur J Surg Oncol* (2018) 44 (8):1170–6. doi: 10.1016/j.ejso.2018.05.006

17. Pan R, Ryan J, Pan D, Wucherpfennig KW, Letai A. Augmenting NK cell-based immunotherapy by targeting mitochondrial apoptosis. *Cell* (2022) 185 (9):1521–1538.e18. doi: 10.1016/j.cell.2022.03.030

18. Marshall RS, Hua Z, Mali S, McLoughlin F, Vierstra RD. ATG8-binding UIM proteins define a new class of autophagy adaptors and receptors. *Cell* (2019) 177(3):766–81.e24. doi: 10.1016/j.cell.2019.02.009

19. Chen X, Zeh HJ, Kang R, Kroemer G, Tang D. Cell death in pancreatic cancer: from pathogenesis to therapy. *Nat Rev Gastroenterol Hepatol* (2021) 18 (11):804–23. doi: 10.1038/s41575-021-00486-6

20. Xi H, Zhang Y, Xu Y, Yang WY, Jiang X, Sha X, et al. Caspase-1 inflammasome activation mediates homocysteine-induced pyrop-apoptosis in endothelial cells. *Circ Res* (2016) 118(10):1525–39. doi: 10.1161/circresaha.116.308501

21. Luo W, Dai W, Li Q, Mo S, Han L, Xiao X, et al. Ferroptosis-associated molecular classification characterized by distinct tumor microenvironment profiles in colorectal cancer. *Int J Biol Sci* (2022) 18(5):1773–94. doi: 10.7150/ijbs.69808

22. Passeri G, Northcote-Smith J, Suntharalingam K. Delivery of an immunogenic cell death-inducing copper complex to cancer stem cells using polymeric nanoparticles. *RSC Adv* (2022) 12(9):5290–9. doi: 10.1039/d1ra08788f

23. Członkowska A, Litwin T, Dusek P, Ferenci P, Lutsenko S, Medici V, et al. Wilson Disease. *Nat Rev Dis Primers* (2018) 4(1):21. doi: 10.1038/s41572-018-0018-3

24. Armer J, De Goede C. How to use tests for disorders of copper metabolism. *Arch Dis Child Educ Pract Ed* (2017) 102(6):319–27. doi: 10.1136/archdischild-2016-310960

25. Ding X, Jiang M, Jing H, Sheng W, Wang X, Han J, et al. Analysis of serum levels of 15 trace elements in breast cancer patients in Shandong, China. *Environ Sci pollut Res Int* (2015) 22(10):7930–5. doi: 10.1007/s11356-014-3970-9

26. Aubert L, Nandagopal N, Steinhart Z, Lavoie G, Nourreddine S, Berman J, et al. Copper bioavailability is a KRAS-specific vulnerability in colorectal cancer. *Nat Commun* (2020) 11(1):3701. doi: 10.1038/s41467-020-17549-y

27. Michniewicz F, Saletta F, Rouaen JRC, Hewavisenti RV, Mercatelli D, Cirillo G, et al. Copper: An intracellular achilles' heel allowing the targeting of epigenetics, kinase pathways, and cell metabolism in cancer therapeutics. *ChemMedChem* (2021) 16(15):2315–29. doi: 10.1002/cmdc.202100172

28. Chen F, Wang J, Chen J, Yan L, Hu Z, Wu J, et al. Serum copper and zinc levels and the risk of oral cancer: A new insight based on large-scale case-control study. *Oral Dis* (2019) 25(1):80–6. doi: 10.1111/odi.12957

29. Ge EJ, Bush AI, Casini A, Cobine PA, Cross JR, DeNicola GM, et al. Connecting copper and cancer: From transition metal signalling to metalloplasia. *Nat Rev Cancer* (2022) 22(2):102–13. doi: 10.1038/s41568-021-00417-2

30. Li Y. Copper homeostasis: Emerging target for cancer treatment. *IUBMB Life* (2020) 72(9):1900–8. doi: 10.1002/iub.2341

31. Risom T, Glass DR, Averbukh I, Liu CC, Baranski A, Kagel A, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* (2022) 185(2):299–310.e18. doi: 10.1016/j.cell.2021.12.023

32. Zhao L, Liu Y, Zhang S, Wei L, Cheng H, Wang J, et al. Impacts and mechanisms of metabolic reprogramming of tumor microenvironment for immunotherapy in gastric cancer. *Cell Death Dis* (2022) 13(4):378. doi: 10.1038/s41419-022-04821-w

33. Thakur A, Johnson A, Jacobs E, Zhang K, Chen J, Wei Z, et al. Energy sources for exosome communication in a cancer microenvironment. *Cancers (Basel)* (2022) 14(7):10. doi: 10.3390/cancers14071698

34. Jiang T, Yang T, Chen Y, Miao Y, Xu Y, Jiang H, et al. Emulating interactions between microorganisms and tumor microenvironment to develop cancer theranostics. *Theranostics* (2022) 12(6):2833–59. doi: 10.7150/thno.70719

35. Chen C, Guan J, Gu X, Chu Q, Zhu H. Prostaglandin E2 and receptors: Insight into tumorigenesis, tumor progression, and treatment of hepatocellular carcinoma. *Front Cell Dev Biol* (2022) 10:834859. doi: 10.3389/fcell.2022.834859

36. Gupta YH, Khanom A, Acton SE. Control of dendritic cell function within the tumour microenvironment. *Front Immunol* (2022) 13:733800. doi: 10.3389/fimmu.2022.733800

37. Wang X, Luo G, Zhang K, Cao J, Huang C, Jiang T, et al. Hypoxic tumor-derived exosomal miR-301a mediates M2 macrophage polarization *via* PTEN/PI3Kγ to promote pancreatic cancer metastasis. *Cancer Res* (2018) 78(16):4586–98. doi: 10.1158/0008-5472.Can-17-3841

38. Affo S, Nair A, Brundu F, Ravichandra A, Bhattacharjee S, Matsuda M, et al. Promotion of cholangiocarcinoma growth by diverse cancer-associated fibroblast subpopulations. *Cancer Cell* (2021) 39(6):866–82.e11. doi: 10.1016/j.ccell.2021.03.012

39. Desbois M, Udyavar AR, Ryner L, Kozlowski C, Guan Y, Dürrbaum M, et al. Integrated digital pathology and transcriptome analysis identifies molecular mediators of T-cell exclusion in ovarian cancer. *Nat Commun* (2020) 11(1):5583. doi: 10.1038/s41467-020-19408-2

40. Hatthakarnkul P, Quinn JA, Ammar A, Lynch G, Van Wyk H, McMillan DC, et al. Molecular mechanisms of tumour budding and its association with microenvironment in colorectal cancer. *Clin Sci (Lond)* (2022) 136(8):521–35. doi: 10.1042/cs20210886

41. Lelliott EJ, Sheppard KE, McArthur GA. Harnessing the immunotherapeutic potential of CDK4/6 inhibitors in melanoma: is timing everything? *NPJ Precis Oncol* (2022) 6(1):26. doi: 10.1038/s41698-022-00273-9

42. Zhang B, Wu Q, Li B, Wang D, Wang L, Zhou YL. m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. *Mol Cancer* (2020) 19(1):53. doi: 10.1186/s12943-020-01170-0

43. Zhao S, Mi Y, Zheng B, Wei P, Gu Y, Zhang Z, et al. Highly-metastatic colorectal cancer cell released miR-181a-5p-rich extracellular vesicles promote liver metastasis by activating hepatic stellate cells and remodelling the tumour microenvironment. *J Extracell Vesicles* (2022) 11(1):e12186. doi: 10.1002/jev2.12186

44. Mancarella S, Serino G, Coletta S, Armentano R, Dituri F, Ardito F, et al. The tumor microenvironment drives intrahepatic cholangiocarcinoma progression. *Int J Mol Sci* (2022) 23(8):10. doi: 10.3390/ijms23084187

45. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus immunoscore for the classification of colon cancer: A prognostic and accuracy study. *Lancet* (2018) 391(10135):2128–39. doi: 10.1016/s0140-6736(18)30789-x

46. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* (2016) 17:13. doi: 10.1186/s13059-016-0881-8

47. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst* (2015) 1(6):417–25. doi: 10.1016/j.cels.2015.12.004

48. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* (2021) 2 (3):100141. doi: 10.1016/j.xinn.2021.100141

49. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* (2009) 462(7269):108–12. doi: 10.1038/nature08460

50. Song Q, Zhou R, Shu F, Fu W. Cuproptosis scoring system to predict the clinical outcome and immune response in bladder cancer. *Front Immunol* (2022) 13:958368. doi: 10.3389/fimmu.2022.958368

51. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* (2006) 98(4):262–72. doi: 10.1093/jnci/djj052

52. Cui L, Gouw AM, LaGory EL, Guo S, Attarwala N, Tang Y, et al. Mitochondrial copper depletion suppresses triple-negative breast cancer in mice. *Nat Biotechnol* (2021) 39(3):357–67. doi: 10.1038/s41587-020-0707-9

53. Brady DC, Crowe MS, Greenberg DN, Counter CM. Copper chelation inhibits BRAF(V600E)-driven melanomagenesis and counters resistance to BRAF (V600E) and MEK1/2 inhibitors. *Cancer Res* (2017) 77(22):6240–52. doi: 10.1158/0008-5472.Can-16-1190

54. Huntington ND, Cursons J, Rautela J. The cancer-natural killer cell immunity cycle. *Nat Rev Cancer* (2020) 20(8):437–54. doi: 10.1038/s41568-020-0272-z

55. Terrén I, Orrantia A, Vitallé J, Zenarruzabeitia O, Borrego F. NK cell metabolism and tumor microenvironment. *Front Immunol* (2019) 10:2278. doi: 10.3389/fimmu.2019.02278

56. Payne KK, Mine JA, Biswas S, Chaurio RA, Perales-Puchalt A, Anadon CM, et al. BTN3A1 governs antitumor responses by coordinating αβ and γδ T cells. *Science* (2020) 369(6506):942–9. doi: 10.1126/science.aay2767

57. Morandi F, Yazdanifar M, Cocco C, Bertaina A, Airoldi I. Engineering the bridge between innate and adaptive immunity for cancer immunotherapy: Focus on γδ T and NK cells. *Cells* (2020) 9(8):10. doi: 10.3390/cells9081757

58. Tortora C, Di Paola A, Argenziano M, Creoli M, Marrapodi MM, Cenni S, et al. Effects of CB2 receptor modulation on macrophage polarization in pediatric celiac disease. *Biomedicines* (2022) 10(4):10. doi: 10.3390/biomedicines10040874

59. Shou Y, Wang X, Chen C, Liang Y, Yang C, Xiao Q, et al. Exosomal miR-301a-3p from esophageal squamous cell carcinoma cells promotes angiogenesis by inducing M2 polarization of macrophages *via* the PTEN/PI3K/AKT signaling pathway. *Cancer Cell Int* (2022) 22(1):153. doi: 10.1186/s12935-022-02570-6

60. Jianfeng W, Yutao W, Jianbin B. TACR2 is associated with the immune microenvironment and inhibits migration and proliferation *via* the wnt/β-catenin signaling pathway in prostate cancer. *Cancer Cell Int* (2021) 21(1):415. doi: 10.1186/s12935-021-02126-0

61. Abramowitz LK, Hanover JA. Chronically elevated O-GlcNAcylation limits nitric oxide production and deregulates specific pro-inflammatory cytokines. *Front Immunol* (2022) 13:802336. doi: 10.3389/fimmu.2022.802336

62. Song W, Ren J, Xiang R, Kong C, Fu T. Identification of pyroptosis-related subtypes, the development of a prognosis model, and characterization of tumor microenvironment infiltration in colorectal cancer. *Oncoimmunology* (2021) 10 (1):1987636. doi: 10.1080/2162402x.2021.1987636

63. Ho KH, Huang TW, Shih CM, Lee YT, Liu AJ, Chen PH, et al. Glycolysis-associated lncRNAs identify a subgroup of cancer patients with poor prognoses and a high-infiltration immune microenvironment. *BMC Med* (2021) 19(1):59. doi: 10.1186/s12916-021-01925-6

64. Oliveri V. Selective targeting of cancer cells by copper ionophores: An overview. *Front Mol Biosci* (2022) 9:841814. doi: 10.3389/fmolb.2022.841814

65. Xia QD, Sun JX, Liu CQ, Xu JZ, An Y, Xu MY, et al. Ferroptosis patterns and tumor microenvironment infiltration characterization in bladder cancer. *Front Cell Dev Biol* (2022) 10:832892. doi: 10.3389/fcell.2022.832892

66. Chong W, Shang L, Liu J, Fang Z, Du F, Wu H, et al. m(6)A regulator-based methylation modification patterns characterized by distinct tumor microenvironment immune profiles in colon cancer. *Theranostics* (2021) 11 (5):2201–17. doi: 10.7150/thno.52717

67. Lei L, Li N, Yuan P, Liu D. A new risk model based on a 11-m(6)A-related lncRNA signature for predicting prognosis and monitoring immunotherapy for gastric cancer. *BMC Cancer* (2022) 22(1):365. doi: 10.1186/s12885-021-09062-2

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fimmu.2022.1019386

Check for updates

OPEN ACCESS

EDITED BY
George Bebis,
University of Nevada, United States

REVIEWED BY
Lichun Wei,
The First Affiliated Hospital of Air Force Medical University, China
Guo Xin Li,
Southern Medical University, China

*CORRESPONDENCE
Junwei Bai
Baijw@zzu.edu.cn
Chao Zhang
Zchao@zzu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 15 August 2022
ACCEPTED 23 September 2022
PUBLISHED 13 October 2022

CITATION
Li J, Zhang C, Guo H, Li S, You Y, Zheng P, Zhang H, Wang H and Bai J (2022) Non-invasive measurement of tumor immune microenvironment and prediction of survival and chemotherapeutic benefits from ^{18}F fluorodeoxyglucose PET/CT images in gastric cancer.
Front. Immunol. 13:1019386.
doi: 10.3389/fimmu.2022.1019386

COPYRIGHT
© 2022 Li, Zhang, Guo, Li, You, Zheng, Zhang, Wang and Bai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Non-invasive measurement of tumor immune microenvironment and prediction of survival and chemotherapeutic benefits from ^{18}F fluorodeoxyglucose PET/CT images in gastric cancer

Junmeng Li[1†], Chao Zhang[1*†], Huihui Guo[2], Shuang Li[3], Yang You[4], Peiming Zheng[5], Hongquan Zhang[6], Huanan Wang[7] and Junwei Bai[1*]

[1]Department of Gastrointestinal Surgery, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, China, [2]Department of Radiology, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, Henan, China, [3]Department of Pathology, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, China, [4]Department of Nuclear Medicine, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, Henan, China, [5]Department of Clinical Laboratory, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, China, [6]Department of Thoracic Surgery, The First Hospital Affiliated of Xinxiang Medical University, Xinxiang, China, [7]Department of Gastrointestinal Surgery, The First Hospital Affiliated of Zhengzhou University, Zhengzhou, China

Background: The tumor immune microenvironment could provide prognostic and predictive information. It is necessary to develop a noninvasive radiomics-based biomarker of a previously validated tumor immune microenvironment signature of gastric cancer (GC) with immunohistochemistry staining.

Methods: A total of 230 patients (training (n = 153) or validation (n = 77) cohort) with gastric cancer were subjected to (Positron Emission Tomography-Computed Tomography) radiomics feature extraction (80 features). A radiomics tumor immune microenvironment score (RTIMS) was developed to predict the tumor immune microenvironment signature with LASSO logistic regression. Furthermore, we evaluated its relation with prognosis and chemotherapy benefits.

Results: A 8-feature radiomics signature was established and validated (area under the curve=0.692 and 0.713). The RTIMS signature was significantly associated with disease-free survival and overall survival both in the training and validation cohort (all P<0.001). RTIMS was an independent prognostic factor in the Multivariate analysis. Further analysis revealed that high RTIMS patients benefitted from adjuvant chemotherapy (for DFS, stage II: HR 0.208

(95% CI 0.061-0.711), p=0.012; stage III: HR 0.321(0.180-0.570), p<0.001, respectively); while there were no benefits from chemotherapy in a low RTIMS patients.

Conclusion: This PET/CT radiomics model provided a promising way to assess the tumor immune microenvironment and to predict clinical outcomes and chemotherapy response. The RTIMS signature could be useful in estimating tumor immune microenvironment and predicting survival and chemotherapy benefit for patients with gastric cancer, when validated by further prospective randomized trials.

KEYWORDS

Gastric cancer, tumor immune microenvironment, chemotherapy, predictive, Positron Emission Tomography-Computed Tomography (PET/CT), radiomics

Introduction

Extensive studies have suggested tumor immune microenvironment are of clinical importance in cancer progression, metastasis, therapeutic response (1–4). The type, density, and location of immune cells in multiple cancers had a prognostic value that was superior to and independent of those of the TNM stage (1, 5, 6). An international consortium of 14 centers in 13 countries assessed the Immunoscore assay in patients with TNM stage I-III colon cancer, and the results supported the implementation of the consensus Immunoscore as a new component of a TNM-Immune classification of cancer (1). For gastric cancer (GC), a tumor immune microenvironment (TME) signature (7) of gastric cancer based on seven features, including CD3 $_{\text{invasive margin (IM)}}$, CD8 $_{\text{IM}}$, CD45RO $_{\text{center of tumor (CT)}}$, CD66b $_{\text{IM}}$, CD34, periostin, and cyclooxygenase-2, which could be important for predicting survival and selecting appropriate patients for chemotherapy. However, the TME signature was mainly determined on postoperative tissue specimens, it is necessary to develop a noninvasive pretreatment tools for prediction of immune infiltrates.

Computational medical imaging, known as radiomics, is an emerging field that converts medical images into a high dimensional quantitative feature space using a large number of automatically extracted data-characterization algorithms (8–10). These imaging features may capture in-depth characterization of tumor distinct phenotypes, with the underlying hypothesis that imaging reflects not only macroscopic but also the cellular and molecular properties of tissues. The goal of radiomics is to develop image-driven biomarkers that serve as instruments that generate a further understanding of cancer biology to facilitate better clinical decision-making (10–12). Radiomics features are complementary to biopsies and have the advantages of being non-invasive and repeated during treatment in routine practice, contrary to genomics or proteomics, which are still challenging to apply in clinical routine (12, 13). Connecting radiomics features to the molecular biological processes active in a tumor could provide deeper information that may complement the molecular data (14). Hence, in some circumstances, the radiomics features could be apply to infer the molecular biological underpinnings of tumor in individual patients.

The association between imaging features and tumor infiltrating immune cell density has been explored (13, 15–17). Many image features extracted by radiomics, not visually observed, were closely related to specific microscopic features at the molecular level and could characterize the tumor and its tumor microenvironment (13, 18–20). Several studies also found that pre-existing tumor immune infiltration correlates with patient response to anti-programmed cell death protein (PD)-1 and anti-programmed cell death ligand 1 (PD-L1) immunotherapy (21, 22).

Gastric cancer is one of the most common cancer and leading cause of cancer death worldwide (23). Imaging with fluorine 18 (18F) fluorodeoxyglucose (FDG) positron emission tomography (PET) was used to detect of distant and lymph nodes metastases, which could stage GC accurately (24). PET/CT could reflect the tumor metabolically active. Chemotherapy have improved survival of GC patients (25, 26). However, for most of GC patients the survival rates were still limited despite initial high response rates (26, 27). Therefore, precise classification of GC that could be necessary to predict patient survival and chemotherapy responses. Radiomics could allow evaluation of a tumor and its microenvironment, thus, may lead to the identification of novel predictors for prognosis and treatment efficiency. We aimed to develop a PET/CT image based radiomics signature of tumor immune microenvironment and to assess the ability of this signature to predict survival and adjuvant chemotherapy benefits.

Patients and methods

Study design and patients

Inclusion criteria were pathology-confirmed gastric cancer, 18 to 80 years old, PET/CT performed before surgery, complete follow-up data and clinicopathologic characteristics, no treatment cancer, and patient informed consent. Excluded criteria was received any anticancer therapy previous. Under approval from the institutional ethics committee, we retrospectively collected patient data and PET/CT images in the training cohort of 153 patients with GC as the inclusion criteria at Henan Provincial People's Hospital (Zhengzhou, China) between January 2012 and December 2020. As the same inclusion criteria, we included 77 patients in the independent validation cohort at the First Affiliated Hospital of Zhengzhou University (Zhengzhou, China) between July 2010 to December 2020 (Figure 1). Tumor staging was reclassified with the American Joint Committee on Cancer (AJCC) TNM Staging Manual, 8th Edition. All of the included patients accepted standard gastrectomy and D2 lymph node dissection accordance with the No.5 version guideline of the Japanese Gastric Cancer Association (JGCA). Disease-free survival (DFS) was defined as the time from the time of surgery until either the date of disease progression, which refers to tumor relapse, distal metastasis, or death, or until the date that the patient was last known to be free of progression. Overall survival (OS) was defined as the time to death from any cause. Clinicopathologic information for each patient with GC, including age, sex, TNM stage, tumor size, location, differentiation, lauren type, CEA, CA199, postoperative chemotherapy and follow-up data (follow-up duration and survival), and time of baseline PET/CT imaging and surgery, were collected from the clinical medical records (Table 1). The two institutional ethics committees approved the retrospective study respectively.

Immunohistochemistry (IHC) staining and classification of tumor microenvironment (TME) signature

An support vector machine (SVM) based tumor microenvironment signature integrating seven features, including CD3 $_{\text{invasive margin (IM)}}$, CD8 $_{\text{IM}}$, CD45RO $_{\text{center of tumor (CT)}}$, CD66b $_{\text{IM}}$, CD34, periostin, and cyclooxygenase-2, was previously developed and validated (7). Formalin-fixed paraffin-embedded (FFPE) samples were processed for IHC staining as previously described (7, 28–30). Following incubation with an antibody against human CD3 (pan T lymphocytes; NeoMarker, clone SP7), CD34 (Abcam, ab81289), CD8 (cytotoxic T lymphocytes; NeoMarker, clone SP16), CD45RO (memory T lymphocytes; Invitrogen, clone UCHL1) and CD66b (neutrophils; BD Pharmingen), periostin (Abcam,ab92460), and cyclooxygenase-2 (Abcam, Cambridge, MA), the sections were stained in an EnVision System (Dako) (Table S1). Two pathologists who were blinded to clinical outcomes independently scored all samples. If there was a difference opinion between the two primary pathologists, the third pathologist was consulted to give the final decision. As the previously described (7) and the result of IHC, every patient was classified into a high-SVM group and a low-SVM group. Detailed information is provided in the Supplementary Materials.

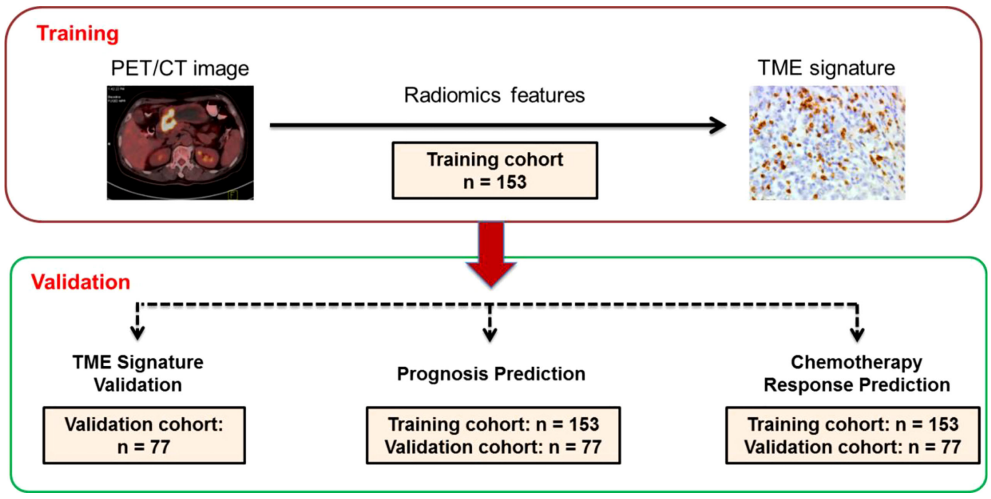


FIGURE 1
Study design for the discovery and validation of a PET/CT radiomic signature for the tumor immune microenvironment signature in gastric cancer. PET/CT image: positron emission tomography/computer tomography image. IH,: immunohistochemistry. Training cohort (n=153); Validation cohort (n=77).

PET/CT imaging

All patients underwent contrast-enhanced ^{18}F fluorodeoxyglucose (FDG) PET/CT scans before surgery. Details about image acquisition and processing procedures are presented in the Supplementary Materials. Tumor segmentation was then performed based on agreement reached by two expert radiology physicians, and checked by authors Yang You and Huihui Guo on the PET/CT images with *ITK-SNAP* software (www.itksnap.org) (31, 32).

Image feature extraction

We calculated a total of 80 quantitative features from each volume of interest (VOI) of each patient's PET/CT image to characterize intratumor heterogeneity and complexity. The feature pool included 14 first-order intensity features, 9 shape features, and 57 second- and higher-order textural features. In this work, we investigated four types of texture features on the basis of gray-level co-occurrence matrices (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix wavelet decompositions (NGTDM). 26, 13, 13 and 5 features were extracted from GLCM, GLRLM, GLSZM and NGTDM, respectively. The detailed mathematical definitions of all imaging features listed in Supplementary Materials. All radiomic features were extracted in Matlab R2012a (The MathWorks Inc.) using an available radiomic analysis package (https://github.com/mvallieres/radiomics/).

The SUV image was discretized by 0.1 SUV unit bin width according to the following equation (33): $SUV_{Dis}(x) = \lfloor SUV(x)/0.1 \rfloor - \min(\lfloor SUV(x)/0.1 \rfloor) + 1$, where $SUV(x)$ is the SUV of voxel x and $SUV_{Dis}(x)$ is the discretized value of voxel x. The discretization step is necessary to generate matrices whose size (defined by the maximum $SUV_{Dis}(x)$) highly impacts computation, and is used to reduce image noise and generate a constant intensity resolution so that textural features from different patients are comparable.

Construction of a radiomics tumor immune microenvironment score

We developed a logistic regression model to predict the IHC-based tumor immune microenvironment score (radiomics tumor immune microenvironment score, RTIMS) *via* a linear combination of selected image features weighted by their respective coefficients. We used the least absolute shrinkage and selection operator (LASSO) method to select the most useful predictive features from the training cohort. The diagnostic ability of the model was assessed with the area under the characteristics operating curves (AUC). The optimal cutoff value for RTIMS was determined using Youden's index in the training cohort, which maximizes the sum of sensitivity and specificity. This cutoff value was fixed and then applied to the validation cohort. The "glmnet" package was used to perform the LASSO regression model analysis (34, 35). Complete details are provided in Supplementary Materials.

Association with prognosis and chemotherapy benefits

The potential association of the RTIMS with DFS and OS was first assessed in the training cohort and then validated in the validation cohorts by using Kaplan-Meier survival analysis. Stratified analyses were performed to explore the potential association of RTIMS with DFS and OS using subgroups within TNM stage. The association between RTIMS and adjuvant chemotherapy response was assessed in patients with stage II and III GC.

Statistical analysis

We compared two groups using the *t*-test for continuous variables and the chi-square test or Fisher exact test for categorical variables, as appropriate. Survival curves were generated using the Kaplan-Meier method and compared using the log-rank test. Univariate and multivariate analyses were performed using the Cox proportional hazards model. Variables that achieved statistical significance at $P < 0.05$ were entered into the multivariate Cox regression analyses. Interactions between the RTIMS and chemotherapy were detected using the Cox model as well. All the statistical tests were done with the SPSS software (version 21.0) and R software (version 3.5.1). A two-sided P value < 0.05 was considered significant.

Results

Clinical characteristics

Table 1 list the detailed clinicopathological characteristics of the patients in the training (n=153) and validation (n=77) cohorts. All of the 230 patients included in the study, 163 (70.87%) were men, and the median (interquartile range, IQR) age was 58.0 (51.0-68.0) years. The number of patients with stage II or stage III GC who received adjuvant chemotherapy was 92 (60.1%) in the training cohort, 39 (50.6%) in the validation cohort.

Development and validation of Radiomics Tumor Immune Microenvironment Score (RTIMS)

In the training cohort, a LASSO logistic regression model was built. The final radiomics signature (RTIMS, Radiomics Tumor Immune Microenvironment Score) included 8 features

(Figure S1). The RTIMS calculation formula is RTIMS = -0.006 * SUV_SD - 0.134 * Hist_Energy + 0.152 * InVar_GLCM + 0.132 * LRHGE_GLRLM + 0.019 * SZLGE_GLSZM + 0.005 * ZSV_GLSZM + 0.195 * Complexity_NGTDM + 0.003 *Contrast_NGTDM. The ability of the radiomics signature to classify high versus low-SVM was shown to have an AUC of 0.692 (95% CI 0.6077-0.777) in the training cohort (Figure 2A). The radiomics signature showed similar accuracy in the validation cohort with AUC: 0.713 (95% CI 0.594-0.833) (Figure 2A). We further confirmed that the RTIMS score was significantly higher in the high TME group than the RTIMS in the low TME group in the training cohort and validation cohort, respectively (Figure 2B). The optimum cutoff of RTIMS determined by the ROC curve analysis in the training cohort was -0.069 (Figure 2A). Accordingly, patients were classified into a low-RTIMS group (RTIMS < -0.069) and a high-RTIMS group (RTIMS ≥-0.069). Table 1 lists the relationships between the RTIMS and clinicopathological characteristics.

Prognostic value of RTIMS

We first assessed the prognostic value of the RTIMS in the training cohort. For the low- RTIMS group, the 5-year DFS and OS were 16.39% and 22.73%; for the high- RTIMS group, the 5-year DFS and OS were 51.67% and 57.58% (hazard ratios (HRs) 0.361 (95%CI 0.239-0.547) and 0.291 (0.164-0.517), all $P<$ 0.0001; Figure 3A). We then performed the same analyses in the validation cohort and found similar results. The 5-year DFS and OS were 7.8% and 13.74% for the low-RTIMS group compared with 50.94% and 59.53% for the high-RTIMS group (HRs 0.339(0.219-0.525) and 0.232 (0.129-0.419), all $P<$ 0.0001; Figure 3B).

Univariate Cox regression analysis and multivariate Cox regression analysis were performed adjusting for clinicopathological variables. The RTIMS remained a powerful and independent prognostic factor for predicting DFS and OS in the training and validation cohorts (Table 2 and Table S2). Finally, we performed additional analyses within subgroups of GC patients who are stratified by stage. High-RTIMS patients had a longer DFS and OS than patients with low-RTIMS within each stage II or III (Figure 4).

Predictive value of RTIMS for chemotherapy benefits

Then, we evaluated the relationship between RTIMS status and survival among stage II and III patients who either received or did not receive postoperative chemotherapy. The characteristics of patients who received adjuvant chemotherapy were similar to those of patients who did not receive adjuvant chemotherapy (Table S3). The results showed that adjuvant chemotherapy was associated with an improved prognosis in the high-RTIMS group for both stage II and III disease, e.g., for DFS, stage II: HR 0.208 (95% CI 0.061-0.711), p=0.005; stage III: HR 0.321(0.180-0.570), p<0.001 (Figure 5). However, for patients in the low-RTIMS group, chemotherapy did not affect survival in either stage II or III disease: for DFS, stage II: HR 1.799(0.646-5.009), p=0.240; stage III: HR 1.302(0.744-2.277), p=0.340. Then, a formal interaction test was performed between the RTIMS signature and chemotherapy, which confirmed a significant interaction regarding the impact on DFS and OS in stage II disease (p for interaction: DFS, p=0.008; OS, p=0.034) as well as in stage III disease (p for interaction: DFS, p=0.001; OS, p=0.004). This analysis suggests a predictive effect of RTIMS for the benefit of adjuvant chemotherapy.

Discussion

Radiomics approaches, when combined with tumor biopsies and genomics, could improve treatment selection. Radiomic features from tumor can provide information on both the tumor and its microenvironment (18, 36). In our study, we developed and validated a radiomic signature (RTIMS) of the tumor immune microenvironment, and the RTIMS signature was able to predict survival. Moreover, the RTIMS signature might help to identify stage II and III patients who could benefit from adjuvant chemotherapy.

The importance of radiomics is shown by the increasing number of oncological clinical trials being done that use radiomics. To date, more than 30 clinical studies are registered in ClinicalTrials.gov, including one prospective study of pembrolizumab (NCT02644369). Notably, Braman et al. (18) that evaluated radiomic features in the context of neoadjuvant chemotherapy for breast cancer found radiomics features could strongly predict pCR independent of choice of classifier, suggesting their robustness as response predictors. Moreover, pathological complete response was associated with infiltration of lymphocytes into tumors (18) Jiang et al. measure immune score based on the radiomics features from CT image (17). And, the CT image signature could predict survival and treatment response, which is concordant with our results.

Notably, several points strengthening the biological and clinical relevance of this radiomic signature have been identified. Grossmann et al. discovered a connection between the radiomic phenotype of a tumor, the signaling pathways inside cells that drive how cancer develops, and clinical treatment outcomes (14). They also did IHC staining of CD3 in 22 tumors that were predicted to show relatively high or low immune response based on one radiomic feature, and they found agreement between radiomic features and pathology. Thus, they deemed that radiomic approaches permit noninvasive assessment of both molecular and clinical characteristics of

TABLE 1 Clinical characteristics of patients according to the radiomics signature in the training and external validation cohorts.

Variables	Training cohort (n =152)			Validation cohort (n = 77)		
	low RTIMS (%)	high RTIMS (%)	*P*	lowRTIMS (%)	high RTIMS (%)	*P*
Gender			0.772			0.93
Female	18 (38.3%)	29 (61.7%)		9 (45.0%)	11 (55.0%)	
Male	38 (35.8%)	68 (64.2%)		25 (43.9%)	32 (56.1%)	
Age (years), median (IQR)	61 (56-68)	57 (48-66)		57.5 (49.5-70.25)	58 (54-69)	
Age (years)			0.02			0.962
<60	22 (27.8%)	57 (72.2%)		18 (43.9%)	23 (56.1%)	
≧60	34 (45.9%)	40 (54.1%)		16 (44.4%)	20 (55.6%)	
Tumor size (cm)			0.518			0.072
<4	19 (33.3%)	38 (66.7%)		9 (31.0%)	20 (69.0%)	
≧4	37 (38.5%)	59 (61.5%)		25 (52.1%)	23 (47.9%)	
Tumor location			0.023			0.442
Cardia	26 (48.1%)	28 (51.9%)		14 (48.3%)	15 (51.7%)	
Body	6 (22.2%)	21 (77.8%)		2 (25.0%)	6 (75.0%)	
Antrum	15 (27.3%)	40 (72.7%)		10 (38.5%)	16 (61.5%)	
Whole	9 (52.9%)	8 (47.1%)		8 (57.1%)	6 (42.9%)	
Differentiation status			0.459			0.297
Well	11 (47.8%)	12 (52.2%)		5 (33.3%)	10 (66.7%)	
Moderate	10 (37.0%)	17 (63.0%)		7 (63.6%)	4 (36.4%)	
Poor and undifferentiated	35 (34.0%)	68 (66.0%)		22 (43.1%)	29 (56.9%)	
Lauren type			0.24			0.842
Intestinal type	28 (41.8%)	39 (58.2%)		15 (45.5%)	18 (54.5%)	
Diffuse or mixed type	28 (32.6%)	58 (67.4%)		19 (43.2%)	25 (56.8%)	
CEA			0.615			0.971
Normal	49 (37.4%)	82 (62.6%)		27 (40.3%)	34 (55.7%)	
Elevated	7 (31.8%)	15 (68.2%)		7 (43.8%)	9 (56.3%)	
CA199			0.232			0.638
Normal	43 (34.4%)	82 (65.6%)		22 (42.3%)	30 (57.7%)	
Elevated	13 (46.4%)	15 (53.6%)		12 (48.0%)	13 (52.0%)	
Depth of invasion			0.007			0.103
T1	1 (6.7%)	14 (93.3%)		0 (0.0%)	4 (100.0%)	
T2	6 (28.6%)	15 (71.4%)		3 (27.3%)	8 (72.7%)	
T3	2 (14.3%)	12 (85.7%)		2 (40.0%)	3 (60.0%)	
T4a	35 (43.2%)	46 (26.8%)		23 (46.9%)	26 (53.1%)	
T4b	12 (54.5%)	10 (45.5%)		6 (75.0%)	2 (25.0%)	
Lymph node metastasis			0.285			0.263
N0	12 (26.1%)	34 (73.9%)		8 (36.4%)	14 (63.6%)	
N1	10 (45.5%)	12 (54.5%)		7 (77.8%)	2 (22.2%)	
N2	11 (39.3%)	17 (60.7%)		5 (35.7%)	9 (64.3%)	
N3a	14 (35.0%)	26 (65.0%)		10 (41.7%)	14 (58.3%)	
N3b	9 (52.9%)	8 (47.1%)		4 (50.0%)	4 (50.0%)	
Stage			0.017			0.075
II	15 (25.0%)	45 (75.0%)		7 (29.2%)	17 (70.8%)	
III	41 (44.1%)	52 (55.9%)		27 (50.9%)	26 (49.1%)	
Chemotherapy			0.022			0.919
No	16 (25.8%)	46 (74.2%)		17 (44.7%)	21 (55.3%)	
Yes	41 (44.0%)	51 (56.0%)		17 (43.6%)	22 (56.4%)	

RTIMS, radiomics tumor immune microenvironment score.

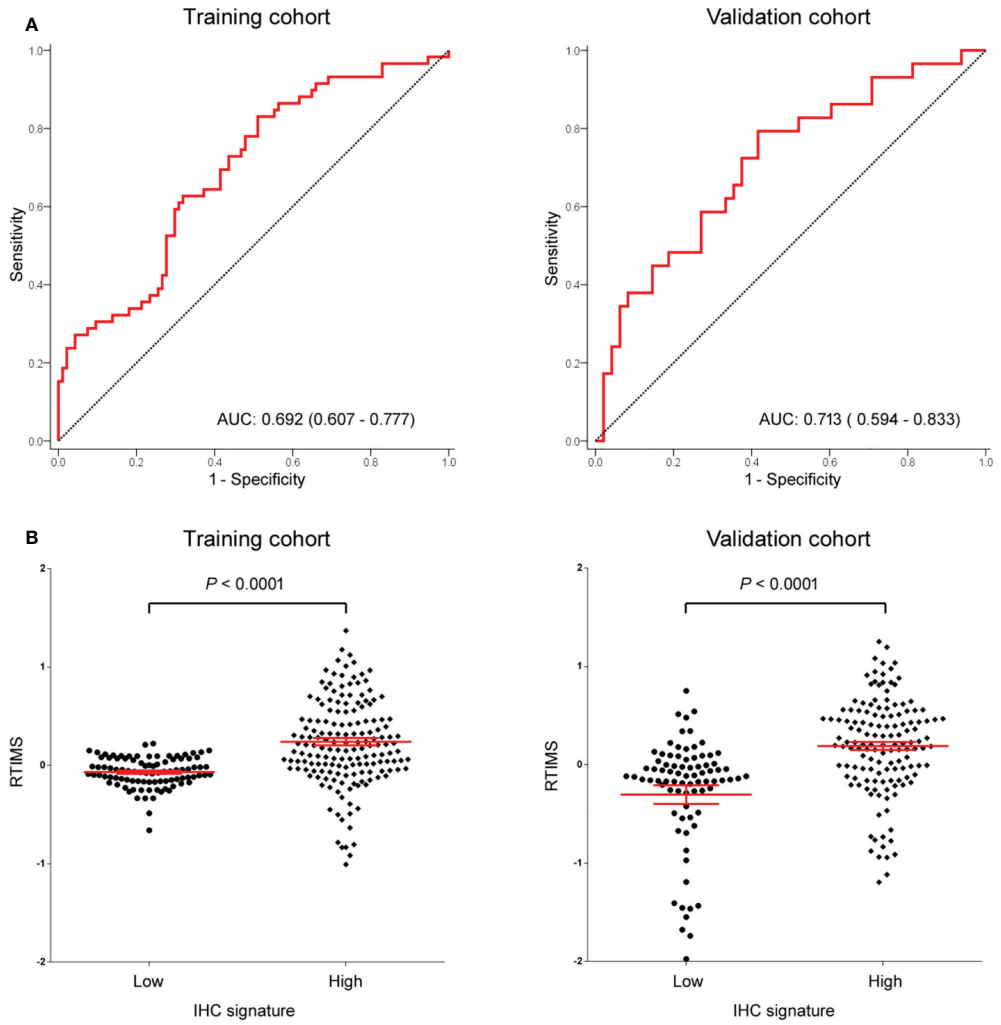


FIGURE 2
(A) AUC of the receiver operator characteristic of RTIMS predicting the tumor immune microenvironment signature in the training cohort and validation cohort. **(B)** RTIMS by high and low IHC signature. RTIMS, radiomics tumor immune microenvironment score. AUC, area under the curves.

tumors, however, the results still should be interpreted with caution. Ferté et al. also developed a radiomic signature for the gene expression signature of CD8 cells, which could predict clinical outcomes in patients treated with anti-PD-1 or anti-PD-L1 immunotherapy (15).

At present, the standard treatment for advanced gastric cancer includes adjuvant chemotherapy after surgery to prevent disease recurrence and improve survival, however, many studies have reported that a subgroup of patients could not benefit from adjuvant chemotherapy (25, 26, 28). Moreover, the criteria for the selection of candidates who are more likely to benefit from adjuvant chemotherapy remain controversial. Thus, the accurate identification of subgroups of patients will improve the prognostic system and lead to more personalized therapy. Recently, several studies reported that radiomics signatures based on CT/MRI/PET images were associated with chemotherapy response in several types of cancers (37–39). Besides, Ferté et al. and Jiang et al. previous study showed that imaging biomarkers could be used to estimate tumor-infiltrating lymphocytes (15, 16), which were associated with chemotherapy response (28, 40). In this study, we found that adjuvant chemotherapy provided a better survival benefit to patients with stage II and III GC patients classified as high-RTIMS, whereas low-RTIMS patients did not obtain benefits from adjuvant chemotherapy; further use of the radiomics signature might allow for better identification of patients who are most likely to benefit from adjuvant therapy. Thus, we suggest that patients with low-RTIMS should be treated with new combinations of more tolerable medication as an adjunct to potentiate the efficacy of systemic approaches. Therefore, our

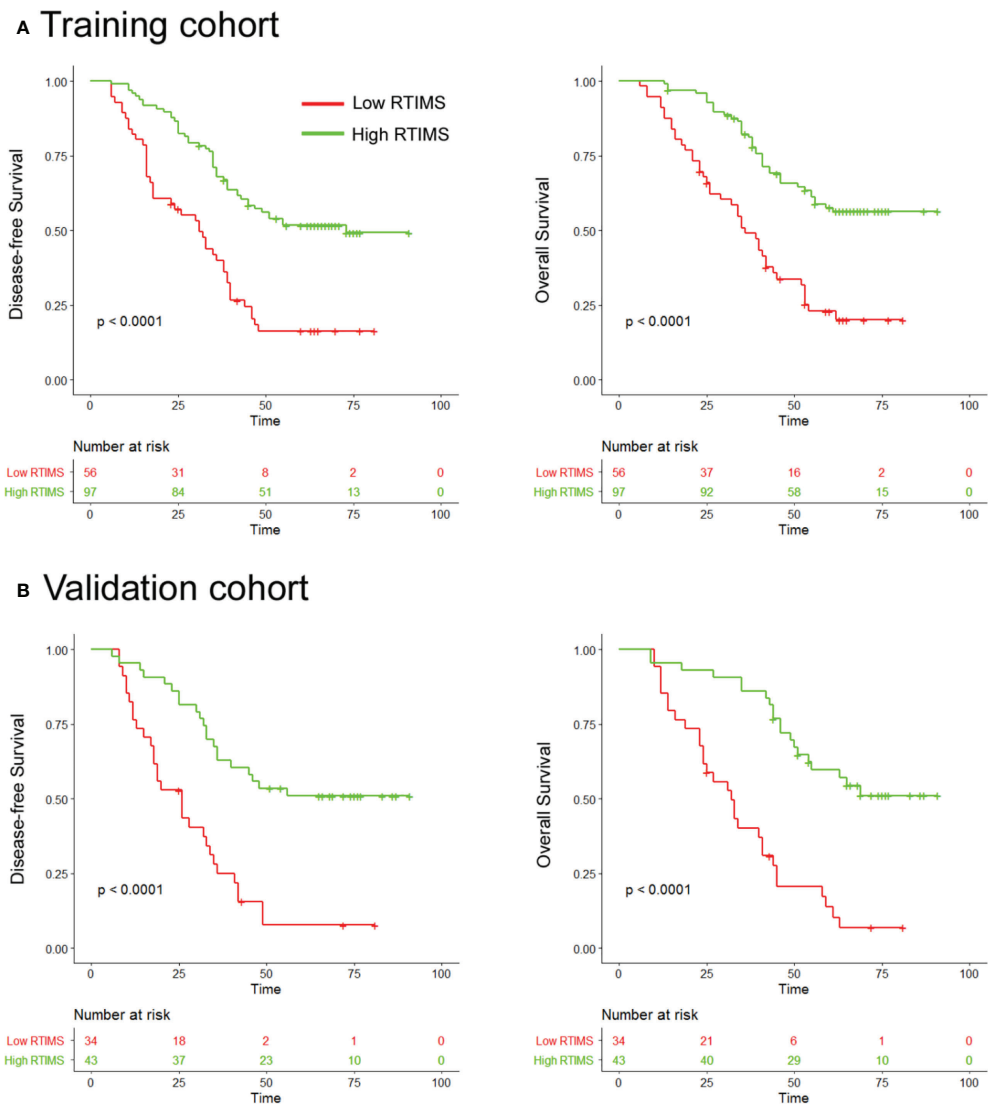


FIGURE 3
Kaplan-Meier analyses of disease-free survival (DFS) and overall survival (OS) according to dichotomized RTIMS signature in patients with gastric cancer. **(A)** Training cohort (n=153), **(B)** Validation cohort (n=77). Left panel: DFS; right panel: OS.

TABLE 2 Multivariate Cox Regression analyses for disease-free survival and overall survival in the training cohort of patients with gastric cancer.

Variables	Disease-free survival		Overall survival	
	HR (95%CI)	*p*	HR (95%CI)	*p*
Training cohort				
RTIMS(High vs. low)	0.404(0.266-0.615)	<0.0001	0.336(0.215-0.526)	<0.0001
Stage(III vs. II)	2.145(1.346-3.420)	0.001	2.036(1.219-3.403)	0.007
Differentiation	/	/	1.413(1.009-1.979)	0.044
Validation cohort				
RTIMS (High vs. low)	0.275(0.150-0.506)	<0.0001	0.198(0.107-0.367)	<0.0001
Stage(III vs. II)	2.066(1.020-4.184)	0.044	/	/
CEA	2.103(1.072-4.124)	0.031	2.446(1.264-4.731)	0.008

RTIMS, radiomics tumor immune microenvironment score.

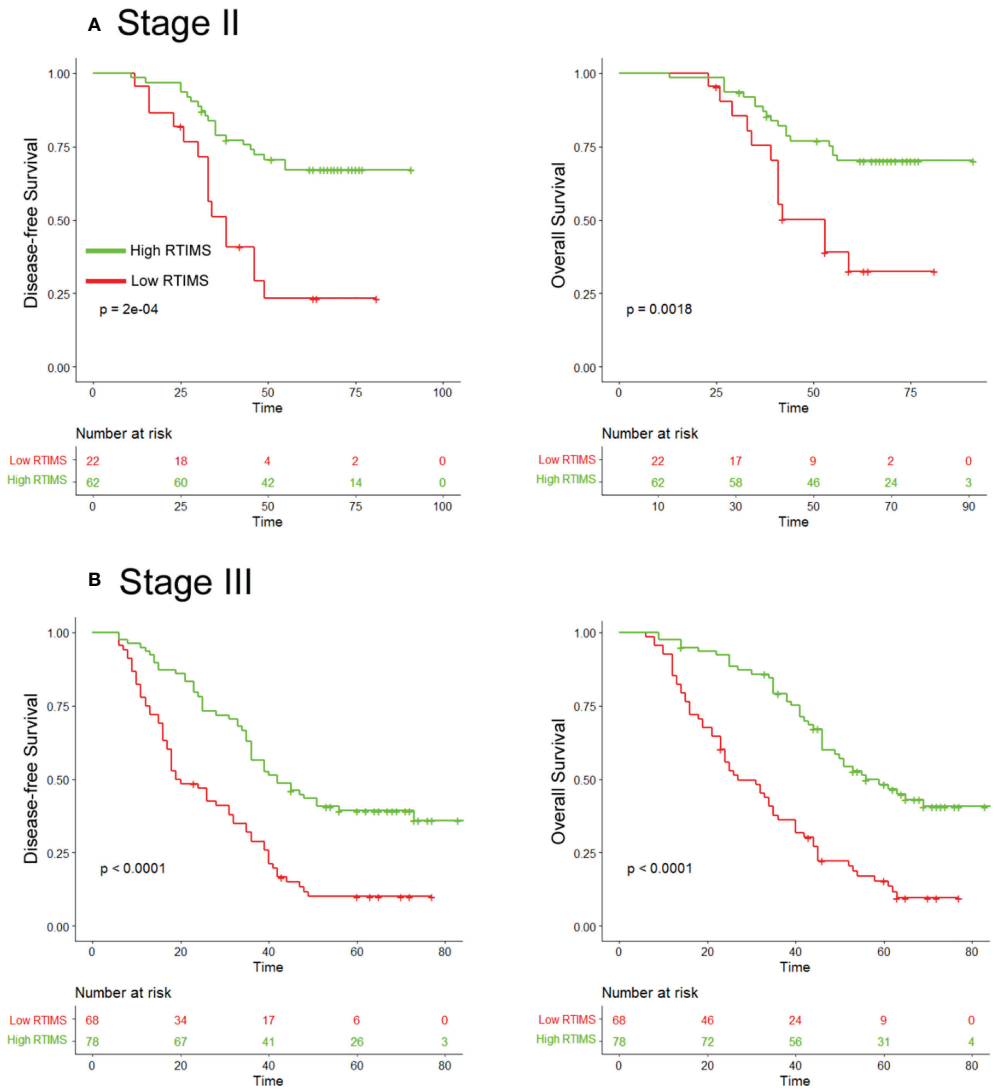


FIGURE 4
Kaplan-Meier survival analysis of disease-free survival and overall survival according to the RTIMS signature in subgroups of stage II and III GC patients. Disease-free survival (left pane) and overall survival (right pane): **(A)** Stage II (n = 84), **(B)** Stage III (n = 146).

PET/CT image-based RTIMS signature for patients with stage II and III gastric cancer is both a prognostic and predictive method, in that patients with a high- RTIMS have a clear benefit from adjuvant chemotherapy.

Interpretation of complex images by radio-pathology/genomic approaches is currently changing several fields in medical imaging, but clinical application of this method is still in its infancy. Validating this approach in a separate cohort, we confirmed the prognostic power of this approach. Thus, A novel biomarker that can be incorporated into existing clinical workflows because it only relies on PET/CT images, which are non-invasive and widely available. Compared to the gene-based signatures, a major advantage of our radiomics method is the ubiquitous availability of PET/CT images, which are available for every cancer patient, and analyzing them is not very costly. Besides, abdominal PET/CT scans are effective tools to diagnose and guide treatment for patients with gastric cancer. Whereas, those gene-based signatures have not been widely led into clinical application as initially expected owing to the variability of measurements in microarray and sequencing assays, inconsistencies in assay platforms, and the requirement for analytical expertise (41–43). In addition, our radiomic approach of the TME signature is reproducible: If presented with the same image twice, the algorithm will export the same result. These points make this new approach well suited for a clinical application.

10.3389/fimmu.2022.1019386

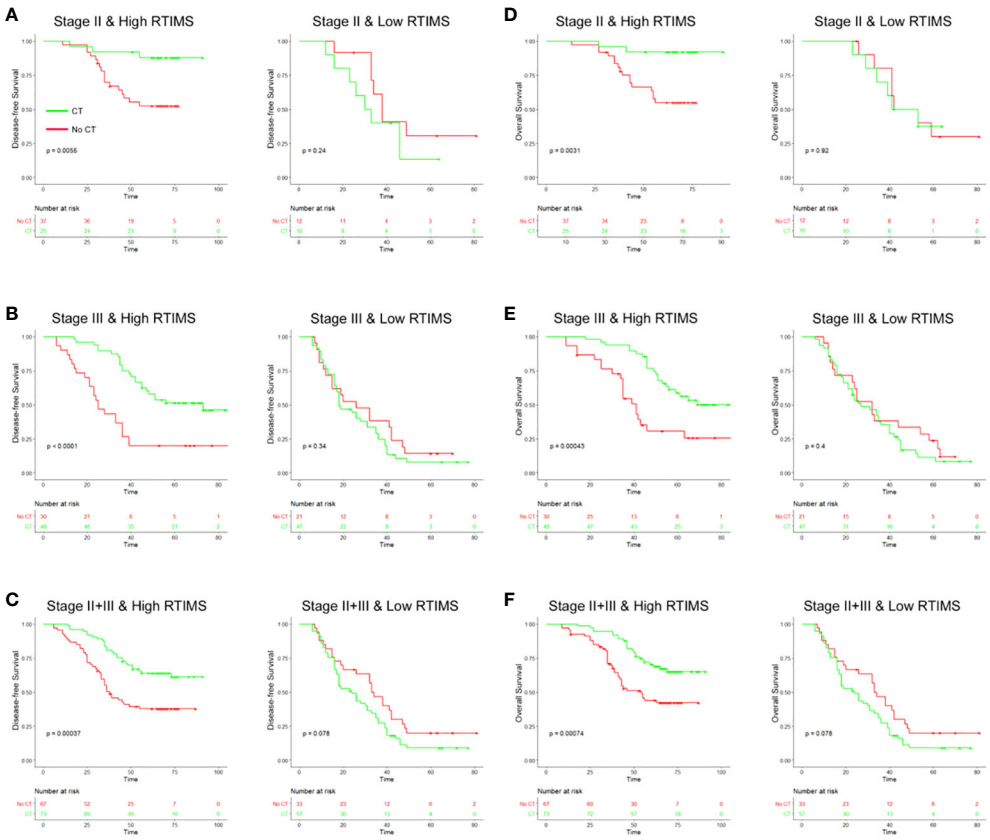


FIGURE 5
Relationship between the RTIMS signature and survival benefit from adjuvant chemotherapy in patients with stage II and III gastric cancer. In stage II, stage III and stage II+III patients, adjuvant chemotherapy improved DFS **(A–C)** and OS **(D–F)** for patients with high RTIMS, whereas no effect on survival was observed in patients with low RTIMS. RTIMS, radiomics tumor immune microenvironment score.

According to the molecular classification of The Cancer Genome Atlas (TCGA) project, gastric cancer was divided into four subtypes: Epstein-Barr virus (EBV)-positive, microsatellite instability (MSI), genomic stability, and chromosomal instability (44). High lymphocytes infiltration were frequently observed in certain molecular subtypes of GC, such as EBV-positive and MSI subtypes (45, 46). According to a phase II study, mismatch repair deficiency or MSI-high renders different solid tumors highly sensitive to immune checkpoint blockade with the PD-1 inhibitor pembrolizumab, and these tumors contain prominent lymphocytes infiltration (47, 48). Recently, several studies showed that the tumor imaging biomarkers could provide a promising way to predict the immune phenotype of tumors and to infer clinical outcomes for patients with cancer who had been treated with PD-1 and PD-L1 inhibitor (15, 16). Therefore, future studies should investigate the association between the radiomic signature of tumor immune microenvironment and molecular classification, and explore whether the radiomic signature could predict the responses of patients with gastric cancer to immunotherapy.

There are some limitations to this study. First, the study was conducted retrospectively, which was susceptible to the inherent biases of such a study format. Secondly, the decision to accept postoperative chemotherapy or not was made by the patients and clinicians together, that was not within a randomized assignment. Thirdly, the model was developed and validated using data from East Asian patients, and its generalizability in western populations remains to be determined. Therefore, a prospective, international, multicenter clinical trial will be needed to further validate our findings.

In conclusion, we developed a PET/CT image based radiomic signature that allows the noninvasive evaluation of

tumor immune microenvironment. Moreover, the RTIMS might be a useful predictive tool to identify stage II and III patients benefit from adjuvant chemotherapy. Thus, the RTIMS potentially may offer clinical value in directing individualized therapeutic regimen selection for patients with stage II and III gastric cancer.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of Henan Provincial Hospital. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

Conception and design: JL and JB. Collection and assembly of data: JL, HW, HZ, and SL. Data analysis and interpretation: CZ, YY, HG, and PZ. Manuscript writing: all authors. Final approval of manuscript: all authors.

Funding

This study was supported by the Project of Tackle Key Problems in Science and Technology of Henan Province (No.212102310679), and the Project Construction of National Key Disciplines (General Surgery Department, 2021).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.1019386/full#supplementary-material

References

1. Pages F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* (2018) 391(10135):2128–39. doi: 10.1016/S0140-6736(18)30789-X

2. Li TJ, Jiang Y, Hu YF, Huang L, Yu J, Zhao LY, et al. Interleukin-17-producing neutrophils link inflammatory stimuli to disease progression by promoting angiogenesis in gastric cancer. *Clin Cancer Res* (2017) 23(6):1575–85. doi: 10.1158/1078-0432.CCR-16-0617

3. Powell DR, Huttenlocher A. Neutrophils in the tumor microenvironment. *Trends Immunol* (2016) 37:41–52. doi: 10.1016/j.it.2015.11.008

4. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* (2012) 12:298–306. doi: 10.1038/nrc3245

5. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J Pathol* (2014) 232:199–209. doi: 10.1002/path.4287

6. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* (2006) 313:1960–4. doi: 10.1126/science.1129139

7. Jiang Y, Xie J, Huang W, Chen H, Xi S, Han Z, et al. Tumor immune microenvironment and chemosensitivity signature for predicting response to chemotherapy in gastric cancer. *Cancer Immunol Res* (2019) 7:2065–73. doi: 10.1158/2326-6066.CIR-19-0311

8. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuze S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol* (2017) 28:1191–206. doi: 10.1093/annonc/mdx034

9. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* (2014) 5:4006. doi: 10.1038/ncomms5006

10. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology* (2016) 278:563–77. doi: 10.1148/radiol.2015151169

11. Kumar V, Gu YH, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging* (2012) 30:1234–48. doi: 10.1016/j.mri.2012.06.010

12. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* (2017) 14:169–86. doi: 10.1038/nrclinonc.2016.162

13. Braman N, Prasanna P, Whitney J, Singh S, Beig N, Etesami M, et al. Association of peritumoral radiomics with tumor biology and pathologic response to preoperative targeted therapy for HER2 (ERBB2)-positive breast cancer. *JAMA Netw Open* (2019) 2:e192561. doi: 10.1001/jamanetworkopen.2019.2561

14. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife* (2017) 6:e23421. doi: 10.7554/eLife.23421

15. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* (2018) 19:1180–91. doi: 10.1016/S1470-2045(18)30413-3

16. Wu J, Li X, Teng X, Rubin DL, Napel S, Daniel BL, et al. Magnetic resonance imaging and molecular features associated with tumor-infiltrating lymphocytes in breast cancer. *Breast Cancer Res* (2018) 20:101. doi: 10.1186/s13058-018-1039-2

17. Jiang Y, Wang H, Wu J, Chen C, Yuan Q, Huang W, et al. Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer. *Ann Oncol* (2020) 31:760–8. doi: 10.1016/j.annonc.2020.03.295

18. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res* (2017) 19:57. doi: 10.1186/s13058-017-0846-1

19. Wu J, Li B, Sun X, Cao G, Rubin DL, Napel S, et al. Heterogeneous enhancement patterns of tumor-adjacent parenchyma at MR imaging are associated with dysregulated signaling pathways and poor survival in breast cancer. *Radiology* (2017) 285:401–13. doi: 10.1148/radiol.2017162823

20. Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur J Radiol* (2009) 70:232–41. doi: 10.1016/j.ejrad.2009.01.050

21. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. *Nature* (2017) 541:321–30. doi: 10.1038/nature21349

22. Kim JM, Chen DS. Immune escape to PD-L1/PD-1 blockade: seven steps to success (or failure). *Ann Oncol* (2016) 27:1492–504. doi: 10.1093/annonc/mdw217

23. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2018) 68:394–424. doi: 10.3322/caac.21492

24. De Raffele E, Mirarchi M, Cuicchi D, Lecce F, Cola B. Evolving role of FDG-PET/CT in prognostic evaluation of resectable gastric cancer. *World J Gastroenterol* (2017) 23:6923–6. doi: 10.3748/wjg.v23.i38.6923

25. Noh SH, Park SR, Yang HK, Chung HC, Chung IJ, Kim SW, et al. Adjuvant capecitabine plus oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): 5-year follow-up of an open-label, randomised phase 3 trial. *Lancet Oncol* (2014) 15:1389–96. doi: 10.1016/S1470-2045(14)70473-5

26. Jiang Y, Li T, Liang X, Hu Y, Huang L, Liao Z, et al. Association of adjuvant chemotherapy with survival in patients with stage II or III gastric cancer. *JAMA Surg* (2017) 152:e171087. doi: 10.1001/jamasurg.2017.1087

27. Razzak M. Genetics: new molecular classification of gastric adenocarcinoma proposed by the cancer genome atlas. *Nat Rev Clin Oncol* (2014) 11:499. doi: 10.1038/nrclinonc.2014.138

28. Jiang Y, Zhang Q, Hu Y, Li T, Yu J, Zhao L, et al. ImmunoScore signature: A prognostic and predictive tool in gastric cancer. *Ann Surg* (2018) 267:504–13. doi: 10.1097/SLA.0000000000002116

29. Jiang Y, Xie J, Han Z, Liu W, Xi S, Huang L, et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. *Clin Cancer Res* (2018) 24:5574–84. doi: 10.1158/1078-0432.CCR-18-0848

30. Jiang Y, Liu W, Li T, Hu Y, Chen S, Xi S, et al. Prognostic and predictive value of p21-activated kinase 6 associated support vector machine classifier in gastric cancer treated by 5-fluorouracil/Oxaliplatin chemotherapy. *EBioMedicine* (2017) 22:78–88. doi: 10.1016/j.ebiom.2017.06.028

31. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015

32. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of radiomic features in [(11)C]Choline and [(18)F]FDG PET/CT imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization. *Mol Imaging Biol* (2016) 18:935–45. doi: 10.1007/s11307-016-0973-6

33. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep* (2015) 5:11075. doi: 10.1038/srep11075

34. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models *via* coordinate descent. *J Stat Softw* (2010) 33:1–22. doi: 10.18637/jss.v033.i01

35. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model *via* coordinate descent. *J Stat Softw* (2011) 39:1–13. doi: 10.18637/jss.v039.i05

36. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* (2014) 515:568–71. doi: 10.1038/nature13954

37. Peng H, Dong D, Fang M, Li L, Tang LL, Chen L, et al. Prognostic value of deep learning PET/CT-based radiomics: Potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. *Clin Cancer Res* (2019) 25(14): 4271–9. doi: 10.1158/1078-0432.CCR-18-3065

38. Jiang Y, Yuan Q, Lv W, Xi S, Huang W, Sun Z, et al. Radiomic signature of (18)F fluorodeoxyglucose PET/CT for prediction of gastric cancer survival and chemotherapeutic benefits. *Theranostics* (2018) 8:5915–28. doi: 10.7150/thno.28018

39. Beukinga RJ, Hulshoff JB, Mul VEM, Noordzij W, Kats-Ugurlu G, Slart R, et al. Prediction of response to neoadjuvant chemotherapy and radiation therapy with baseline and restaging (18)F-FDG PET imaging biomarkers in patients with esophageal cancer. *Radiology* (2018) 287:983–92. doi: 10.1148/radiol.2018172229

40. Mlecnik B, Van den Eynde M, Bindea G, Church SE, Vasaturo A, Fredriksen T, et al. Comprehensive intrametastatic immune quantification and major impact of immunoscore on survival. *J Natl Cancer Inst* (2018) 110 (1): 97–108. doi: 10.1093/jnci/djx123

41. Zhu ZH, Sun BY, Ma Y, Shao JY, Long H, Zhang X, et al. Three immunomarker support vector machines-based prognostic classifiers for stage IB non-Small-Cell lung cancer. *J Clin Oncol* (2009) 27:1091–9. doi: 10.1200/JCO.2008.16.6991

42. Ueda T, Volinia S, Okumura H, Shimizu M, Taccioli C, Rossi S, et al. Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol* (2010) 11:136–46. doi: 10.1016/S1470-2045(09)70343-2

43. Widschwendter M, Jones A, Evans I, Reisel D, Dillner J, Sundstrom K, et al. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat Rev Clin Oncol* (2018) 15:292–309. doi: 10.1038/nrclinonc.2018.30

44. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* (2014) 513:202–9. doi: 10.1038/nature13480

45. Tan P, Yeoh KG. Genetics and molecular pathogenesis of gastric adenocarcinoma. *Gastroenterology* (2015) 149:1153–62.e1153. doi: 10.1053/j.gastro.2015.05.059

46. Wang W, Wang K, Chen Z, Chen L, Guo W, Liao P, et al. Immunoclassification characterized by CD8 and PD-L1 expression is associated with the clinical outcome of gastric cancer patients. *Oncotarget* (2018) 9:12164–73. doi: 10.18632/oncotarget.24037

47. Lee V, Murphy A, Le DT, Diaz LA Jr. Mismatch repair deficiency and response to immune checkpoint blockade. *Oncologist* (2016) 21:1200–11. doi: 10.1634/theoncologist.2016-0046

48. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* (2017) 357:409–13. doi: 10.1126/science.aan6733

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 07 November 2022
DOI 10.3389/fonc.2022.1034810

Check for updates

OPEN ACCESS

EDITED BY
Dinler Amaral Antunes,
University of Houston, United States

REVIEWED BY
Iria Gomez-Tourino,
University of Santiago de Compostela, Spain
Massimo Andreatta,
Université de Lausanne, Switzerland

*CORRESPONDENCE
Arnav Solanki
solan053@umn.edu

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Oncology

RECEIVED 02 September 2022
ACCEPTED 17 October 2022
PUBLISHED 07 November 2022

CITATION
Solanki A, Riedel M, Cornette J, Udell J and Vasmatzis G (2022) Hydrophobicity identifies false positives and false negatives in peptide-MHC binding. *Front. Oncol.* 12:1034810. doi: 10.3389/fonc.2022.1034810

COPYRIGHT
© 2022 Solanki, Riedel, Cornette, Udell and Vasmatzis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Hydrophobicity identifies false positives and false negatives in peptide-MHC binding

Arnav Solanki[1]*, Marc Riedel[1], James Cornette[2], Julia Udell[1,3] and George Vasmatzis[3]

[1]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, United States, [2]Department of Mathematics, Iowa State University, Ames, IA, United States, [3]Biomarker Discovery Group, Mayo Clinic, Center for Individualized Medicine, Rochester, MN, United States

Major Histocompability Complex (MHC) Class I molecules allow cells to present foreign and endogenous peptides to T-Cells so that cells infected by pathogens can be identified and killed. Neural networks tools such as NetMHC-4.0 and NetMHCpan-4.1 are used to predict whether peptides will bind to variants of MHC molecules. These tools are trained on data gathered from binding affinity and eluted ligand experiments. However, these tools do not track hydrophobicity, a significant biochemical factor relevant to peptide binding, in their predictions. A previous study had concluded that the peptides predicted to bind to HLA-A*0201 by NetMHC-4.0 were much more hydrophobic than expected. This paper expands that study by also focusing on HLA-B*2705 and HLA-B*0801, which prefer binding hydrophilic and balanced peptides respectively. The correlation of hydrophobicity of 9-mer peptides with their predicted binding strengths to these various HLAs was investigated. Two studies were performed, one using the data that the two neural networks were trained on, and the other using a sample of the human proteome. NetMHC-4.0 was found to have a statistically significant bias towards predicting highly hydrophobic peptides as strong binders to HLA-A*0201 and HLA-B*2705 in both studies. Machine Learning metrics were used to identify the causes for this bias: hydrophobic false positives and hydrophilic false negatives. These results suggest that the retraining the neural networks with biochemical attributes such as hydrophobicity and better training data could increase the accuracy of their predictions. This would increase their impact in applications such as vaccine design and neoantigen identification.

KEYWORDS

MHC class I, peptide, hydrophobicity, machine learning, neural networks

1 Introduction

The Human Leukocyte Antigen (HLA) gene system encodes cell-surface proteins that play a key role in the immune system. HLA proteins of Major Histocompatibility Complex (MHC) Class I allow nucleated cells to present peptides from within the cell (1). In these cells, endogenous proteins are eventually broken down into small peptides, 8-15 amino acids long, by the proteasome. These antigens are then trafficked to and loaded onto MHC Class I molecules. If sufficient binding affinity is achieved then a stable peptide-MHC (pMHC) complex is formed and transported to the cell surface. Self-peptides, antigens encoded in the human proteome, and foreign peptides, derived from pathogenic proteins, can thus be presented. By surveilling these extracellular pMHCs, $CD8^+$ T-cells can distinguish normal cells from pathogen-infected cells, and kill the latter.

The mechanics of peptide binding are specific to a given MHC variant. The HLA genes are among the most diverse in the human population (2). Thus the set of all antigens presented by a person's MHCs, labelled as their *immunopeptidome*, is unique and determines the capacity of their immune system. Since the immune response of a person to a viral infection like COVID-19, for instance, is dependent on whether the foreign antigens presented by their MHCs are distinguishable from self-peptides, understanding and predicting pMHC binding is an important topic. In this paper, we have focused on NetMHC-4.0 (3) and NetMHCpan-4.1 (4), two state-of-the-art neural network based methods that predict pMHC binding. Both software tools have been applied in predicting cancer immune escape mechanisms (5), checkpoint blockade immunotherapy for tumors (6), and identifying COVID-19 T-cell response targets (7).

While these tools provide valuable pMHC predictions, they do not model pMHC binding at the molecular level or capture the entire antigen presentation pathway's effects. Hydrophobicity is a measure of how repulsive a molecule is to water, often a consequence of nonpolarity. It plays a vital role in protein binding – for example, the MHC molecule HLA-A*0201 (A2) contains hydrophobic binding pockets that bind to correspondingly hydrophobic amino acids (8). In contrast, the MHC HLA-B*2705 (B27) prefers to bind peptides with a hydrophilic amino acid in one of its pockets (9). Historically, immunopeptidomes have been predicted by modelling the interaction of the MHC binding pocket and peptide, particularly focusing on biochemical attributes such as sidechain conformations, solvation energies, electrostatic interactions, and hydrophobicity (10, 11). However with improved computing power, larger datasets, and the need for interpolation due to the high polymorphism in MHC Class I alleles (12), artificial intelligence based methods have become popular over such mechanistic means of prediction. As NetMHC-4.0 and NetMHCpan-4.1 are trained with sequence data and binding scores only, they lack the means of modelling these biochemical attributes. Other software tools such as ANN-Hydro (13) have utilized hydrophobicity in their immunogenic predictions, but do not predict binding affinity and are outperformed by NetMHCpan (14). In our use of NetMHC-4.0 we had observed a prevalence of highly hydrophobic peptides in the predicted A2 immunopeptidome. We had found this contrary to our expectations, since peptides in which all amino acids are hydrophobes would not dissolve in the aqueous cytosol within the cell and would thus likely not be available for binding with the MHC. We had therefore sought to investigate the possibility that these tools were over-estimating binding scores for such hydrophobic peptides. In a previous study (15), we had tested these tools' predictions on A2 and observed hydrophobic biases that suggested a false positive problem in NetMHC-4.0. Here, we expanded that study to look at multiple HLAs with different binding preferences in more detail. Once again, we conducted two analyses on both NetMHC-4.0 and NetMHCpan-4.1, one using training data and the other using a sample of the human proteome, to investigate the correlation of predicted strong binders and hydrophobicity. We present our results and highlight the unintended bias within NetMHC-4.0 for predicting hydrophobic peptides as strong binders, and for predicting hydrophilic peptides as non-binders.

2 Methods

NetMHC-4.0 and NetMHCpan-4.1 allow users to input a list of peptides or whole proteins, and test the binding of all peptides with a chosen MHC molecule. Both tools return an adjusted score between 0 (for non-binders) and 1 (for strong binders) for all peptides. A notable distinction between the two is that NetMHC is limited to predicting binding for MHC variants it is trained on, i.e. curated MHCs. In contrast, NetMHCpan is capable of interpolating predictions for uncurated MHCs if users provide the MHC amino acid sequence. This is achieved through the integration of MHC sequence as a data feature in training, and by a larger training dataset generated using a sophisticated machine learning method called NNAlign_MA (16). NetMHCpan-4.1 consists of an ensemble of 50 neural networks, each with hidden layers containing 55 and 66 neurons, that were trained using 5-fold cross validation. NetMHC-4.0 consists of 20 neural networks, each with a single hidden layer of 5 neurons, that were trained using a nested 5-fold cross validation approach (3).

2.1 Data mining

NetMHC-4.0 was trained on $CD8^+$ epitope binding affinity (BA) data from the Immune Epitope Database. This data provides binding scores for peptides to single allele MHCs,

with a score that is scaled between 0 and 1 that measures how strongly the peptide binds. NetMHCpan-4.1 was trained on BA data and additional eluted ligand (EL) data from mass spectrometry experiments from multiple sources (4). The EL data includes multi-allele information that was deconvoluted into single allele datapoints using NNAlign_MA. EL score is binary (either 0 or 1) since it checks if a peptide is present in a MHC's immunopeptidome. The combined BA and EL dataset contained more than 13 million pMHC data points spread across numerous HLAs. For all our analyses, we focused on peptides of length 9, i.e. 9-mers, as these are the most frequent length of antigens in human immunopeptidomes. Also, we chose to analyze the 3 MHC molecules HLA-A02:01 (A2), HLA-B27:05 (B27), and HLA-B08:01 (B8). We picked these HLAs because they were highly represented in the training set (A2 ranked 1st, B27 ranked 11th, and B8 ranked 8th based on number of training datapoints), they are HLA supertypes (they represent the behavior of numerous less frequent HLA types), and they have different binding motifs (discussed in section 2.2).

Our first analysis was the training data analysis. For each HLA, we collected all its 9-mers that were reported in the training dataset. A2 had 52569 9-mers, B27 had 17422, and B8 had 19448. The distributions of the experimentally obtained training scores for these HLAs are shown in Figure 1. We ran NetMHC-4.0 and NetMHCpan-4.1 on these 9-mers to gather each neural network's predicted binding scores with the corresponding HLAs. These scores are shown in Figure 1. Furthermore, both tools classified 9-mers with large enough predicted binding scores (using a 0.5% rank to be precise) as strong binders for a tested HLA. We measured these thresholds by finding the lowest predicted binding score for a strong binder identified by these tools. Figure 1 also shows these measured thresholds. Please note that this component of the training data approach had been used in our previous publication (15) for analyzing A2, so only B27 and B8 results are shown in this Figure.

As the training scores were available for all 9-mers we tested in the training data analysis, we also calculated confusion matrices (i.e. we counted the number of positives and negatives, both true and false) for both neural network tools. We used the previously measured strong binding thresholds on the actual training scores for each 9-mer to identify actual strong binders and non-binders in the context of each neural network tool. For example, in Figure 1 all 9-mers on the blue plot above the red dashed line were classified as actual strong binders when testing NetMHC-4.0. The results of the confusion matrices are shown in Table S1. We also plotted the receiver operating characteristic (ROC) curve for each neural network tool for all

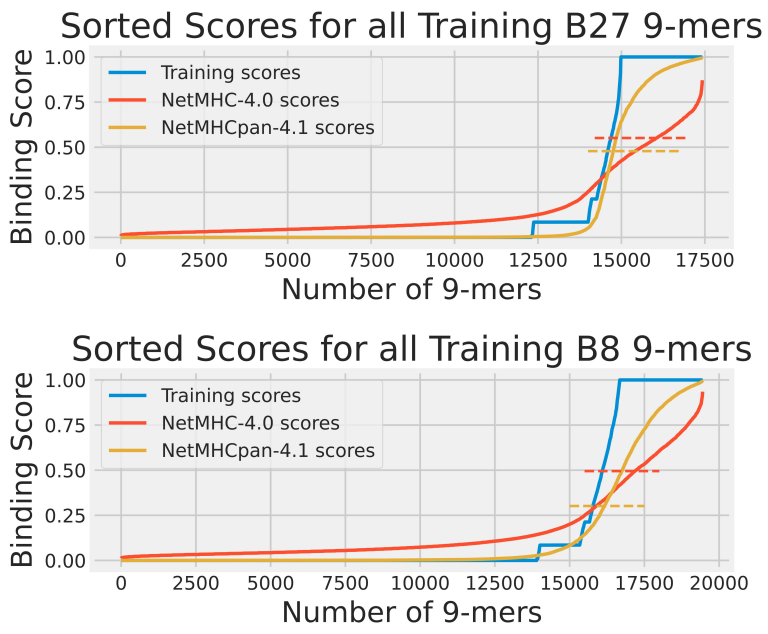


FIGURE 1
The cumulative distribution of the experimental training scores (blue), NetMHC-4.0 predicted scores (red), and NetMHCpan-4.1 predicted scores (yellow) for peptides in the training dataset for HLAs A2, B27, and B8. The strong binder thresholds for NetMHC-4.0 and NetMHCpan-4.1 are shown as dashed lines of the corresponding colors. For B27, these were 0.551 and 0.478, and for B8 these were 0.495 and 0.301 respectively. Each plot of scores was independently sorted. Consequently, the order of peptides is not conserved across the 3 plots in each subfigure. Note that the A2 results can be accessed from our previous study (15). For A2, the NetMHC-4.0 and NetMHCpan-4.1 thresholds were 0.659 and 0.419 respectively.

3 HLAs, as shown in Figure S1. We also computed the accuracy, precision, recall, and F1 score from the confusion matrices (17).

While the training data analysis was useful for identifying prediction biases, it alone was not a sufficient means for comparing NetMHC-4.0 and NetMHCpan-4.1. As NetMHC-4.0 was only trained on BA data while NetMHCpan-4.1 was trained on BA and EL data, NetMHCpan-4.1 had an advantage of having "seen" the EL peptides in its training over NetMHC-4.0. Therefore, we performed a human proteome analysis. We gathered the protein sequences for all reviewed human proteins from Uniprot (18), randomly sampled 100 of them, and fragmented them to create a set of 50804 9-mers. These peptides were also passed through NetMHC-4.0 and NetMHCpan-4.1 for all 3 HLAs to gather their predicted scores. These scores are shown in Figure 2. Since no experimentally obtained binding scores were available for these peptides, the Pearson correlations and the confusion matrices were not calculated. Again, this sampled human proteome approach had also been used in our previous publication (15) for analyzing A2 so only B27 and B8 results are shown here.

2.2 Hydrophobicity

As noted in section 2.1, one of the reasons we chose A2, B27, and B8 as our 3 target HLAs was their different binding motifs (19). A2 has a strong affinity for 9-mers with hydrophobic amino acids such as L, V, M, and I in positions 2 and 9. B27, on the other hand, binds 9-mers with hydrophilic R at position 2. In between these two, B8 prefers to bind 9-mers with both hydrophobic amino acids L, V, M, and I at position 2 and 9, but also hydrophilic amino acids R and K at position 3 and 5. Clearly hydrophobicity plays a crucial role in distinguishing the binding preferences of different HLAs. For our analyses, we decided to investigate the role of hydrophobicity in NetMHC-4.0's and NetMHCpan-4.1's predictions.

Hydrophobicity scales assign hydrophobicity values to single amino acids. They are designed so the hydrophobicity of long peptides or protein chains can be estimated by simply linearly adding up the scores of their constituent amino acids. Scales such as Kyte-Doolittle (20), Cornette (21), and Hopp-Woods (22) are commonly used. However, we settled on the Moon scale (23) for calculating hydrophobicity in our analyses as it specifically focuses on the sidechain hydrophobicity and polarity of single amino acids. Unlike the other scales, which are well suited for protein folding problems that do not correlate with sidechain hydrophobicity (24), the Moon scale is more representative of how small peptides would behave in an aqueous solution. The scale ranks the 20 amino acids in decreasing order of hydrophobicity as follows: F (1.43), L (1.26), I (1.15), P (1.13), Y (0.94), V (0.80), M (0.79), W (0.63), A (0.46), C (0.24), E (-0.27), G (-0.30), T (-0.33), S

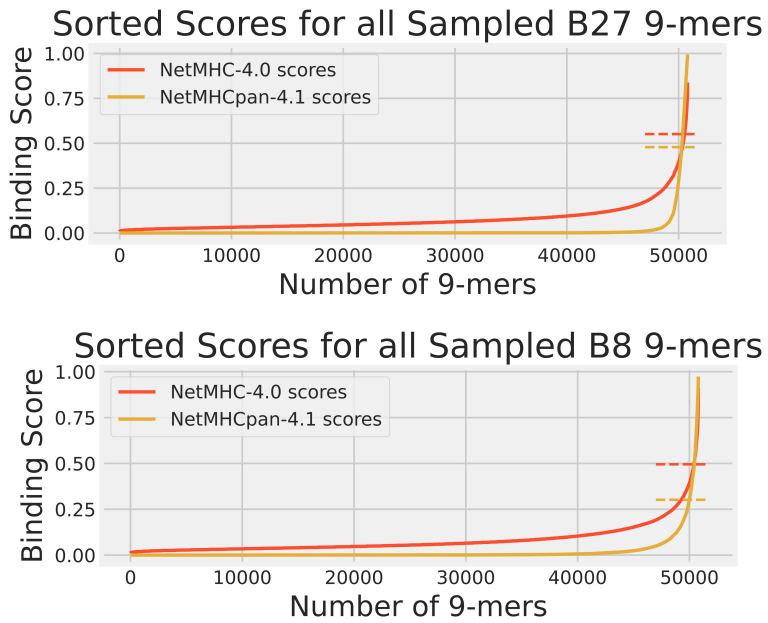


FIGURE 2
The cumulative distribution of NetMHC-4.0 predicted scores (red) and NetMHCpan-4.1 predicted scores (yellow) for peptides in the human proteome dataset for HLAs A2, B27, and B8. The strong binder thresholds for NetMHC-4.0 and NetMHCpan-4.1 are shown as dashed lines of the corresponding colors. These thresholds are the same as those in Figure 1. Each plot of scores was independently sorted. Consequently, the order of peptides is not conserved across the 2 plots in each subfigure. Note that the A2 results can be accessed from our previous study (15).

(-0.35), D (-0.85), Q (-0.88), N (-1.08), R (-1.19), H (-1.65), K (-1.93).

For any given 9-mer, we calculated its total hydrophobicity by adding up the Moon scale values for each of its 9 amino acids. For any given set of peptides, we measured the mean and standard deviation of the hydrophobicity scores of all peptides in it. Furthermore, we classified any given peptide into 1 of 3 classes: Hydrophobic (total hydrophobicity greater than 3), Hydrophilic (total hydrophobicity less than -3), or Balanced (total hydrophobicity between -3 and 3). This classification distinguished peptides based on their net hydrophobicity, and allowed us to investigate the impact of hydrophobicity on the differential prediction of MHC binding for different classes of peptides. We added these categories in Figure S1 and Tables S1 and Table 1 to highlight any prominent trends specific to a peptide category.

It was possible the smaller BA training dataset for NetMHC-4.0 was biased or unrepresentative of the numerous possible binding peptides. This bias could also be caused due to the binding affinity assays used to obtain BA scores, since these experiments only measure MHC-peptide affinity and do not account for the rest of the antigen presentation pathway or physiological conditions. Therefore, we also compared the hydrophobicities of the set of BA training data points (i.e. peptides NetMHC-4.0 was trained on) to the set of EL training data points (i.e. more than 80% of the peptides NetMHCpan-4.1 was trained on) for all 3 HLAs.

3 Results

From the scores shown in Figure 1, it was clear that the pMHC binding data fit a mostly binary data classification problem, since only 15% of the analyzed peptides had a training score not equal to 0 or to 1. This was mostly due to the addition of EL data which provided a binary "yes" or "no" answer to whether a given peptide was found attached to our chosen HLAs through mass spectroscopy. NetMHC-4.0's predicted scores were dispersed smoothly between 0 and 1. In contrast, NetMHCpan-4.1 had more lopsided predictions with more non-binders assigned a binding score of 0. However, neither neural network tool predicted a definitive score of 1 to strong binders and instead used their thresholds to identify binders. NetMHCpan-4.1 predicted more strong binders than NetMHC-4.0 in all 3 cases. A2 results can be accessed from our previous publication (15).

These results of NetMHC-4.0 and NetMHCpan-4.1 on the sample human proteome are shown in Figure 2. Note that no experimental binding data was available for these peptides, and that the same set of these peptides was used for each HLA's predictions when comparing Figure 2 with 1. Again, NetMHCpan-4.1 seemed more stringent in predicting non-zero binding scores. NetMHCpan-4.1 also predicted slightly more strong binders than NetMHC-4.0 for all 3 HLAs.

The performances of NetMHC-4.0 and NetMHCpan-4.1 as binary classifiers are shown in Figure S1 as ROC curves. The

TABLE 1 Various binary classification metrics on the training data analysis for NetMHC-4.0 (N-4.0) and NetMHCpan-4.1 (NP-4.1).

HLA	Peptide Case	Accuracy		Precision		Recall		F1 score	
		N-4.0	NP-4.1	N-4.0	NP-4.1	N-4.0	NP-4.1	N-4.0	NP-4.1
A2	All	0.918	0.936	0.882	0.918	0.618	0.756	0.727	0.829
	Hydrophobic only	0.875	0.871	0.861	0.922	0.745	0.735	0.745	0.818
	Hydrophilic only	0.986	0.991	0.786	0.826	0.134	0.613	0.229	0.704
	Balanced only	0.924	0.952	0.909	0.915	0.514	0.776	0.657	0.840
B27	All	0.909	0.965	0.914	0.917	0.460	0.865	0.612	0.890
	Hydrophobic only	0.926	0.969	0.954	0.964	0.535	0.832	0.685	0.893
	Hydrophilic only	0.916	0.957	0.793	0.832	0.362	0.814	0.497	0.823
	Balanced only	0.903	0.966	0.918	0.919	0.454	0.878	0.608	0.898
B8	All	0.927	0.951	0.925	0.915	0.625	0.818	0.746	0.864
	Hydrophobic only	0.915	0.935	0.910	0.930	0.571	0.728	0.702	0.816
	Hydrophilic only	0.960	0.965	0.921	0.892	0.728	0.832	0.813	0.861
	Balanced only	0.924	0.953	0.930	0.914	0.627	0.840	0.749	0.876

figure also includes the area under the curve (AUC) for each classifier. It breaks down the performance across all training peptides and even the 3 peptide categories defined in 2.2. For A2 and B27, NetMHC-4.0 and NetMHCpan-4.1 had similar AUC values (no more than 1% apart), while for B8 NetMHC-4.0 under-performed by 3%. Across all HLAs, both tools reported AUC values higher than 95%. The different peptide categories did not highlight any notable trends on the ROC plots. It is interesting to note that both tools had different performances across the 3 peptide cases in each HLA. To investigate this observation in detail, we used violin plots to visualize the predicted immunopeptidomes.

The distributions of the hydrophobicity scores of 9-mers in the training data analysis are shown in Figure 3, and those in the human proteome data analysis in Figure 4. For both analyses, we used the 2 sample t-test to compare the immunopeptidomes predicted by NetMHC-4.0 and NetMHCpan-4.1, and to identify any discrepancies in their predictions on the basis of hydrophobicity (all values used in the t-tests are listed in Table S2).

Strong binders to A2 were expected to have two hydrophobic amino acids (L, V, M, or I) at positions 2 and 9, and thus the expected A2 immunopeptidome would be more hydrophobic than the training or sampled data (expected to be approximately centered about a Moon score of 2). In both analyses, NetMHCpan-4.1 predicted strong binders with a closer hydrophobicity score to our expected value than NetMHC-4.0 did. This difference in predictions was extremely statistically significant in both analyses (p-values less than 0.0001). That is, NetMHC-4.0's predicted strong binders for A2 were more hydrophobic than NetMHCpan-4.1's.

Strong binders to B27 were expected to have one hydrophilic amino acid (R) at position 2, and thus the expected B27 immunopeptidome would be slightly more hydrophilic than the training or sampled data (expected to be approximately centered about a Moon score of -1). Neither tool exhibited this hydrophilic shift in its predictions on the training dataset, but with the human proteome, NetMHCpan-4.1 did predict strong binders centered at a hydrophobicity score of -0.775; NetMHC-4.0's mean was -0.155. The difference in predictions was statistically significant in both analyses (p-values no larger than 0.002). Again, NetMHC-4.0's predicted strong binders for B27 were more hydrophobic than NetMHCpan-4.1's.

Strong binders to B8 were expected to have two hydrophobic amino acids (L, V, M, or I) at positions 2 and 9, and two

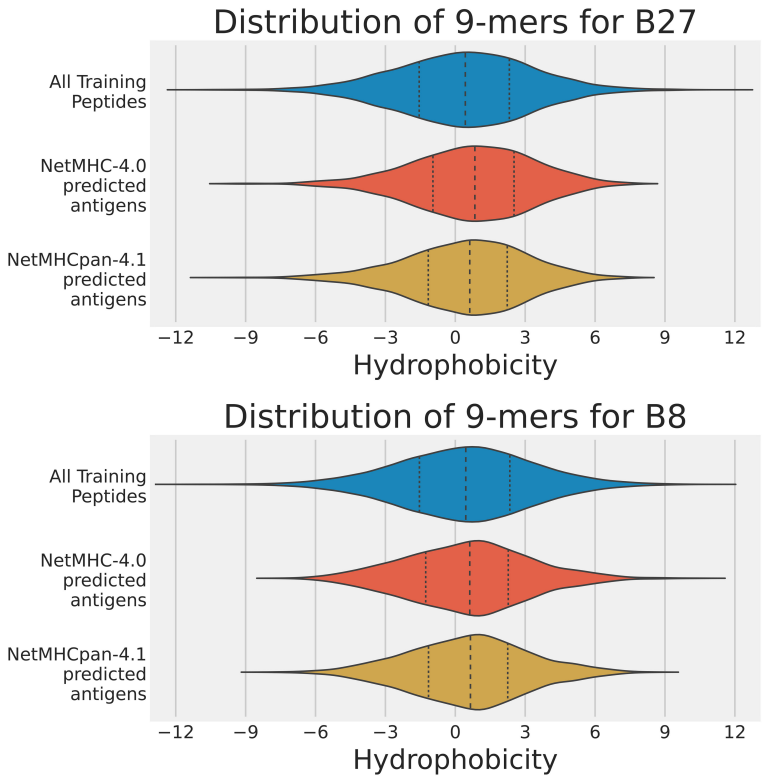


FIGURE 3
Violin plots of the hydrophobicity of the sets of strong binders predicted by NetMHC-4.0 and NetMHCpan-4.1 on the training dataset for A2, B27, and B8. The x-axis represents the hydrophobicity of a 9-mer, and the y-axis represents the frequency. Note that the A2 results can be accessed from our previous study (15). The mean and two quartiles are also depicted in each distribution.

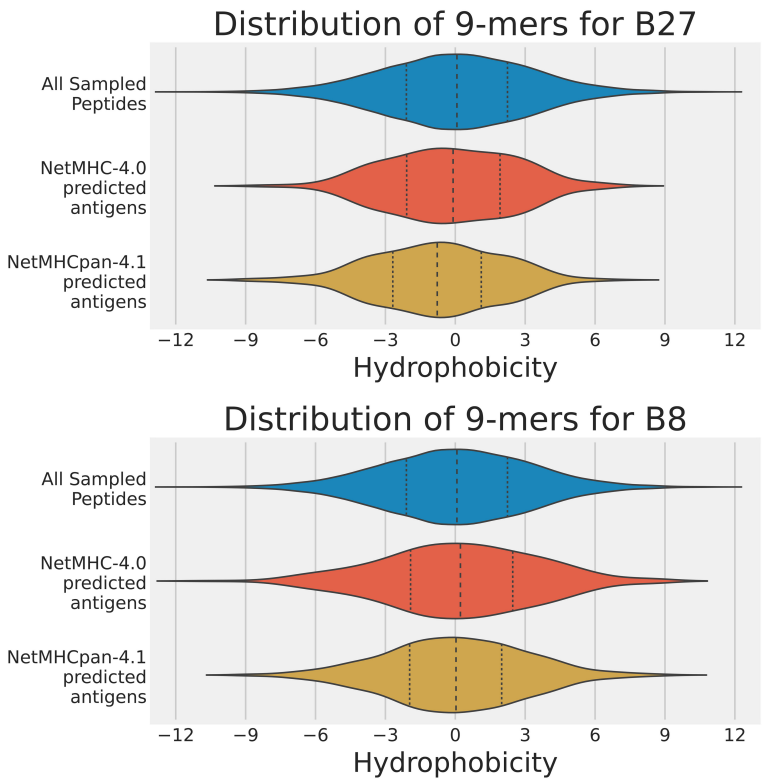


FIGURE 4
Violin plots of the hydrophobicity of the sets of strong binders predicted by NetMHC-4.0 and NetMHCpan-4.1 on the human proteome dataset for A2, B27, and B8. The x-axis represents the hydrophobicity of a 9-mer, and the y-axis represents the frequency. The distributions of all sampled peptides (blue), strong binders predicted by NetMHC-4.0 (red), and those predicted by NetMHCpan-4.1 (yellow) are shown. The mean and two quartiles are also depicted in each distribution. Note that the A2 results can be accessed from our previous study (15).

hydrophilic amino acids (R or K) at positions 3 and 5. Consequently, no major shift in hydrophobicity expected in the B8 immunopeptidomes predicted by either neural network tool. This was indeed the result observed in both analyses, and no significant difference was observed between the predictions of NetMHC-4.0 and NetMHCpan-4.1 (p-values were 0.716 and 0.425 for the training dataset analysis and the human proteome analysis respectively). In this case, NetMHC-4.0's predicted set of binders for B8 were not distinguishable from NetMHCpan-4.1's in terms of hydrophobicity.

Overall, the trend observed from these violin plots seemed to be that NetMHC-4.0 was incorrectly accounting for hydrophobicity when predicting strong binders for A2 and B27. In contrast, NetMHCpan-4.1 was predicting less hydrophobic strong binders for allHLAs in the human proteome analysis. As NetMHCpan-4.1 more closely matched our expectations for A2 and B27, we reasoned that its predictions were more accurate. Since NetMHCpan-4.1 also predicted more strong binders, we hypothesized that the new strong binders gained in NetMHCpan-4.1's immunopeptidome were slightly hydrophilic (with respect to NetMHC-4.0's immunopeptidome) and therefore skewing the mean hydrophobicity lower. To investigate this, we referred to Table 1 in which we tracked each neural network tool's performance on the training data. In particular, we broke down the performances of these tools in our 3 specific peptide cases using 4 different classification metrics discussed below.

The Accuracy metric tracks the number of true negatives and true positives identified by a classifier relative to all the tested data points. Across all HLAs, both neural network tools maintained high accuracy, though NetMHCpan-4.1 performed slightly better (by roughly 2%). For B27 in particular, NetMHCpan-4.1 had a notably higher accuracy (by about 6%) in all peptides cases. No specific improvement was observed in any individual peptide category.

The Precision metric inversely measures the number of false positives identified by a classifier. NetMHCpan-4.1 exhibited higher precision for A2 (by 3%), roughly equivalent precision for B27, and slightly lower precision for B8 (by about 2%) in all peptide cases. The highlight here was that NetMHC-4.0 had low precision (lower than 80%) when dealing with hydrophilic peptides for A2 and B27.

The Recall metric inversely represents the number of false negatives not identified by a classifier. NetMHCpan-4.1 showed significant improvement (consistently higher than 10%) in recall for all HLAs. For A2, these improvements were observed in classifying hydrophilic and balanced peptides. For B27 and B8, these improvements were observed in all peptide categories.

The F1 score is a combination of precision and recall, and tracks overall performance of a classifier. For all HLAs, NetMHCpan-4.1 outperformed NetMHC-4.0 (by at least 10%) when considering all peptides categories.

The observations from these data, in particular the accuracy and F1 score, support our initial assumption that NetMHCpan-4.1 had stronger predictions than NetMHC-4.0 when focusing on hydrophobicity. The precision and recall scores elucidate the reasons behind this improvement: NetMHCpan-4.1 predicted fewer false positives, and much fewer false negatives for all HLAs. The sources of these false positives and negatives in NetMHC-4.0's predictions varied across the different HLAs. For A2, most false positives were found in non-balanced (hydrophobic and hydrophilic) peptides cases, while the majority of false negatives came from non-hydrophobic peptides. For B27, a few false positives were observed in the hydrophilic peptides, but most notably numerous false negatives were found across all types of peptides. For B8, false negatives in all peptides cases lowered the performance of NetMHC-4.0.

It is also important to acknowledge that NetMHCpan-4.1 has an unfair advantage over NetMHC-4.0 – the newer tool was trained on a much larger dataset. Furthermore, NetMHC-4.0 was trained on only peptide-MHC binding affinity data, while NetMHCpan-4.1 was trained on eluted ligand data that was representative of the entire antigen presentation pathways. We investigated the possibility of the small BA data in the training dataset being biased towards being hydrophobic. These values were contrasted to the mean hydrophobicity values of the EL dataset in Figure S2. In each case, the BA training data was more hydrophobic than the EL training data set. This bias was the most prominent in the A2 peptides, and least prominent in B27. This discrepancy in training data could be one of the causes for why NetMHC-4.0's predicted strong binders contained many hydrophilic false negatives.

4 Conclusion

In our previous study, we had identified a significant preference for hydrophobic peptides in NetMHC-4.0's predicted immunopeptidome for A2 (15). We had argued that highly hydrophobic peptides were being classified by NetMHC-4.0 as false positives. We had suggested that highly hydrophobic peptides would never be trafficked in the aqueous cytosol of cells and were therefore obvious false positives.

In this study, we expanded our previous research to focus on more HLA types – i.e. A2, B27, and B8. These HLAs prefer to bind hydrophobic, hydrophilic, and balanced (neither hydrophobic nor hydrophilic) peptides respectively. By comparing the predictions by NetMHC-4.0 and NetMHCpan-4.1 on both the training dataset (see Figure 3) and the sampled human proteome (see Figure 4), we confirmed NetMHC-4.0's hydrophobicity bias for A2 and B27. In these cases, NetMHC-4.0's predicted immunopeptidome was much more hydrophobic than NetMHCpan-4.1's predictions. This hydrophobic bias was not statistically significant in the B8 immunopeptidome. These results suggest that NetMHC-4.0 struggles to predict strong binders correctly in HLAs with strong hydrophobic or hydrophilic binding motifs.

We used several machine learning metrics, such as accuracy, and recall, on the training dataset analysis (see Table 1). From these results, we discovered the improvement in NetMHC-4.0's predictions (over NetMHC-4.0's) stemmed from fewer false negatives in the non-balanced peptide cases, and fewer false positives in general. In particular, the biased immunopeptidome predicted by NetMHC-4.0 was not just a consequence of overestimating the binding of hydrophobic peptides, but also due to overlooking binders that were hydrophilic.

A key takeway of our analyses is that this hydrophobicity bias could only be discovered and expounded upon by focusing on hydrophobicity of peptides as a core factor in pMHC binding. Merely using machine learning metrics without accounting for such biochemical attributes would have been insufficient in capturing this bias. This is evident from how both neural network tools had similar performances across all 3 HLAs in Figure S1. Just as understanding the erroneous predictions from NetMHC-4.0 required the use of hydrophobicity as a metric, we believe that mechanistically modelling the biochemistry (to some extent) improves upon a purely data-driven artificial intelligence's prediction.

We conclude that NetMHCpan-4.1 is the more reliable of the two neural network tools. It had stronger results across the various metrics we used, and the hydrophobicity of its predicted immunopeptidome matched our expected hydrophobicity values. In contrast, NetMHC-4.0 struggled to predict all strong binders for HLAs that had notable hydrophobic or hydrophilic preferences. There could be several reasons why NetMHCpan-4.1 outperformed NetMHC-4.0. NetMHCpan-4.1 had a larger training set. Binding Affinity data alone could not model the effects of the entire antigen presentation pathway as Eluted Ligand data could have. For example, the binary values of eluted ligand data might have trained NetMHCpan-4.1 to be more decisive in its predictions. Tracking the MHC sequence could have allowed NetMHCpan-4.1 to model binding mechanics of MHC binding pockets. Eluted ligand data might have set NetMHCpan to capture aspects of the entire antigen presentation pathway instead of estimating pMHC binding strength alone. The generation of negative training data (16) for NetMHCpan-4.1 could have resolved false positives that NetMHC-4.0 was vulnerable to.

In future work, we will focus on identifying more significant structural and mechanistic attributes that pose hurdles for AI-based methods. We are developing a structural prediction tool capable of predicting peptide binding with uncurated MHC molecules. Since we were limited to using the training dataset, and the sample human proteome dataset without binding data, it would be interesting to expand upon this study with a large evaluation dataset to test the predictions of NetMHCpan-4.1 and NetMHC-4.0 as well.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

Author contributions

AS, MR, JC, JU, and GV led discussions on this research. AS conducted the data analyses and wrote the manuscript. AS, MR, and GV reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the NSF Grant 2036064.

Acknowledgments

We thank Chad Myers for recommending using machine learning metrics to score neural networks.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.1034810/full#supplementary-material.

SUPPLEMENTARY FIGURE 1
Receiver Operating Characteristic curves for NetMHC-4.0 (left) and NetMHCpan-4.1 (right) for A2, B27, and B8 based on the training dataset. In each subfigure, plots are drawn for all peptides (blue), hydrophobic peptides only (red), hydrophilic peptides only (yellow), and balanced peptides only (green). A completely random classifier is also plotted for reference (dashed black). For each plot, the Area Under the Curve (AUC) is also noted in the legend.

SUPPLEMENTARY FIGURE 2
Violin plots of the hydrophobicity of all the training 9-mers, split based on experimental source: Binding Affinity (in blue) and Eluted Ligand (in red). For A2, there were 7940 BA peptides and 44719 EL peptides, with mean hydrophobicities of 2.227 and 0.667 respectively. For B27, there were 2683 BA peptides and 14739 EL peptides, with mean hydrophobicities of 0.597 and 0.322 respectively. For B8, there were 2777 BA peptides and 16671 EL peptides, with mean hydrophobicities of 0.999 and 0.292 respectively.

References

1. Neefjes J, Jongsma ML, Paul P, Bakke O. Towards a systems understanding of mhc class i and mhc class ii antigen presentation. *Nat Rev Immunol* (2011) 11:823–36. doi: 10.1038/nri3084

2. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. *PLoS One* (2014) 9:e97282. doi: 10.1371/journal.pone.0097282

3. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* (2016) 32:511–7. doi: 10.1093/bioinformatics/btv639

4. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* (2020) 48:W449–54. doi: 10.1093/nar/gkaa379

5. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TB, Wilson GA, et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* (2017) 171:1259–71. doi: 10.1016/j.cell.2017.10.001

6. Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* (2017) 551:517–20. doi: 10.1038/nature24473

7. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* (2020) 181:1489–501. doi: 10.1016/j.cell.2020.05.015

8. Doytchinova IA, Flower DR. A comparative molecular similarity index analysis (comsia) study identifies an hla-a2 binding supermotif. *J computer-aided Mol design* (2002) 16:535–44. doi: 10.1023/A:1021917203966

9. Jardetzky T, Lane W, Robinson R, Madden D, Wiley D. Identification of self peptides bound to purified hla-b27. *Nature* (1991) 353:326–9. doi: 10.1038/353326a0

10. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* (1997) 267:707–26. doi: 10.1006/jmbi.1996.0859

11. Vasmatzis G, Zhang C, Cornette JL, DeLisi C. Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol Immunol* (1996) 33:1231–9. doi: 10.1016/S0161-5890(96)00090-9

12. Nielsen M, Andreatta M, Peters B, Buus S. Immunoinformatics: predicting peptide–MHC binding. *Annu Rev Biomed Data Sci* (2020) 3:191–215. doi: 10.1146/annurev-biodatasci-021920-100259

13. Chowell D, Krishna S, Becker PD, Cocita C, Shu J, Tan X, et al. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc Natl Acad Sci* (2015) 112:E1754–62. doi: 10.1073/pnas.1500973112

14. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, NP C, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Briefings Bioinf* (2020) 21:1119–35. doi: 10.1093/bib/bbz051

15. Solanki A, Riedel M, Cornette J, Udell J, Koratkar I, Vasmatzis G. The role of hydrophobicity in peptide-MHC binding. In: *International symposium on mathematical and computational oncology*. (Springer Nature Switzerland AG, Switzerland: Springer) (2021) 24–37. Available at: https://link.springer.com/chapter/10.1007/978-3-030-91241-3_3#chapter-info.

16. Alvarez B, Reynisson B, Barra C, Buus S, Ternette N, Connelley T, et al. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Mol Cell Proteomics* (2019) 18:2459–77. doi: 10.1074/mcp.TIR119.001658

17. Powers DM. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv* (2020) 37–63. doi: 10.48550/arXiv.2010.16061

18. Consortium U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* (2019) 47:D506–15. doi: 10.1093/nar/gky1049

19. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* (2017) 13:e1005725. doi: 10.1371/journal.pcbi.1005725

20. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* (1982) 157:105–32. doi: 10.1016/0022-2836(82)90515-0

21. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol* (1987) 195:659–85. doi: 10.1016/0022-2836(87)90189-6

22. Hopp TP, Woods KR. A computer program for predicting protein antigenic determinants. *Mol Immunol* (1983) 20:483–9. doi: 10.1016/0161-5890(83)90029-9

23. Moon CP, Fleming KG. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc Natl Acad Sci* (2011) 108:10174–7. doi: 10.1073/pnas.1103979108

24. Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodges RS. Relationship of sidechain hydrophobicity and *α*-helical propensity on the stability of the single-stranded amphipathic *α*-helix. *J Pept science: an Off Publ Eur Pept Soc* (1995) 1:319–29. doi: 10.1002/psc.310010507

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 28 November 2022
DOI 10.3389/fimmu.2022.1024707

Check for updates

OPEN ACCESS

EDITED BY
Jinzhuang Dou,
University of Texas MD Anderson Cancer Center,
United States

REVIEWED BY
Albino Eccher,
Integrated University Hospital Verona, Italy
Haiyu Zhou,
Guangdong Provincial People's Hospital, China

*CORRESPONDENCE
Jinghui Wang
jinghuiwang2006@163.com
Zhidong Liu
liuzhidong@bjxkyy.cn
Yuanming Pan
peter.f.pan@hsc.pku.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 22 August 2022
ACCEPTED 07 November 2022
PUBLISHED 28 November 2022

CITATION
Li S, Li W, Ma T, Fu S, Gao X, Qin N, Wu Y, Zhang X, Wang J, Pan Y and Liu Z (2022) Assessing the efficacy of immunotherapy in lung squamous carcinoma using artificial intelligence neural network.
Front. Immunol. 13:1024707.
doi: 10.3389/fimmu.2022.1024707

COPYRIGHT
© 2022 Li, Li, Ma, Fu, Gao, Qin, Wu, Zhang, Wang, Pan and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Assessing the efficacy of immunotherapy in lung squamous carcinoma using artificial intelligence neural network

Siqi Li[1†], Wei Li[1†], Tianyu Ma[1†], Siyun Fu[2], Xiang Gao[2], Na Qin[2], Yuhua Wu[2], Xinyong Zhang[2], Jinghui Wang[2,3*], Yuanming Pan[3*] and Zhidong Liu[1*]

[1]Department of Thoracic Surgery, Beijing Chest Hospital/Beijing Tuberculosis and Thoracic Tumor Research Institute, Capital Medical University, Beijing, China, [2]Department of Medical Oncology, Beijing Tuberculosis and Thoracic Tumor Research Institute/Beijing Chest Hospital, Capital Medical University, Beijing, China, [3]Cancer Research Center, Beijing Tuberculosis and Thoracic Tumor Research Institute/Beijing Chest Hospital, Capital Medical University, Beijing, China

Background: At present, immunotherapy is a very promising treatment method for lung cancer patients, while the factors affecting response are still controversial. It is crucial to predict the efficacy of lung squamous carcinoma patients who received immunotherapy.

Methods: In our retrospective study, we enrolled lung squamous carcinoma patients who received immunotherapy at Beijing Chest Hospital from January 2017 to November 2021. All patients were grouped into two cohorts randomly, the training cohort (80% of the total) and the test cohort (20% of the total). The training cohort was used to build neural network models to assess the efficacy and outcome of immunotherapy in lung squamous carcinoma based on clinical information. The main outcome was the disease control rate (DCR), and then the secondary outcomes were objective response rate (ORR), progression-free survival (PFS), and overall survival (OS).

Results: A total of 289 patients were included in this study. The DCR model had area under the receiver operating characteristic curve (AUC) value of 0.9526 (95%CI, 0.9088–0.9879) in internal validation and 0.9491 (95%CI, 0.8704–1.0000) in external validation. The ORR model had AUC of 0.8030 (95%CI, 0.7437–0.8545) in internal validation and 0.7040 (95%CI, 0.5457–0.8379) in external validation. The PFS model had AUC of 0.8531 (95%CI, 0.8024–0.8975) in internal validation and 0.7602 (95%CI, 0.6236–0.8733) in external validation. The OS model had AUC of 0.8006 (95%CI, 0.7995–0.8017) in internal validation and 0.7382 (95%CI, 0.7366–0.7398) in external validation.

Conclusions: The neural network models show benefits in the efficacy evaluation of immunotherapy to lung squamous carcinoma patients, especially the DCR and ORR models. In our retrospective study, we found

that neoadjuvant and adjuvant immunotherapy may bring greater efficacy benefits to patients.

KEYWORDS

immunotherapy, lung squamous carcinoma, neural network, deep learning, predictive model

Highlights

- The expression level of PD-L1 is not an ideal tool for evaluating efficacy.
- The neural network models show benefits in the efficacy evaluation of immunotherapy to lung squamous carcinoma patients, especially the DCR and ORR models.
- Neoadjuvant and adjuvant immunotherapy may bring greater efficacy benefits to lung squamous carcinoma patients.

Background

Lung cancer occupies first place in mortality and second place in morbidity globally (1, 2). Non-small cell lung cancer (NSCLC) accounts for about 85% of lung cancer cases, and the 5-year survival rate is only 18% (3, 4). Recently, the advent of targeted drugs and the emergence of immunotherapy have significantly prolonged the survival of lung cancer patients and improved their quality of life (3). Lung squamous carcinoma accounts for 25%–30% of lung cancer cases, while the occurrence rate of common driver gene mutation is less than 7% (5–9). Patients with lung squamous carcinoma have few chances to receive targeted therapy. Fortunately, immunotherapy brings a new light to patients with lung squamous carcinoma, which can significantly improve objective response rate (ORR) and prolong progression-free survival (PFS) and overall survival (OS) (10–12). However, it also brings challenges to the selection of biomarkers for predicting efficacy, the treatment plan for lung cancer, and treatment-related adverse events.

At present, the expression level of programmed cell death ligand 1 (PD-L1) is an indicator that may predict the effectiveness of anti-programmed cell death 1 (anti-PD-1)/PD-L1 immunotherapy and screen the population sensitive to them (13). However, studies also found that some patients with high PD-L1 expression have a poor immune response. On the contrary, up to 10% of patients with negative PD-L1 expression have a good immune response (14). It suggested that the expression level of PD-L1 is not an ideal tool for evaluating efficacy. Several studies showed that traditional clinical or pathological features, including smoking status, age, pathological type, and tumor grade, are associated with immunotherapy for NSCLC (15–17). Therefore, it is necessary to use a new method to evaluate the efficacy of immunotherapy and identify the dominant population sensitive to immunotherapy.

Deep learning neural network, as a subdiscipline of artificial intelligence (AI), has shown good performance in predicting and monitoring treatment response, which is also gradually gaining the attention of clinicians (18, 19). Convolutional neural network is currently used to diagnose solid tumors (lung cancer, melanoma, gastrointestinal tumors, etc.) through automatic quantification of radiological images, digital histopathological image interpretation, or biomarker analysis (18, 20–22). However, little research focused on the evaluation of immunotherapy efficacy in NSCLC based on AI.

To better predict the efficacy of the immunotherapy of lung squamous carcinoma patients and thus further provide more optimal treatment strategies, we introduced the neural network algorithm to build a fully connected neural network (also known as a dense neural network (DNN)) based on clinical information of the above patients. The original codes and data have been uploaded for use by clinicians and future visualization platforms.

Method

Study design and clinical information

This study was designed as a retrospective cohort study. Eligible patients aged ≥18 years were diagnosed with lung

Abbreviations: AJCC, American Joint Committee on Cancer; AI, artificial intelligence; ALK, anaplastic lymphoma kinase; AUC, area under the receiver operating characteristic curve; BOR, best overall response; CI, confidence interval; CR, complete response; DCR, disease control rate; DNN, dense neural network; EGFR, epidermal growth factor receptor; IHC, immunohistochemistry; MPR, major pathological response; NSCLC, non-small cell lung cancer; ORR, objective response rate; OS, overall survival; PD, progressive disease; PD-1, programmed cell death 1; PD-L1, programmed cell death ligand 1; PFS, progression-free survival; pCR, pathological complete response; PR, partial response; SD, stable disease, TPS, tumor proportion score.

squamous carcinoma pathologically and received immunotherapy in Beijing Chest Hospital affiliated with Capital Medical University between 16 January 2017 and 10 December 2021. Patients with active autoimmune disease, symptomatic interstitial lung disease, multiple primary pulmonary carcinomas, or missing any of the included clinical characteristics, like status or follow-up records, were excluded. Tumor PD-L1 expression was assessed using the PD-L1 immunohistochemistry 22C3 pharmDx kit (Agilent Technologies, Carpinteria, CA, USA) at the Pathology Department of Beijing Chest Hospital. The expression level of PD-L1 protein of archival tumor tissue or tissue obtained through biopsy was determined by the tumor proportion score (TPS). Then, all patients were grouped into two cohorts randomly, the training cohort (80% of the total) and the test cohort (20% of the total). The training cohort was used to build neural network models to assess the efficacy and outcome of immunotherapy in lung squamous carcinoma based on clinical information, with evaluated by internal validation (using training cohort data) and external validation (using test cohort data) (Figure 1). The last follow-up date was 30 April 2022.

This research has been approved by the Ethics Committee of Beijing Chest Hospital affiliated with Capital Medical University. Given that this was a retrospective analysis, individual consent was waived.

Potential predictive variables

The potential clinical predictive variables were as follows: age, sex, smoking status, performance status (PS) score before receiving immunotherapy, PD-L1 expression, TNM and clinical stage, vascular invasion, pleural metastasis, extra-thoracic metastasis, brain metastasis, liver metastasis, bone metastasis, adrenal metastasis, received chemoimmunotherapy or not, received immunotherapy plus antiangiogenic therapy or not, neoadjuvant immunotherapy, immunotherapy lines, and gene mutations. Considering that gene mutations contained much missing value, we did not include them in the final models. Supplementary Figure 1 shows the distribution of missing values. TNM and clinical stage were evaluated by at least two senior clinicians, referring to the 8th edition of the American Joint Committee on Cancer (AJCC) stage. All patients have received immunotherapy, of whom some accepted another therapy such as chemotherapy or antiangiogenic medicine. Partial patients in this study were treated with neoadjuvant immunotherapy, which meant they would undergo operations later.

Multiple imputations of missing values have been performed before the above predictive variables were included in models, with the help of the R package (23).

Outcome

The main outcome was the disease control rate (DCR), and secondary outcome indicators were ORR, PFS, and OS. The imaging manifestations have been interpreted by local researchers according to the Response Evaluation Criteria in Solid Tumors (RECIST version 1.1). The best overall response (BOR) was assessed including complete response (CR), partial response (PR), or stable disease (SD) after immunotherapy, regarded as DCR. ORR included CR or PR patients. PFS was defined as the time from the day of receiving immunotherapy to objective tumor progression, surgery, or death. We processed PFS as a binary variable, PFS ≤6 months or >6 months.

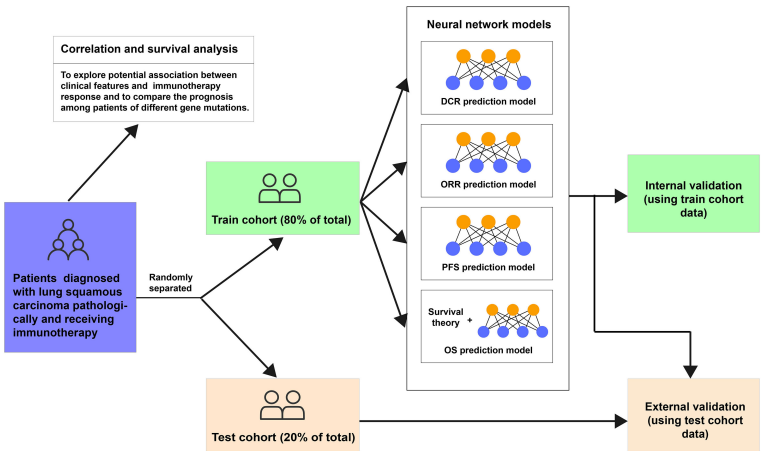


FIGURE 1
Flowchart of this study. DCR, disease control rate; ORR, objective response rate; PFS, progression-free survival; OS, overall survival.

Data pre-processing

Standardization of data is a common process in many machine learning situations, which implies numerical variables subtracting their means and dividing by their standard deviations. Categorical variables have been converted into dummy variables, such as replacing the sex variable with two dummy variables (female = 0 or male = 1). Age and PS score before receiving immunotherapy were standardized, while sex, smoking status, PD-L1 expression, TNM and clinical stage, vascular invasion, pleural metastasis, extra-thoracic metastasis, brain metastasis, liver metastasis, bone metastasis, adrenal metastasis, received chemoimmunotherapy or not, received immunotherapy plus antiangiogenic therapy or not, neoadjuvant immunotherapy, and immunotherapy lines were transformed to dummy variables. Data from the test cohort were standardized according to the training cohort, and Supplementary Table 1 exhibits the mean and standard deviations of numerical variables.

Python package pandas and scikit-learn helped us to achieve the above processing (24, 25).

Model training and validation

To assess the DCR, ORR, and PFS, three dense neural networks have been built. To obtain accurate predictions, batch training and normalization were used. Dropout layers were used to avoid overfitting (which means performing well in the training cohort but badly in the test cohort), and early stopping functions were used to end training epochs if necessary.

To predict the OS possibility of lung squamous carcinoma patients after immunotherapy, we conducted a neural network survival model based on Katzman's DeepSurv (26). It has to be mentioned that the neural network was designed to tackle traditional classification issues instead of time-dependent tasks, so the performance of OS prediction might be moderate.

As mentioned above, since our models were applied to solve classification problems, we used the area under the receiver operating characteristic curve (AUC) to evaluate their performance. The closer the AUC is to 1, the better the performance of the model. In order to obtain AUC with detailed 95% confidence interval (CI), we run bootstrap 1,000 times.

Python and its packages PyTorch, torchtuples, NumPy, pycox, and matplotlib and the R package pROC helped us in these analyses (27–32).

Correlation and survival analysis

We used correlation analysis to explore the potential association between clinical features and patients' immunotherapy response with Spearman's rank correlation, visualized by heatmap and chord diagram. Finally, survival analysis was performed to compare the prognoses of patients with different gene mutations. R and packages corrplot, circlize, and survminer were used during these procedures (33–35).

Statistical analysis

The numerical data of skewed distribution were analyzed by the Wilcoxon test, while categorical data were compared by chi-square or Fisher's exact test. A two-sided p-value of less than 0.05 was considered statistically significant. These analyses and relevant plotting were completed using R software and packages epiDisplay, ggplot2, and ggridges (36, 37).

Result

Characteristics of patients

A total of 289 patients were included; the PFS of 138 patients was shorter or equal to 6 months, and the PFS of 151 patients was longer than 6 months. In the group with PFS ≤ 6 months, there were 11 women and 127 men. As for the expression of PD-L1, 30 patients had <1% expression, 37 patients had 1%–49% expression, and 42 patients had ≥50% expression. Among them, four patients had CR, 44 patients had PR, 65 patients had SD, and 24 patients were diagnosed with PD. The group with PFS > 6 months included 12 women and 139 men. The PD-L1 expression of 29 patients was <1%; 36 patients, between 1% and 49%; and 55 patients, ≥50%. Among them, five had CR, 108 had PR, and 38 had PD. The median follow-up time was 7.75 months in the group with PFS ≤ 6 months and 12.8 months in the group with PFS > 6 months. The detailed information of patients is illustrated in Table 1 and Figure 2.

Model training

We conducted four neural networks to assess the efficacy and outcome of immunotherapy in lung squamous carcinoma (Figure 1). The DCR model was designed to return the probabilities of patients showing DCR after immunotherapy, with the ORR model predicting the ORR possibilities and the PFS model judging their PFS longer or shorter than 6 months. The OS model was used to predict patients' OS possibility based on a neural network survival algorithm. The training curves are shown in Supplementary Figure 2.

Model performance and inference

The DCR model had AUC of 0.9526 (95%CI, 0.9088–0.9879) in internal validation and 0.9491 (95%CI, 0.8704–

10.3389/fimmu.2022.1024707

TABLE 1 The clinical features of patients.

	PFS		Statistical method	p-Value
	≤6 months (N = 138)	>6 months (N = 151)		
	N (%)			
Age			Wilcoxon	0.3461
Median (IQR)	64 (58, 68.75)	65 (58.5, 69)		
Sex			Chi-square	0.9940
Female	11 (7.97)	12 (7.95)		
Male	127 (92.03)	139 (92.05)		
Smoke			Chi-square	0.3291
No	27 (19.57)	23 (15.23)		
Yes	110 (79.71)	127 (84.11)		
Unknown	1 (0.72)	1 (0.66)		
PS score			Wilcoxon	0.0435*
Median (IQR)	1 (1,1)	1 (1,1)		
PD-L1 expression			Chi-square	0.5360
<1%	30 (21.74)	29 (19.21)		
1%–49%	37 (26.81)	36 (23.84)		
>50%	42 (30.43)	55 (36.42)		
Unknown	29 (21.01)	31 (20.53)		
T			Fisher's exact	0.1763
1	3 (2.17)	2 (1.32)		
1b	3 (2.17)	0 (0)		
1c	4 (2.90)	1 (0.66)		
2	22 (15.94)	34 (22.52)		
2a	7 (5.07)	11 (7.28)		
2b	7 (5.07)	12 (7.95)		
3	25 (18.12)	32 (21.19)		
3b	1 (0.72)	0 (0)		
4	66 (47.83)	59 (39.07)		
N			Chi-square	0.1234
0	19 (13.77)	34 (22.52)		
1	11 (7.97)	18 (11.92)		
2	68 (49.28)	63 (41.72)		
3	37 (26.81)	33 (21.85)		
Unknown	3 (2.17)	3 (1.99)		
M			Chi-square	0.0261*
0	69 (50.00)	98 (64.90)		
1	7 (5.07)	3 (1.99)		
1a	19 (13.77)	22 (14.57)		
1b	12 (8.70)	5 (3.31)		
1c	27 (19.57)	18 (11.92)		
Unknown	4 (2.90)	5 (3.31)		
Clinical stage			Fisher's exact	0.0335*
IA	2 (1.45)	0 (0)		
IB	1 (0.72)	5 (3.31)		
IIA	4 (2.90)	8 (5.30)		
IIB	3 (2.17)	11 (7.28)		
IIIA	25 (18.12)	36 (23.84)		

(Continued)

TABLE 1 Continued

	PFS		Statistical method	p-Value
	≤6 months (N = 138)	>6 months (N = 151)		
	N (%)			
IIIB	29 (21.01)	30 (19.87)		
IIIC	4 (2.90)	7 (4.64)		
IV	70 (50.72)	54 (35.76)		
Vascular invasion			Chi-square	0.9061
No	132 (95.65)	144 (95.36)		
Yes	6 (4.35)	7 (4.64)		
Pleural metastasis			Chi-square	0.0169*
No	113 (81.88)	138 (91.39)		
Yes	25 (18.12)	13 (8.61)		
Extra-thoracic metastasis			Chi-square	0.0052**
No	96 (69.57)	126 (83.44)		
Yes	42 (30.43)	25 (16.56)		
Brain metastasis			Chi-square	0.0192*
No	124 (89.86)	146 (96.69)		
Yes	14 (10.14)	5 (3.31)		
Liver metastasis			Chi-square	0.6595
No	128 (92.75)	142 (94.04)		
Yes	10 (7.25)	9 (5.96)		
Bone metastasis			Chi-square	0.0088**
No	109 (78.99)	136 (90.07)		
Yes	29 (21.01)	15 (9.93)		
Adrenal metastasis			Chi-square	0.2243
No	128 (92.75)	145 (96.03)		
Yes	10 (7.25)	6 (3.97)		
Gene mutation			Fisher's exact	0.1202
EGFR	5 (3.62)	2 (1.32)		
KRAS	1 (0.72)	7 (4.64)		
TP53	15 (10.87)	26 (17.22)		
Uncommon	7 (5.07)	5 (3.31)		
Negative	40 (28.99)	47 (31.13)		
Unknown	70 (50.72)	64 (42.38)		
Therapy			Fisher's exact	0.4352
Immu	23 (16.67)	18 (11.92)		
Immu + Antiangio	3 (2.17)	2 (1.32)		
Immu + Chemo	112 (81.16)	131 (86.75)		
Neoadjuvant immunotherapy			Chi-square	<0.001***
No	115 (83.33)	97 (64.24)		
Yes	23 (16.67)	54 (35.76)		
Immunotherapy lines			Chi-square	<0.001***
First-line	94 (68.12)	131 (86.75)		
Second-line and above	44 (31.88)	20 (13.25)		
BOR			Fisher's exact	<0.001***
CR	4 (2.90)	5 (3.31)		
PR	44 (31.88)	108 (71.52)		
SD	65 (47.10)	38 (25.17)		

(Continued)

TABLE 1 Continued

	PFS		Statistical method	p-Value
	≤6 months (N = 138)	>6 months (N = 151)		
	N (%)			
PD	24 (17.39)	0 (0)		
Unknown	1 (0.72)	0 (0)		
ORR			Chi-square	<0.001***
No	89 (64.49)	38 (25.17)		
Yes	48 (34.78)	113 (74.83)		
Unknown	1 (0.72)	0 (0)		
DCR			Chi-square	<0.001***
No	24 (17.39)	0 (0)		
Yes	113 (81.88)	151 (100.00)		
Unknown	1 (0.72)	0 (0)		
PFS			Wilcoxon	<0.001***
Median (IQR)	3.26 (2, 4.55)	10.07 (7.68, 13.33)		
OS			Wilcoxon	<0.001***
Median (IQR)	7.75 (4.45, 13.2)	12.8 (10.67, 18.57)		

PFS, progression-free survival; IQR, interquartile range; PS score, performance status score; PD-L1, programmed cell death ligand 1; Immu, immunotherapy; Antiangio, antiangiogenic therapy; Chemo, chemotherapy; BOR, best overall response; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; ORR, objective response rate; DCR, disease control rate; OS, overall survival.
* P<0.05, ** P<0.01, ***P<0.001.

1.0000) in external validation. The ORR model had AUC of 0.8030 (95%CI, 0.7437–0.8545) in internal validation and 0.7040 (95%CI, 0.5457–0.8379) in external validation. The PFS model had AUC of 0.8531 (95%CI, 0.8024–0.8975) in internal validation and 0.7602 (95%CI, 0.6236–0.8733) in external validation. The receiver operating characteristic curve valuess of the DCR model, ORR model, and PFS model are shown in Figure 3, and their original codes are shown in Supplementary Material File 1. The OS model had AUC of 0.8006 (95%CI, 0.7995–0.8017) in internal validation and 0.7382 (95%CI, 0.7366–0.7398) in external validation (Table 2). The codes of the OS model are shown in Supplementary Material File 2. The weights and hyper-parameters of the above four models are shown in Supplementary Material File 3.

Age, sex, smoking status, PS score, PD-L1 expression, TNM and clinical stage, vascular invasion, pleural metastasis, extra-thoracic metastasis, brain metastasis, liver metastasis, bone metastasis, adrenal metastasis, received chemoimmunotherapy or not, received immunotherapy plus antiangiogenic therapy or not, neoadjuvant immunotherapy, and immunotherapy lines were the predictive clinical features. When these codes are open in a python environment (https://www.python.org/) and jupyter notebook software (https://www.jupyter.org/), with predictive clinical features inputted, models will return the predictive possibility of DCR, ORR, PFS, or OS.

Correlation analysis

General overview, PS score, PD-L1 expression, TNM stage, distant metastasis, and invasion (vascular, pleural, brain, liver, and bone) had statistical relation with immunotherapy (BOR, DCR, ORR, PFS, or OS) (Figure 4). We also compared the OS of patients with different gene mutations but did not find some statistical discrepancies (Supplementary Figure 3).

Discussion

Immunotherapy significantly improves the prognosis of lung cancer patients, but not everyone who receives immunotherapy can benefit from it (38). Hence, it is essential to choose a better method to predict immunotherapy efficacy and screen the population sensitive to immunotherapy. We developed and validated deep learning neural network models based on clinical data for immune efficacy prediction for lung squamous carcinoma. The results of our study verified that the deep learning model showed good predictive performance in these patients.

Currently, multiple clinical trials have shown that the expression level of PD-L1 not only could decide who to treat but also could hint at whom to benefit (16, 39–42). Interestingly, several trials have examined PD-L1 as a viable biomarker to

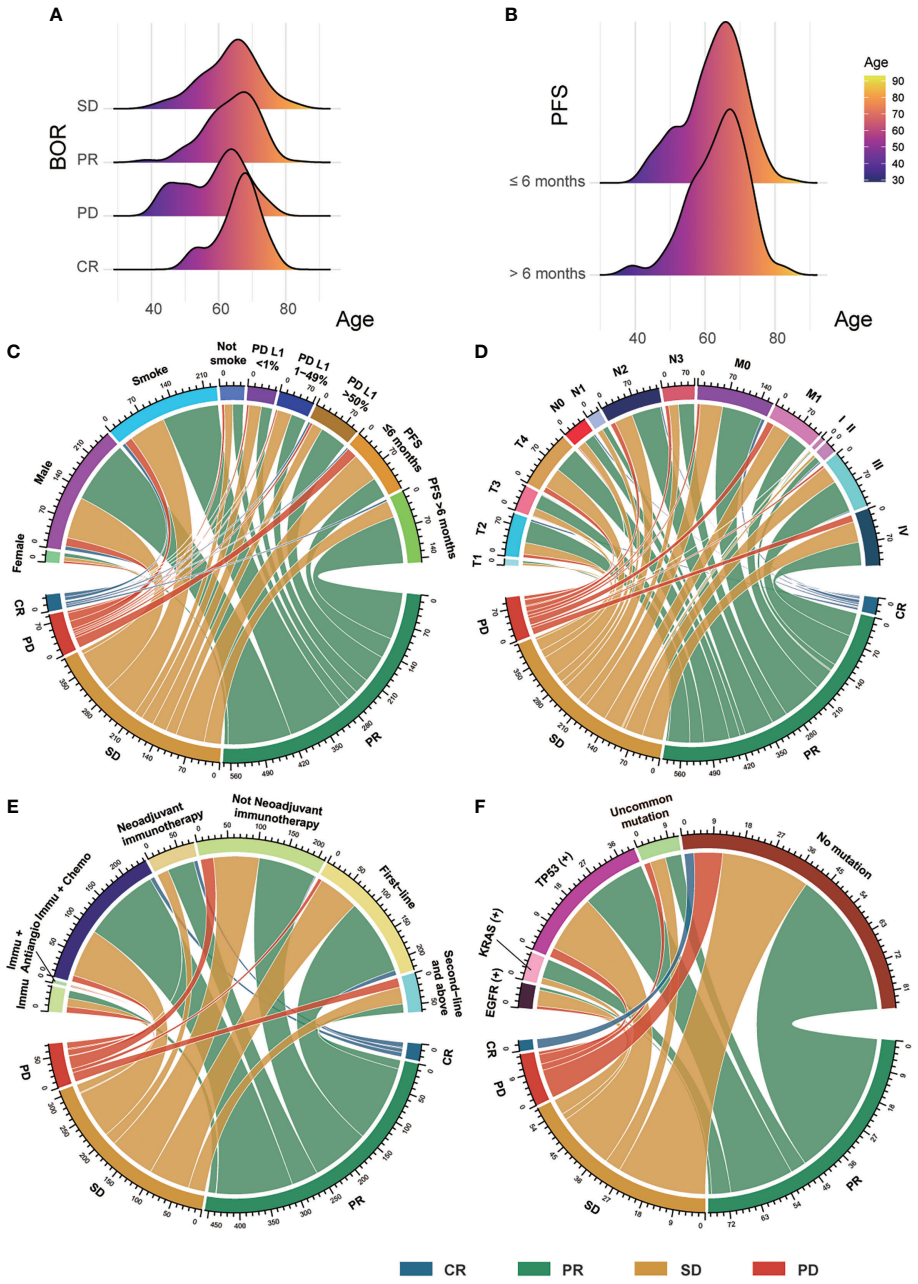


FIGURE 2
The clinical features of patients. **(A)** The distribution between BOR and patients' age. **(B)** The distribution between PFS and patients' age. **(C)** The connections among patients' sex, smoking status, PD-L1 expression, and PFS with their BOR. **(D)** The connections among patients' TNM stage with their BOR. **(E)** The correlations between patients' therapy and their BOR. **(F)** The correlations between patients' gene mutation and their BOR. BOR, best overall response; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; PFS, progression-free survival; PD-L1, programmed cell death ligand 1; Immu, immunotherapy; Antiangio, antiangiogenic therapy; Chemo, chemotherapy; uncommon gene mutation.

predict response to immune checkpoint inhibitors (ICIs) (43–45). Moreover, due to the heterogeneity of PD-L1 expression in tumors, there is a certain difference between puncture biopsy specimens and surgical resection specimens (46, 47). Some studies have shown that the positive rate of PD-L1 was related to clonal selection, biopsy sites, and detection time, and researchers should further coordinate the harmonization of utilized clones, scores, and interobserver variability (48). It is suggested that the independent predictive effect of tumor PD-L1 expression is still imperfect.

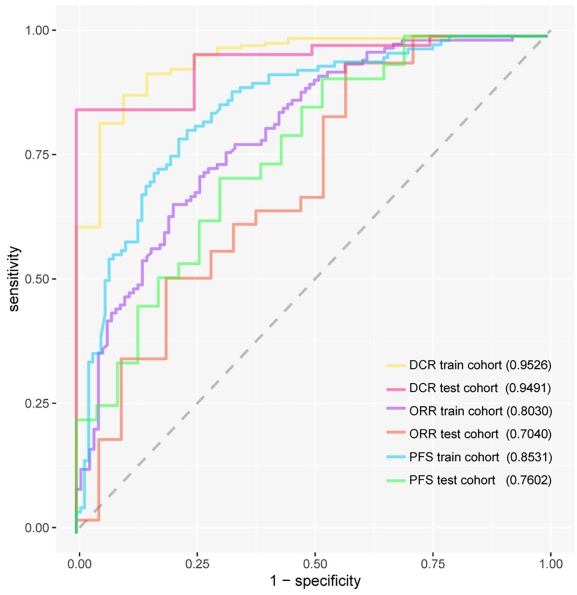


FIGURE 3
The receiver operating characteristic curves of three models. DCR, disease control rate; ORR, objective response rate; PFS, progression-free survival.

We developed and validated the immunotherapy predictive deep learning models using clinical information in lung squamous carcinoma. Raw data were divided into two independent groups, the training cohort (80% of the total) and the test cohort (20% of the total). Age, sex, smoking status, PS score before receiving immunotherapy, PD-L1 expression, TNM and clinical stage, vascular invasion, pleural metastasis, extra-thoracic metastasis, brain metastasis, liver metastasis, bone metastasis, adrenal metastasis, received chemoimmunotherapy or not, received immunotherapy plus antiangiogenic therapy or not, neoadjuvant immunotherapy, and immunotherapy lines were chosen as the predictive variables. The train cohort was used to conduct the ORR model, DCR model, PFS model, and OS model, which were validated using both train cohort and test cohort. Teo avoid overfitting, early stopping function and dropout layers were adopted after numerical variables were standardized and categorical variables were converted into dummy variables. Finally, the above models showed satisfactory performances. The DCR model had AUC of 0.9526 (95%CI, 0.9088–0.9879) in internal validation and 0.9491 (95%CI, 0.8704–1.0000) in external validation. The ORR model had AUC of 0.8030 (95%CI, 0.7437–0.8545) in internal validation and 0.7040 (95%CI, 0.5457–0.8379) in external validation. The PFS model had AUC of 0.8531 (95%CI, 0.8024–0.8975) in internal validation and 0.7602 (95%CI, 0.6236–0.8733) in external validation. The benefit of immunotherapy can be predicted by deep learning models that integrate patient clinical information. Compared with PD-L1 expression, the efficacy indicators DCR and ORR predicted it more accurately.

As a new efficacy prediction model, the deep learning model will have the potential to support clinical decision-making more accurately. She et al. evaluated the use of deep learning algorithms to evaluate the specific survival of NSCLC patients and concluded that deep learning was significantly better than previous models in lung cancer prognosis assessment and treatment recommendations (18). Mu et al. used PET/CT image deep learning to measure the PD-L1 state to predict immunotherapy response non-invasively and then found that deep learning can replace PD-L1 detected by immunohistochemistry (IHC) (49). In our study, we enrolled immunotherapy patients with lung squamous carcinoma, including stage I–IV, for model training and testing. The deep learning models have multiple hidden layers, each of which contains multiple nodes. The node weights between different layers are updated in time according to the loss function and the reverse propagation of the optimizer. Coupled with the application of activation functions, deep learning can better learn and simulate the non-linear relationship between predictive variables and outcomes than traditional statistical models and some machine learning algorithms. Realistic data, especially clinical data, have a relationship that is not simple linear but intricate. Deep learning is more suitable for the analysis and modeling of clinical data. Therefore, our research found that this will be an interesting attempt, and the models also showed satisfactory performance and could precisely predict the efficacy of immunotherapy.

Of note, ICIs have contributed to improving the survival of patients with lung squamous carcinoma. Our results found that first-line immunotherapy can increase DCR to 86.7% and ORR to 46.7%, which is similar to other clinical studies. In KEYNOTE-024 (50), pembrolizumab, compared with chemotherapy, can remarkably improve the ORR (44.8% *vs.* 27.8%).

With the development of immunotherapy, many studies have found that chemo-immunotherapy strategy can significantly improve the response of NSCLC patients (51–55). The KEYNOTE-407 (53) study found that first-line immunotherapy combined with chemotherapy was better than chemotherapy; the ORR was 66.6% versus 38.4%, respectively. A plethora of phase III clinical trials such as IMpower110 (54),

TABLE 2 The performance of OS model.

	AUC	95%CI
Training cohort	0.8006	0.7995–0.8017
Test cohort	0.7382	0.7366–0.7398

OS, overall survival; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

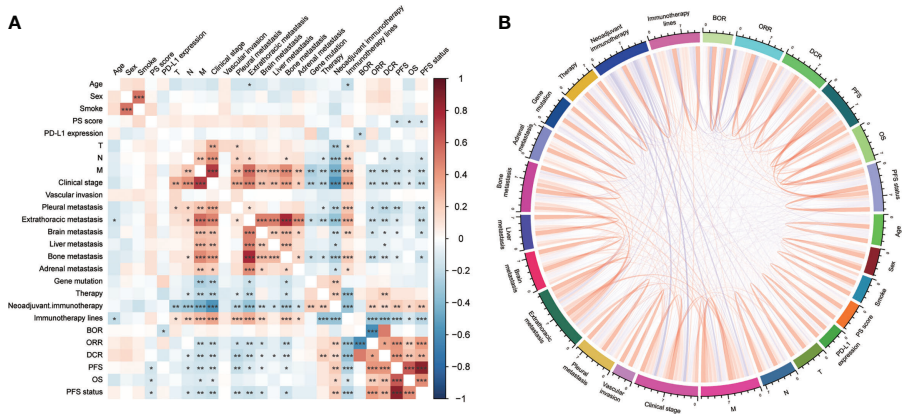


FIGURE 4
The correlation among clinical features with immunotherapy was visualized by **(A)** heatmap and **(B)** chord diagram. PD-L1, programmed cell death ligand 1; BOR, best overall response; ORR, objective response rate; DCR, disease control rate; PFS, progression-free survival; OS, overall survival. *P<0.05, **P<0.01, ***P<0.001

CameL-Sq (52), ORIT-12 (51), and GEMSTONE-302 (55) showed superior efficacy with ICIs plus chemotherapy compared with chemotherapy alone. In our study, we observed that the DCR of patients who received chemoimmunotherapy was 97.7% and ORR was 62.93%, which were similar to the above studies. To be brief, first-line immunotherapy combined with chemotherapy has rapidly expanded first-line treatment options for advanced NSCLC patients without sensitizing epidermal growth factor receptor (EGFR) mutations or anaplastic lymphoma kinase (ALK) fusions.

The success of ICIs in NSCLC has expanded to unresectable stage III and more recently to resectable stage II–IIIA disease. The NADIM study supported the addition of neoadjuvant nivolumab to platinum-based chemotherapy in patients with resectable stage IIIA NSCLC; the major pathological response (MPR) rate was 83%, pathological complete response (pCR) rate was 71%, and 90% of patients had tumor stage decline (33 cases) (56). In CheckMate 816, the neoadjuvant treatment of nivolumab plus chemotherapy improved the pCR rate (24.0% *vs.* 2.2%) in resectable NSCLC patients, and median OS showed a beneficial trend (57). In a neoadjuvant study of sintilimab combined with chemotherapy, 40.5% and 10.8% of patients attained MPR and pCR, respectively, and 3-year OS and disease-free survival (DFS) rates were 95.5% and 81.8%, respectively (58). In our study, we found that the MPR rate of patients with neoadjuvant immunotherapy was 58.4% and the pCR rate was 38.9%, which was similar to the above studies. Neoadjuvant chemoimmunotherapy could change the perception of locally advanced lung cancer from being a potentially lethal disease to one that is curable.

In summary, it is found that immunotherapy can improve the efficacy of patients to a certain extent, but the factors affecting immune response and specific immune resistance mechanisms are multifaceted.

However, this study still has some limitations. Neural network algorithms have the disadvantages of black boxes, which are complex to explain with time-consuming training. Biomarkers such as tumor mutation burden (TMB) were not included. Additionally, this model requires more multi-center prospective data to validate. Further study is needed to validate the advantages of deep learning networks in immunotherapy predictive models.

Conclusion

In conclusion, this study found that the neural network model based on clinical information can accurately predict the efficacy benefits of ICI therapies for lung squamous carcinoma patients, especially DCR and ORR. This novel predictive model may provide reliable individual response information and treatment recommendations. In our retrospective study, we found that neoadjuvant and adjuvant immunotherapy may bring greater efficacy benefits to patients.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This research has been approved by the Ethics Committee of Beijing Chest Hospital affiliated with Capital Medical University.

Given that this was a retrospective analysis, individual consent was waived.

Author contributions

JW, SL, WL and TM came up with the idea for the study. SL, TM, SF, XG, NQ, YW, XZ and JW collected the data and conducted the follow-up. WL carried out the statistical analysis. SL, TM and WL wrote the manuscript. JW, ZL and YP edited and reviewed the manuscript. All authors discussed the results and commented on the manuscript. All the authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.1024707/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The distribution of missing values of clinical features in this study. PD-L1, programmed cell death ligand 1. BOR, best of response. ORR, objective response rate. DCR, disease control rate. PFS, progression-free survival. OS, overall survival.

SUPPLEMENTARY FIGURE 2
The training curves of **(A)** DCR model, **(B)** ORR model, **(C)** PFS model and **(D)** OS model. DCR, disease control rate. ORR, objective response rate. PFS, progression-free survival. OS, overall survival.

SUPPLEMENTARY FIGURE 3
The Kaplan-Meier curve compared the overall survival of patients with different gene mutations. Uncommon gene mutation.

ADDITION FILE 1
The original codes of DCR model, ORR model and PFS model. DCR, disease control rate. ORR, objective response rate. PFS, progression-free survival.

ADDITION FILE 2
The original codes of OS model. OS, overall survival.

ADDITION FILE 3
The weights and hyper-parameters of DCR model, ORR model, PFS model and OS model. DCR, disease control rate. ORR, objective response rate. PFS, progression-free survival. OS, overall survival.

ADDITION FILE 4
The clinical data used by this study.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. cancer statistics, 2021. *CA Cancer J Clin* (2021) 71(1):7–33. doi: 10.3322/caac.21654

2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin* (2022) 72(1):1–27. doi: 10.3322/caac.21708

3. Duma N, Santana-Davila R, Molina JR. Non-small cell lung cancer: epidemiology, screening, diagnosis, and treatment. *Mayo Clin Proc* (2019) 94 (8):1623–40. doi: 10.1016/j.mayocp.2019.01.013

4. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC lung cancer staging project: Proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. *J Thorac Oncol* (2016) 11(1):39–51. doi: 10.1016/j.jtho.2015.09.009

5. Wang J, Shen Q, Shi Q, Yu B, Wang X, Cheng K, et al. Detection of ALK protein expression in lung squamous cell carcinomas by immunohistochemistry. *J Exp Clin Cancer Res* (2014) 33:109. doi: 10.1186/s13046-014-0109-2

6. Houston KA, Henley SJ, Li J, White MC, Richards TB. Patterns in lung cancer incidence rates and trends by histologic type in the United States, 2004-2009. *Lung Cancer* (2014) 86(1):22–8. doi: 10.1016/j.lungcan.2014.08.001

7. Caliò A, Nottegar A, Gilioli E, Bria E, Pilotto S, Peretti U, et al. ALK/EML4 fusion gene may be found in pure squamous carcinoma of the lung. *J Thorac Oncol* (2014) 9(5):729–32. doi: 10.1097/JTO.0000000000000109

8. Youlden DR, Cramb SM, Baade PD. The international epidemiology of lung cancer: geographical distribution and secular trends. *J Thorac Oncol* (2008) 3 (8):819–31. doi: 10.1097/JTO.0b013e31818020eb

9. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* (2008). doi: 10.1002/0471142905.hg1011s57

10. Mazieres J, Rittmeyer A, Gadgeel S, Hida T, Gandara DR, Cortinovis DL, et al. Atezolizumab versus docetaxel in pretreated patients with nsclc: Final results from the randomized phase 2 poplar and phase 3 oak clinical trials. *J Thorac Oncol* (2021) 16(1):140–50. doi: 10.1016/j.jtho.2020.09.022

11. Herbst RS, Garon EB, Kim D-W, Cho BC, Gervais R, Perez-Gracia JL, et al. Five year survival update from keynote-010: Pembrolizumab versus docetaxel for previously treated, programmed death-ligand 1-positive advanced nsclc. *J Thorac Oncol* (2021) 16(10):1718–32. doi: 10.1016/j.jtho.2021.05.001

12. Borghaei H, Gettinger S, Vokes EE, Chow LQM, Burgio MA, de Castro Carpeno J, et al. Five-year outcomes from the randomized, phase iii trials checkmate 017 and 057: Nivolumab versus docetaxel in previously treated non-small-cell lung cancer. *J Clin Oncol* (2021) 39(7):723–33. doi: 10.1200/JCO.20.01605

13. Bie F, Tian H, Sun N, Zang R, Zhang M, Song P, et al. Research progress of anti-pd-1/pd-l1 immunotherapy related mechanisms and predictive biomarkers in nsclc. *Front Oncol* (2022) 12:769124. doi: 10.3389/fonc.2022.769124

14. Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Mol Cancer Ther* (2015) 14(4):847–56. doi: 10.1158/1535-7163.MCT-14-0983

15. Shimoda Y, Yoshida T, Shirasawa M, Mizuno T, Jo H, Matsumoto Y, et al. Smoking history predicts high presence of tils and efficacy of pd-1 blockade in pd-l1 expression-negative non-small cell lung cancer patients. *Anticancer Res* (2021) 41(11):5739–47. doi: 10.21873/anticanres.15390

16. Nosaki K, Saka H, Hosomi Y, Baas P, de Castro G, Reck M, et al. Safety and efficacy of pembrolizumab monotherapy in elderly patients with PD-L1-positive advanced non-small-cell lung cancer: Pooled analysis from the KEYNOTE-010, KEYNOTE-024, and KEYNOTE-042 studies. *Lung Cancer* (2019) 135:188–95. doi: 10.1016/j.lungcan.2019.07.004

17. Gettinger SN, Horn L, Gandhi L, Spigel DR, Antonia SJ, Rizvi NA, et al. Overall survival and long-term safety of nivolumab (anti-programmed death 1 antibody, bms-936558, ono-4538) in patients with previously treated advanced non-small-cell lung cancer. *J Clin Oncol* (2015) 33(18):2004–12. doi: 10.1200/JCO.2014.58.3708

18. She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. *JAMA Netw Open* (2020) 3(6):e205842. doi: 10.1001/jamanetworkopen.2020.5842

19. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res* (2019) 25(11):3266–75. doi: 10.1158/1078-0432.CCR-18-2495

20. Nam JG, Park S, Hwang EJ, Lee JH, Jin K-N, Lim KY, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* (2019) 290(1):218–28. doi: 10.1148/radiol.2018180237

21. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* (2019) 25(7):1054–6. doi: 10.1038/s41591-019-0462-y

22. Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* (2019) 2(3):e191095. doi: 10.1001/jamanetworkopen.2019.1095

23. Van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. *J Stat software* (2011) 45:1–67. doi: 10.18637/jss.v045.i03

24. McKinney W. (2010). Data structures for statistical computing in python, in: *Proceedings of the 9th Python in Science Conference, Austin, TX*. 51-56

25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.

26. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* (2018) 18(1):24. doi: 10.1186/s12874-018-0482-1

27. vanRossum G. *Python reference manual*. Department of computer science [CS] (1995). R 9525.

28. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. *Adv Neur In* (2019), 32.

29. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature* (2020) 585 (7825):357–62. doi: 10.1038/s41586-020-2649-2

30. Hunter JD. Matplotlib: A 2D graphics environment. *Computing Sci Eng* (2007) 9(03):90–5. doi: 10.1109/MCSE.2007.55

31. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *arXiv preprint arXiv:190700825* (2019).

32. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf* (2011) 12(1):1–8. doi: 10.1186/1471-2105-12-77

33. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. Package 'corrplot'. *Statistician* (2017) 56(316):e24.

34. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics* (2014) 30(19):2811–2. doi: 10.1093/bioinformatics/btu393

35. Kassambara A, Kosinski M, Biecek P, Fabian S. Package 'survminer'. *Drawing Survival Curves using 'ggplot2'(R Package version 03 1)* (2017).

36. Team RC. *R: A language and environment for statistical computing*. (2013).

37. Villanueva RAM, Chen ZJ. *ggplot2: elegant graphics for data analysis*. Taylor & Francis (2019).

38. Sunshine J, Taube JM. PD-1/PD-L1 inhibitors. *Curr Opin In Pharmacol* (2015) 23:32–8. doi: 10.1016/j.coph.2015.05.011

39. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* (2017) 389(10066):255–65. doi: 10.1016/S0140-6736(16)32517-X

40. Mamdani H, Matosevic S, Khalid AB, Durm G, Jalal SI. Immunotherapy in lung cancer: Current landscape and future directions. *Front Immunol* (2022) 13:823618. doi: 10.3389/fimmu.2022.823618

41. Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet* (2016) 387(10030):1837–46. doi: 10.1016/S0140-6736(16)00587-0

42. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *New Engl J Med* (2018) 378(22):2078–92. doi: 10.1056/NEJMoa1801005

43. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *New Engl J Med* (2015) 373(2):123–35. doi: 10.1056/NEJMoa1504627

44. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *New Engl J Med* (2015) 373(17):1627–39. doi: 10.1056/NEJMoa1507643

45. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, et al. First-line nivolumab in stage iv or recurrent non-small-cell lung cancer. *New Engl J Med* (2017) 376(25):2415–26. doi: 10.1056/NEJMoa1613493

46. Satturwar S, Girolami I, Munari E, Ciompi F, Eccher A, Pantanowitz L. Program death ligand-1 immunocytochemistry in lung cancer cytological samples: A systematic review. *Diagn Cytopathol* (2022) 50(6):313–23. doi: 10.1002/dc.24955

47. Paolino G, Pantanowitz L, Barresi V, Pagni F, Munari E, Moretta L, et al. PD-L1 evaluation in head and neck squamous cell carcinoma: Insights regarding specimens, heterogeneity and therapy. *Pathol Res Pract* (2021) 226:153605. doi: 10.1016/j.prp.2021.153605

48. Marletta S, Fusco N, Munari E, Luchini C, Cimadamore A, Brunelli M, et al. Atlas of pd-l1 for pathologists: indications, scores, diagnostic platforms and reporting systems. *J Pers Med* (2022) 12(7):1–16. doi: 10.3390/jpm12071073

49. Mu W, Jiang L, Shi Y, Tunali I, Gray JE, Katsoulakis E, et al. Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images. *J Immunother Cancer* (2021) 9(6):1–15. doi: 10.1136/jitc-2020-002118

50. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus chemotherapy for pd-l1-positive non-small-cell lung cancer. *New Engl J Med* (2016) 375(19):1823–33. doi: 10.1056/NEJMoa1606774

51. Zhou C, Wu L, Fan Y, Wang Z, Liu L, Chen G, et al. Sintilimab plus platinum and gemcitabine as first-line treatment for advanced or metastatic squamous nsclc: Results from a randomized, double-blind, phase 3 trial (orient-12). *J Thorac Oncol* (2021) 16(9):1501–11. doi: 10.1016/j.jtho.2021.04.011

52. Ren S, Chen J, Xu X, Jiang T, Cheng Y, Chen G, et al. Camrelizumab plus carboplatin and paclitaxel as first-line treatment for advanced squamous nsclc (camel-sq): A phase 3 trial. *J Thorac Oncol* (2022) 17(4):544–57. doi: 10.1016/j.jtho.2021.11.018

53. Paz-Ares L, Vicente D, Tafreshi A, Robinson A, Soto Parra H, Mazieres J, et al. A randomized, placebo-controlled trial of pembrolizumab plus chemotherapy in patients with metastatic squamous nsclc: Protocol-specified final analysis of keynote-407. *J Thorac Oncol* (2020) 15(10):1657–69. doi: 10.1016/j.jtho.2020.06.015

54. Herbst RS, Giaccone G, de Marinis F, Reinmuth N, Vergnenegre A, Barrios CH, et al. Atezolizumab for first-line treatment of pd-l1-selected patients with nsclc. *N Engl J Med* (2020) 383(14):1328–39. doi: 10.1056/NEJMoa1917346

55. Zhou C, Wang Z, Sun Y, Cao L, Ma Z, Wu R, et al. Sugemalimab versus placebo, in combination with platinum-based chemotherapy, as first-line treatment of metastatic non-small-cell lung cancer (GEMSTONE-302): interim and final analyses of a double-blind, randomised, phase 3 clinical trial. *Lancet Oncol* (2022) 23(2):220–33. doi: 10.1016/S1470-2045(21)00650-1

56. Provencio M, Nadal E, Insa A, García-Campelo MR, Casal-Rubio J, Dómine M, et al. Neoadjuvant chemotherapy and nivolumab in resectable non-small-cell lung cancer (NADIM): an open-label, multicentre, single-arm, phase 2 trial. *Lancet Oncol* (2020) 21(11):1413–22. doi: 10.1016/s1470-2045(20)30453-8

57. Forde PM, Spicer J, Lu S, Provencio M, Mitsudomi T, Awad MM, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. *N Engl J Med* (2022) 386(21):1973–85. doi: 10.1056/NEJMoa2202170

58. Zhang F, Guo W, Zhou B, Wang S, Li N, Qiu B, et al. Three-year follow-up of neoadjuvant programmed cell death protein-1 inhibitor (sintilimab) in nsclc. *J Thorac Oncol* (2022) 17(7):909–20. doi: 10.1016/j.jtho.2022.04.012

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 06 December 2022
DOI 10.3389/fonc.2022.955056

Check for updates

OPEN ACCESS

EDITED BY
Ken Chen,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Wei Wei,
Yale University, United States
Thomas Murray,
University of Minnesota Twin Cities, United States

*CORRESPONDENCE
Emily C. Zabor
zabore2@ccf.org

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Oncology

RECEIVED 28 May 2022
ACCEPTED 15 November 2022
PUBLISHED 06 December 2022

CITATION
Zabor EC, Kaizer AM, Pennell NA and Hobbs BP (2022) Optimal predictive probability designs for randomized biomarker-guided oncology trials. *Front. Oncol.* 12:955056. doi: 10.3389/fonc.2022.955056

COPYRIGHT
© 2022 Zabor, Kaizer, Pennell and Hobbs. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author (s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Optimal predictive probability designs for randomized biomarker-guided oncology trials

Emily C. Zabor[1]*, Alexander M. Kaizer[2], Nathan A. Pennell[3] and Brian P. Hobbs[4]

[1]Lerner Research Institute & Taussig Cancer Institute, Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, United States, [2]Colorado School of Public Health, Department of Biostatistics and Informatics, University of Colorado, Aurora, CO, United States, [3]Taussig Cancer Institute, Department of Hematology and Medical Oncology, Cleveland Clinic, Cleveland, OH, United States, [4]Department of Population Health, University of Texas-Austin, Austin, TX, United States

Introduction: Efforts to develop biomarker-targeted anti-cancer therapies have progressed rapidly in recent years. With efforts to expedite regulatory reviews of promising therapies, several targeted cancer therapies have been granted accelerated approval on the basis of evidence acquired in single-arm phase II clinical trials. And yet, in the absence of randomization, patient prognosis for progression-free survival and overall survival may not have been studied under standard of care chemotherapies for emerging biomarker subpopulations prior to the submission of an accelerated approval application. Historical control rates used to design and evaluate emerging targeted therapies often arise as population averages, lacking specificity to the targeted genetic or immunophenotypic profile. Thus, historical trial results are inherently limited for inferring the potential "comparative efficacy" of novel targeted therapies. Consequently, randomization may be unavoidable in this setting. Innovations in design methodology are needed, however, to enable efficient implementation of randomized trials for agents that target biomarker subpopulations.

Methods: This article proposes three randomized designs for early phase biomarker-guided oncology clinical trials. Each design utilizes the optimal efficiency predictive probability method to monitor multiple biomarker subpopulations for futility. Only designs with type I error between 0.05 and 0.1 and power of at least 0.8 were considered when selecting an optimal efficiency design from among the candidate designs formed by different combinations of posterior and predictive threshold. A simulation study motivated by the results reported in a recent clinical trial studying atezolizumab treatment in patients with locally advanced or metastatic urothelial carcinoma is used to evaluate the operating characteristics of the various designs.

Results: Out of a maximum of 300 total patients, we find that the enrichment design has an average total sample size under the null of 101.0 and a total average sample size under the alternative of 218.0, as compared to 144.8 and 213.8 under the null and alternative, respectively, for the stratified control arm design. The pooled control arm design enrolled a total of 113.2 patients under the null and 159.6 under the alternative, out of a maximum of 200. These average sample sizes that are 23-48% smaller under the alternative and 47-64% smaller under the null, as compared to the realized sample size of 310 patients in the phase II study of atezolizumab.

Discussion: Our findings suggest that potentially smaller phase II trials to those used in practice can be designed using randomization and futility stopping to efficiently obtain more information about both the treatment and control groups prior to phase III study.

KEYWORDS

predictive probability, oncology, clinical trial, phase II, randomized, futility monitoring, two-sample, Bayesian

Introduction

Over the last decade the focus of drug discovery in oncology has shifted away from cytotoxic treatments and toward biomarker-targeted agents. For these types of drugs, such as small molecule inhibitors, antibody drug conjugates, immune checkpoint inhibitors, and monoclonal antibodies, the traditional approach to clinical trial design is not always appropriate. Traditionally, new chemotherapeutic treatments were evaluated in phase I dose-escalation trials to assess safety and identify the maximum tolerated dose. Next, the maximum tolerated dose would be tested for preliminary efficacy in single-arm phase II trials, with a historical control rate forming the basis of comparison. Finally, successful drugs would proceed to phase III, where randomized trials would be used to directly compare efficacy against a standard of care treatment. But the historical control rates used in single-arm phase II studies may not be valid in the context of biomarker-targeted agents. Historical control rates used to design and evaluate emerging targeted therapies often arise as population averages, lacking specificity to the targeted genetic or immunophenotypic profile of interest. Patient prognosis for objective response, progression-free survival, and overall survival may not have been studied under standard of care chemotherapies for emerging biomarker subpopulations prior to phase III. Other factors, such as patient population drift or stage shift, add heterogeneity and bias (1). Consequently, expectations for response and survival for the current biomarker delineated patient populations may differ meaningfully from population averages observed in prior studies of current standard of care therapies. Additionally, in the specific context of biomarker-targeted agents, heterogeneity of response to standard of care treatments based on the biomarker of interest is also possible, so that the historical control rate may represent an averaging of effect across levels of the biomarker of interest. If the biomarker of interest is prognostic, then response to standard of care treatment in the biomarker-targeted subgroup will differ from the population-averaged response regardless of the treatment being given (2).

Consider the recent clinical trial of atezolizumab for use in metastatic urothelial carcinoma (NCT01375842). Atezolizumab is a programmed death-ligand 1 (PD-L1) blocking monoclonal antibody that was given accelerated approval by the U.S. Food and Drug Administration in May 2016 for the treatment of patients with locally advanced or metastatic urothelial carcinoma who had disease progression following platinum-containing chemotherapy. The approval was based on the results of a single-arm phase II study in 310 patients (3). The phase II study used a hierarchical fixed-sequence testing procedure to test increasingly broad subgroups of patients based on PD-L1 status, and found overall response rates of 26% (95% CI: 18-36), 18% (95% CI: 13-24), and 15% (95% CI 11-19) in patients with ≥5% PD-L1-positive immune cells (IC2/3 subgroup), in patients with ≥1% PD-L1-positive immune cells (IC1/2/3 subgroup), and in all patients, respectively (3). All three rates exceeded the historical control rate of 10%. Then, in March 2021, the approval in this indication was voluntarily withdrawn by the sponsor following negative results from a randomized phase III study (NCT02302807) (4). In the phase III study, 931 patients were randomly assigned to receive atezolizumab or chemotherapy in a 1:1 ratio, and the same hierarchical fixed-

sequence testing procedure as in the phase II study was used. The phase III study found that overall survival did not differ significantly between the atezolizumab and chemotherapy groups of the IC2/3 subgroup (median survival 11.1 months [95% CI: 8.6-15.5] versus 10.6 months [95% CI: 8.4-12.2]), so no further testing was conducted for the primary endpoint (4). Further analyses revealed that while the response rates to atezolizumab were comparable to those seen in the phase II study, the response rates to chemotherapy were much higher than the historical control rate of 10%. The overall response rates to chemotherapy were 21.6% (95% CI: 14.5-30.2), 14.7% (95% CI: 10.9-19.2), and 13.4% (95% CI: 10.5-16.9) for the IC2/3 subgroup, IC1/2/3 subgroup, and all patients, respectively. The overall response rates to atezolizumab were 23% (95% CI: 15.6-31.9), 14.1% (95% CI: 10.4-18.5), and 13.4% (95% CI: 10.5-16.9) for the IC2/3 subgroup, IC1/2/3 subgroup, and all patients, respectively. These results indicate that PD-L1 status is a prognostic biomarker, with higher response rates to both the standard of care chemotherapies that comprised the control arm and to atezolizumab treatment in the biomarker-enriched subgroup (2).

The example of atezolizumab in metastatic urothelial carcinoma is one of many. Between 2015 and 2021, the U.S. Food and Drug Administration (FDA) approved six antibodies against PD-L1 or programmed death 1 (PD-1) for 75 cancer indications, and 35 of these approvals were accelerated based on early phase trial results (5). This extremely rapid pace of development within a single drug class was unprecedented, and led to ten such "dangling" accelerated approvals, which are approved indications for which the confirmatory trial showed no benefit, yet the drug remained on the market for that indication (5). Other voluntary withdrawals following confirmatory trial results include durvalumab treatment for metastatic urothelial carcinoma, and nivolumab and pembrolizumab treatments for metastatic small-cell lung cancer (5–9). These failed confirmatory phase III trials highlight both the need for rapid development of new treatments in patient populations with few therapeutic options, and the need for innovations that facilitate more rigorous designs of phase II trials for targeted therapies. To overcome many issues, including those associated with the use of historical control rates, randomization may be unavoidable in this setting. Arguments for the use of randomization in the phase II setting have been prominent for over a decade (10–13). In addition to addressing the inconvenient reality that historical control rates often used in single arm studies may have limited value for novel targets, randomized phase II trials can also overcome issues of selection bias and patient heterogeneity. Randomized designs that incorporate futility stopping can provide information on current control rates to the treatment under study while also stopping inefficacious treatments early. The FDA's Project Frontrunner (https://www.fda.gov/about-fda/oncology-center-excellence/project-frontrunner) encourages the use of randomized controlled trials earlier in the drug development process, especially if accelerated approval is one of the trial goals.

This article proposes three different biomarker-guided randomized phase II trial designs with optimal efficiency predictive probability monitoring for futility. Using the trial of atezolizumab for metastatic urothelial carcinoma as a case study and motivating example, we compare designs based on their traditional statistical properties of type I error and power through simulation study. The designs are also evaluated based on the number of patients enrolled, the number of patients treated, the number of patients who undergo biomarker testing, and accurate estimation of the response rates of interest. Our findings suggest that potentially smaller phase II trials to those used in practice can be designed using randomization and futility stopping to efficiently obtain more information about both the treatment and control groups prior to phase III study.

Materials and methods

This paper focuses on the setting of a two-sample randomized trial with a binary outcome. We will refer to the binary outcome as "response" and use "response rate" to describe the probability of a response throughout the article, in line with the motivating example of the phase II study of atezolizumab in metastatic urothelial carcinoma, which estimated response rates among biomarker subpopulations and compared to the historical average in the primary analysis. Any hypothetical measure of efficacy, however, such as progression-free survival, could be used with the design methodology proposed. Each patient enrolled in the trial is denoted by i, and they either have a response such that x_i=1 or do not have a response such that x_i=0. Then $X = \sum_{i=1}^{n} x_i$ represents the total number of responses out of n currently observed patients, up to a maximum sample size of N total patients. The probability of response is denoted p , where p_0 represents the null response rate under the standard of care treatment and p_1 represents the alternative response rate under the experimental treatment. We wish to test the null hypothesis H_0:$p_1 \leq p_0$ versus the alternative hypothesis H_1:$p_1 > p_0$.

The Bayesian statistical paradigm is based on a mathematical approach to combine prior distributions, which reflect prior beliefs about parameters such as the true response rate, with observed data (e.g., the observed number of responses in a given trial) to obtain posterior distributions of the model parameters. Here we assume a beta-binomial model, based on the computational ease and its popularity for use in the context of sequential trial monitoring. The prior distribution of the response rate has a beta distribution $Beta(a_0, b_0)$. We specifically use a $Beta(0.5, 0.5)$ prior distribution, which reflects the effective information of a single patient's observation. We also perform sensitivity analyses using $Beta(1, 1)$, $Beta(0, 0)$, $Beta(2, 2)$, $Beta(0.75, 0.25)$, and $Beta(0.25, 0.75)$

priors, to examine the variation in operating characteristics across a range of priors. Our data X follow a binomial distribution $bin(n, p)$. We combine the likelihood function for the observed data $L_x(p) \propto p^x(1-p)^{n-x}$ with the prior to obtain the posterior distribution of the response rate, which follows the beta distribution $p|x \sim Beta(a_0+x,\ b_0+n-x)$. Posterior probabilities represent the probability that the experimental response rate exceeds the null response rate based on the data accrued so far in the trial. Posterior decision can be obtained by applying a clinically relevant threshold, θ, to the posterior distribution. We would declare a treatment efficacious if the posterior probability exceeded the posterior threshold, i.e. $Pr(p>p_0|X)>\theta$.

Bayesian predictive probability monitoring has been a popular approach for designing clinical trials with sequential futility monitoring (14–17). It is a natural fit for this type of trial, as it allows for flexibility in both the timing and the number of looks. In addition, predictive probability is an intuitive interim monitoring strategy because it tells the investigator what the chances are of declaring the treatment efficacious at the end of the trial if enrollment is continued to the maximum planned sample size. At any given interim look, the posterior predictive distribution of the number of future responses $X^\star$ in the remaining $n^\star=N-n$ future patients follows a beta-binomial distribution $Beta\text{–}binomial(n^\star,\ a_0+x,\ b_0+n-x)$. The posterior predictive probability (PPP) represents the probability that the experimental treatment will be declared efficacious at the end of the trial when full enrollment is reached, conditional on the currently observed data and the specified priors. The posterior predictive probability is calculated as $PP = \sum_{x^\star=0}^{n^\star} \Pr(X^\star = x^\star|x) \times I(\Pr(p > p_0|X,\ X^\star = x^\star) > \theta)$. A second predictive threshold $\theta^\star$ is defined, and we would stop the trial early for futility if the predictive probability dropped below the given threshold, i.e. $PPP<\theta^\star$. Predictive thresholds closer to 0 lead to less frequent stopping for futility whereas predictive thresholds closer to 1 lead to frequent stopping in the absence of almost certain probability of success.

When designing a trial with sequential predictive probability monitoring for futility, it is essential to ensure the trial conforms to traditional standards for type I error control and power. To do so, we must examine the operating characteristics of a variety of designs based on combinations of the posterior threshold θ and the predictive threshold $\theta^\star$ and select a single design for use in the trial. In earlier work, we proposed two optimization criteria to help select from among a variety of designs in the setting of a one-sample study (18). Here we focus on the optimal efficiency design, defined as the combination of posterior and predictive thresholds with minimal average sample size under the null and maximal average sample size under the alternative, subject to constraints on the type I error and power, and extend the approach to the setting of a two-sample study for targeted therapy.

The simulation study is based on the phase II trial of atezolizumab in metastatic urothelial carcinoma. There are three independent biomarker subgroups based on the percentage of PD-L1-expressing immune cells: IC0 (<1%), IC1 (≥1% and<5%), and IC2/3 (≥5%). The subgroups have equal prevalence of 33% in the study population. We consider a standard of care arm denoted "chemotherapy" and an experimental treatment arm denoted "atezolizumab". The null response rate was based on the stated historical control rate of 10% (3). As no specific alternative was specified, we examine subtype-specific alternative rates of 10%, 20%, and 30% in the IC0, IC1, and IC2/3 subgroups, respectively, in line with what we expect for a predictive biomarker, for which the treatment effect differs according to levels of the biomarker of interest (2). We further investigate a simulation setting where the treatment effect is a homogeneous 30% across the three biomarker subgroups. Interim looks for futility are planned after every 10 patients. A random number of responses was generated for every 10 patients up to the maximum sample size, based on a binomial distribution with the setting-specific response rate. 1000 simulated datasets were generated under the null and 1000 simulated datasets were generated under the alternative. We considered posterior thresholds θ of 0.7, 0.74, 0.78, 0.82, 0.86, 0.9, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, and 0.99, and predictive thresholds $\theta^\star$ of 0.05, 0.1, 0.15, and 0.2. For each combination of posterior and predictive threshold, the predictive probability that the experimental treatment arm response rate exceeds the standard of care arm response rate at the end of the trial is calculated at each interim look until it either fell below the given predictive threshold or the end of the trial was reached, whichever came first. If the end of the trial was reached, the trial was considered positive if the predictive probability was greater than the given posterior threshold and negative otherwise. If halted early for futility, the trial was considered negative. We propose and compare three strategies for conducting randomized two-sample biomarker-guided designs that use optimal efficiency predictive probability monitoring for futility: a pooled control arm design, a stratified control arm design, and an enrichment design.

The pooled control arm design is depicted in Figure 1A. In this design, patients are randomized to atezolizumab or chemotherapy in a 3:1 ratio. PD-L1 testing is performed only on patients randomized to receive atezolizumab. The atezolizumab arm is separated into three biomarker-specific treatment subgroups. This design has a maximum sample size of 200: n=50 patients for the pooled chemotherapy control arm and n=50 for each PD-L1 biomarker-specific atezolizumab arm.

The stratified control arm design is depicted in Figure 1B. In this design, PD-L1 testing is conducted on all patients. Then, within each subgroup, patients are randomized to atezolizumab or chemotherapy in a 1:1 ratio. This design has a maximum sample size of 300: n=50 for each PD-L1 biomarker-specific chemotherapy and atezolizumab arm.

The enrichment design is depicted in Figure 1C. This design is equivalent to the pooled design at stage 1. If all subgroups stop for futility in stage 1, then the trial is stopped. Otherwise, at the

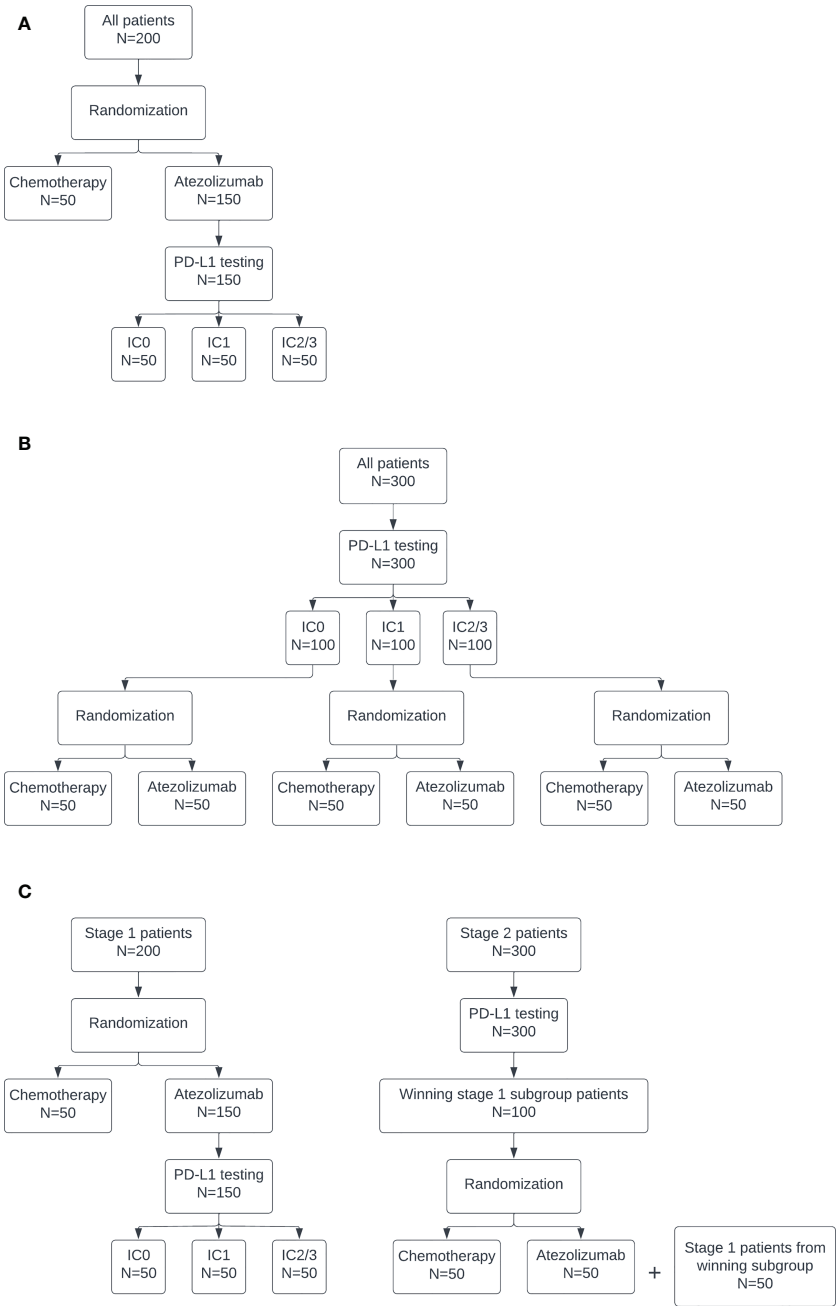


FIGURE 1
Diagrams of the **(A)** pooled control arm, **(B)** stratified control arm, and **(C)** enrichment designs.

end of stage 1, the subgroup with the highest posterior predictive probability, subject to some lower bound, continues to stage 2. The lower bound was selected as the 80th percentile of maximum posterior predictive probability across the three subgroups at stage 1 under the null. This percentile was used to target a 20% rate of moving a subgroup forward when all of the subgroups are truly null. This higher rate of stage 1 type I error is consistent with the phase objective, which emphasizes acquiring more data on safety and efficacy for promising treatments in early phase trial designs of this type. The actual type I error at stage 1 was calculated as the proportion of simulated trials under the global null, i.e. if all three biomarker-specific subgroups had a true response rate of 10%, in which the subgroup with maximum posterior predictive probability exceeded the lower bound and did not stop early for futility, so was selected to continue to stage 2. The power at

stage 1 was calculated as the proportion of simulated trials under the alternative in which the IC2/3 subgroup was selected as having the maximum posterior predictive probability, subject to the lower bound, and did not stop early for futility. In stage 2, PD-L1 testing is conducted on all patients. Only those patients belonging to the subgroup selected in stage 1 are enrolled on the trial and randomized 1:1 to atezolizumab or chemotherapy. The stage 1 treatment group results for the selected subgroup, if any, are carried forward into stage 2. An additional n=100 biomarker-specific patients are enrolled at stage 2, for a total maximum sample size of 300.

For the pooled and stratified designs, the type I error was calibrated in the IC0 subgroup null setting as the proportion of simulated trials in which the IC0 subgroup was declared positive as compared to the control group. The power was calibrated in the IC2/3 subgroup alternative setting as the proportion of simulated trials in which the IC2/3 subgroup was declared positive as compared to the control group. The IC1 subgroup was considered an intermediate setting and no results were calibrated based on this subgroup. For the enrichment design, the type I error was calibrated based on the stage 2 results as the proportion of simulated trials under the null in which the selected subgroup, if any, was declared positive. The power was calibrated based on the stage 2 results as the proportion of simulated trials under the alternative in which the selected subgroup, if any, was declared positive. The stage 2 calibration occurs in fewer than 1000 simulated trials, as only the specific trials in which a subgroup was selected to continue to stage 2 were used. The resulting design options were limited to those that resulted in a type I error rate between 0.05 and 0.1 and a power of at least 0.8. Then, the efficiency distance metric was calculated as described in Zabor et al. (18) using total trial-level sample sizes. The design with the minimal efficiency distance metric was identified as the optimal design.

All results were generated using R software version 4.2.0 (19) along with the 'ppseq' R package (20).

Results

The accuracy and efficiency results for all 56 possible posterior and predictive threshold combinations are plotted in Figure 2. Each point represents the combination of one posterior threshold and one predictive threshold, and the orange diamond on each plot indicates the design that was identified to have optimal efficiency while maintaining type I error between 0.05 and 0.1 and with power of at least 0.8. For the enrichment design, only threshold combinations that ever proceeded to stage 2 are plotted so there are only 36 points, as 20 threshold combinations never resulted in designs that continued to stage 2. The optimal efficiency pooled control arm design had posterior threshold 0.9 and predictive threshold 0.1, the optimal efficiency stratified control arm design had posterior threshold 0.9 and predictive threshold 0.2, and the optimal efficiency enrichment design had posterior threshold 0.96 and predictive threshold 0.15. We see that the different threshold combinations result in a wide range of results, some with low power < 0.5 or high type I error > 0.2, and some with too low average sample size under the alternative or too high average sample size under the null. By applying the optimal efficiency criteria, we are able to identify a design that seeks to maximize sample size under the alternative and minimize sample size under the null, within the pre-specified range of type I error and minimum power.

The type I error and power for each biomarker-specific subgroup under the pooled control arm and stratified control arm designs, and the overall type I error and power for the enrichment design, are presented in Table 1. We see that both the pooled control arm and stratified control arm designs result in reasonable power to detect an effect for the IC2/3 subgroup, with slightly higher power of 0.82 in the stratified control arm design as compared to 0.8 in the pooled control arm design. The pooled control arm design and stratified control arm design both have type I error of 0.07 for the IC2/3 subgroup. Both the pooled control arm and stratified control arm designs have very low power< 0.5 to detect the IC1 subgroup and< 0.1 to detect the IC0 subgroup. Only overall results are available for the enrichment design, which results in a type I error and power of 0.09 and 0.73, respectively, for stage 1. The type 1 error rate of 0.09 for stage 1 means that under the null 91% of simulated trials did not proceed to stage 2; however, 4.2% proceeded to stage 2 with the IC2/3 subgroup, 2.9% proceeded to stage 2 with the IC1 subgroup, and 1.9% proceeded to stage 2 with the IC0 subgroup. The power of 0.73 for stage 1 means that under the alternative 73% of simulated trials proceeded to stage 2, and all of them did so with the IC2/3 subgroup; the remaining 27% of simulated trials did not proceed to stage 2. The overall type I error rate for stage 2 of the enrichment design was 0.09 and the overall power for stage 2 of the enrichment design was 0.86. Since the IC2/3 subgroup was exclusively carried forward to stage 2 of the enrichment design under the alternative, this could also be considered the power for the IC2/3 subgroup, and it exceeds the power of 0.82 of the stratified control arm design and the power of 0.8 of the pooled control arm design.

The resulting average sample sizes under the null and alternative for each selected optimal efficiency design are presented in Table 2. We can directly compare the total sample sizes between the stratified control arm and enrichment designs, which can each enroll a maximum of 300 patients. We find that the enrichment design has a lower average total sample size under the null of 101.0 and a higher total average sample size under the alternative of 218.0, as compared to 144.8 and 213.8 under the null and alternative, respectively, for the stratified control arm design. This occurs because the use of the pooled control arm design at stage 1 combined with the low stage 1 type 1 error rate and high stage 2 power means that

10.3389/fonc.2022.955056

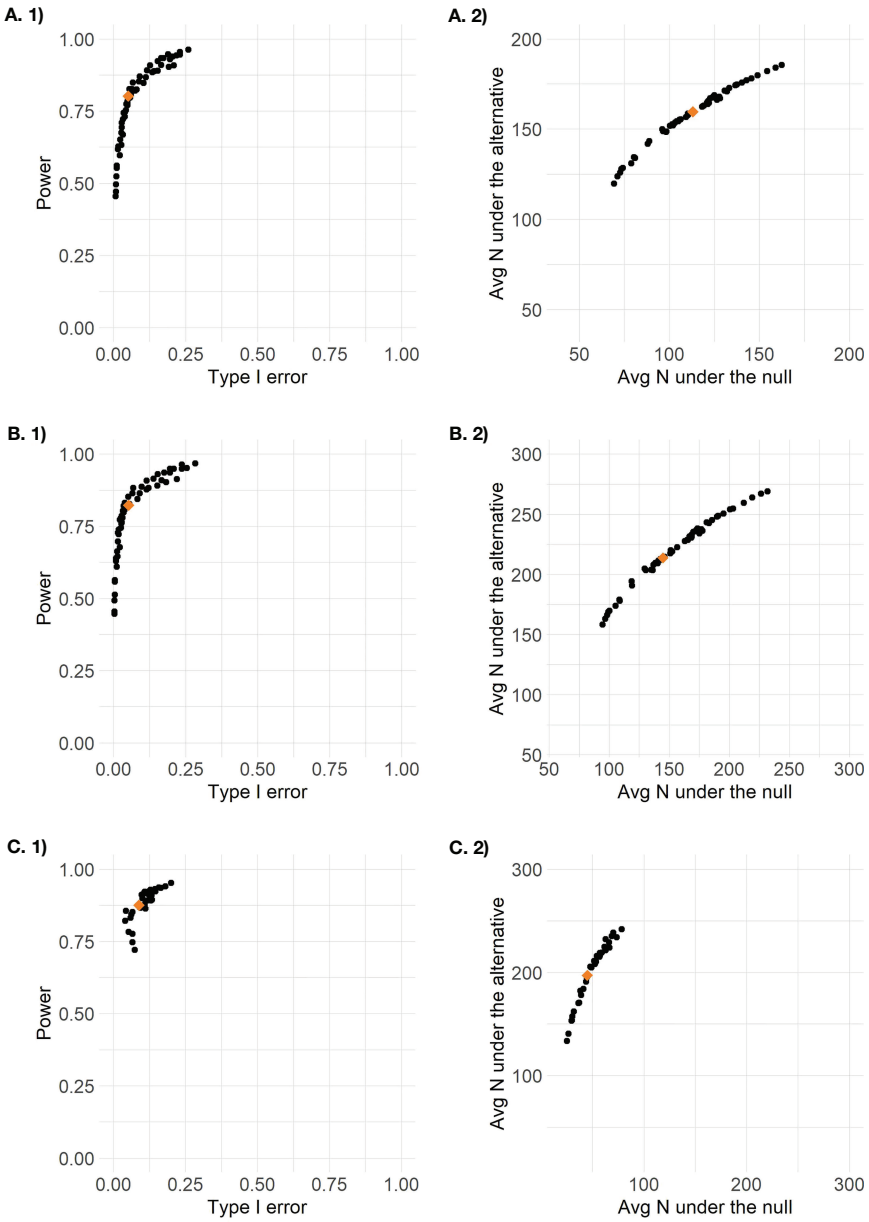


FIGURE 2
Plots of design options for **(A)** the pooled control design, **(B)** the stratified control design, and **(C)** the enrichment design (stage 2 results only) based on 1) accuracy defined as type I error by power and 2) efficiency defined as average total sample size under the null versus average total sample size under the alternative.

the enrichment design frequently stops after stage 1 under the null and frequently continues to full enrollment in stage 2 under the alternative. The pooled control arm design enrolled a total of 113.2 patients under the null and 159.6 patients under the alternative, out of a maximum of 200. The sample sizes of patients treated with atezolizumab are directly comparable for the pooled control arm and stratified control arm designs, which can each treat a maximum of 150 patients with atezolizumab, 50 per biomarker subgroup. The pooled control arm design treated more patients with atezolizumab on average as compared to the stratified control arm design, 78.2 and 111.7 under the null and alternative, respectively, as compared to 72.4 and 106.9. The enrichment design can treat a maximum of 200 patients with atezolizumab across the two stages, and treats an average of 68.0 patients under the null and 137.0 patients under the alternative. We find that there is minimal variation in operating characteristics across the six prior distributions examined in sensitivity analysis (Supplemental Table 1).

TABLE 1 Type I error and power for each biomarker-specific subgroup under the pooled control arm and stratified control arm designs, and the overall type I error and power for the enrichment design.

	Pooled		Stratified		Enrichment			
					Stage 1		Stage 2	
	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power
IC0	0.05	0.05	0.05	0.08	–	–	–	–
IC1	0.05	0.40	0.08	0.45	–	–	–	–
IC2/3	0.07	0.80	0.07	0.82	–	–	–	–
Overall	–	–	–	–	0.09	0.73	0.09	0.86

For the simulation setting where the treatment effect was homogeneous across the biomarker subgroups, we find that both the pooled control arm and stratified control arm designs perform well, with consistent type I error, power, and sample size across the three subgroups (Supplemental Tables 2 and 3). The enrichment design, however, has low stage 1 power (Supplemental Table 2). This is because there is no distinction between the three biomarker subgroups with respect to which should be the winning subgroup, so it is easy for the method to make the "wrong" choice, as we defined power in this setting based on a predictive biomarker as the proportion of simulated trials under the alternative in which the IC2/3 subgroup was selected as having the maximum posterior predictive probability, subject to the lower bound, and did not stop early for futility. However, the enrichment design performs well at stage 2, since any subgroup selected to continue would have equal chance of success in the second stage of the trial, given the homogeneous true response rates (Supplemental Table 2).

The enrichment design requires testing the largest number of patients, with 450 patients requiring PD-L1 testing if the design proceeds to stage 2, whereas the pooled design only tests 150 patients and the stratified design tests 300 patients. The pooled control arm and enrichment designs cannot address the question of whether the biomarker is predictive of response to the standard of care treatment, since they do not estimate response rates separately within each biomarker subgroup, though the enrichment design can fully characterize the response rate to standard of care treatment within the selected stage 2 biomarker subgroup. Only the stratified control arm design fully characterizes the response rates to standard of care treatment within each biomarker subgroup, and can therefore address the question of whether the biomarker is predictive of response for both standard of care and targeted therapies.

Discussion

This article presented three different optimal efficiency predictive probability designs for randomized biomarker-guided oncology clinical trials. A simulation study was conducted to demonstrate that posterior and predictive thresholds can be selected to maintain appropriate levels of type I error between 5% and 10% and power of at least 80% in all three designs. This work was motivated by the case study of atezolizumab for the treatment of patients with locally advanced or metastatic urothelial carcinoma who had disease progression following platinum-containing chemotherapy. In the phase II trial that led to accelerated approval, 310 patients were enrolled and treated. The trial did not include any futility stopping rules and planned for a total sample size of 300 patients, expecting about 100 patients in each biomarker subgroup. So in practice the realized and expected sample sizes are broadly equivalent. By

TABLE 2 Average sample size under the null ("Avg N Null") and average sample size under the alternative ("Avg N Alt") by design and treatment subgroup.

	Pooled		Stratified		Enrichment	
	Avg N Null	Avg N Alt	Avg N Null	Avg N Alt	Avg N Null	Avg N Alt
Control	35.0	48.0	72.4	106.9	33.1	81.3
IC0	25.9	26.7	23.6	23.8	–	–
IC1	26.3	38.8	24.4	37.5	–	–
IC2/3	26.1	46.2	24.3	45.6	–	–
Total Atezolizumab	78.2	111.7	72.4	106.9	68.0	137.0
Total Enrolled	113.2	159.6	144.8	213.8	101.0	218.0

comparison, the three proposed designs, which incorporate futility stopping and randomization, result in average phase II sample sizes that are 23-48% smaller under the alternative and 47-64% smaller under the null, and therefore represent a more efficient use of both human and financial resources.

At the same time, the three proposed designs provide additional information about response rates to standard of care treatment in the control arms, thus potentially avoiding the pitfall of the atezolizumab trial, in which the historical control rate used to show efficacy in the phase II trial proved to be far below the actual response rate to standard of care treatment in the biomarker-targeted subgroup of patients. The stratified control arm design results in the most information, allowing one to determine if the biomarker of interest is predictive of response to either the standard of care treatment or the experimental treatment or both. The enrichment design only characterizes the response rate to the standard of care treatment in the biomarker subgroup that is selected to continue to stage 2. But both the pooled control arm design and stage 1 of the enrichment design are superior to use of a historical control rate, since the patient population of the control group is identical to that of the treatment group in both timing and characteristics as a result of randomization. So these designs not only have lower average expected sample sizes than the 310 used in the atezolizumab phase II trial, but also have properties such as control groups and sequential futility monitoring that facilitate valid inference of comparative effectiveness and improve decision-making for continuation to phase III. The phase III trial of atezolizumab for this patient population randomized 931 patients who could have been available to enroll in trials of more promising treatments, or could have avoided the rigors of a clinical trial altogether in favor of the established standard of care treatment.

The decision of which design to select will depend on a number of factors. One is the costs of biomarker testing, including invasiveness of the testing procedure, turnaround time, and actual financial cost. The enrichment design tests the most patients whereas the pooled control arm design tests the fewest patients. So in the case of extremely invasive or expensive tests, the pooled control arm design may be preferred. Another consideration is the prevalence of the biomarker in the population. The enrichment design in stage 2 requires testing all patients in order to identify and enroll only patients with the biomarker of interest, which could be prohibitively expensive or time consuming in the setting of a rare biomarker. In that case, the pooled control arm design may be preferable since all patients are enrolled and the control group will more easily reach full enrollment by containing a mix of patients regardless of biomarker status. But any of the proposed designs could result in a more efficient use of resources in the setting of a rare biomarker, considering both the ability to stop the trial early for futility, and the potential to avoid embarking on a confirmatory trial without adequate information about the population under study. A third consideration is clinical evidence for the biomarker being prognostic in nature, leading to differential response rates across biomarker subgroups on even standard of care therapies. If there is preliminary evidence or biological plausibility that such an effect might exist, the stratified control arm design may be preferable since it fully characterizes the response rates of the control groups within each biomarker subgroup. While the trial would "fail" in this setting, since it would find no difference between the control and treatment arms within the biomarker subgroups, the information from the stratified control arm design would be useful for planning future studies. And a final consideration is clinical evidence for the biomarker being predictive of experimental treatment response. If there is a strong belief that only biomarker positive patients will benefit from the treatment under study, then the enrichment design may be best as it enrolls more patients in only the selected subgroup at stage 2. Additional sample size savings could be achieved by eliminating the control group at stage 1 of the enrichment design, though the properties of such a design were not investigated in detail here.

The main limitation to the use of these designs is the computational intensity required to perform calibration across a variety of posterior and predictive thresholds for the setting of interest in order to identify a design with the desired operating characteristics of type I error and power. While we have developed open-source R software for the design of single-arm and two-arm optimal sequential predictive probability designs, specialized programming using the functions from the 'ppseq' R package would be required to design a pooled control arm, stratified control arm, or enrichment design of the type presented here. Moreover, a large memory server is needed to complete the computations in any reasonable time span. However, once the thresholds have been selected, decision rule tables for early stopping can be generated so that no mid-trial computations would be necessary.

As rapid development of biomarker-targeted agents in oncology continues, new implementations of existing statistical methods such as those presented here will represent the most nimble way for the statistical design of trials to keep up with the changing context of cancer treatment. Randomization is an old statistical tool that has not traditionally been employed in early phase oncology clinical trials due to the sample size requirements. But in the context of increasingly large early phase clinical trials that can include hundreds of patients across multiple cancer types or multiple biomarker levels or both, randomization is no longer the constraint that it once was. This kind of efficient design also stresses the importance of mandating that all patients enrolled to biomarker-targeted trials have the biomarker of interest tested at enrollment (as opposed to only a subset of those with tissue available) so that the most

accurate information about efficacy within biomarker groups can be obtained. Here we have demonstrated that it is possible to conduct randomized phase II trials with smaller sample sizes than those being used in practice for single-arm trials. Moreover, Bayesian sequential design with predictive probability yields more efficient and informative early phase clinical trial results than the standard frequentist approaches commonly implemented in practice.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/zabore/article-code-repository/tree/master/Zabor_Randomized-Biomarker-Guided-Designs/data.

Author contributions

EZ, AK, and BP contributed to conception and design of the study. EZ performed the statistical analysis and wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

AK is supported by NHLBI K01HL151754.

Conflict of interest

BH declares advising/consulting for Amgen, Bayer HealthCare Pharmaceuticals Inc., Telperian, and STCube Pharmaceuticals. NP declares advising/consulting for AstraZeneca, Merck, Pfizer, Eli Lilly/LOXO, Genentech, BMS, Amgen, Mirati, Inivata, G1 Therapeutics, Viosera, Xencor, Janssen, Boehringer Ingelheim, and Sanofi-Genzyme.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.955056/full#supplementary-material

References

1. Cannistra SA. Phase II trials in journal of clinical oncology. *J Clin Oncol Off J Am Soc Clin Oncol* (2009) 27(19):3073–6. doi: 10.1200/JCO.2009.23.1811

2. Ballman KV. Biomarker: Predictive or prognostic? *J Clin Oncol* (2015) 33 (33):3968–71. doi: 10.1200/JCO.2015.63.3651

3. Rosenberg JE, Hoffman-Censits J, Powles T, van der Heijden MS, Balar AV, Necchi A, et al. Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet* (2016) 387 (10031):1909–20. doi: 10.1016/S0140-6736(16)00561-4

4. Powles T, Durán I, van der Heijden MS, Loriot Y, Vogelzang NJ, De Giorgi U, et al. Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *Lancet* (2018) 391(10122):748–57. doi: 10.1016/S0140-6736(17)33297-X

5. Beaver JA, Pazdur R. "Dangling" accelerated approvals in oncology. *New Engl J Med* (2021) 384(18):e68. doi: 10.1056/NEJMp2104846

6. Powles T, van der Heijden MS, Castellano D, Galsky MD, Loriot Y, Petrylak DP, et al. Durvalumab alone and durvalumab plus tremelimumab versus chemotherapy in previously untreated patients with unresectable, locally advanced or metastatic urothelial carcinoma (DANUBE): a randomised, open-label, multicentre, phase 3 trial. *Lancet Oncol* (2020) 21(12):1574–88. doi: 10.1016/S1470-2045(20)30541-6

7. Owonikoko TK, Park K, Govindan R, Ready N, Reck M, Peters S, et al. Nivolumab and ipilimumab as maintenance therapy in extensive-disease small-cell lung cancer: CheckMate 451. *J Clin Oncol Off J Am Soc Clin Oncol* (2021) 39 (12):1349–59. doi: 10.1200/JCO.20.02212

8. Spigel DR, Vicente D, Ciuleanu TE, Gettinger S, Peters S, Horn L, et al. Second-line nivolumab in relapsed small-cell lung cancer: CheckMate 331(☆). *Ann Oncol Off J Eur Soc Med Oncol* (2021) 32(5):631–41. doi: 10.1016/j.annonc.2021.01.071

9. Rudin CM, Awad MM, Navarro A, Gottfried M, Peters S, Csőszi T, et al. Pembrolizumab or placebo plus etoposide and platinum as first-line therapy for extensive-stage small-cell lung cancer: Randomized, double-blind, phase III KEYNOTE-604 study. *J Clin Oncol Off J Am Soc Clin Oncol* (2020) 38(21):2369–79. doi: 10.1200/JCO.20.00793

10. Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: inevitable or inadvisable? *J Clin Oncol Off J Am Soc Clin Oncol* (2010) 28(15):2641–7. doi: 10.1200/JCO.2009.26.3343

11. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *J Clin Oncol Off J Am Soc Clin Oncol* (2001) 19(1):265–72. doi: 10.1200/JCO.2001.19.1.265

12. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomisation. *Eur J Cancer* (2009) 45(2):275–80. doi: 10.1016/j.ejca.2008.10.029

13. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol Off J Am Soc Clin Oncol* (2005) 23(28):7199–206. doi: 10.1200/JCO.2005.01.149

14. Dmitrienko A, Wang MD. Bayesian Predictive approach to interim monitoring in clinical trials. *Stat Med* (2006) 25(13):2178–95. doi: 10.1002/sim.2204

15. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin trials (London England)* (2008) 5(2):93–106. doi: 10.1177/1740774508089279

16. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin trials (London England)* (2014) 11(4):485–93. doi: 10.1177/1740774514531352

17. Hobbs BP, Chen N, Lee JJ. Controlled multi-arm platform design using predictive probability. *Stat Methods Med Res* (2018) 27(1):65–78. doi: 10.1177/0962280215620696

18. Zabor EC, Kaizer AM, Garrett-Mayer E, Hobbs BP. Optimal sequential predictive probability designs for early-phase oncology expansion cohorts. *JCO Precis Oncol* (2022) 6:e2100390. doi: 10.1200/PO.21.00390

19. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (2022).

20. Zabor EC, Hobbs BP, Kane MJ. *Ppseq: Design clinical trials using sequential predictive probability monitoring*. R package (2022). Available at: https://CRAN.R-project.org/package=ppseq

frontiers | Frontiers in Molecular Biosciences

TYPE Methods
PUBLISHED 23 December 2022
DOI 10.3389/fmolb.2022.1056461

Check for updates

OPEN ACCESS

EDITED BY
Guido Tiana,
University of Milan, Italy

REVIEWED BY
Mathias Luidor Heltberg,
University of Copenhagen, Denmark
Nenad Filipovic,
University of Kragujevac, Serbia

*CORRESPONDENCE
Trachette L. Jackson,
tjacks@umich.edu

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 30 September 2022
ACCEPTED 09 December 2022
PUBLISHED 23 December 2022

CITATION
Jain HV, Norton K-A, Prado BB and Jackson TL (2022), SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth. *Front. Mol. Biosci.* 9:1056461. doi: 10.3389/fmolb.2022.1056461

COPYRIGHT
© 2022 Jain, Norton, Prado and Jackson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

SMoRe ParS: A novel methodology for bridging modeling modalities and experimental data applied to 3D vascular tumor growth

Harsh Vardhan Jain[1], Kerri-Ann Norton[2], Bernardo Bianco Prado[3] and Trachette L. Jackson[3]*

[1]Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN, United States, [2]Reem and Kayden Center for Science and Computation, Computational Biology Laboratory, Computer Science Program, Bard College, Annandale-on-Hudson, NY, United States, [3]Department of Mathematics, University of Michigan, Ann Arbor, MI, United States

Multiscale systems biology is having an increasingly powerful impact on our understanding of the interconnected molecular, cellular, and microenvironmental drivers of tumor growth and the effects of novel drugs and drug combinations for cancer therapy. Agent-based models (ABMs) that treat cells as autonomous decision-makers, each with their own intrinsic characteristics, are a natural platform for capturing intratumoral heterogeneity. Agent-based models are also useful for integrating the multiple time and spatial scales associated with vascular tumor growth and response to treatment. Despite all their benefits, the computational costs of solving agent-based models escalate and become prohibitive when simulating millions of cells, making parameter exploration and model parameterization from experimental data very challenging. Moreover, such data are typically limited, coarse-grained and may lack any spatial resolution, compounding these challenges. We address these issues by developing a first-of-its-kind method that leverages explicitly formulated surrogate models (SMs) to bridge the current computational divide between agent-based models and experimental data. In our approach, Surrogate Modeling for Reconstructing Parameter Surfaces (SMoRe ParS), we quantify the uncertainty in the relationship between agent-based model inputs and surrogate model parameters, and between surrogate model parameters and experimental data. In this way, surrogate model parameters serve as intermediaries between agent-based model input and data, making it possible to use them for calibration and uncertainty quantification of agent-based model parameters that map directly onto an experimental data set. We illustrate the functionality and novelty of Surrogate Modeling for Reconstructing Parameter Surfaces by applying it to an agent-based model of 3D vascular tumor growth, and experimental data in the form of tumor volume time-courses. Our method is broadly applicable to situations where preserving underlying mechanistic information is of interest, and where computational complexity and sparse, noisy calibration data hinder model parameterization.

KEYWORDS

agent-based model, parameter identifiability, surrogate model, uncertainty quantification, vascular tumor growth

1 Introduction

Validated mathematical models of tumor growth mediated by complex microenvironmental interactions and signals are increasingly being recognized as an invaluable aid for elucidating mechanisms underpinning experimental and clinical observations (Byrne, 2010; Franssen et al., 2019; Butner et al., 2020; Butner et al., 2021). These models often use continuum ordinary or partial differential equations (ODEs/PDEs) to predict cancer cell number (or densities) in time and/or space. Continuum approaches are a common choice because they allow for rapid simulation and open the door to advanced analyses (global sensitivity, structural and practical identifiability, bifurcations, etc.) that reveal key parameter relationships. They also enable the use of time-course experimental data for parameter estimation and model validation (Brouwer et al., 2017; Eisenberg and Jain, 2017).

An alternative approach is a discretized method that models cells as autonomous, decision making "agents," each with their own set of properties and behaviors. These agent-based models (ABMs) have become a valuable tool in translational systems oncology, which has goals of predicting the effects of novel drugs and drug combinations on difficult-to-treat tumors (Altrock et al., 2015; Wang et al., 2015; Bergman et al., 2022). ABMs provide a logical structure for capturing the multiple time and spatial scales associated with cancer growth and progression because they allow for the characterization of tumor heterogeneity at an individual cell level that better reflects the complexity seen *in vivo* (Bergman et al., 2022). One major advantage of ABMs over traditional continuum ODE/PDE models is that they can generate realistic 3-dimensional virtual tumors that current state-of-art imaging technologies cannot infer from patient scans [for a discussion on limitations of imaging in cancer, see for instance (Bogdanovic et al., 2021; Ding et al., 2021; Martinez-Heras et al., 2021)]. However, to make useful, reliable quantitative predictions, ABMs need to relate to real-world data through model parameterization and calibration (Byrne, 2010; Eisenberg and Jain, 2017). Unfortunately, a significant limitation of these models is that they can be computationally expensive, especially as the number of agents (cells) expands. Computational times and memory requirements can become prohibitive when simulating upwards of 10^6–10^7 agents (Ghaffarizadeh et al., 2018). This is in direct opposition to the fact that just one cubic centimeter of tissue will contain 10^8–10^9 cells and many *in vivo* experiments begin with 10^4–10^6 cells (Del Monte, 2009). These computational costs are exacerbated when ABMs include molecular level details of cell signaling or targeted therapeutics Ghaffarizadeh et al. (2018). The inherent stochasticity and heavy computational requirements of an ABM are significant obstacles for data-driven parameterization and for conducting rigorous parameter space exploration and sensitivity analyses (Norton and Popel, 2016; Zhang et al., 2020; Broniec et al., 2021). Moreover, experimental data is typically limited, coarse-grained and may lack any spatial resolution, resulting in issues of parameter identifiability (Eisenberg and Jain, 2017).

There is hence a need for developing new theoretical and computational frameworks that can bridge this gap between ABM parameters and real-world data. Estimating ABM parameters from noisy experimental data is particularly challenging because ABM behavior emerges from interactions among many individuals and the computational expense scales with the number of parameters (Broniec et al., 2021). One approach for exploring ABMs is to run extensive Monte Carlo simulations, but this is infeasible for complex models (Nardini et al., 2021). Bayesian methods are not ideal because they rely on prior knowledge about the probability distributions of the components being modeled, which is rarely available (Broniec et al., 2021). Some researchers have used genetic algorithms (GA) together with agent-based models for parameter space exploration and parameter estimation (Calvez and Hutzler, 2005; Lee et al., 2015); however, GAs require a very large number of iterations to converge, thus exacerbating computational expense issues (Broniec et al., 2021). Yet another approach entails the derivation of coarse-grained ODE/PDEs (mean-field models) to predict average outputs of the ABM. However, such mean-field models typically fail to accurately describe ABM dynamics in certain parameter regimes (Klank et al., 2018; Nardini et al., 2021).

To address some of these challenges, we develop an approach that uses an explicitly formulated surrogate model (SM) that will bridge ABM simulations and experimental data. Surrogate models (also called metamodels or response surfaces) are computationally cheaper models designed to approximate the dominant features of a complex model, here, the ABM (Blanning, 1975; Regis and Shoemaker, 2005; O'Hagan, 2006; Asher et al., 2015). They have been used extensively in engineering applications (see (Palar et al., 2019) for a review) and weather forecasting [see (Vlahogianni, 2015; Schultz et al., 2021) for recent reviews]. Specifically, we employ model selection to infer an SM directly from both ABM output and experimental data so that we accurately capture aggregate ABM dynamics. In our approach, Surrogate Modeling for Reconstructing Parameter Surfaces (SMoRe ParS), we quantify the relationship between parameter values across the two types of models (ABM and SM)

and between SM parameters and experimental data. Thus, SM parameters act as interlocutors between ABM inputs and data that can be used for calibration and uncertainty quantification of ABM parameters.

Indeed, parameterizing ABMs with SMs that use machine learning algorithms, where the SM does not have a closed form, is becoming increasingly popular. Using examples from finance, (Lamperti et al., 2018; Zhang et al., 2020) describe a surrogate modeling method for ABM calibration that combines supervised machine-learning and iterative sampling. These methods can learn a surrogate model as the approximation of the original system with a relatively small number of training points by using an iterative sampling algorithm that intelligently searches the response surface. In (Perumal and van Zyl, 2020), different sampling methods and SMs derived from machine learning algorithms are integrated with a temporal ABM that describes infectious disease epidemiology to test how these strategies affect parameter space exploration. They show that surrogate assisted methods perform better than standard sampling methods in that they better identify the most likely parameter vector by matching the synthetic data distribution it generates with a real data distribution.

Our method differs from the approaches mentioned above in several ways. Two major differentiators of our approach are: 1) The SM equations are explicitly formulated, this formulation being informed by the experimental data; and 2) SM parameters are distinctly mapped to both, the ABM the input parameters, as well as the calibration data set. In this way, our strategy enables the SM to be informed by both the ABM output, and the experimental data. We also infer ABM parameter regions that correspond to the data and propagate uncertainty *via* SM parameters to ABM parameters. Finally, by making such an explicit connection between ABM input, SM parameters and the data, we can account for inherent differences in dimensionality or physical units between ABM output and experimental data.

In the sections below we describe the details of our new methods for narrowing the current divide between computationally intense, difficult to analyze/parameterize computational modeling approaches and experimental data. We then demonstrate the usefulness and novelty of our approach by applying it to an ABM of vascular tumor growth and experimental data in the form of tumor volume time-courses.

2 Methods

2.1 Surrogate modeling for reconstructing parameter surfaces (SMoRe ParS)

To accurately compare ABM output with real world data, we propose our novel methodology, SMoRe ParS. A schematic diagram of the full approach is provided in Figure 1. SMoRe ParS is a six-step strategy that users can implement as follows.

Step 1: Use real-world data to inform SM formulation and variables

First, determine the formulation of the SM from a real-world (experimental) data set. In particular, the goal is to determine both, the type of model to use (ODE, PDE, Boolean, etc.), and the variables needed for the model formation. For instance, time-course data would suggest a system of ODEs, whereas spatially resolved data might accommodate a PDE SM. Additionally, the quantities measured in the data set should inform the choice of SM variables. For instance, tumor volume measurements would suggest tumor cell numbers as a SM variable.

Step 2: Generate ABM data

In this step, identify a subset of ABM parameters of interest, say $\vec{p}_{\mathrm{ABM}}$, based on some predetermined criteria. For instance, in a model of chemotherapy, one might select parameters such as cancer cell proliferation rate and death rate, that is the input parameters that are directly relevant to the treatment of interest. Next, generate ABM output for a broad range of the chosen parameter values. Specifically, vary ABM parameters one at a time to sample along the boundary of the parameter space, and also select several parameter combinations at non-boundary points, to generate reference points in the interior. For each parameter combination, the ABM should be simulated multiple times to get meaningful average behavior. Finally, process the generated ABM output for inherent differences in dimensionality or physical units between ABM output and SM variables, if necessary. For instance, if the ABM output is a spatially resolved time-course of a growing tumor and a variable in the SM is total number of tumor cells as a function of time alone, then the number of tumor cells in the ABM simulations should be integrated over its spatial domain.

Step 3: Perform SM model selection

Select several potential models as SM candidates and test their ability to capture both the experimental data and the ABM output. Then perform model selection to arrive at a "most likely" SM. There are numerous model selection approaches to choose from when selecting the best model to move forward with, including probabilistic Information Criteria (Anderson and Burnham, 2004; Burnham and Anderson, 2004) or resampling methods (Efron, 1983; Shao, 1996). Others (Nardini et al., 2021) have proposed learning equations directly from data as a method to arrive at a consensus model.

Step 4: Reconstruct SM parameter surfaces from ABM output

Next, infer a quantitative relationship between each of the SM input parameters, $\vec{p}_{\mathrm{SM}} = \langle p_{\mathrm{SM},1}, \ldots, p_{\mathrm{SM},i}, \ldots, p_{\mathrm{SM},n} \rangle$ and selected ABM parameters, $\vec{p}_{\mathrm{ABM}}$. This is done by fitting SM parameters to ABM output generated in Step 2, for instance by performing maximum likelihood estimation (MLE) (Millar, 2011). A key advantage of our method is that any uncertainty in SM parameters is also quantified in this step. For example, if

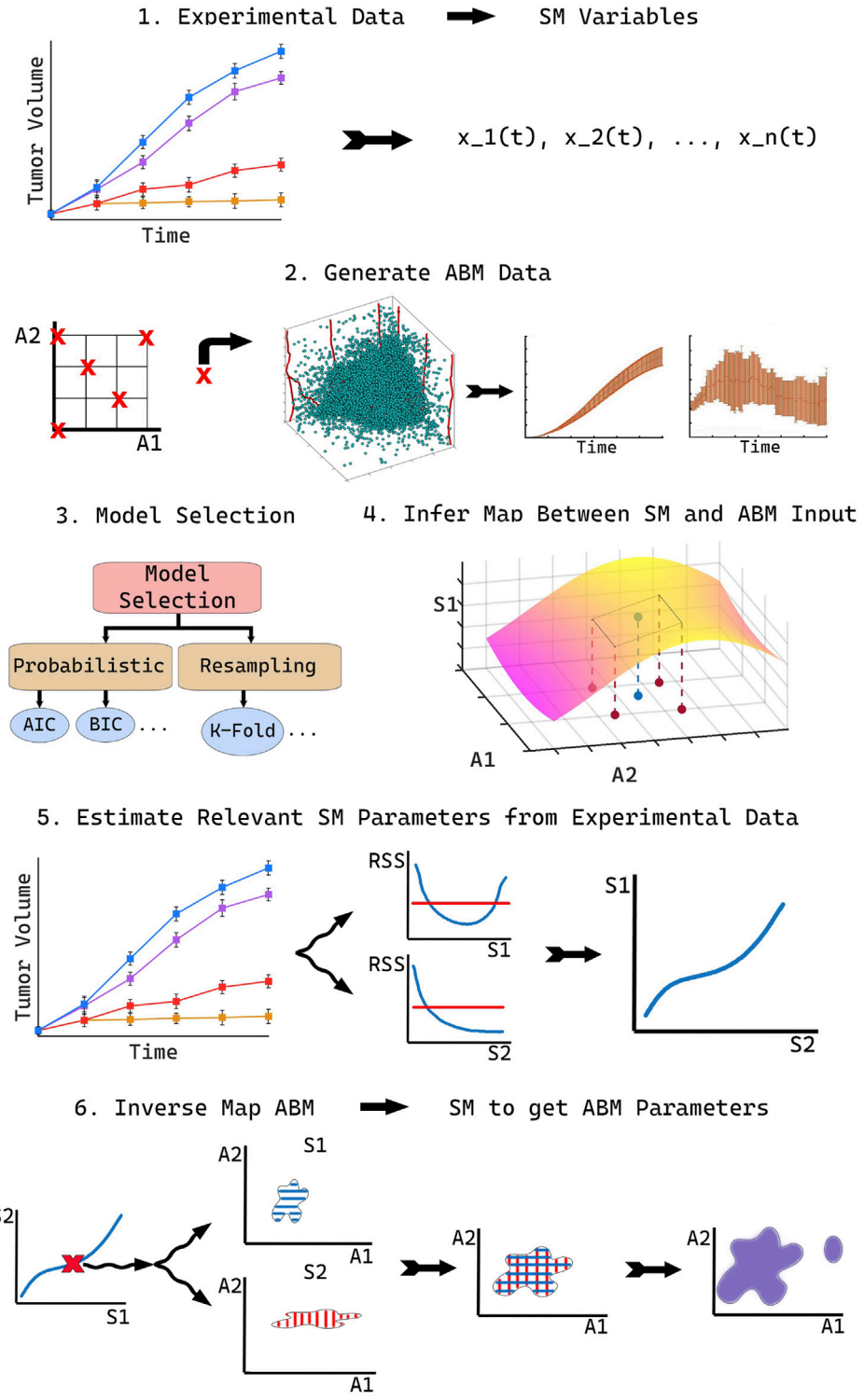


FIGURE 1
Schematic for implementing the SMoRe ParS method.

MLE is used to estimate SM parameter values, then the profile likelihood approach Eisenberg and Jain (2017) can be employed to quantify this uncertainty.

At this stage, for a given SM parameter $p_{SM,i}$, estimates for its appropriate range of values (e.g., 95% confidence bounds) should be calculated at each of the sampled ABM parameter combinations. Assuming that $p_{SM,i}$ and its confidence bounds (C.B.$_i$) are continuous but unknown functions of the ABM parameters $\vec{p}_{ABM}$, reconstruct these functions—or hypersurfaces—as follows. The 95% confidence bound estimates found above correspond to discrete points on the upper and lower 95% confidence hypersurfaces (see Step 4 in Figure 1). Now, "fill in" the unknown upper and lower hypersurfaces, for instance, using polynomial or quadratic interpolation [see (Smith, 2013) for an overview of these methods]. That is, reconstruct parameter response surfaces that $p_{SM,i}$ lies within. The completion of this step will result in an explicit (numerical) relationship between SM parameters and ABM parameters, that also preserves information on uncertainty in the SM parameters. That is:

$$p_{SM,i} = f_i\left(\vec{p}_{ABM}\right) \pm \text{C.B.}_i\left(\vec{p}_{ABM}\right). \quad (1)$$

In the above hypersurface relationship, the function f_i is not explicitly determined, rather, it is numerically approximated.

Step 5: Estimate SM parameters from real-world data

In the next step, estimate values of SM input parameters $\vec{p}_{SM}$ from the real-world data, for instance by performing maximum likelihood estimation (MLE) as in the previous step. Ideally at this stage, practical identifiability of the SM model should be investigated to arrive at identifiable combinations of SM input parameters. Practical identifiability examines how real-world considerations, such as noise or sampling frequency, affect one's ability to uniquely estimate model parameters from a given data set (Jacquez and Greif, 1985). This additional step of finding the practically identifiable combinations of SM parameters will help constrain the desired ABM parameter space that maps to real-world data in the next step.

Step 6: Infer regions of ABM parameter space that correspond to real-world data

In the final step of SMoRe ParS, overlay the ranges on data-derived SM parameters in the previous step on the inferred relationship between SM parameters and ABM parameters found in Step 4. This yields regions of ABM parameter space that correspond to experimental data. Specifically, for each data-informed choice of SM parameter vector $\vec{p}_{SM}$, regions in ABM parameter-hyperspace are obtained *via* projection mapping for all its components, $p_{SM,i}$. The intersection of these regions yields ABM parameter ranges that correspond to that specific choice of $\vec{p}_{SM}$. Repeat this for several choices of $\vec{p}_{SM}$—constrained by the practical identifiability information from Step 5—and take the union of the resulting ABM regions to arrive at the desired ranges in parameter values that match with the experimental data.

2.2 Proof of concept: Vascular tumor growth

In this section we work through the set up of a detailed proof of concept of our new method. To demonstrate the functionality and originality of our approach, we apply it to a 3D, multi-grid, on-lattice ABM of stem-cell driven vascular tumor growth.

2.2.1 SM formulation and variables

We use experimental data from the breast cancer literature in the form of tumor volume time-courses [see Figure 5A in (Zhou et al., 2019)]. These data suggest that a single ODE tracking the number of tumor cells over time is an appropriate formulation for the SM.

2.2.2 ABM formulation

The ABM of vascular tumor growth implemented here is a simplified form of previous models (Norton et al., 2018; Ventoso and Norton, 2020; Fischel et al., 2021). The simplified version consists of two modules: a tumor module and a vasculature module. Both modules are on-lattice, but they occupy different grids. The overall simulation domain is a cube of side 1 mm. As in (Norton et al., 2018), the tumor is initiated with cancer cells, progenitor and stem, placed in one corner of the grid. Cells cannot leave the boundary of the simulation domain. The vascular network at tumor initiation consists of mature vessels, each comprised of individual segments located along the grid boundaries closest to the initial tumor. This initial set up is visualized in Figure 2. The ABM simulates a tumor growing on the surface of healthy, vascularized tissue, which acts as an additional source of oxygen. The simulation is run for 300 iterations, each iteration corresponding to ~ 6 h. For more information on model assumptions, setup and simulation methodology, we refer the reader to (Norton et al., 2017; Ventoso and Norton, 2020; Fischel et al., 2021). A list of parameter values used for baseline simulations of the ABM is provided in Table 1 and an algorithm for simulating the ABM is outlined in the Appendix and Figure 9 (adapted from Ghaffarizadeh et al., 2018).

2.2.2.1 Tumor module

The tumor module resides on a 50 × 50 × 50 lattice, in which each cancer cell can only occupy one lattice point. The cancer cells have two proliferative phenotypes: stem cells and progenitor cells, and two migratory phenotypes: high and low migration. Proliferating stem cells have a certain probability of division that remains fixed at predetermined values throughout our simulations. Cancer stem cells are also assumed to have limitless replicative potential (Hanahan and Weinberg, 2000). Progenitor cell behavior is determined by two main input parameters: p_{div}, the division

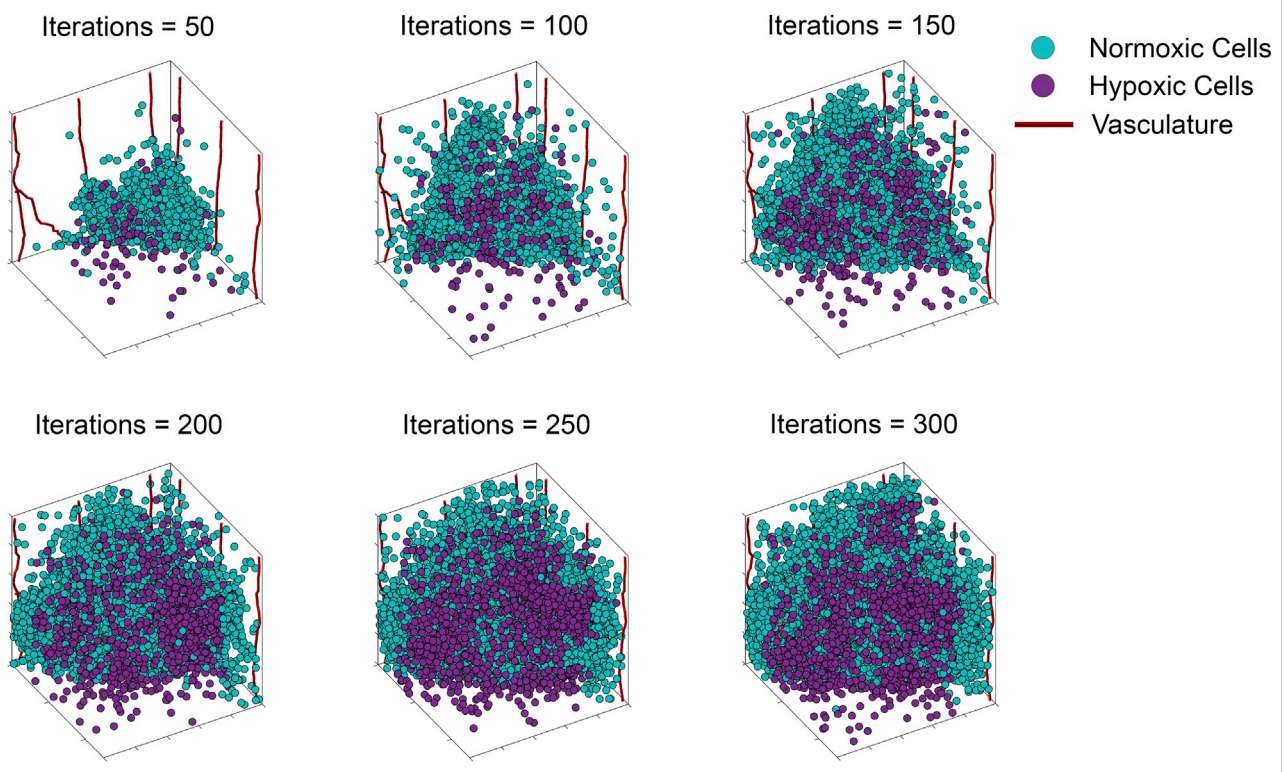


FIGURE 2
Storyboard showing a typical ABM simulation of vascular tumor growth, depicting the locations in space of normoxic (teal circles) and hypoxic (purple circles) tumor cells, along with vasculature (red curves).

TABLE 1 Baseline parameter values used for ABM simulations. For a complete list, see (Norton et al., 2017).

Parameter description	Parameter value	Source
Progenitor cell division limit (div_{lim})	8–15	See text
Progenitor division probability (p_{div})	0.05–0.245 per iteration	See text
Stem cell division probability	0.05 per iteration	Norton et al. (2017)
Stem cell symmetric division probability	0.05	Norton et al. (2017)
Initial number of mature vessels	8	Ghaffarizadeh et al. (2018)
Initial number of stem cells	20	Norton et al. (2017)
Initial number of progenitor cells	80	Norton et al. (2017)
High migration rate	8.3 μm per hour	Norton et al. (2017)
Low migration rate	0.83 μm per hour	Norton et al. (2017)
Probability of daughter cell to be a high migratory cell	5%	Norton et al. (2017)
Maximum vessel branching probability	0.2 per iteration	Norton and Popel (2016)
Senescent cell death probability	0.1 per iteration	Norton et al. (2017)

probability of the cell; and div_{lim}, the number of times a cell can divide before becoming senescent. A stem cell proliferates less often than a progenitor cell and can divide symmetrically into two stem cells or asymmetrically into a stem cell and a progenitor cell. Progenitor cells can only divide symmetrically into two progenitor cells. Each daughter cell, whether stem or progenitor, has a certain fixed probability of being a high migratory cell.

2.2.2.2 Vasculature module

The vasculature module resides on a 500 × 500 × 500 lattice, which is 10-fold finer than the tumor module lattice because microvessel diameter is typically smaller than the size of a tumor cell (Tsuji et al., 2002; Hao et al., 2018). The initial vasculature is made up of mature segments which are oxygenated. In each simulation step, a new branch or sprout can form at a random location along a mature segment with a certain probability, if there is a hypoxic tumor cell within a certain distance of the vessel. The sprout's movement is dictated by a tip cell and its length, by proliferating stalk cells. Tip cells migrate towards the nearest source of vascular endothelial growth factor (VEGF) (Gerhardt et al., 2003), which in our model are breast cancer cells (Linderholm et al., 2009). A sprout can fuse with another sprout if the two tips cells are close to one another, or with a stalk cell if the tip cell is close enough to it, through a process called anastomosis. Blood can only flow in new vasculature when such loops are completed (Chaplain et al., 2006). Blood-bearing vessels release oxygen and thus govern normoxic and hypoxic regions within the tumor. Cancer cells in hypoxic regions have a reduced proliferation probability and an increased migration rate (Lin et al., 2012).

2.2.2.3 ABM parameters of interest

Although the ABM has a number of input parameters, those governing progenitor cell proliferation emerge as a natural choice for several reasons. The experimental data comprises tumor volume time-courses, and the bulk of a growing tumor is due to non-stem cancer cells (Morton et al., 2011). Further, since this is a proof of concept study, we wanted to minimize the degrees of freedom, and emphasize ease of visualization. We therefore select p_{div} and div_{lim}, defined earlier, as ABM parameters of interest.

2.2.2.4 ABM output

From each ABM simulation run at specific values of p_{div} and div_{lim}, we record the number of cancer cells, the number of hypoxic cells, the number of stem cells, and the number of cell divisions, at each iteration. We also collect additional information at the final iteration of the simulation including the locations of all cancer cells and the location of the vasculature within their respective grids. Values of ABM parameters at which we generate output are all possible pairwise combinations from p_{div} = {0.05, 0.125, 0.245} and div_{lim} = {8, 12, 15}.

2.2.3 Model selection

We consider three classical models of tumor growth as candidates for our SM, namely, generalized Gompertz, generalized logistic, and Von Bertalanffy. The equations for each of these models can be found in Table 2. To assess goodness of fit and parsimony for each of the models we use AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) (Burnham and Anderson, 2004). These are statistical techniques that involve a scoring method that uses the maximum of a log-likelihood function or the residual sum of squares (RSS) to choose the best among candidate models.

2.2.4 SM parameter surface reconstruction

For every sampled combination of ABM parameters p_{div} and div_{lim}, we fit the SM model parameters to ABM output by minimizing the weighted sum of squared residuals:

$$\chi^2\left(\vec{p}_{\text{SM}}\right) = \sum_i \left(\frac{z_i - y_i\left(\vec{p}_{\text{SM}}\right)}{\sigma_i} \right)^2, \tag{2}$$

where: z_i denotes averaged ABM output generated at time point i; σ_i, the corresponding standard error; and $y_i(\vec{p}_{\text{SM}})$, the SM output at time point i as predicted by parameters $\vec{p}_{\text{SM}}$. We then use the profile likelihood method outlined in (Eisenberg and Jain, 2017), which exploits uncertainty in data (here, ABM output) to infer information on estimated parameters. Specifically, each estimated SM parameter $p_{\text{SM},i}$ is "profiled" by fixing it across a range of values and the remaining parameters are estimated for each fixed value of $p_{\text{SM},i}$ (Venzon and Moolgavkar, 1988; Murphy and Van der Vaart, 2000). The maximum value of the likelihood function for each parameter value yields the likelihood profile for that parameter (Eisenberg and Jain, 2017). The likelihood profiles are also used to calculate confidence bounds based on a likelihood threshold. The parameter values at which the profile crosses the threshold (on either side of the optimal parameter value) define the confidence interval at a particular level of significance (Eisenberg and Jain, 2017), here taken to be 95%. Bilinear interpolation—followed by a coordinate transformation for ease of visualization—is used to infer upper and lower bounding hypersurfaces as functions of ABM parameters, for each SM parameter $p_{\text{SM},i}$.

2.2.5 Estimate SM parameters from experimental data

We now fit the SM model parameters to the xenograft time-course data in (Zhou et al., 2019) by once again minimizing a weighted sum of squared residuals as described in the previous step. Next, we repeat the profile likelihood method described above, but now, with the experimental data. We additionally uncover practically identifiable combinations of SM input parameters, following the approach outlined in (Eisenberg and Jain, 2017). This is done by fitting rational functions (for instance) to the

parameter relationships inferred from the profile likelihood graphs (Eisenberg and Hayashi, 2014).

2.2.6 Infer regions of ABM parameters space that correspond to experimental data

Lastly, the identifiable ranges for the data-derived SM parameters found in Step 5 are overlaid on the interpolated map between SM and ABM parameters generated in Step 4 giving us the specific regions of ABM parameter space that correlate with the experimental data. Specifically, for each of our chosen SM parameters, we generate regions in the p_{div}–div_{lim} (ABM) parameter-plane. The intersection of these regions yield ranges for p_{div} and div_{lim} that correspond to a specific choice of our SM parameters. We repeat this process for multiple choices of our SM parameters and take the union of the resulting ABM regions to arrive at the desired ranges for p_{div} and div_{lim} that match with the experimental data.

2.2.7 Applying knowledge gained from SMoRe ParS to compare inferred tumor characteristics

Two distinct sets of ABM parameters are chosen from the experimental data-informed region computed in the previous step, namely, $p_{div} = 0.18$, $div_{lim} = 9$ and $p_{div} = 0.24$, $div_{lim} = 11$. At each parameter combination, ABM simulations are repeated six times, and used to calculate several metrics to distinguish between the resulting virtual tumors: 1) The Euclidean distance of the farthest cancer cell from the tumor at initiation; 2) the fractal dimension of the tumor vasculature [using MATLAB Central File Exchange file boxcount from F. Moisy (Moisy, 2008)]; 3) the surface area to volume ratio of the tumor; and 4) the compactness of the tumor [using the formula $Comp = (Vol)^{1/3} * (36\pi)^{1/6} / \sqrt{SA}$ from (Limkin et al., 2019)]. Here, SA refers to the surface area of the tumor and Vol refers to the volume of the tumor, calculated as follows. We use the Matlab function alphaShape to find the volume and surface area that encloses all points at which tumor cells are located in the 3D simulation domain. To eliminate confounding effects from tumor cells that have migrated away from the primary tumor mass, any regions of tumor cells smaller than a cutoff threshold of pixel volume = 50 are first removed using the Matlab function RegionThreshold.

3 Results

3.1 ABM simulations

To illustrate our ABM of 3D vascular tumor growth, we select representative values of p_{div} and div_{lim} at which we generate ABM output. Figure 2 depicts the progression over time of the resultant tumor, showing normoxic (cyan) and hypoxic (purple) tumor cells. Starting from a few cells in the corner of the grid, the tumor expands within the simulation domain as cells proliferate and tumor vasculature evolves. Figure 3 shows the concomitant evolution of tumor vasculature.

3.2 Surrogate model selection

The candidate surrogate models are shown in Table 2 along with their information criterion (AIC/BIC) values

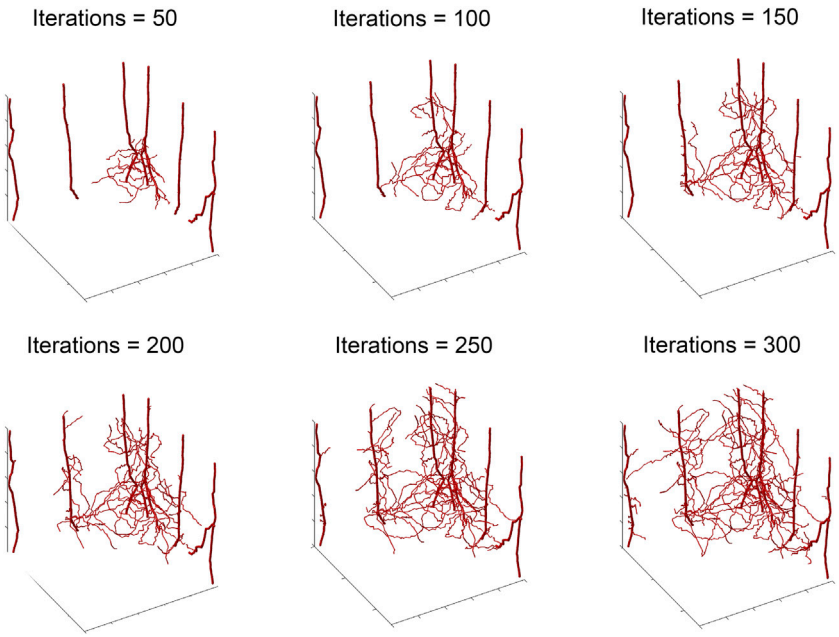


FIGURE 3
Storyboard showing vasculature evolution within the tumor shown in Figure 2.

TABLE 2 Information Criteria (AIC and BIC) for candidate surrogate models. Exp refers to experimental data.

Model	Equation	ABM		Exp	
		AIC	BIC	AIC	BIC
Generalized Gompertz	$\frac{dN}{dt} = N^{\lambda}(\delta - \gamma \ln N)$	45,696	61,629	5483.8	6793.1
Generalized Logistic	$\frac{dN}{dt} = \gamma N(1 - \frac{N^{\lambda}}{K})$	43,839	59,771	5483.3	6792.6
von Bertalanffy	$\frac{dN}{dt} = \alpha N^{\gamma} - \beta N$	41,947	57,880	5483.7	6793.0

associated with both ABM output and experimental data. These results show that experimental data alone may not distinguish between the models. However, when fitting to ABM output, the generalized Gompertz (GG) and Logisitic (GL) equations are e^{-1875} and e^{-942} times as probable as the von Bertalanffy (vB) model to minimize information loss, respectively. This means that GG and GL cannot explain the ABM data better than vB. Our results confirm that comparatively the vB growth model provides a better fit to the ABM data. Therefore we select the vB model as our surrogate. This agrees with findings in (Ghaffari Laleh et al., 2022) where these and other test models were fit to tumor volume time-courses from five different data sets.

The vB model has three input parameters (α, β, γ) of which α is related to the environmental carrying capacity. This differs significantly between the ABM (1 mm^3) and the experimental system (~2,000 mm^3). Since the two carrying capacities are uncorrelated, α cannot function as an interlocutor between

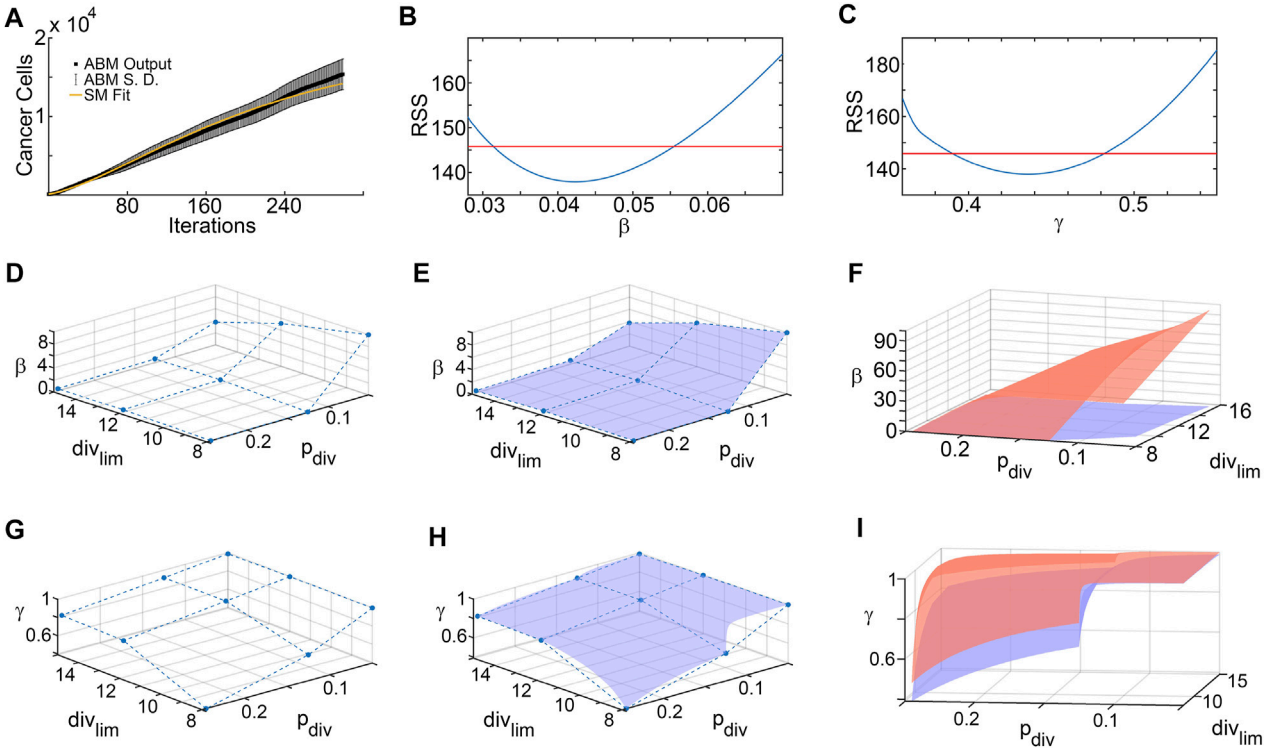


FIGURE 4
Surrogate model parameter surface reconstruction from ABM output. **(A)** Best fit of surrogate model to averaged ABM output generated with p_{div} = 0.245 and div_{lim} = 8. **(B,C)** Profile likelihoods (solid blue lines) for estimated surrogate model parameters: **(B)** β, and **(C)** γ. Thresholds for the 95% confidence intervals are shown as red lines and RSS stands for residual sum of squares. The left and right points of intersection of the blue and red curves give the lower and upper bounds, respectively, for the estimated surrogate model parameter, corresponding to these specific values of ABM parameters (p_{div} = 0.245 and div_{lim} = 8). **(D–F)** Lower and upper surface reconstruction for β. **(D)** Lower bounds for β obtained from 95% confidence thresholds like those shown in panel **(B)**, for various choice of ABM parameters p_{div} and div_{lim}. **(E)** Lower bound surface for β reconstructed from the discrete points shown in panel **(D)**. **(F)** Final lower (blue) and similarly reconstructed upper (orange) surfaces for β. **(G–I)** Lower and upper surface reconstruction for γ, following similar steps.

the ABM and the experimental data. Therefore, we select β and γ as our SM parameters of interest.

3.3 Parameter surface reconstruction

Figure 4 shows the results of the SM parameter surface reconstruction from ABM output. Figure 4A depicts the best fit of SM output (time-course of # of tumor cells) to ABM output (time-course of # of tumor cells integrated over space) for one specific combination of p_{div} and div_{lim}. The results of the profile likelihood analysis, quantifying uncertainty in SM parameters β and γ for this choice of p_{div} and div_{lim}, are shown in Figures 4B,C. Both parameters are identifiable from the ABM output, as evidenced by u-shaped profiles. The 95% confidence bounds for these fits correspond to discrete points on the upper and lower 95% confidence hypersurfaces in (p_{div}, div_{lim}, β) and (p_{div}, div_{lim}, γ) space. Repeating this for all sampled combinations of p_{div} and div_{lim} yields the sets of discrete points that lie on the upper and lower hypersurfaces of each SM parameter. As an illustration, Figures 4D,G show the discrete points on the lower hypersurfaces for β and γ, respectively. Next, the surfaces are "filled in" using interpolation, as shown in Figures 4E,H. Finally, Figures 4F,I show the fully reconstructed upper (orange) and lower (blue) hypersurfaces for β and γ, respectively. For this region of ABM parameter space, we are 95% confident that the SM parameters lie in between these hypersurfaces.

3.4 SM parameter estimation from experimental data

Figure 5A shows the results from fitting the SM parameters β and γ to the experimental data for breast cancer xenografts taken from (Zhou et al., 2019). From the subset profiles for each parameter graphed in Figures 5C,D, we see that both parameters are practically unidentifiable (or inestimatable) from the experimental data set. Although each parameter on its own is not estimable, the following practically identifiable combination is inferred from parameter relationships between β and γ, shown in Figure 5B:

$$\gamma = \frac{\beta + 0.0164}{\beta + 0.0392}. \tag{3}$$

We remark that the values for β and γ depicted in Figure 5B are from within their respective 95% confidence bounds inferred from Figures 5C,D.

3.5 Inferring regions of ABM parameter space that correspond to experimental data

For any pair of admissible values of β and γ as determined by Eq. 3, a corresponding region of ABM parameter space is inferred from Figure 4F for β, and Figure 4I for γ. The intersection of these two regions gives the region of ABM

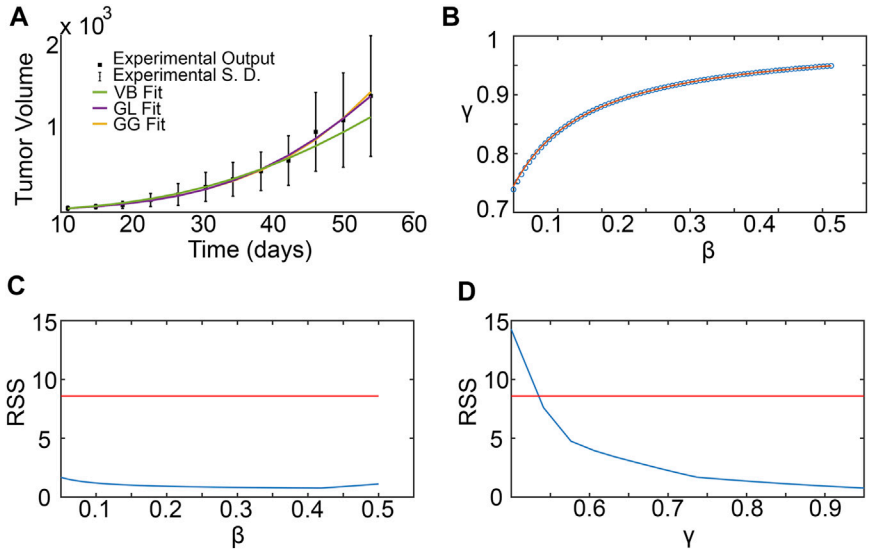


FIGURE 5
Surrogate model parameter estimation and practical identifiability analysis using breast cancer xenograft data from (Zhou et al., 2019). **(A)** Surrogate model fit to experimental tumor volume time-courses. **(B)** Inferred relationship between γ and β using the profile-likelihood method (Eisenberg and Jain, 2017), with combinations plotted as blue squares, and potential combination form plotted as a red curve. **(C,D)** Profile likelihoods (solid blue lines) for estimated surrogate model parameters: **(C)** β, and **(D)** γ. Thresholds for the 95% confidence intervals are shown as red lines and RSS stands for residual sum of squares.

10.3389/fmolb.2022.1056461

parameter space that maps to the experimental data for this specific β-γ combination. Figure 6 shows these inferred regions for three representative pairs of values of β and γ. Along each row, the first panel shows the ABM parameter region corresponding to that value of β, the second panel shows the ABM parameter region corresponding to that value of γ, and the third panel shows the intersection of these two regions. Finally, Figure 7 shows the union of several such common regions, for the range of possible values of β and γ. This is the desired region of ABM parameter space inferred from the experimental data.

3.6 Comparing metrics from the ABM parameter space

Two distinct sets of parameters from within the inferred ABM parameter space are chosen to illustrate how the same xenograft volume time-course may come from tumors with very different spatial microstructure. For each parameter set, ABM simulations are repeated six times, and the number of tumor cells, the number of hypoxic cells and the number of cancer cell divisions are recorded at each time step. Additionally, we also calculate the compactness of the tumor, the surface area to volume ratio of the tumor, the fractal

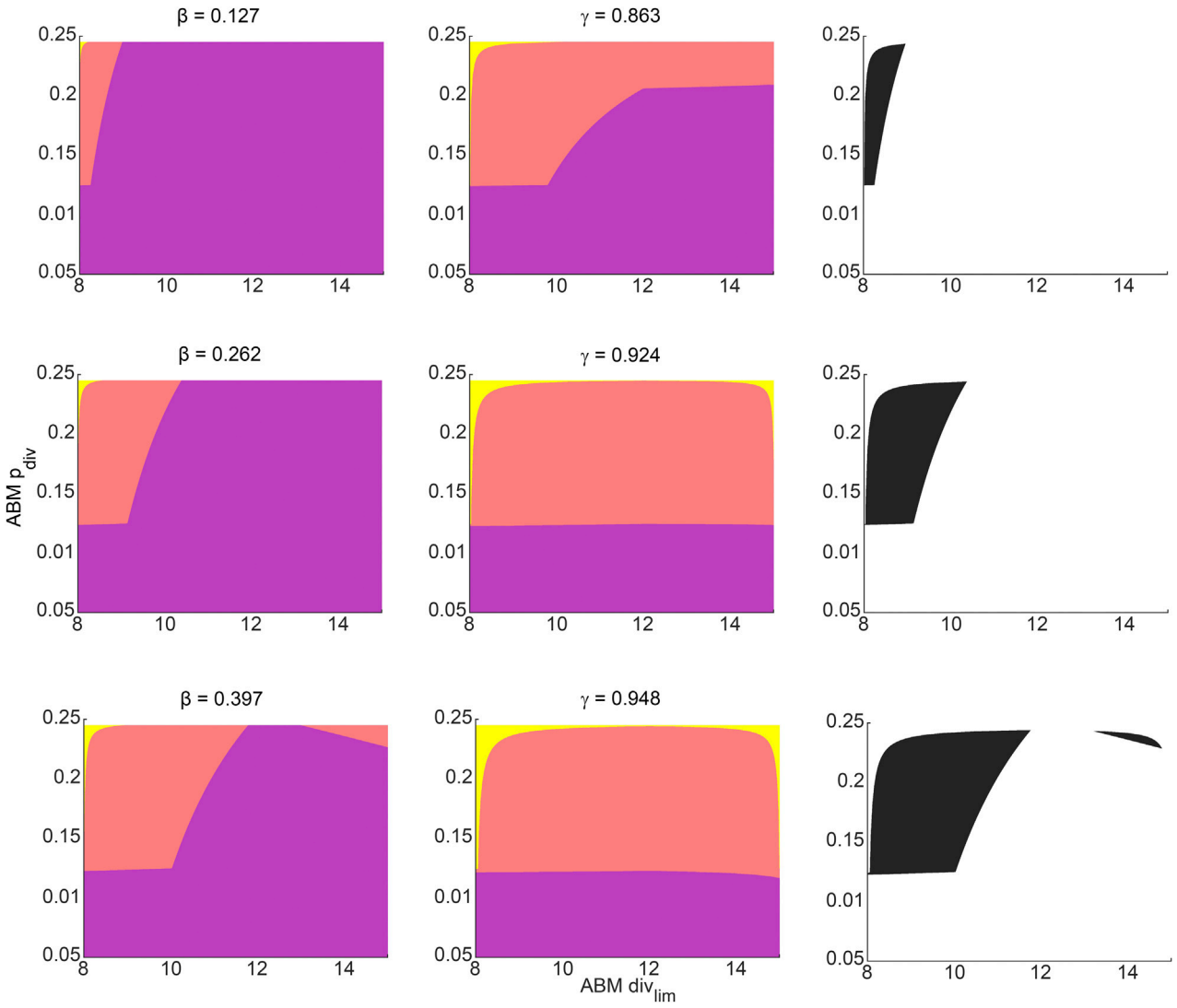


FIGURE 6
Regions in ABM parameter space corresponding to various choices of surrogate model parameter combinations that were inferred by fitting to experimental data. Orange tinted areas represent regions in ABM parameter space for which that surrogate model parameter lies between its upper and lower reconstructed surfaces. Purple and yellow tinted areas represent (inadmissible) regions when the surrogate model parameter is outside these bounds. The first and second columns represent ABM regions corresponding to various choices of β and γ, respectively. The third column graphs the intersection of the admissible ABM parameter regions, with each entry corresponding to the pair of β-γ values from that row. The β-γ pairs in each row are points that lie on the practically identifiable combination form plotted in Figure 5B.

10.3389/fmolb.2022.1056461

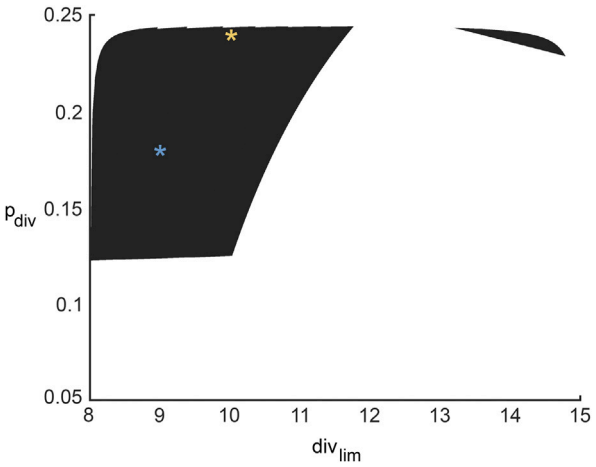


FIGURE 7
Region in ABM parameter space inferred from all possible surrogate model parameter combinations that fit experimental data equally well. The yellow and blue stars denote parameter sets 1 and 2 (p_{div} = 0.18, div_{lim} = 9 and p_{div} = 0.24, div_{lim} = 11; respectively, as discussed in Section 3.6 and Figure 8.

dimension of the 3D vasculature, and the distance of the farthest cancer cell from the original tumor, at the end of simulations (iterations = 300). Figure 8 shows how these features compare between the two sets of simulations.

As can be seen from Figure 8A, the mean number of cancer cells of parameter set 2 is consistently larger than parameter set 1, with a difference of about a thousand cells. Parameter set 1 has a relatively small variation between runs as compared to parameter space 2. Similarly, Figure 8B shows that the mean number of hypoxic cells is consistently larger for parameter set 2 than 1. On the other hand, although the number of cancer cell divisions is initially higher for parameter set 2, both sets of simulations stabilize at a similar number (Figure 8C). These findings are unsurprising, given that parameter set 2 allows for a higher probability of division as well as number of allowed divisions, than parameter set 1.

Interestingly, tumors generated from parameter set 2 are significantly more compact than those generated from parameter set 1 (p-value = .0357 using a two sample t-test, see Figure 8D). This makes sense as parameter set 2 has a larger division probability and cells can divide more times than parameter set 1. Therefore, they should generally reproduce more often and longer before they become senescent, creating a larger, more compact tumor. Although we do see that the variance for parameter space 2 is much larger than parameter space 1, meaning that while they tend to be more compact, there are also cases in which they are less compact, similarly to parameter space 1. In contrast, the surface area to volume ratio is significantly lower for tumors from parameter set 1 than parameter set 2, with a p-value of .0149 (see Figure 8E). The average fractal dimensions of the final tumor vasculature are

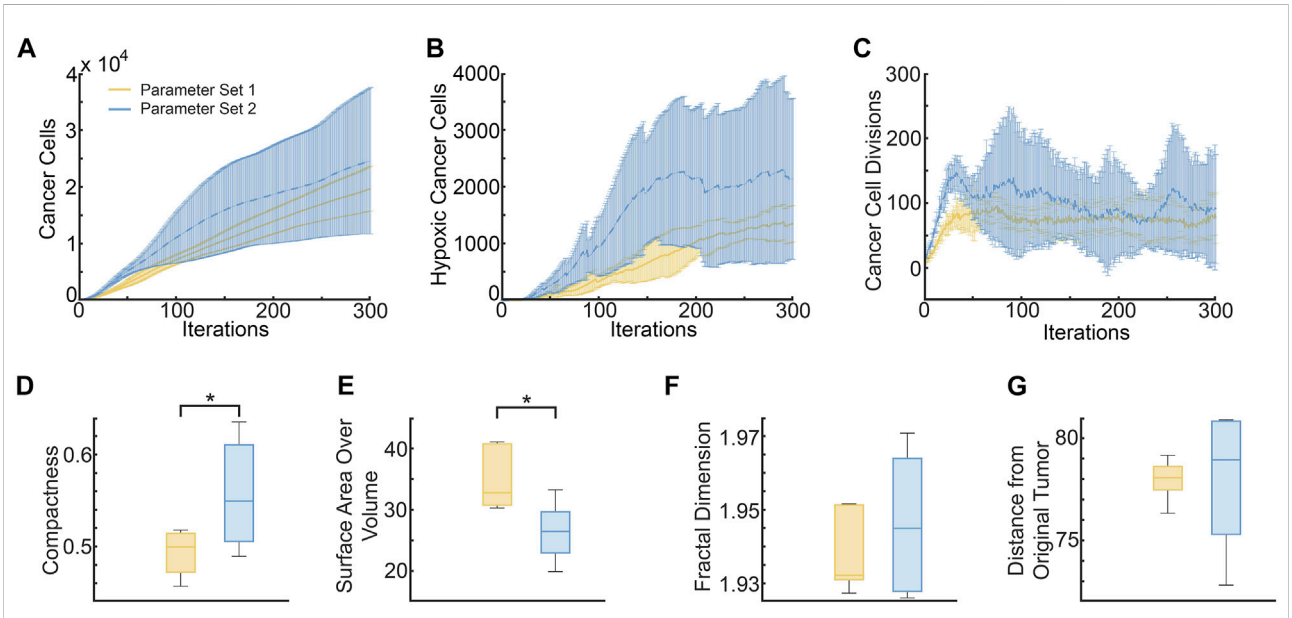


FIGURE 8
Metrics distinguishing ABM-simulated tumors using parameter set 1 (p_{div} = 0.18 and div_{lim} = 9, yellow curves and bars) and parameter set 2 (p_{div} = 0.24 and div_{lim} = 11, blue curves and bars). **(A)** Mean and standard deviation of total cancer numbers at each iteration. **(B)** Mean and standard deviation of total number of hypoxic cells at each iteration. **(C)** Mean and standard deviation of total number of cancer cell divisions at each iteration. **(D–G)** Metrics calculated at simulation end-point (iterations = 300), with statistically significant differences indicated. **(D)** Compactness of the simulated tumors. **(E)** Surface area-to-Volume ratio of the simulated tumors. **(F)** Fractal dimension of tumor vasculature. **(G)** Distance of the farthest cancer cell from the origin (location of tumor cells at iteration 0).

similar between the two parameter sets, with values within 1.93 and 1.97 (see Figure 8F). This is in line with experimental results that found vessels from whole tumor xenografts had fractal dimensions between 1.94 and 2.04 (Kim et al., 2012). Finally, the distance from the original tumor of the farthest cell at the end of the simulations, is also similar between the two parameter sets (see Figure 8G).

4 Discussion

There is an unmet need to develop new theoretical and computational frameworks that advance current efforts for making critical connections between computationally complex model (CCM) parameters and real-world data, which can be sparse and highly variable. To that end, we developed SMoRe ParS, which is a potentially paradigm-shifting method for parameter surface reconstruction that tackles this problem. SMoRe ParS envisages an explicitly formulated, data-informed, simpler, surrogate model (SM) as an intermediary that is used to quantify the uncertainty in the relationship between CCM inputs and SM parameters, and also between SM parameters and real-world data. SM parameters, thus, serve as a link between difficult-to-estimate CCM inputs and noisy data and enable calibration and uncertainty quantification of CCM parameters that map directly onto an experimental data set.

To illustrate the capability of SMoRe ParS to connect CCM output and real-world data, we applied it to an ABM of 3D vascular tumor growth as the CCM, and data from tumor xenograft growth experiments as real-world data. Our method allowed us to

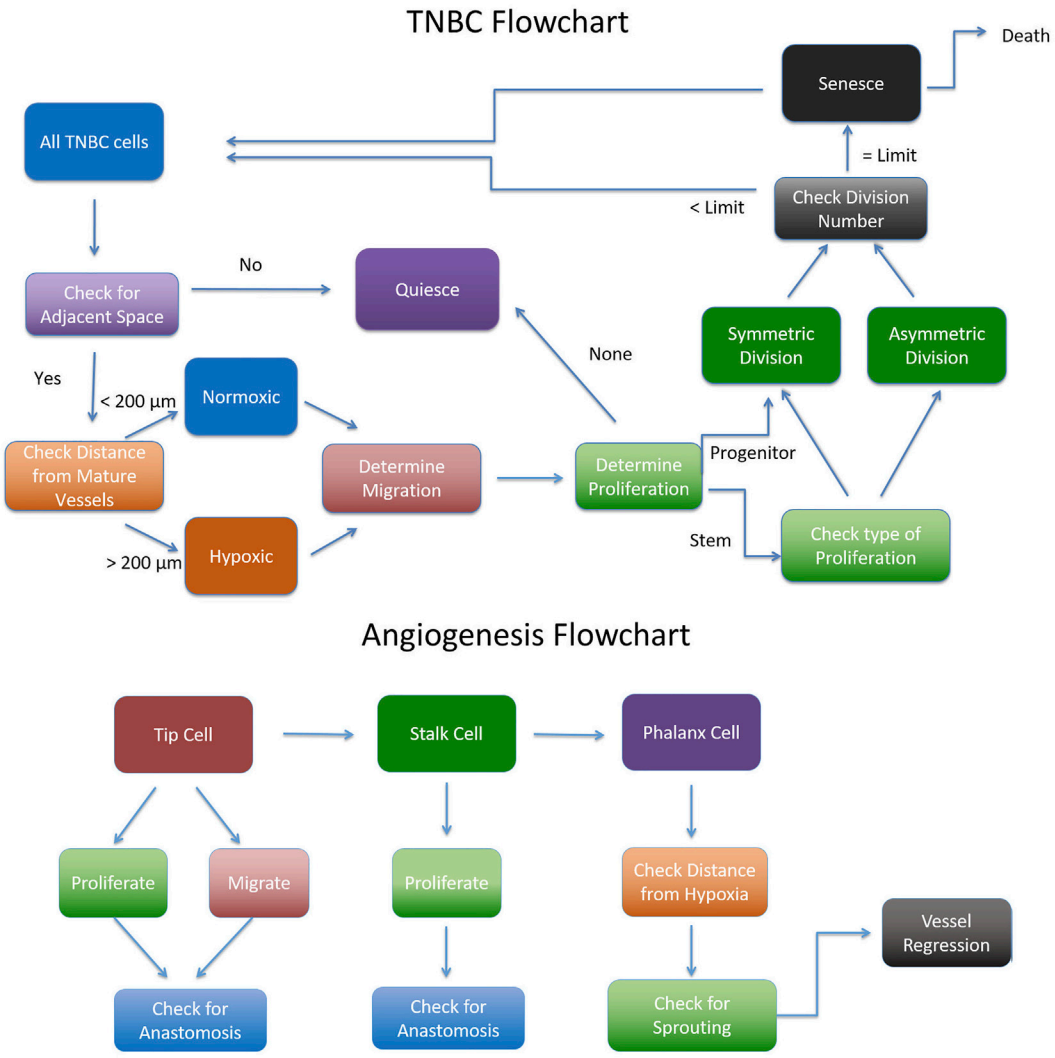


FIGURE 9
Flowchart for implementing the Agent Based Model adapted from (Norton, 2018).

construct an explicit mapping between ABM parameters and tumor volume time-courses, which encodes within it information on uncertainty in inferred parameter values. We then selected two distinct sets of ABM parameters that map onto the same data set, to investigate any differences between the resultant simulated tumors. Indeed, several trends distinguished the two sets of simulations. Variances in tumor cell number time-courses shown in Figure 8 suggest that parameter set 1 (lower probability of cell division and fewer number of allowed cell divisions) is consistent with a slow growing tumor, whereas parameter set 2 (higher probability of cell division and greater number of allowed cell divisions) allows for both slower and faster growing tumors. In fact, the variation within parameter set 1 was consistently smaller than within parameter set 2 across all metrics, except surface area to volume ratio (Figure 8). This suggests that while in parameter set 1 all tumors grew relatively similarly, in parameter set 2 the randomness of which cells could proliferate or move could lead to a substantial difference in the growth rate of the tumor. This is consistent with previous results that showed if cells are surrounded by other cells, even if their proliferation probability is high, they will not be able to divide because there is not enough space, thus limiting the overall growth of the tumor (Norton et al., 2017). Therefore, tumors that have the capability of excessive growth may not be able to do so under certain conditions where their growth is limited by spatial inhibition. This also explains the trends in compactness and surface area to volume ratio of the parameter sets. Tumors generated from parameter set 1 were less compact than those from parameter set 2, with a higher surface area-to-volume ratio, indicating more space to grow. Both these metrics have been suggested as predictors of malignancy in lung and head and neck cancers Aerts et al. (2014), Bogowicz et al. (2017), He et al. (2014), Wang et al. (2016). Our results suggest that tumors with very distinct malignant potential could be "hiding" within aggregate, macroscopic data.

In this paper, we chose to select the SM from a set of phenomenological models because our main concern was providing an easy to follow proof of concept example for cellular-level tumor growth. In other applications, where for example molecular or microenvironmental drivers of tumor progression and treatment response are of interest, it is possible to choose a mechanistic formulation of the surrogate. There are several advantages to doing so, including being able to more fully leverage the SM's ability to directly connect the ABM to the experimental data. A mechanistic SM would also have stand alone value as it provides a more complete characterization of the system and can be used for long term forecasting with greater confidence. We remark that in our approach, only a handful of ABM parameters can be considered at a time. In general, the precise number would depend on the computational complexity of the ABM and SM and how much experimental data is available. Further, deriving a mechanistic SM that can match both the experimental data and the ABM output may prove time-consuming. While we use cancer as an illustrative example and as the subject of our future studies, SMoRe ParS can easily be applied to a wide range of CCMs for basic biology and translational systems biology investigations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HJ, K-AN, and TJ conceived the original idea. HJ, K-AN, and TJ developed the mathematical methods. BP, HJ, and K-AN wrote the code, analyzed the output, and performed the numerical simulations. HJ, K-AN, BP, and TJ wrote the manuscript. HJ and K-AN contributed equally to this work. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by NIH/NCI U01CA243075 (TJ). Part of the research for this paper was carried out while the authors were in residence at the Mathematisches Forschungsinstitut Oberwolfach, supported by their "Research in Pairs" program.

Acknowledgments

We thank Marisa Eisenberg for valuable conversations and helpful suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* 5, 4006–4009. doi:10.1038/ncomms5006

Altrock, P. M., Liu, L. L., and Michor, F. (2015). The mathematics of cancer: Integrating quantitative models. *Nat. Rev. Cancer* 15, 730–745. doi:10.1038/nrc4029

Anderson, D., and Burnham, K. (2004). *Model selection and multi-model inference*. New York: Springer-Verlag.

Asher, M. J., Croke, B. F., Jakeman, A. J., and Peeters, L. J. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resour. Res.* 51, 5957–5973. doi:10.1002/2015wr016967

Bergman, D., Sweis, R. F., Pearson, A. T., Nazari, F., and Jackson, T. L. (2022). A global method for fast simulations of molecular dynamics in multiscale agent-based models of biological tissues. *iScience* 25, 104387. doi:10.1016/j.isci.2022.104387

Blanning, R. W. (1975). The construction and implementation of metamodels. *simulation* 24, 177–184. doi:10.1177/003754977502400606

Bogdanovic, B., Solari, E. L., Asiares, A. V., McIntosh, L., van Marwick, S., Schachoff, S., et al. (2021). Pet/mr technology: Advancement and challenges. *Seminars Nucl. Med.* 52, 340–355. doi:10.1053/j.semnuclmed.2021.11.014

Bogowicz, M., Riesterer, O., Ikenberg, K., Stieb, S., Moch, H., Studer, G., et al. (2017). Computed tomography radiomics predicts hpv status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. *Int. J. Radiat. Oncology*Biology*Physics* 99, 921–928. doi:10.1016/j.ijrobp.2017.06.002

Broniec, W., An, S., Rugaber, S., and Goel, A. K. (2021). "Guiding parameter estimation of agent-based modeling through knowledge-based function approximation," in *Proceedings of the AAAI 2021 spring symposium on combining machine learning and knowledge engineering (AAAI-MAKE 2021)*. Editors A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, et al. (Palo Alto, California: Stanford University).

Brouwer, A. F., Meza, R., and Eisenberg, M. C. (2017). A systematic approach to determining the identifiability of multistage carcinogenesis models. *Risk Anal.* 37, 1375–1387. doi:10.1111/risa.12684

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociol. Methods & Res.* 33, 261–304. doi:10.1177/0049124104268644

Butner, J. D., Elganainy, D., Wang, C. X., Wang, Z., Chen, S.-H., Esnaola, N. F., et al. (2020). Mathematical prediction of clinical outcomes in advanced cancer patients treated with checkpoint inhibitor immunotherapy. *Sci. Adv.* 6, eaay6298. doi:10.1126/sciadv.aay6298

Butner, J. D., Martin, G. V., Wang, Z., Corradetti, B., Ferrari, M., Esnaola, N., et al. (2021). Early prediction of clinical response to checkpoint inhibitor therapy in human solid tumors through mathematical modeling. *Elife* 10, e70130. doi:10.7554/eLife.70130

Byrne, H. M. (2010). Dissecting cancer through mathematics: From the cell to the animal model. *Nat. Rev. Cancer* 10, 221–230. doi:10.1038/nrc2808

Calvez, B., and Hutzler, G. (2005). "Automatic tuning of agent-based models using genetic algorithms," in *International workshop on multi-agent systems and agent-based simulation* (Cham: Springer), 41.

Chaplain, M. A., McDougall, S. R., and Anderson, A. (2006). Mathematical modeling of tumor-induced angiogenesis. *Annu. Rev. Biomed. Eng.* 8, 233–257. doi:10.1146/annurev.bioeng.8.061505.095807

Del Monte, U. (2009). Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle* 8, 505–506. doi:10.4161/cc.8.3.7608

Ding, H., Velasco, C., Ye, H., Lindner, T., Grech-Sollars, M., O'Callaghan, J., et al. (2021). Current applications and future development of magnetic resonance fingerprinting in diagnosis, characterization, and response monitoring in cancer. *Cancers* 13, 4742. doi:10.3390/cancers13194742

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* 78, 316–331. doi:10.1080/01621459.1983.10477973

Eisenberg, M. C., and Hayashi, M. A. (2014). Determining identifiable parameter combinations using subset profiling. *Math. Biosci.* 256, 116–126. doi:10.1016/j.mbs.2014.08.008

Eisenberg, M. C., and Jain, H. V. (2017). A confidence building exercise in data and identifiability: Modeling cancer chemotherapy as a case study. *J. Theor. Biol.* 431, 63–78. doi:10.1016/j.jtbi.2017.07.018

Fischel, H., Giorgadze, T., Tessier, A., and Norton, K.-A. (2021). "Computational modeling of chimeric antigen receptor (car) t-cell therapy of a binary model of antigen receptors in breast cancer," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), Houston, TX, USA, 09-12 December 2021, 3267–3274.

Franssen, L. C., Lorenzi, T., Burgess, A. E., and Chaplain, M. A. (2019). A mathematical framework for modelling the metastatic spread of cancer. *Bull. Math. Biol.* 81, 1965–2010. doi:10.1007/s11538-019-00597-x

Gerhardt, H., Golding, M., Fruttiger, M., Ruhrberg, C., Lundkvist, A., Abramsson, A., et al. (2003). Vegf guides angiogenic sprouting utilizing endothelial tip cell filopodia. *J. cell Biol.* 161, 1163–1177. doi:10.1083/jcb.200302047

Ghaffari Laleh, N., Loeffler, C. M. L., Grajek, J., Staňková, K., Pearson, A. T., Muti, H. S., et al. (2022). Classical mathematical models for prediction of response to chemotherapy and immunotherapy. *PLoS Comput. Biol.* 18, e1009822. doi:10.1371/journal.pcbi.1009822

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M., and Macklin, P. (2018). Physicell: An open source physics-based cell simulator for 3-d multicellular systems. *PLoS Comput. Biol.* 14, e1005991. doi:10.1371/journal.pcbi.1005991

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi:10.1016/s0092-8674(00)81683-9

Hao, S.-J., Wan, Y., Xia, Y.-Q., Zou, X., and Zheng, S.-Y. (2018). Size-based separation methods of circulating tumor cells. *Adv. Drug Deliv. Rev.* 125, 3–20. doi:10.1016/j.addr.2018.01.002

He, X., Sahiner, B., Gallas, B. D., Chen, W., and Petrick, N. (2014). Computerized characterization of lung nodule subtlety using thoracic ct images. *Phys. Med. Biol.* 59, 897–910. doi:10.1088/0031-9155/59/4/897

Jacquez, J. A., and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Math. Biosci.* 77, 201–227. doi:10.1016/0025-5564(85)90098-7

Kim, E., Stamatelos, S., Cebulla, J., Bhujwalla, Z. M., Popel, A. S., and Pathak, A. P. (2012). Multiscale imaging and computational modeling of blood flow in the tumor vasculature. *Ann. Biomed. Eng.* 40, 2425–2441. doi:10.1007/s10439-012-0585-5

Klank, R. L., Rosenfeld, S. S., and Odde, D. J. (2018). A brownian dynamics tumor progression simulator with application to glioblastoma. *Converg. Sci. Phys. Oncol.* 4, 015001. doi:10.1088/2057-1739/aa9e6e

Lamperti, F., Roventini, A., and Sani, A. (2018). Agent-based model calibration using machine learning surrogates. *J. Econ. Dyn. Control* 90, 366–389. doi:10.1016/j.jedc.2018.03.011

Lee, J.-S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., et al. (2015). The complexities of agent-based modeling output analysis. *J. Artif. Soc. Soc. Simul.* 18. doi:10.18564/jasss.2897

Limkin, E. J., Reuzé, S., Carré, A., Sun, R., Schernberg, A., Alexis, A., et al. (2019). The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. *Sci. Rep.* 9, 4329–4412. doi:10.1038/s41598-019-40437-5

Lin, S., Wan, S., Sun, L., Hu, J., Fang, D., Zhao, R., et al. (2012). Chemokine c-c motif receptor 5 and c-c motif ligand 5 promote cancer cell migration under hypoxia. *Cancer Sci.* 103, 904–912. doi:10.1111/j.1349-7006.2012.02259.x

Linderholm, B., Hellborg, H., Johansson, U., Elmberger, G., Skoog, L., Lehtiö, J., et al. (2009). Significantly higher levels of vascular endothelial growth factor (vegf) and shorter survival times for patients with primary operable triple-negative breast cancer. *Ann. Oncol.* 20, 1639–1646. doi:10.1093/annonc/mdp062

Martinez-Heras, E., Grussu, F., Prados, F., Solana, E., and Llufriu, S. (2021). Diffusion-weighted imaging: Recent advances and applications. *Seminars Ultrasound, CT MRI* 42, 490–506. doi:10.1053/j.sult.2021.07.006

Millar, R. B. (2011). *Maximum likelihood estimation and inference: With examples in R, SAS and ADMB*. New York, United States: John Wiley & Sons.

Moisy, F. (2008). *boxcount*. Matlab Central File Exchange.

Morton, C. I., Hlatky, L., Hahnfeldt, P., and Enderling, H. (2011). Non-stem cancer cell kinetics modulate solid tumor progression. *Theor. Biol. Med. Model.* 8, 48–13. doi:10.1186/1742-4682-8-48

Murphy, S. A., and Van der Vaart, A. W. (2000). On profile likelihood. *J. Am. Stat. Assoc.* 95, 449–465. doi:10.1080/01621459.2000.10474219

Nardini, J. T., Baker, R. E., Simpson, M. J., and Flores, K. B. (2021). Learning differential equation models from stochastic agent-based model simulations. *J. Roy. Soc. Interface* 18, 20200987. doi:10.1098/rsif.2020.0987

Norton, K.-A., Jin, K., and Popel, A. S. (2018). Modeling triple-negative breast cancer heterogeneity: Effects of stromal macrophages, fibroblasts and tumor vasculature. *J. Theor. Biol.* 452, 56–68. doi:10.1016/j.jtbi.2018.05.003

Norton, K.-A., and Popel, A. S. (2016). Effects of endothelial cell proliferation and migration rates in a computational model of sprouting angiogenesis. *Sci. Rep.* 6, 36992–37010. doi:10.1038/srep36992

Norton, K.-A., Wallace, T., Pandey, N. B., and Popel, A. S. (2017). An agent-based model of triple-negative breast cancer: The interplay between chemokine receptor ccr5 expression, cancer stem cells, and hypoxia. *BMC Syst. Biol.* 11, 68–15. doi:10.1186/s12918-017-0445-x

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Saf.* 91, 1290–1300. doi:10.1016/j.ress.2005.11.025

Palar, P. S., Liem, R. P., Zuhal, L. R., and Shimoyama, K. (2019). "On the use of surrogate models in engineering design optimization and exploration: The key issues," in Proceedings of the Genetic and Evolutionary Computation Conference Companion, Prague Czech, 13-17 July, 2019, 1592.

Perumal, R., and van Zyl, T. L. (2020). "Surrogate assisted methods for the parameterisation of agent-based models," in 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI) (IEEE), Stockholm, Sweden, 14-15 November 2020, 78–82.

Regis, R. G., and Shoemaker, C. A. (2005). Constrained global optimization of expensive black box functions using radial basis functions. *J. Glob. Optim.* 31, 153–171. doi:10.1007/s10898-004-0570-0

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., et al. (2021). Can deep learning beat numerical weather prediction? *Philosophical Trans. R. Soc. A* 379, 20200097. doi:10.1098/rsta.2020.0097

Shao, J. (1996). Bootstrap model selection. *J. Am. Stat. Assoc.* 91, 655–665. doi:10.1080/01621459.1996.10476934

Smith, R. C. (2013). *Uncertainty quantification: Theory, implementation, and applications, vol. 12*. SIAM.

Tsuji, T., Sasaki, Y., Tanaka, M., Hanabata, N., Hada, R., and Munakata, A. (2002). Microvessel morphology and vascular endothelial growth factor expression in human colonic carcinoma with or without metastasis. *Lab. Investig.* 82, 555–562. doi:10.1038/labinvest.3780450

Ventoso, M., and Norton, K.-A. (2020). "Simulating an immune response with a combined agent-based model of a triple-negative breast cancer tumor and vascular network," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 16-19 December 2020, 1303.

Venzon, D., and Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *J. R. Stat. Soc. Ser. C Appl. Statistics)* 37, 87–94. doi:10.2307/2347496

Vlahogianni, E. I. (2015). Optimization of traffic forecasting: Intelligent surrogate modeling. *Transp. Res. Part C Emerg. Technol.* 55, 14–23. doi:10.1016/j.trc.2015.03.016

Wang, J., Liu, X., Dong, D., Song, J., Xu, M., Zang, Y., et al. (2016). "Prediction of malignant and benign of lung tumor using a quantitative radiomic method," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), Orlando, FL, USA, 16-20 August 2016, 1272–1275.

Wang, Z., Butner, J. D., Kerketta, R., Cristini, V., and Deisboeck, T. S. (2015). Simulating cancer growth with multiscale agent-based modeling. *Semin. Cancer. Biol.* 30, 70–78. doi:10.1016/j.semcancer.2014.04.001

Zhang, Y., Li, Z., and Zhang, Y. (2020). Validation and calibration of an agent-based model: A surrogate approach. *Discrete Dyn. Nat. Soc.* 2020, 1–9. doi:10.1155/2020/6946370

Zhou, R., Yazdanifar, M., Roy, L. D., Whilding, L. M., Gavrill, A., Maher, J., et al. (2019). Car t cells targeting the tumor muc1 glycoprotein reduce triple-negative breast cancer growth. *Front. Immunol.* 10, 1149. doi:10.3389/fimmu.2019.01149

Appendix A

Agent based model

The cancer agent-based model is made up of two main modules: the vascular module and the cancer cell module, as described in Section 2.2.2. Figure 9 shows a flow chart of how the ABM is implemented.

Model setup

The cancer module is initialized with 100 cells, of which 20 are stem cells and 80 are progenitor cells. 5% of all cells are assumed to have a higher migratory speed. Each progenitor cell may divide at most div_{lim} times before undergoing senescence, and the initial pool of progenitor cells are randomly assigned a division cycle count between 1 and div_{lim}. The vascular module is initialized with eight capillaries that are aligned along the edges of the simulation grid, with two capillaries branching off of another. The floor of the simulation domain is assumed to rest on healthy tissue and acts as a constant source of oxygen.

Vascular module

The vascular module starts to evolve once hypoxic cancer cells appear in the simulation. These are an assumed source of angiogenic factors such as VEGF. A cancer cell becomes hypoxic once it is 200 microns away from a source of oxygen, namely a mature capillary or the floor of the simulation domain. The vascular network evolves as follows. In each iteration, a cell lining a capillary has a chance to generate a new tip cell, determined by local hypoxic conditions. Each active tip cell determines whether it migrates or proliferates. A tip cell can only proliferate if there is no stalk cell in the sprout, in which case the tip cell proliferates to produce a stalk cell behind it. Tip cells with stalk cells behind them do not proliferate. Once a tip cell has proliferated, we test whether it is adjacent to another tip cell or vascular segment and if so, the two tip cells or the tip cell and vascular segment anastomose. This results in the formation of a closed loop which represents a blood-bearing vessel that is a source of oxygen. All segments in such a vessel are then marked as mature and can no longer proliferate or migrate. If the tip cell does not proliferate, it checks whether it can migrate. We introduce a variable *migdist*, which determines how far the tip cell migrates. *migdist* cannot exceed more than 1.5 times the length of the tip cell, and is calculated based on the local VEGF concentration. This, in turn, is a function of the number of neighboring cancer cells. Details on how *migdist* is computed are provided in (Norton and Popel, 2016). The tip cell randomly migrates towards one of its neighboring cancer cells, excluding cells that would cause the tip cell to migrate backwards. The tip cell does not migrate if it would cause it to leave the vascular grid. After migration, the tip cell checks whether it can anastomose, as described previously.

The second step of the vascular module involves stalk cell decisions. Stalks cells' main function is to proliferate and push the tip cell forward. A stalk cell only proliferates when it reaches the end of its cell cycle, and if there is enough space. If a stalk cell proliferates, a new stalk cell is created replacing the old tip cell. Afterwards a new tip cell is created of 1 micron length in the direction of the old tip cell. The old stalk cell then becomes a quiescent phalanx cell which cannot proliferate or migrate. The new stalk cell resets its cell cycle and the tip cell checks if it should anastomose.

The last step of the vascular module allows for vessel branching of phalanx cells. Neither tip cells nor stalk cells are allowed to branch. Branching occurs due to the presence of nearby hypoxic cells. Specifically, the phalanx cell can only branch if there are hypoxic cells within 250 microns of it. The new branch creates a tip cell that is extended in the direction of the nearest hypoxic cancer cell. Once a phalanx cell has branched the two cells next to it cannot branch.

Cancer module

The cancer module runs through each cancer cell in a random order so as not to introduce bias. Each cancer cell can migrate, proliferate, quiesce, senesce and/or die in each iteration. First, the cell determines whether it is normoxic or hypoxic by checking whether it is less than 200 microns from a mature vessel. Hypoxic cells are more migratory, increasing the speed they migrate 3-fold, and are less proliferative, decreasing their chances to proliferate by half. In order for the cancer cell to migrate or proliferate there must be available space. If there is no available space, the cell becomes quiescent. The cell decides whether it will migrate based on its migration probability. The number of voxels the cell migrates is based on its migration speed. Therefore, each migrating cell randomly chooses an open space to migrate into and repeats this as many times as its migration speed.

The next decision the cancer cell makes is whether it can proliferate. Each cancer cell has its own proliferation rate

10.3389/fmolb.2022.1056461

depending on whether it is a stem cell or a progenitor cell and whether it is hypoxic or not. If it is a progenitor cell, it can only proliferate if it has not reached its division limit, div_{lim}. In this case, the progenitor cell decides whether it will divide based on its proliferation probability p_{div}. If it decides to proliferate, the progenitor creates a new progenitor cell in a random adjacent grid space and increases its division number by 1. The new progenitor cell inherits the parent cell's division number and has a 5% probability of being highly migratory. Once a progenitor cell has reached its division limit, it becomes senescent. Alternatively, if the current cancer cell is a stem cell, it first decides whether it will divide based on its proliferation probability. Then it decides whether it will divide symmetrically into another stem cell or asymmetrically into a progenitor cell. Stem cells have no division limit but if a stem cell creates a new progenitor cell, the new progenitor cell has a full division limit of div_{lim}. At the end of the simulation, any cell that has been hypoxic for 40 iterations dies. Each senescent cell has a 10% probability of dying in each iteration.

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 06 January 2023
DOI 10.3389/fimmu.2022.1011247

Check for updates

OPEN ACCESS

EDITED BY
Dinler Amaral Antunes,
University of Houston, United States

REVIEWED BY
Estrella Mariel Levy,
(CONICET), Argentina
Kelly Quek,
Accenture, United States

*CORRESPONDENCE
Carolyn Shembrey
Carolyn.Shembrey@petermac.org
Frédéric Hollande
frederic.hollande@unimelb.edu.au

†PRESENT ADDRESSES
Carolyn Shembrey,
Rosie Lew Program in Immunotherapy and Cancer Cell Death Laboratory, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia
Momeneh Foroutan,
Department of Biochemistry and Molecular Biology, oNKo-innate Pty. Ltd, Moonee Ponds, VIC, Australia

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 04 August 2022
ACCEPTED 30 November 2022
PUBLISHED 06 January 2023

CITATION
Shembrey C, Foroutan M and Hollande F (2023) A new natural killer cell-specific gene signature predicting recurrence in colorectal cancer patients. *Front. Immunol.* 13:1011247. doi: 10.3389/fimmu.2022.1011247

COPYRIGHT
© 2023 Shembrey, Foroutan and Hollande. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A new natural killer cell-specific gene signature predicting recurrence in colorectal cancer patients

Carolyn Shembrey[1,2,3*†], Momeneh Foroutan[1,2,4†] and Frédéric Hollande[1,2*]

[1]Department of Clinical Pathology, The University of Melbourne, Victorian Comprehensive Cancer Centre, Melbourne, VIC, Australia, [2]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Melbourne, VIC, Australia, [3]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC, Australia, [4]Biomedicine Discovery Institute, Monash University, Melbourne, VIC, Australia

The protective role of Natural Killer (NK) cell tumour immunosurveillance has long been recognised in colorectal cancer (CRC). However, as most patients show limited intra-tumoral NK cell infiltration, improving our ability to identify those with high NK cell activity might aid in dissecting the molecular features which underlie NK cell sensitivity. Here, a novel CRC-specific NK cell gene signature that infers NK cell load in primary tissue samples was derived and validated in multiple patient CRC cohorts. In contrast with other NK cell gene signatures that have several overlapping genes across different immune cell types, our NK cell signature has been extensively refined to be specific for CRC-infiltrating NK cells. The specificity of the signature is substantiated in tumour-infiltrating NK cells from primary CRC tumours at the single cell level, and the signature includes genes representative of NK cells of different maturation states, activation status and anatomical origin. Our signature also accurately discriminates murine NK cells, demonstrating the applicability of this geneset when mining datasets generated from preclinical studies. Differential gene expression analysis revealed tumour-intrinsic features associated with NK cell inclusion versus exclusion in CRC patients, with those tumours with predicted high NK activity showing strong evidence of enhanced chemotactic and cytotoxic transcriptional programs. Furthermore, survival modelling indicated that NK signature expression is associated with improved survival outcomes in CRC patients. Thus, scoring CRC samples with this refined NK cell signature might aid in identifying patients with high NK cell activity who could be prime candidates for NK cell directed immunotherapies.

KEYWORDS

NK cells, gene signature, cancer immunology, colorectal cancer, immunotherapy

Introduction

Despite improvements in the surgical and medical management late-stage colorectal cancer (CRC), 5-year survival rates for patients with late-stage colorectal cancer (CRC) remain extremely poor and innovative treatment strategies are needed. Although T-cell directed immunotherapies are strikingly effective in several solid cancer types, durable responses are limited to colorectal tumours exhibiting defective mismatch repair processes (dMMR), thus benefiting approximately 5% of all CRC patients (1). Downregulation of major histocompatibility complex class I (MHC-I), whose expression is a pre-requisite for T-cell-mediated immune killing, has been reported as a major cause of secondary resistance to checkpoint blockade therapy (2, 3). Thus, there is an urgent need to explore alternate immunotherapeutic strategies which harness other antigen-independent cell types.

In recent years, natural killer (NK) cells have emerged as a promising candidate for immunotherapeutic development. Adoptive transfer of primary NK cells suppresses T-cell mediated graft-versus-host disease and exacerbates graft-versus-tumour responses (4, 5). Similarly, a myriad of alloreactive NK cell lines - primarily NK92, but also NKL, KHYG01 and YTS – have demonstrated safety and efficacy in HLA-mismatched recipients in early-stage clinical trials (reviewed in (6)). This suggests that NK cell therapeutics may be harnessed in an "off-the-shelf" manner in the future, circumventing the difficulty associated with the *ex vivo* expansion of antigen-specific T-cell clones. Likewise, chimeric antigen receptor (CAR) NK cells have shown promising results in xenograft models (7, 8), and monoclonal antibodies which neutralise NK cell inhibitory receptors such as the KIR family (9, 10) and NKG2A (11) have entered early-stage clinical trials.

NK cell activity is inversely correlated with cancer incidence (12) and there is a wealth of evidence supporting the role of NK cells in controlling both spontaneous and experimental metastasis (13, 14). In CRC, NK cell infiltration has been identified as a positive prognostic marker in both primary (15, 16) and metastatic (16, 17) disease. NK cells differ from conventional lymphocytes in that they function in an MHC-I unrestricted manner in accordance with the "missing-self" hypothesis (18, 19). In this manner, NK cell-directed immunotherapies may overcome the restricted benefit of antigen-specific T-cell responses in tumours with high mutational diversity. Moreover, harnessing NK cell cytotoxicity is a promising opportunity in the treatment of immunologically "cold" tumours such as CRC that undergo loss of MHC-I expression. Indeed, recent studies have reported that 40% of patient-derived CRC organoids exhibit MHC-I loss which could not be rescued by IFN stimulation (20), concordant with clinical reports of MHC-I loss in approximately 60% of MSI CRCs (21). Yet, clinical enthusiasm has been tempered due to the scant NK load reported in most CRC tumors, despite high levels of chemokines and cytokines (22). Thus, novel means of determining which patients show high NK cell infiltration and activity and might therefore benefit from NK cell directed immunotherapies is needed.

Transcriptomic signatures are sets of genes whose coordinated expression has a verified association with specific biological parameters – such as cell type and phenotypic state – or clinical measures, including disease subtype, survival outcome or therapeutic response. Deconvoluting the relative abundance of immune cell subsets using transcriptomic signatures offers many benefits over the comparatively low-throughput methods of immunophenotyping which are currently available. In CRC alone, a multitude of prognostic signatures have been reported (23, 24) alongside several signatures which predict response to 5-FU based chemotherapies in Stage II-III disease (25–27). Additionally, RNAseq data from a cohort of 40 CRC patients with metastatic or relapsed disease was used to derive a 27-gene signature able to discriminate responders and non-responders to the FOLFOX6 chemotherapy regimen with accuracy of 92.5%, demonstrating the powerful role which signature analysis may play in personalising patient care (28).

In their simultaneous assessment of multiple markers, transcriptome-wide approaches circumvent the limitations of single-marker or low-throughput phenotypic assessments. This may also allow for previously unidentified markers or associations to be identified which may provide important insights into immune cell biology. However, a major consideration when employing such computational approaches is the suitability of the reference profiles, as the gene signatures derived from sorted cell types in healthy individuals may not accurately reflect those of the potentially dysregulated cells in a diseased individual. This is particularly pertinent in the case of NK cells, as activated NK cells have unique transcriptional profiles as compared with resting NKs from healthy volunteers (29). Additionally, as signatures often rely on highly expressed genes and thus often incorporate genes that are not that specific of a given cell type, the accuracy of cell type detemination may be compromised. Although prognostic NK cell gene signatures exist for highly immunogenic tumour types such as melanoma (30) and renal cell carcinoma (31), it is currently unclear how appropriate these are for use in tumours such as CRC which traditionally show a poor immune cell infiltration. Other works which have looked at the transcriptomic profiles of lymphocytes in CRC have focussed on defining residency versus exhaustion states (32). where the genesets used to define CD4+ and CD8+ T-cells had several genes overlapping with the NK cell geneset.

Here, we present a novel NK cell transcriptomic signature which can be used to infer NK cell abundance from bulk RNA sequencing data of primary CRC samples. Candidate NK cell-related genes were pooled from previously published works and in-house differential expression (DE) analyses and sequentially filtered to ensure their fidelity as NK cell markers with minimal

off-target expression in tumour, stromal and other immune cell types. Single-cell RNAseq (scRNAseq) data from primary CRC samples were then used to validate each signature gene in tumour-infiltrating NK cells. We then show that high NK cell score is associated with upregulation of cytolytic and chemotactic transcriptional processes, and survival analysis revealed that patients with higher evidence of NK cell activity demonstrate significantly longer recurrence and disease-free intervals. Collectively, the NK cell signature allows for the identification of CRC patients with high NK cell activity, which may aid in defining the molecular characteristics associated with strong response to NK cell targeting immunotherapies.

Methods

Preparation of publicly available RNAseq data

Datasets used in the present study are listed in Table S4. Raw counts files from publicly available datasets were downloaded rom Gene Expression Omnibus (GEO) using the NCBI portal (http://www.ncbi.nlm.nih.gov/geo/). CCLE data was downloaded as a PharmacoSet (PSet) through the PharmacoGx R/Bioconductor package (version 1.6.1). For in-house data, 3' RNA-seq reads were aligned using HISAT2 against human genome GRCh37 (release 75) and the featureCounts tool from the RSubread package was used to quantify the number of reads for each gene per sample. The filterByExpr function from the edgeR package (v 3.28.1) was used to filter lowly expressed genes and calculated count- or transcript-per-million (CPM/TPM) values. The MyGeneset application of the ImmGen online databrower (http://www.immgen.org/Databrowser19/DatabrowserPage.html) was used for analysis of GSE15907.

The GSE107011 dataset is composed of 29 FACS-isolated cell types from the blood of healthy individuals. For this project, The NK cell, basophils, and neutrophil samples were annotated as per the authors' annotation file. However, to circumvent the known issues associated with multiple comparisons, and to facilitate comparisons with data sets offering less cellular subtype resolution, the various maturation and functional states for other cell types were merged to form larger groups: naïve, switched, non-switched and exhausted B-cell samples were merged and annotated as "B-cells" ($n_{samples}$ = 16); naïve, central memory, effector memory and terminal effector CD8+ T-cell samples were merged and annotated as "CD8+ T-cells ($n_{samples}$ = 16); classical, intermediate and non-classical monocytes were merged and annotated as "monocytes" ($n_{samples}$ = 12); myeloid and plasmacytoic dendritic cell samples were merged and annotated as "dendritic cells" ($n_{samples}$ = 8). Plasma and progenitor cells were excluded from downstream analyses. Given the transcriptional overlap between NK- and T-cells, T- cell subsets (namely CD8+, CD4+, MAIT cells, Tfh, TH1, TH17, Th2, Treg, γδ+ and γδ -) were interrogated both individually and as part of a larger, merged group annotated as "pan-T-cells". Resultingly, 17 discrete immune cell subtypes from this dataset were interrogated. For GSE60424, only blood samples from healthy individuals were included.

Differential expression analysis, gene set testing and survival modelling

The singscore (v1.6.0) R package was used for Single-sample gene set enrichment analysis against various molecular signatures. Depending on the direction of gene sets, we used different settings of simpleScore function as specified in the documentation. NK-high and NK-low groupings were defined as the top and bottom 10% of samples, respectively, when ranked by NK scores. DE analysis between NK-high and NK-low samples was performed using the voom-limma (v 3.42.2) pipeline (Law et al., 2014). After running eBayes, we considered genes with absolute log2FC > 1 (for ovexpression) or log2FC < −1 (for repression) and adjusted p-value < 0.05 as DEGs. The goana function from the limma package was used to perform gene ontology (GO) analysis and camera gene set testing was performed using MSigDB signatures retrieved from the WEHI bioinformatics portal (http://bioinf.wehi.edu.au/software/MSigDB/). Survival modelling was performed using the survminer (v0.4.8) and survival (v3.2-7) packages using the clinical annotation files provided from the sources listed in Table S4.

Data wrangling and visualization

All computational analyses were performed using R (version 3.6.1). For data wrangling and visualization, base R functions were using alongside several core packages from the tidyverse (v 1.3.0) R package. tidyr (v 1.1.2) and dplyr (v 1.0.2) were used for reading and manipulating the data, as well as ggplot2 (v 3.2.1) and cowplot (v 1.0.0) for data visualization. Heatmaps were generated using the complexHeatmap (v 2.2.0) or pheatmap (v 1.0.12) R packages.

Statistical analysis

All statistical analyses were performed using R. Data is expressed as mean ± standard deviation (SD) unless otherwise indicated. The minimum threshold for rejecting the null hypothesis was $p<0.05$. For results where statistics are shown, significance is denoted as: * = $p<0.05$; ** = $p< 0.01$; *** = $p<0.001$.

Code availability

The code used throughout this study is available on Github (https://github.com/cshembrey/NK_Signature_CRC).

Results

Collation of candidate NK cell signature genes

To identify a CRC-specific gene expression signature associated with NK cell abundance, we implemented a novel pipeline that involved curation and sequential refinement of putative NK cell genes against immune, tumour and stromal cells in multiple bulk and single cell RNAseq datasets (workflow outlined in Figure S1). Firstly, 605 unique genes were collated from eight partially overlapping sources (Figure 1). Four previously curated NK cell signatures were compiled: "CIBERSORT Active" (n_{genes} = 56) and "CIBERSORT Resting" (ngenes = 56) refer to the gene sets corresponding to activated and resting NK cells, respectively, as reported in the LM22 signature matrix used for the CIBERSORTx algorithm (33); The "Cursons Extended NK cell Geneset" gene set (n_{genes} = 112 genes) was derived from the supplementary data table of a melanoma-specific NK cell signature previously reported by Cursons and colleagues (30); The "Wang NK cell marker" gene set (ngenes = 13; 34) is composed of markers used to guide the immunophenotyping of different cell subsets in bulk RNAseq data from CRC cell lines and primary samples, and the "receptors" gene set (n_{genes} = 43) was compiled by mining the literature for various receptor subsets (eg. activating, inhibitory, chemokine or cytokine receptors) with documented expression on NK cells.

Subsequently, three novel gene sets were derived from the results of differential expression (DE) analysis, where putative genes enriched in NK cells were identified by pairwise comparison of NK cells and at least one other immune cell type. DE analysis of the GSE60424 bulk RNAseq dataset (35), composed of six immune FACS-isolated cell types (NK cells, B-cells, CD4+ T-cells, CD8+ T-cells, monocytes and neutrophils), identified 280 DEGs ("GSE60424 DEGs"). Similarly, DE analysis of GSE107011 (36), an RNA-seq dataset composed of 29 FACS-isolated immune cell types, yielded 427 DEGs ("GSE107011 DEGs"). Finally, to enhance our resolution when discriminating between NK and T-cells, a subsequent "pan-T DEGs" (ngenes = 23; from GSE107011) was constructed from the DE genes when only NK versus T-cell comparisons were considered; here, the selected genes were those upregulated in NK cells relative to all T-cell subsets (eg. upregulated in NKs relative to CD8+ and CD4 + and MAIT etc.), irrespective of their expression level in other innate immune cell types.

Following compilation, 605 unique candidate NK genes were identified. To examine the representation of traditional NK cell markers versus potentially novel NK cell-related genes, the interconnectedness of this gene set was compared across the eight sources (Figure 1). Most genes identified (451/608; 74%) were uniquely found from our DE analyses. Surprisingly, there were zero genes which were conserved across all eight sources and only two genes, KLRD1 and KIR3DL2, were conserved across 7/8 sources. We hypothesize that the lack of consensus between these datasets may reflect the multiple discrepancies between studies in terms of criteria chosen for candidate gene selection and filtering, combined with variations in the NK subtypes present in each dataset.

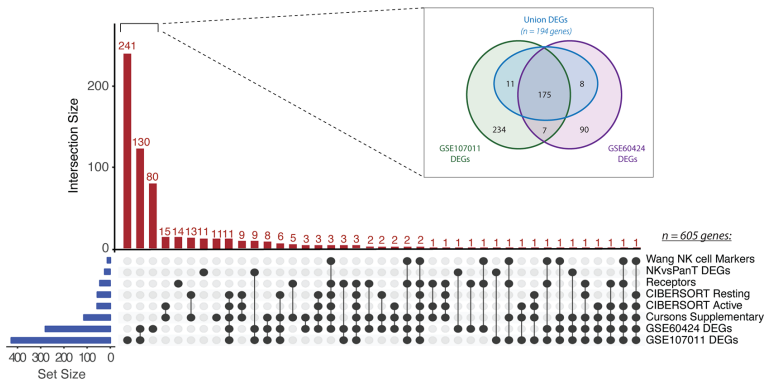


FIGURE 1
Collation of putative NK cell signature genes. UpSet plot of gene intersects from eight partially overlapping sources (listed at right). Gene set (blue bars) and intersection (red bars) sizes are indicated. Inset: Venn diagram of DEGs identified from pairwise comparisons of NK cells versus other immune cell types in GSE60424 and GSE107011. Only the "Union DEGs" (n_{genes} = 194) were retained as candidate signature genes.

To further refine the comparatively large GSE60424 and GSE107011 DEG sets, a "Union DEGs" gene set was created (Figure 1; inset). For inclusion in this merged gene set, a given candidate gene needed to be differentially upregulated in NK cells relative to all other immune cell types in one or both data sets (eg. upregulated in NK cells versus T-cells and B-cells and neutrophils) differing from the previous DEG analyses where the gene in question need only be upregulated in NK cells on a pairwise basis (eg. upregulated in NK cells versus T-cells, but not necessarily B-cells nor neutrophils). Of the 525 unique DEGs identified across GSE60424 and GSE107011, 194 genes fit this criterion. After reintegration of these "Union DEGs" with the genes derived from other sources, 295 candidate genes remained.

Although DE analysis identifies genes which are preferentially expressed by NK cells as compared with other cell types, the expression of individual DEGs by NK cells may still be very low. This is problematic when aiming to identify NK cell signals from tumour sequencing data, particularly given that NK cells represent a small proportion of the total cell number. Thus, we further refined out geneset by retaining only those candidates with higher median in NK cells relative to other immune cell subsets in GSE107011 (Figure 2A; Figure S2A; purple boxes) and GSE60424 (Figure S2B; purple boxes). Sets of passing genes for each dataset were derived by taking the intersect of the specific genes (n_{genes} = 49) from GSE107011 (Figure 2A) and GSE60424 (Figure S2B). Of these, 10 genes (ADGRB2, B4GALT6, LDB2, LIM2, LINC01451, LRRC43, PCDH1, PRSS57, RAMP1 and RNF165) were identified in both data sets.

Interestingly, this analysis highlighted that many of the "classical" NK cell markers that are in multiple sources (Figure 2A; yellow/green rings) are also expressed at very high levels in "unconventional" T-cell subsets such as MAIT, γδ+ and γδ- T-cells. As illustrated by GSE60424 (Figure S2B), where T-cells are exclusively grouped as CD8+ or CD4+, many FACS-based studies do not include these relatively niche T-cell subsets, possibly leading to the identification of putative NK cell-specific genes which are in fact highly expressed by unconventional T-cell subsets.

The two sorted NK cell-containing data sets interrogated thus far (GSE60424 and GSE107011) were derived from the peripheral blood of healthy individuals; this approach has limitations, as the gene expression profiles of such cells may not necessarily reflect those of tissue-infiltrating NK cells nor NK cells in the context of cancer. To address this concern, we performed an independent DE analysis (see methods for details) on an NK-cell containing scRNAseq data set (GSE146771 (37), composed of SMART-Seq2 and 10X subseries) generated from a cohort of CRC patients ($n_{samples}$ = 20; from 18 unique patients) with Stage II-III disease of varying pathological grade, MSI status, and extent of nodal involvement. This analysis identified several additional NK cell marker genes, and allowed us to cross-validate the 49 genes derived from bulk RNA-seq analysis (mentioned above) at the single cell level. Following these analysis steps, 82 candidate genes were prioritised.

Signature gene specificity filtration against tumour and stromal cells

As the NK cell signature is designed to resolve the NK cell fraction from bulk RNAseq data of tumour tissue, it is imperative that the genes in the signature should not be expressed by the tumour cells themselves. To determine whether any of the 82 identified genes were expressed by tumour cells, additional specificity thresholding was performed against two sources of CRC cells (Figure 2B). As bulk RNAseq from CRC tumour tissue would be expected to contain immune cell fractions, two cohorts of CRC cell lines were used in this analysis: The Cancer Cell Line Encyclopedia (38) ($n_{samples}$ = 57) and GSE90830 [nsamples = 44 (39)]. These datasets revealed that 12 genes had relatively higher expression in CRC cells compared with NK cells (Figure 2B; blue boxes), and they were therefore excluded from further analyses.

Despite the immune- and tumour-specific filtration steps previously performed, the possibility of contaminant expression of our candidate genes by other non-NK cell types such as including fibroblasts, stromal cells and non-transformed epithelial cells has not been accounted for. To address this, the relative expression of our candidate genes was interrogated in the leukocyte (CD45+), stromal cell (CD31+) and epithelial cell (EpCAM+) fractions isolated from the tumours of 6 patients with CRC (GSE39397 (39); as well as against cultured, normal colon mucosa-derived fibroblasts (CCD-Co-18). Failing genes (ngenes = 14) were defined as those with significantly higher expression relative to leukocytes in a non-leukocyte subset (Figure S3; red headers, summarised in Figure 2C).

Interrogating expression of signature genes in tumour-infiltrating NK cells

Having validated the specificity of our candidate genes at the immune, tumour and stromal cell level in bulk RNAseq, we returned to GSE146771 to confirm this specificity in scRNAseq data. By interrogating the gene expression distribution across Uniform Manifold Approximation and Projection (UMAP) plots (Figure 3), the specificity of each gene for particular NK cell subsets was elucidated. For example, in the SMART-Seq dataset three NK cell subgroups could be discerned based upon marker expression, providing greater cellular resolution: CD16+, a marker expressed on virtually all $CD56^{dim}$ NK cells; GZMK+, a cytotoxicity marker frequently associated with the $CD56^{bright}$ subpopulation; and CD103+, a marker of tissue residency. Certain "NK cell specific" genes such as SH2D1B and KIR2DL3 showed equivalent expression in each of the three

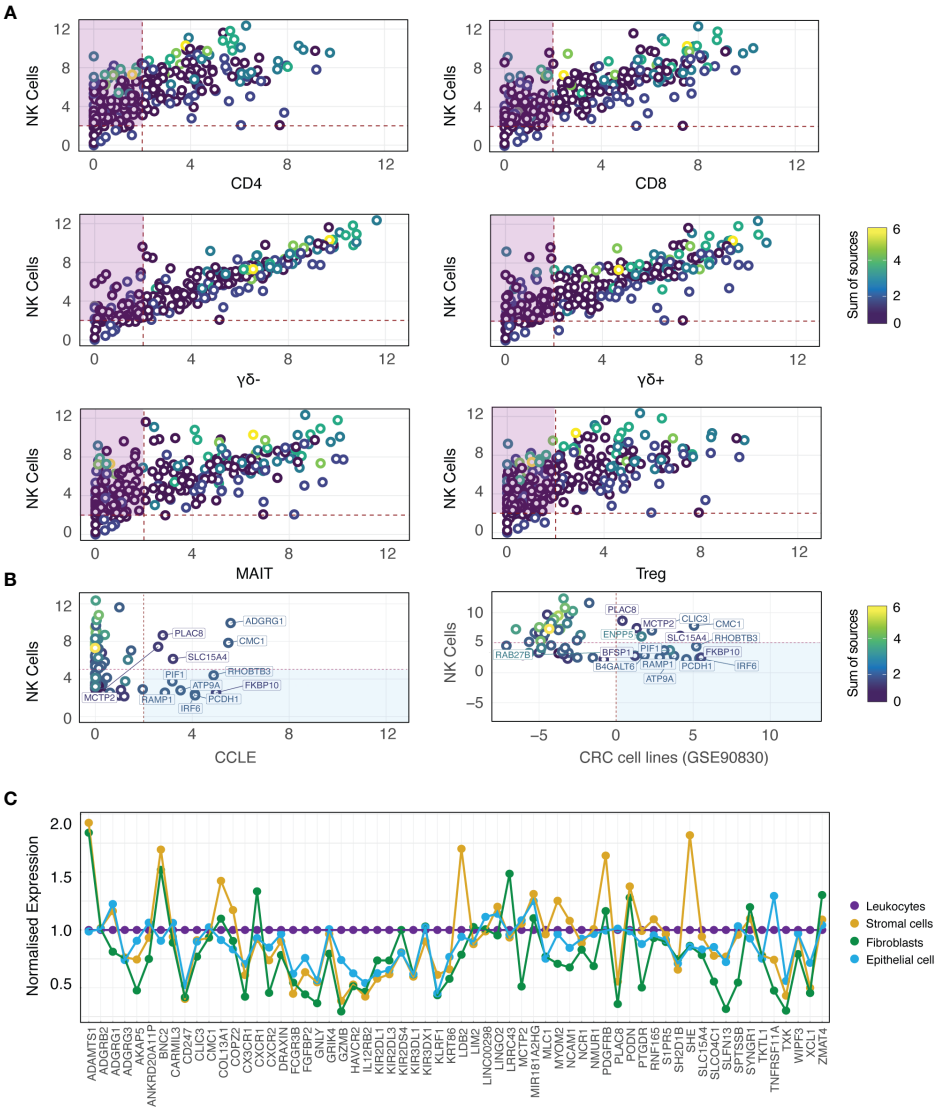


FIGURE 2
NK cell specificity filtration against immune and CRC-supportive cell types. **(A)** Biplots depicting median expression for each candidate gene (rings; coloured by sum of sources) in NK cells from GSE107011 (logTPM) versus other immune cell types. The intersect of passing genes for each pairwise comparison (purple boxes) were retained as candidate genes. **(B)** Biplots depicting median expression for each candidate gene (rings; coloured by sum of sources) in NK cells from GSE107011 (logTPM) versus CCLE CRC cell lines (left column; logRPKM) and GSE90830 (CRC cell lines; right column; logRPKM). The union of failing genes for each pairwise comparison (blue boxes; n_{genes} = 12) were flagged for removal from the candidate geneset. **(C)** Lineplot of candidate gene expression in the stromal cell (CD31+; gold), fibroblast (FAP+; green) and epithelial cell (EpCAM+; light blue) compartments of six primary CRC samples (GSE39396), normalised to leukocyte (CD45+; purple) expression. Genes with normalised expression >1 were flagged for removal. *CCLE, Cancer Cell Line Encyclopedia; CoT, CRC primary tumour; LT, CRC liver metastasis.*

NK subsets, whereas other genes were subtype selective. For example, LINGO2, PRSS57 and RNF165 were preferentially expressed by the CD16+ subgroup, whereas WIPF3 expression was high across both the CD103+ and GZMK+ subgroups but sparse in the CD16+ population.

This approach also allowed us to identify genes which, although differentially upregulated in NK cells, showed moderate basal expression across multiple cell types (Figures S4, S5; "High DEGs"; pale blue headers) as well as genes whose expression was no longer specific to NK cells once evaluated at the single-cell level (Figures S4, S5; "Non-specific"; red headers). For example, many candidate genes - including PLAC8, SLC15A4, SLFN13 and ST8SIA6 – have high expression in NK cells although they are promiscuously expressed by multiple cell types, with particularly high expression in the myeloid subset, warranting their exclusion from the final gene list.

10.3389/fimmu.2022.1011247

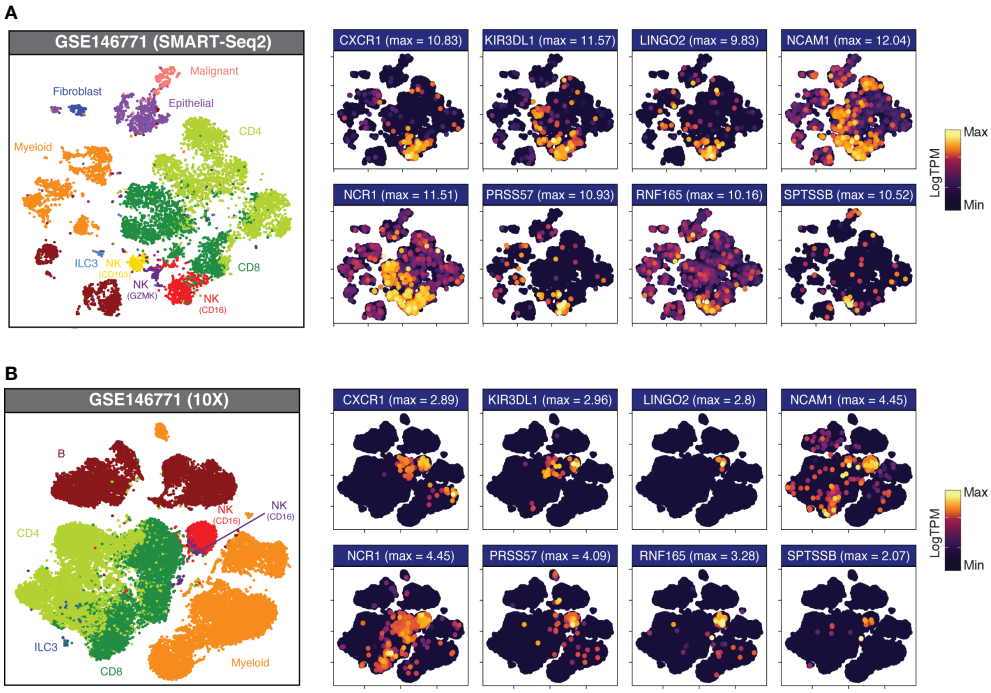


FIGURE 3
Candidate gene expression in CRC-infiltrating immune cells. UMAP plots of dissociated primary CRC samples from **(A)** GSE146771 (SMART-Seq2 scRNAseq; $n_{samples}$ = 10) and **(B)** GSE146771 (10X scRNAseq; $n_{samples}$ = 10) coloured by cell type (at left) and candidate gene expression. Maximum expression (LogTPM) is indicated in parentheses above each plot.

Validating the subtype specificity of the curated NK cell signature

Having refined the NK cell signature using CRC-infiltrating NK cells, we next interrogated whether the signature exhibited any biases towards NK cells from particular subsets or sources. GSE133383 (40) contains transcriptomic data for both the immature CD56^{bright} and mature CD56^{dim} subsets of NK cells isolated from the blood, lymphoid organs (spleen, lymph nodes and bone marrow) and lungs of four healthy donors. As is the case with most transcriptional analyses of NK cells, when interrogated in GSE133383 our signature genes cluster according to NK cell maturation state rather than tissue source (Figure 4A). Whilst a substantial number of genes which appear to have relatively uniform expression across both the CD56 bright and dim subsets (Figure 4A), multiple subtype-enriched markers have also been retained. Enriched markers for the populous CD56^{dim} class include the KIR family members, GZMB (40–42) and the chemokine receptors CXCR1 and CXCR2 (43). Two recently identified markers of terminally mature NK cells, HAVCR2 and CX3CR1 (44) show preferential expression in the CD56^{dim} subset as expected. Analogously, enriched markers for the CD56^{bright} subset include CD56 (NCAM1) itself, XCL1 (44) and KLRC1 (45), whose inclusion indicate that the signature has sufficient resolution to detect this relatively minor population. Collectively, this demonstrates that our signature captures both general and subset-enriched NK cell markers, suggesting that it is representative of all types of NK cell.

Many of the genes which are lowly-expressed in GSE133383 (Figure 4A; LIM2, SPTSSB, LGALS9B, PRSS57, ADGRG3, and, to a lesser extent, DRAXIN), which lacks intestinal NK cells samples, are highly expressed in particular subsets of CRC-infiltrating NK cells (Figures 4A, B), possibly reflecting a role for these genes as CRC-specific NK cell markers induced by the tumour microenvironment. It is noteworthy that the peripheral blood CD56^{bright} subset – the most abundant NK cell subset in humans - do not cluster together, supporting the idea that this signature is more CRC-specific than previously reported NK cell signatures derived from blood-based analyses.

Next, as it is expected that genes which are considered robust markers of a given cell type should have coordinated expression patterns, the correlation between the individual genes in the signature was assessed. Correlation analysis demonstrated that the signature genes cluster into two major blocks where expression is highly cross-correlated, representing the genes selective for CD56^{dim} or CD56^{bright} NK cells (Figure 4B; orange and purple annotations at left). Subtype selectivity for

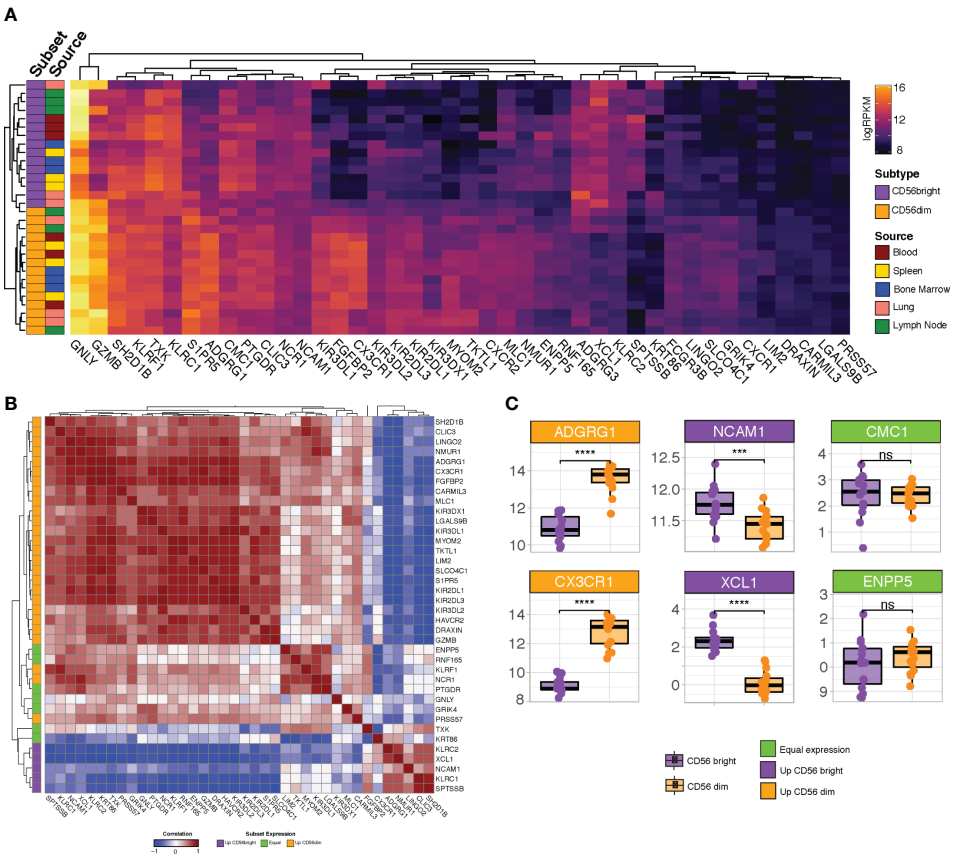


FIGURE 4
Profiling refined gene set expression in NK cell subsets. **(A)** Heatmap of refined gene set expression in CD56^{bright} (orange) and CD56^{dim} (purple) NK cells isolated from the blood (red), spleen (yellow), bone marrow (blue), lung (pink) and lymph node (green) of four healthy donors (GSE133383). **(B)** Correlation matrix of refined gene set. Genes are annotated (at left) according to whether they show preferential expression in the CD56^{dim} (orange), CD56^{bright} (purple) or neither (green) NK cell population **(C)** Boxplot expression of selected genes in CD56^{dim} (orange boxplots) vs CD56^{bright} (purple boxplots) NK cell populations. Headers are coloured based on whether the gene is preferentially expressed in CD56^{dim} (orange headers) vs CD56^{bright} (purple headers) NK cells or shows equivalent expression across both subsets (green headers). Significance was assessed with Student's T-test; ***p-value < 0.001. ; ****p<0.0001; ns: non-significant.

each gene was defined based on its relative expression in each of these subsets (Figure 4C and Figure S6). That the two major blocks are anti-correlated may reflect the progressive loss of the CD56^{bright} markers as NK cells mature and reinforces the idea that our signature allows for pan-NK cell rather than subset-specific detection. Situated between the two major clusters are a subset of genes (including GNLY, TXK and PTGDR) which appear to have equivalent expression in CD56^{bright} and CD56^{dim} cells (Figure 4B; green annotations), likely corresponding to a "transitional" NK cell phenotype reported by several groups (42–44, 46).

As a final visualisation measure, the expression of the NK cell signature in aggregate (ie. profiling expression of the signature as a whole, rather than the expression of each individual gene) was performed in an independent human dataset, GSE22886 (47). This dataset is composed of twelve different leukocyte subsets isolated from the PBMCs of healthy donors and was used to avoid the biases of testing our signature on the GSE60424 and GSE107011 data from which it was partially derived. In GSE22886, our NK cell signature was clearly enriched in the NK cell samples relative to the other cell types (Figure 5A). Due to the high sequence and transcriptomic homology between human and murine NK cells (48), we next interrogated the performance of the human NK cell signature in GSE15907 (Immgen) (49); a microarray dataset generated following the *ex vivo* isolation of multiple immune lineages from adult B6 mice (Figure 5B). Notably, in the murine context our NK cell signature clearly discriminates NK cells from CD8+ T-cells, γδ-T-cells and NKT cells, highlighting the flexibility of this novel NK cell signature to be used in *in vivo* studies to detect either endogenous murine or xeno-transplanted human NK cells.

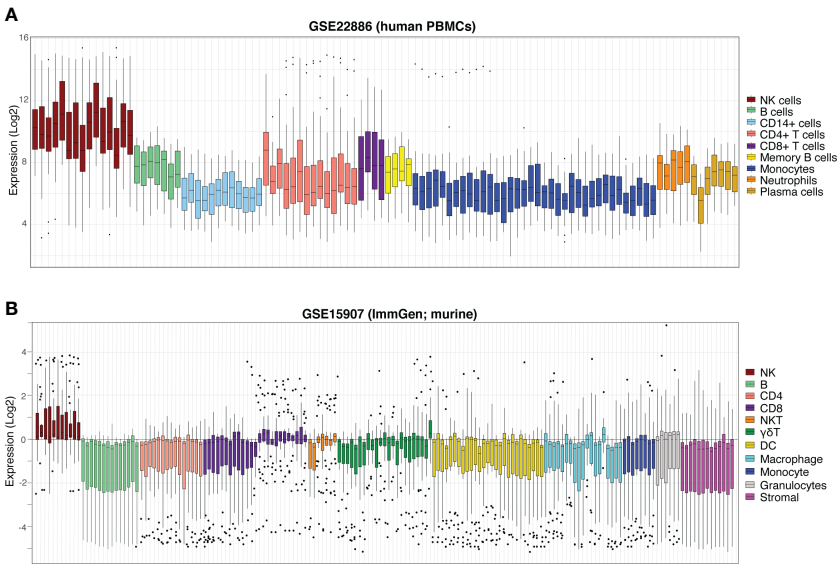


FIGURE 5
NK cell signature efficiently discriminates NK cells from other haematopoietic compartments. Boxplots of NK cell signature expression in sorted peripheral blood cells from **(A)** GSE22886 (Log2) independent human PBMC data and **(B)** GSE15907 (Immgen) murine data. Each boxplot represents one sample (coloured by cell type, at right) and individual points are single signature genes.

Differential gene expression analysis confirms that cytotoxic and migratory programs are associated with high NK score

Having finalised the NK cell signature (Table S1, $n_{genes} = 43$), we next sought to determine which genes were concomitantly up- or downregulated in the samples whose transcriptomic profiles showed strong evidence of this signature. To identify these samples, we used a single-sample, rank-based gene-set scoring method termed *singscore* (50) to score samples against our NK signature. Here, a sample with a high NK score is interpreted as having high evidence of NK cell activity, whereas a sample with low NK score exhibits limited NK cell signature expression. Using samples sourced from two large, publicly available repositories of primary CRCs, the TCGA colorectal adenocarcinomas (TCGA-COAD; $n_{samples} = 454$), and GSE39582 ($n_{samples} = 566$ (51), we defined the NK-high and NK-low groupings based on the top 10% and bottom 10% of scored samples, respectively. We then performed differential expression analysis to compare the gene expression profiles (GEPs) of samples with high scores to those with low scores in each of the two data sets.

A strong positive correlation was observed between the logFC of genes in TCGA and GSE39582 datasets (Figures 6A, B; Spearman correlation R = 0.72, p-value < 2.2 x 10^{-16}), suggesting similar transcriptional programs of NK inclusion/exclusion in both data sets.

48 genes were differentially upregulated in NK-high samples from both datasets (Figure 6B, upper panel, Table S2). Of these, there was a strong enrichment of genes encoding cytotoxic effectors which are critical for NK cell killing, including the granule proteins NKG7 and granulysin (GNLY), as well as multiple members of the granzyme family including GZMB, GZMA and GZMH. Additionally, there was an overrepresentation of genes encoding ligands for chemokines implicated in NK cell trafficking, including CXCL9, CXCL10 and CCL5 (RANTES). Importantly, these ligands are expressed by tumour cells rather than the NK cells, suggesting that the high NK cell density in these tumours is at least partly driven by a tumour-intrinsic factor.

Conversely, 45 genes conserved between the two data sets were differentially repressed in the NK-high group (Table S3). The parallel downregulation of multiple genes which have been associated with increased metastasis and poor prognosis in CRC was observed, including Dachshund family transcription factor 1 (DACH1) (52) and the metalloprotease meprin-α (MEP1A) (53) and CXCL14 (54). Interestingly, recent evidence indicates that CXCL14 is essential for MHC-I upregulation (55) and as such, it stands to reason that CRC samples with high NK cell activity show decreased CXCL14 expression.

To identify biological processes associated with the NK-high phenotype, gene ontology (GO) enrichment analysis was conducted. Competitive gene set testing against the whole transcriptome of NK-high vs NK-low samples confirmed significant enrichment of the GO term “NK cell mediated immunity” in the NK-high group (Figure 6C). Moreover, the

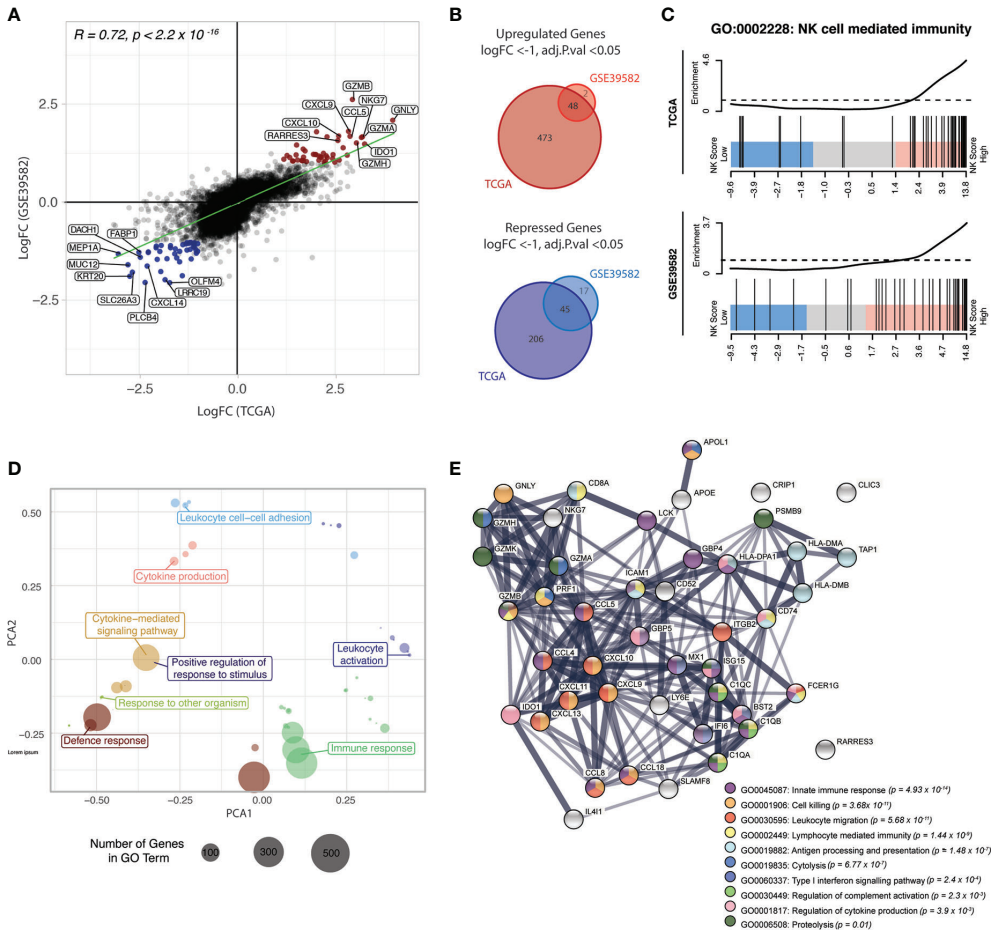


FIGURE 6
NK score is associated with high chemokine and cytolytic activity. **(A)** Scatterplot of the LogFCs of genes comparing NK-high and NK-low groups in TCGA and GSE39582 datasets; Genes that were differentially upregulated (red points) or repressed (blue points) in both datasets are highlighted. DEGs were defined as those with adjusted p-value < 0.05 and absolute LogFC > 1 or LogFC < -1. **(B)** Venn diagram of upregulated (*upper*) and down-regulated (*lower*) DEGs identified in (A). **(C)** Barcode plots showing enrichment of genes in the "NK cell mediated immunity" gene set (GO:0002228) in NK-high samples in both data sets **(D)** PCA plot of top 50 most significant GO terms in the NK-high group, distributed by semantic similarity of genes within each term **(E)** STRING network analysis of protein-protein interactions between the overlapping upregulated DEGs from (A), coloured by biological process. Thickness of the connecting line indicates the strength of evidence for the predicted interaction. *GO, gene ontology*.

top 50 most significantly enriched GO terms in the NK-high group, when clustered by semantic similarity, converged on umbrella terms including "immune response", "leukocyte activation", "leukocyte cell-cell adhesion" and "cytokine production" (Figure 6D).

To visualise the functional synergy of the 45 overlapping upregulated DEGs (from Figure 6B) which define the NK-high group, the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) platform was used to construct a protein-protein interaction (PPI) network (Figure 6E). Consistent with DE results, the PPI network was primarily centred around CC- and CXC-family chemokine ligands (CCL5, CCL4, CCL8, CXCL9, CXCL10, CXCL11, CXCL13, CXCL18) as well as cytolytic effectors (GNLY, GZMH, GZMK, GZMB, GZMK, NKG7, PRF1). A minor node related to antigen processing and presentation (HLA-DMA, HLA-DMB, HLA-DPA1, TAP1, CD74) may point towards a concomitant activation of effector T-cells in the NK-high group; however, chronic antigen stimulation is also known to induce lymphocyte exhaustion, resulting in reduced cytotoxic function (56).

Collectively, these data indicate that the NK-high group is defined by strong chemotactic signalling and high cytolytic activity. As these genes are signposts of an immune-active microenvironment, their upregulation in the NK-high group suggests that the NK cell signature selects for functionally active NK cells, rather than merely NK cell presence.

High NK score is associated with improved survival outcomes and other clinical parameters

Given that NK cell load has been associated with better patient prognosis in CRC (15, 17), we next performed Kaplan-Meier survival analysis to determine whether signature expression was associated with survival probability. Due to significant survival differences between patients with a primary or metastatic disease, we focussed only on those patients with Stage I-III CRC. For all survival analyses, patients were stratified by NK score where "NK-High" and "NK-Low" are defined as samples above and below the median NK score, respectively.

For the TCGA-COAD cohort, the NK-High group had significantly increased disease-free interval (DFI; defined as the period from date of diagnosis until a tumour progression event e.g. locoregional recurrence or distant metastasis) as compared with the NK-Low group (Figure 7A, log rank p-value = 0.0054, p-value from multivariate Cox regression model [accounting for age, stage and MMR status] = 0.02). Similarly, the GSE39582 cohort exhibited a trend towards significantly prolonged recurrence-free survival (RFS), defined as the period between

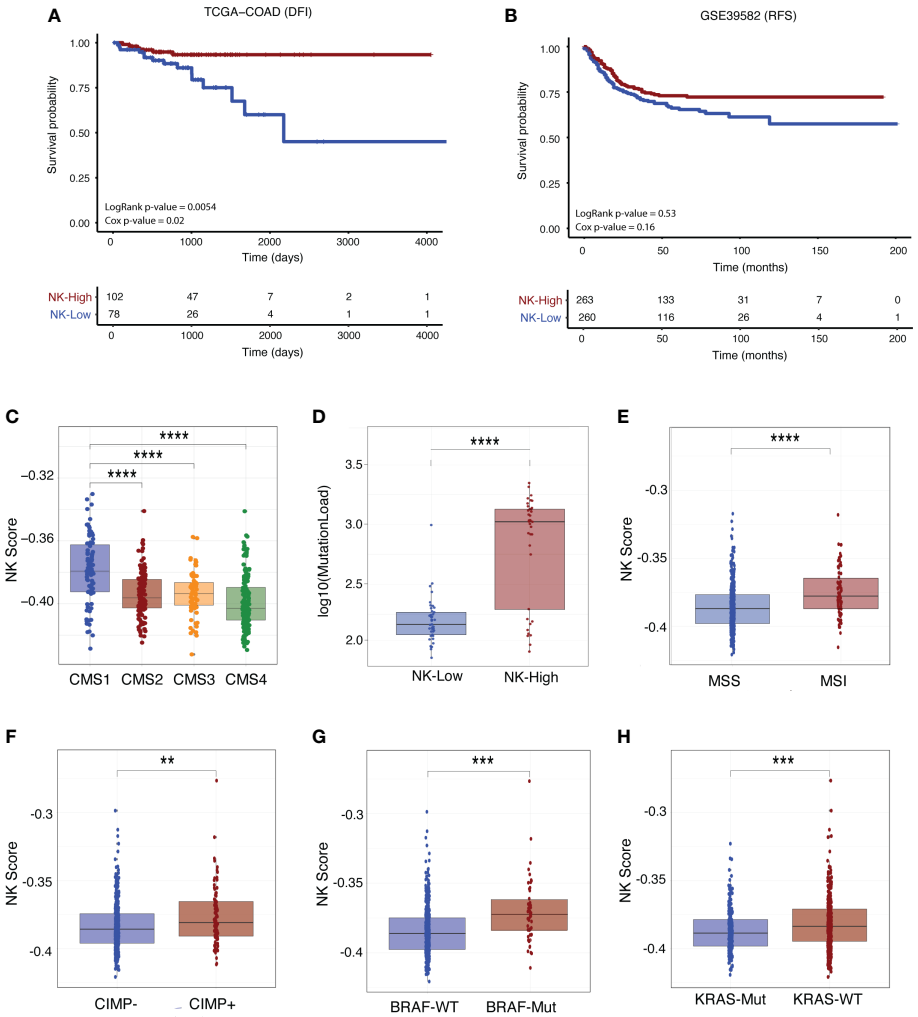


FIGURE 7
NK score is associated with survival outcome and other clinical parameters. Kaplan-Meier survival curves for patients stratified by NK score (where "NK-High" and "NK-Low" are defined as samples above and below the median NK score, respectively) for Stage I-III CRC patients in the **(A)** TCGA-COAD (DFI) and **(B)** GSE39582 (RFS) cohorts. Survival differences were tested using both log-rank and multivariate Cox proportional hazards models (adjusted for age, tumour stage and MMR status) with corresponding p-values indicated. **(C)** Boxplots of association of NK score with CMS status and **(D)** mutational load in TCGA-COAD patients. **(E)** Boxplots of association of NK score with MSI **(F)** CIMP **(G)** BRAF and **(H)** KRAS status in GSE39582 patients. (Student's T-test; **p-value < 0.01; ***p-value < 0.001; ****p < 0.0001). *DFI, disease-free interval; RFS, recurrence-free survival, MMR, mismatch repair; CIMP, CpG island methylator phenotype.*

surgical resection and a tumour progression event, at the univariate level (Figure 7B, log rank p-value = 0.053). There were no significant differences in overall survival (OS; Figures S7A, C) nor progression-free interval (PFI; Figure S7B) between the NK-High and NK-Low groups in either dataset.

We next interrogated the relationship between NK score and various clinical and molecular parameters such as patient history, molecular subtype, and driver mutation status (see additional analyses in Supplementary Data). In both the GSE29582 (Figure S8A) and TCGA (Figures S9B, C) cohorts, NK scores were increased in early-stage disease (Stage I-II) as compared with late-stage disease (Stages III-IV). In the TCGA data, NK scores were considerably enriched in the immunogenic CMS1 subtype (Figure 7C) and in those tumours with higher mutational load (Figure 7D). In GSE29582, high NK score was significantly increased in those patients with MSI (Figure 7E) and CpG Island Methylator Phenotype (CIMP+; Figure 7F) disease. This was corroborated in the TCGA data (Figures S9A, B), where high NK score was also significantly enriched in those with MSI-associated clinical parameters such as tumour hypermutation (Figure S9D) and MLH1 silencing (Figure S9E).

With respect to CRC driver mutations, NK score was significantly enriched in those patients with BRAF mutant (Figure 7G) or KRAS wildtype (Figure 7H) genotypes, although there was no association with TP53 status (Figure S8D). There was no association between NK score and tumour characteristics such as tumour location (Figure S8B; distal versus proximal colon) and histological subtype (Figure S9F; mucinous vs non-mucinous). Similarly, no association was found between NK score and factors related to clinical history including adjuvant chemotherapy status (Figure S8C), prior CRC diagnosis (Figure S9G) or evidence of synchronous disease (Figure S9H). In sum, these results support the clinical utility of using this newly derived NK cell signature in the context of CRC.

Discussion

Bioinformatic approaches which allow for the deconvolution of immune cell subsets from bulk sequencing data have revolutionised our ability to assess the immune landscape of individual tumours. Transcriptomic signatures which predict intra-tumoral NK cell infiltration have been shown to indicate improved patient survival in various cancers (30, 31). However, despite accumulating evidence that NK cell load is prognostic in CRC, there are currently no extensively curated NK cell signatures explicitly designed for use in the context of CRC.

Here, through extensive computational curation of established NK cell-related genes with putative markers discovered through DE analysis of various bulk RNAseq and scRNAseq datasets, we define a comprehensive NK cell signature specific for CRC samples. Survival modelling in two large cohorts of primary CRC patients indicated that NK signature expression is associated with prolonged progression- and recurrence-free intervals, consistent with previous reports on the beneficial prognostic impact of NK cell infiltration in other solid cancers (57). That these metrics are associated with tumour progression (rather than overall survival) may relate to the fact that NK cells are believed to play a greater role in the suppression of metastases rather than the prevention of primary tumours (58, 59).

Although the NK cell signature presented herein is not the first of its kind, we believe that it is the most specific and the most extensively curated NK cell signature for the precise identification of the wide range of NK cell subsets in CRC samples due to its derivation in the context of this tumour type. Practical limitations have meant that most NK cell gene signatures, including that employed in the widely used CIBERSORT immune cell deconvolution tool, were curated using NK cells isolated from the peripheral blood. As NK cells exhibit tissue-specific phenotypes (40, 42, 60), and the tumour microenvironment is known to rewire NK cell transcriptional programs (60–62), it is difficult to assess how accurately these pre-existing NK cell signatures perform when extrapolated for use in the context of cancer. Moreover, striking differences in the transcriptomic profiles of circulating NK cells versus tumour-infiltrating NK cells from the same patients have also been reported (62), emphasising the importance of using scRNAseq data from tumour-infiltrating NK cells to faithfully derive a reference transcriptomic profile for CRC-associated NK cells.

Another issue facing pre-existing NK cell signatures has been a potential lack of specificity. For example, although a four-gene signature (composed of NCR1, PRF1, CX3CL1 and CX3CR1) was shown to accurately distinguish NK-high vs NK-low subgroups in clear cell renal cell carcinoma (31), these genes are ubiquitously expressed by many immune cell types. More broadly, the transcriptional programs of NK cells are highly overlapping with those of multiple T-cell subsets, particularly γδ-T-cells (63). Numerous studies have demonstrated that several archetypical NK cell markers including NKG2A (64), NCRs (65, 66) and various KIRs (67) are expressed by both traditional αβ- and unconventional γδ-T-cells. Our strategy of prioritising genes based on their NK-cell specificity rather than only those with the highest expression may therefore increase precision when teasing apart the NK cell contribution from complex mixtures of other immune cells.

In both the TCGA and GSE39582 cohorts investigated, NK scores were higher in patients with early-stage disease, aligning with results of a large, pan-cancer meta-analysis reporting that NK cell infiltration is lower in advanced-stage tumours (57). Similarly, high NK score was associated with several clinically useful molecular parameters such as MSI disease, CIMP positive status and BRAF mutation which have each been previously linked with high TIL and NK cell infiltration (68, 69). These features are all defining characteristics of the highly

immunogenic CMS1 molecular subtype of CRC (70). Moreover, DE analysis between the NK high- and low-scoring patients identified transcriptional and biological processes associated with high cytotoxic and chemokine activity in samples with high NK scores. As the signature efficiently discriminates murine NK cells from other immune cell types, this signature may also prove useful in prospective *in vivo* studies.

It remains unclear how optimally transcriptional signatures validated in a particular cancer will perform in alternate tumour types. Thus, future work may focus on cross-validating NK cell signatures in contexts outside their tumour-of-derivation, and/or on defining a pan-cancer NK cell signature which is less sensitive to subtype-specific influences. Prospective studies may also aim to dissect the transcriptional profile of intra-metastatic NK cells and determine whether, or in what ways, this diverges from that of primary CRC-infiltrating NK cells.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

CS, FH, and MF contributed to conception and design of the study. FH provided funding to support all of the study. CS sourced data and performed all computational analyses, with guidance from FH and MF. MF wrote sections of the codes used in the analysis. FH and MF supervised CS throughout the study. CS wrote the first draft of the manuscript. MF wrote sections of the manuscript. FH contributed to manuscript revision. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by The National Health and Medical Research Council of Australia (project grant 1164081, FH) and by the Tour de Cure Foundation (Senior research grant, FH).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.1011247/full#supplementary-material

References

1. Birenda K, Hwang JJ. Advances in immunotherapy in the treatment of colorectal cancer. *Am J Hematol/Oncol* (2017) 13:4–8.

2. Rodig SJ, Gusenleitner D, Jackson DG, Gjini E, Giobbie-Hurder A, Jin C, et al. MHC proteins confer differential sensitivity to CTLA-4 and PD-1 blockade in untreated metastatic melanoma. *Sci Trans Med* (2018) 10:3342. doi: 10.1126/scitranslmed.aar3342

3. Lee JH, Shklovskaya E, Lim SY, Carlino MS, Menzies AM, Stewart A, et al. Transcriptional downregulation of MHC class I and melanoma de- differentiation in resistance to PD-1 inhibition. *Nat Commun* (2020) 11:1897. doi: 10.1038/s41467-020-15726-7

4. Olson JA, Leveson-Gower DB, Gill S, Baker J, Beilhack A, Negrin RS. NK cells mediate reduction of GVHD by inhibiting activated, alloreactive T cells while retaining GVT effects. *Blood* (2010) 115:4293–301. doi: 10.1182/blood-2009-05-222190

5. Lundqvist A, McCoy JP, Samsel L, Childs R. Reduction of GVHD and enhanced antitumor effects after adoptive infusion of alloreactive Ly49-mismatched NK cells from MHC-matched donors. *Blood* (2007) 109:3603–6. doi: 10.1182/blood-2006-05-024315

6. Shembrey C, ND; H, Hollande F. Impact of tumor and immunological heterogeneity on the anti-cancer immune response. *Cancers (Basel)* (2019) 11:1217. doi: 10.3390/cancers11091217

7. Liu E, Tong Y, Dotti G, Shaim H, Savoldo B, Mukherjee M, et al. Cord blood NK cells engineered to express IL-15 and a CD19-targeted CAR show long-term persistence and potent antitumor activity. *Leukemia* (2018) 32:520–31. doi: 10.1038/leu.2017.226

8. Han J, Chu J, Keung Chan W, Zhang J, Wang Y, Cohen JB, et al. CAR-engineered NK cells targeting wild-type EGFR and EGFRvIII enhance killing of glioblastoma and patient-derived glioblastoma stem cells. *Sci Rep* (2015) 5:11483. doi: 10.1038/srep11483

9. Bagot M, Porcu P, Marie-Cardine A, Battistella M, William BM, Vermeer M, et al. IPH4102, a first-in-class anti-KIR3DL2 monoclonal antibody, in patients with relapsed or refractory cutaneous T-cell lymphoma: An international, first-in-human, open-label, phase 1 trial. *Lancet Oncol* (2019) 20:1160–70. doi: 10.1016/S1470-2045(19)30320-1

10. Vey N, Dumas P, Recher C, Gastaud L, Lioure B, Bulabois C, et al. Randomized phase 2 trial of lirilumab (anti-KIR monoclonal antibody, mAb) as

maintenance treatment in elderly patients (pts) with acute myeloid leukemia (AML): Results of the effikir trial. *Blood* (2017) 130:889. doi: 10.1182/blood.V130.Suppl_1.889.889

11. Cohen R, Fayette J, Posner M, Surgery G–. Phase II study of monalizumab, a first-in-class NKG2A monoclonal antibody, in combination with cetuximab in previously treated recurrent or metastatic. *Innate-PharmaCom* (2018). doi: 10.1158/1538-7445.AM2018-CT158

12. Imai K, Matsuyama S, Miyake S, Suga K, Nakachi K. Natural cytotoxic activity of peripheral-blood lymphocytes and cancer incidence: An 11-year follow-up study of a general population. *Lancet* (2000) 356:1795–9. doi: 10.1016/S0140-6736(00)03231-1

13. Hanna N, Burton RC. Definitive evidence that natural killer (NK) cells inhibit experimental tumor metastases in vivo. *J Immunol* (1981) 127:1754–8.

14. Sathe P, Delconte RB, Souza-Fonseca-Guimaraes F, Seillet C, Chopin M, Vandenberg CJ, et al. Innate immunodeficiency following genetic ablation of Mcl1 in natural killer cells. *Nat Commun* (2014) 5:4539. doi: 10.1038/ncomms5539

15. Coca S, Perez-Piqueras J, Martinez D, Colmenarejo A, Saez MA, Vallejo C, et al. The prognostic significance of intratumoral natural killer cells in patients with colorectal carcinoma. *Cancer* (1997) 79:2320–8. doi: 10.1002/(SICI)1097-0142(19970615)79:12<2320::AID-CNCR5>3.0.CO;2-P

16. Kondo E, Koda K, Takiguchi N, Oda K, Seike K, Ishizuka M, et al. Preoperative natural killer cell activity as a prognostic factor for distant metastasis following surgery for colon cancer. *Digestive Surgery.* (2003) 20:445–51. doi: 10.1159/000072714

17. Donadon M, Hudspeth K, Cimino M, Di Tommaso L, Preti M, Tentorio P, et al. Increased infiltration of natural killer and T cells in colorectal liver metastases improves patient overall survival. *J Gastrointestinal Surg* (2017) 21:1–11. doi: 10.1007/s11605-017-3446-6

18. Kärre K, Ljunggren HG, Piontek G, Kiessling R. Selective rejection of h-2-deficient lymphoma variants suggests alternative immune defence strategy. *Nature* (1986) 319:675–8. doi: 10.1038/319675a0

19. Karlhofer FM, Ribaudo RK, Yokoyama WM. MHC class I alloantigen specificity of ly-49+ IL-2-activated natural killer cells. *Nature* (1992) 358:66–70. doi: 10.1038/358066a0

20. Dijkstra KK, Cattaneo CM, Weeber F, Chalabi M, van de Haar J, Fanchi LF, et al. Generation of tumor-reactive T cells by Co-culture of peripheral blood lymphocytes and tumor organoids. *Cell* (2018) 174:1586–1598.e12. doi: 10.1016/j.cell.2018.07.009

21. Dierssen JWF, de Miranda NFCC, Ferrone S, van Puijenbroek M, Cornelisse CJ, Fleuren GJ, et al. HNPCC versus sporadic microsatellite-unstable colon cancers follow different routes toward loss of HLA class I expression. *BMC Cancer* (2007) 7:33. doi: 10.1186/1471-2407-7-33

22. Halama N, Braun M, Kahlert C, Spille A, Quack C, Rahbari N, et al. Natural killer cells are scarce in colorectal carcinoma tissue despite high levels of chemokines and cytokines. *Clin Cancer Res* (2011) 17:678–89. doi: 10.1158/1078-0432.CCR-10-2173

23. Tan IB, Tan P. Genetics: An 18-gene signature (ColoPrint®) for colon cancer prognosis. *Nat Rev Clin Oncol* (2011) 8:131–3. doi: 10.1038/nrclinonc.2010.229

24. Chen H, Sun X, Ge W, Qian Y, Bai R, Zheng S. A seven-gene signature predicts overall survival of patients with colorectal cancer. *Oncotarget* (2017) 8:95054–65. doi: 10.18632/oncotarget.10982

25. Gao S, Tibiche C, Zou J, Zaman N, Trifiro M, O'Connor-McCourt M, et al. Identification and construction of combinatory cancer hallmark-based gene signature sets to predict recurrence and chemotherapy benefit in stage II colorectal cancer. *JAMA Oncol* (2016) 2:37–45. doi: 10.1001/jamaoncol.2015.3413

26. Tong M, Zheng W, Li H, Li X, Ao L, Shen Y, et al. Multi-omics landscapes of colorectal cancer subtypes discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. *Oncogenesis* (2016) 5:e242–2. doi: 10.1038/oncsis.2016.51

27. Cao B, Luo L, Feng L, Ma S, Chen T, Ren Y, et al. A network-based predictive gene-expression signature for adjuvant chemotherapy benefit in stage II colorectal cancer. *BMC Cancer* (2017) 17:844. doi: 10.1186/s12885-017-3821-4

28. Watanabe T, Kobunai T, Yamamoto Y, Matsuda K, Ishihara S, Nozawa K, et al. Gene expression signature and response to the use of leucovorin, fluorouracil and oxaliplatin in colorectal cancer patients. *Clin Trans Oncol* (2011) 13:419–25. doi: 10.1007/s12094-011-0676-z

29. Costanzo MC, Kim D, Creegan M, Lal KG, Ake JA, Currier JR, et al. Transcriptomic signatures of NK cells suggest impaired responsiveness in HIV-1 infection and increased activity post-vaccination. *Nat Commun* (2018) 9:1–16. doi: 10.1038/s41467-018-03618-w

30. Cursons J, Souza-Fonseca Guimaraes F, Foroutan M, Anderson A, Hollande F, Hediyeh-Zadeh S, et al. A gene signature predicting natural killer cell infiltration and improved survival in melanoma patients. *Cancer Immunol Res* (2019) 7 (7):1162–74. doi: 10.1158/2326-6066.CIR-18-0500

31. Eckl J, Buchner A, Prinz PU, Riesenberg R, Siegert SI, Kammerer R, et al. Transcript signature predicts tissue NK cell content and defines renal cell carcinoma subgroups independent of TNM staging. *J Mol Med* (2012) 90:55–66. doi: 10.1007/s00109-011-0806-7

32. Foroutan M, Molania R, Pfefferle A, Behrenbruch C, Scheer S, Kallies A, et al. The ratio of exhausted to resident infiltrating lymphocytes is prognostic for colorectal cancer patient outcome. *Cancer Immunol Res* (2021) 9:1125–40. doi: 10.1158/2326-6066.CIR-21-0137

33. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* (2015) 12:453–7. doi: 10.1038/nmeth.3337

34. Wang J, Mouradov D, Wang X, Jorissen RN, Chambers MC, Zimmerman LJ, et al. Colorectal cancer cell line proteomes are representative of primary tumors and predict drug sensitivity. *Gastroenterology* (2017) 153:1082–95. doi: 10.1053/j.gastro.2017.06.008

35. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PloS One* (2014) 9(10):e109760. doi: 10.1371/journal.pone.0109760

36. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep* (2019) 26:1627–1640.e7. doi: 10.1016/j.celrep.2019.01.041

37. Zhang L, Li Z, Skrzypczynska KM, Fang Q, Zhang W, O'Brien SA, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* (2020) 181:442–59.e29. doi: 10.1016/j.cell.2020.03.048

38. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012) 483:603–7. doi: 10.1038/nature11003

39. Calon A, Espinet E, Palomo-Ponce S, Tauriello DVF, Iglesias M, Céspedes MV, et al. Dependency of colorectal cancer on a TGF-β-Driven program in stromal cells for metastasis initiation. *Cancer Cell* (2012) 22:571–84. doi: 10.1016/j.ccr.2012.08.013

40. Dogra P, Rancan C, Ma W, Toth M, Senda T, Carpenter DJ, et al. Tissue determinants of human NK cell development, function, and residence. *Cell* (2020) 180:749–63.e13. doi: 10.1016/j.cell.2020.01.022

41. Fehniger TA, Cai SF, Cao X, Bredemeyer AJ, Presti RM, French ARR, et al. Acquisition of murine NK cell cytotoxicity requires the translation of a pre-existing pool of granzyme b and perforin mRNAs. *Immunity* (2007) 26:798–811. doi: 10.1016/j.immuni.2007.04.010

42. Pfefferle A, Netskar H, Ask EH, Lorenz S, Goodridge JP, Sohlberg E, et al. A temporal transcriptional map of human natural killer cell differentiation. *bioRxiv* (2019), 630657. doi: 10.1101/630657

43. Lima M, Leander M, Santos M, Santos AH, Lau C, Queirós ML, et al. Chemokine receptor expression on normal blood CD56+ NK-cells elucidates cell partners that comigrate during the innate and adaptive immune responses and identifies a transitional NK-cell population. *J Immunol Res* (2015) 2015:839684. doi: 10.1155/2015/839684

44. Yang C, Siebert JR, Burns R, Gerbec ZJ, Bonacci B, Rymaszewski A, et al. Heterogeneity of human bone marrow and blood natural killer cells defined by single-cell transcriptome. *Nat Commun* (2019) 10(1):3931. doi: 10.1038/s41467-019-11947-7

45. Smith SL, Kennedy PR, Stacey KB, Worboys JD, Yarwood A, Seo S, et al. Diversity of peripheral blood human NK cells identified by single cell RNA sequencing. *Blood Adv* (2019) 4:1388–406. doi: 10.1182/bloodadvances.2019000699

46. Yu J, Mao HC, Wei M, Hughes T, Zhang J, Park IK, et al. CD94 surface density identifies a functional intermediary between the CD56bright and CD56dim human NK-cell subsets. *Blood* (2010) 115:274–81. doi: 10.1182/blood-2009-04-215491

47. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): Immune-specific genes identified from a compendium of microarray expression data. *Genes Immunity* (2005) 6:319–31. doi: 10.1038/sj.gene.6364173

48. Crinier A, Milpied P, Escalière B, Piperoglou C, Galluso J, Balsamo A, et al. High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice. *Immunity* (2018) 49:971–986.e5. doi: 10.1016/j.immuni.2018.09.009

49. Bezman NA, Kim CC, Sun JC, Min-Oo G, Hendricks DW, Kamimura Y, et al. Molecular definition of the identity and activation of natural killer cells. *Nat Immunol* (2012) 13:1000–8. doi: 10.1038/ni.2395

50. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. *BMC Bioinf* (2018) 19:1–10. doi: 10.1186/s12859-018-2435-4

51. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PloS Med* (2013) 10:e1001453. doi: 10.1371/journal.pmed.1001453

52. Hu X, Zhang L, Li Y, Ma X, Dai W, Gao X, et al. Organoid modelling identifies that DACH1 functions as a tumour promoter in colorectal cancer by modulating BMP signalling. *EBioMedicine* (2020) 56:102800. doi: 10.1016/j.ebiom.2020.102800

53. Wang X, Chen J, Wang J, Yu F, Zhao S, Zhang Y, et al. Metalloproteases meprin-α (MEP1A) is a prognostic biomarker and promotes proliferation and invasion of colorectal cancer. *BMC Cancer* (2016) 16:383. doi: 10.1186/s12885-016-2460-5

54. Zeng J, Yang X, Cheng L, Liu R, Lei Y, Dong D, et al. Chemokine CXCL14 is associated with prognosis in patients with colorectal carcinoma after curative resection. *J Trans Med* (2013) 11:6. doi: 10.1186/1479-5876-11-6

55. Westrich JA, Vermeer DW, Silva A, Bonney S, Berger JN, Cicchini L, et al. CXCL14 suppresses human papillomavirus-associated head and neck cancer through antigen-specific CD8+ T-cell responses by upregulating MHC-I expression. *Oncogene* (2019) 38:7166–80. doi: 10.1038/s41388-019-0911-6

56. Bucks CM, Norton JA, Boesteanu AC, Mueller YM, Katsikis PD. Chronic antigen stimulation alone is sufficient to drive CD8 + T cell exhaustion. *J Immunol* (2009) 182:6697–708. doi: 10.4049/jimmunol.0800997

57. Nersesian S, Schwartz SL, Grantham SR, MacLean LK, Lee SN, Pugh-Toole M, et al. NK cell infiltration is associated with improved overall survival in solid cancers: A systematic review and meta-analysis. *Trans Oncol* (2021) 14(1):100930. doi: 10.1016/j.tranon.2020.100930

58. Malaise M, Rovira J, Renner P, Eggenhofer E, Sabet-Baktach M, Lantow M, et al. KLRG1+ NK cells protect T-bet-Deficient mice from pulmonary metastatic colorectal carcinoma. *J Immunol* (2014) 192:1954–61. doi: 10.4049/jimmunol.1300876

59. Takeda K, Hayakawa Y, Smyth MJ, Kayagaki N, Yamaguchi N, Kakuta S, et al. Involvement of tumor necrosis factor-related apoptosis-inducing ligand in surveillance of tumor metastasis by liver natural killer cells. *Nat Med* (2001) 7:94–100. doi: 10.1038/83416

60. Platonova S, Cherfils-Vicini J, Damotte D, Crozet L, Vieillard V, Validire P, et al. Profound coordinated alterations of intratumoral NK cell phenotype and function in lung carcinoma. *Cancer Res* (2011) 71:5412–22. doi: 10.1158/0008-5472.CAN-10-4179

61. Gillard-Bocquet M, Caer C, Cagnard N, Crozet L, Perez M, Fridman WH, et al. Lung tumor microenvironment induces specific gene expression signature in intratumoral NK cells. *Front Immunol* (2013) 4:19. doi: 10.3389/fimmu.2013.00019

62. De Andrade LF, Lu Y, Luoma A, Ito Y, Pan D, Pyrdol JW, et al. Discovery of specialized NK cell populations infiltrating human melanoma metastases. *JCI Insight* (2019) 4(23):e133103. doi: 10.1172/jci.insight.133103

63. Pizzolato G, Kaminski H, Tosolini M, Franchini DM, Pont F, Martins F, et al. Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVδ1 and TCRVδ2 γδ T lymphocytes. *Proc Natl Acad Sci USA* (2019) 116:11906–15. doi: 10.1073/pnas.1818488116

64. D'Ombrain MC, Hansen DS, Simpson KM, Schofield L. γδ-T cells expressing NK receptors predominate over NK cells and conventional T cells in the innate IFN-γ response to plasmodium falciparum malaria. *Eur J Immunol* (2007) 37:1864–73. doi: 10.1002/eji.200636889

65. Von Lilienfeld-Toal M, Nattermann J, Feldmann G, Sievers E, Frank S, Strehl J, et al. Activated γδ T cells express the natural cytotoxicity receptor natural killer p44 and show cytotoxic activity against myeloma cells. *Clin Exp Immunol* (2006) 144:528–33. doi: 10.1111/j.1365-2249.2006.03078.x

66. Correia DV, Fogli M, Hudspeth K, Gomes Da Silva M, Mavilio D, Silva-Santos B. Differentiation of human peripheral blood Vδ1+ T cells expressing the natural cytotoxicity receptor NKp30 for recognition of lymphoid leukemia cells. *Blood* (2011) 118:992–1001. doi: 10.1182/blood-2011-02-339135

67. Miyazaki K, Yamaguchi M, Imai H, Kobayashi T, Tamaru S, Nishii K, et al. Gene expression profiling of peripheral T-cell lymphoma including {gamma}{delta} T-cell lymphoma. *Blood* (2009) 113:1071–4. doi: 10.1182/blood-2008-07-166363

68. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol* (2015) 16:64. doi: 10.1186/s13059-015-0620-6

69. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus immunoscore for the classification of colon cancer: A prognostic and accuracy study. *Lancet* (2018) 391:2128–39. doi: 10.1016/S0140-6736(18)30789-X

70. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* (2015) 21:1350–6. doi: 10.1038/nm.3967

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fonc.2022.1091767

Check for updates

OPEN ACCESS

EDITED BY
Jinzhuang Dou,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Shaolong Cao,
Biogen Idec, United States
Minglei Yang,
Department of Medical Genetics, Sun Yat-sen University, China

*CORRESPONDENCE
Zaixiang Tang
tangzx@suda.edu.cn
Ke Lu
sgu8434@sina.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

RECEIVED 07 November 2022
ACCEPTED 19 December 2022
PUBLISHED 10 January 2023

CITATION
Shen J, Li H, Yu X, Bai L, Dong Y, Cao J, Lu K and Tang Z (2023) Efficient feature extraction from highly sparse binary genotype data for cancer prognosis prediction using an auto-encoder. *Front. Oncol.* 12:1091767. doi: 10.3389/fonc.2022.1091767

COPYRIGHT
© 2023 Shen, Li, Yu, Bai, Dong, Cao, Lu and Tang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Efficient feature extraction from highly sparse binary genotype data for cancer prognosis prediction using an auto-encoder

Junjie Shen[1,2], Huijun Li[1,2], Xinghao Yu[2,3], Lu Bai[1,2], Yongfei Dong[1,2], Jianping Cao[4], Ke Lu[5*†] and Zaixiang Tang[1,2*†]

[1]Department of Biostatistics, School of Public Health, Medical College of Soochow University, Suzhou, China, [2]Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, Medical College of Soochow University, Suzhou, China, [3]Center for Genetic Epidemiology and Genomics, School of Public Health, Medical College of Soochow University, Suzhou, China, [4]School of Radiation Medicine and Protection and Collaborative Innovation Center of Radiation Medicine of Jiangsu Higher Education Institutions, Soochow University, Suzhou, China, [5]Department of Orthopedics, Affiliated Kunshan Hospital of Jiangsu University, Suzhou, China

Genomics involving tens of thousands of genes is a complex system determining phenotype. An interesting and vital issue is how to integrate highly sparse genetic genomics data with a mass of minor effects into a prediction model for improving prediction power. We find that the deep learning method can work well to extract features by transforming highly sparse dichotomous data to lower-dimensional continuous data in a non-linear way. This may provide benefits in risk prediction-associated genotype data. We developed a multi-stage strategy to extract information from highly sparse binary genotype data and applied it for cancer prognosis. Specifically, we first reduced the size of binary biomarkers *via* a univariable regression model to a moderate size. Then, a trainable auto-encoder was used to learn compact features from the reduced data. Next, we performed a LASSO problem process to select the optimal combination of extracted features. Lastly, we applied such feature combination to real cancer prognostic models and evaluated the raw predictive effect of the models. The results indicated that these compressed transformation features could better improve the model's original predictive performance and might avoid an overfitting problem. This idea may be enlightening for everyone involved in cancer research, risk reduction, treatment, and patient care *via* integrating genomics data.

KEYWORDS

auto-encoder, highly sparse binary data, feature extraction, risk prediction, LASSO

1 Introduction

Modern omics technologies can generate large-scale molecular data, such as genomic, transcriptomic, proteomic, and metabolomic data, inducing the opportunity to build more accurate predictive and prognostic models (1, 2). These data have been used to provide tailored healthcare and precision medicine for many individuals (3). However, such data also present computational and statistical challenges because the complexity of the algorithms grows fast with the number of variables.

The underlying representation of many real processes is often sparse. It is of benefit to be able to efficiently eliminate features in a pre-processing step. From the perspective of data dimension reduction, it can be classified into feature selection and feature extraction. Most existing work on feature selection are based on a variant of $l1$-norm penalty due to its sparsity-induced property, strong theoretical guarantees, and great empirical success in kinds of applications (4). The paper about the least absolute shrinkage and selection operator (LASSO) has had an enormous influence (5).

Count data are ubiquitous in genetic risk studies, where it is highly possible to observe excessive zero counts in rare mutation loci. In the face of mass mutation loci, many penalty methods have been adopted in GWAS analyses to select key genetic loci (6–8). For example, Yang et al. detected genetic risk factors among millions of single-nucleotide polymorphisms (SNPs) in ADNI whole genome sequencing data *via* the LASSO method along with the EDPP screening rules (9). Another solution lies in reducing the number of markers before employing a shrinkage method in genetic model such as (10). "Clumping and thresholding" is a two-step method that is often used to derive polygenic risk score (PRS) from results of GWAS studies (11).

Genetic variation is considered associated with cancer prognosis. However, there is little literature on the use of genetic omics data to predict cancer outcomes. As a matter of fact, it is well documented that a large number of genetic markers and generally the small size of their effects make much of the heritability hidden, as a mass of variants with weak effects on disease usually fail to reach the prespecified thresholds of significance (12). It is always an interesting issue how to aggregate these small effects. To better utilize big data in reasoning systems, feature extraction rather than feature selection may allow for discovery of new pathways and principles (13). We identified the auto-encoder as a promising tool. The auto-encoder is a derivative of artificial neural networks (ANNs), with the aim of learning compact and efficient representations from the input data (14). Usually, these representations have a much lower dimension. Departing from supervised ANNs whose performance depends on the quality of gold standards, the auto-encoder directly uses unlabeled data, i.e., the input data itself is the target of reconstruction. Compared to commonly used feature extraction approaches like principal component analysis or independent component analysis that linearly map input to features, the auto-encoder extracts features into non-linear space and work much better as a tool to reduce dimensionality of data (13).

To sum up, we identified that the auto-encoder could learn compact and efficient features from highly sparse binary data and accordingly developed a multiple-stage process to extract information from binary genotype data and applied it for cancer prognosis. In the first stage (screening), we reduced the number of markers *via* a univariable regression model to a moderate size. In the second stage (extracting), we used a trainable auto-encoder to extract representations from the reduced data. In the third stage (selecting), we performed a LASSO process over a grid of tuning parameter values to select the optimal combination of the extracted features. Finally, we applied such feature combination to cancer prognostic models, and evaluated the raw predictive effect of the models.

Abbreviations: LASSO, least absolute shrinkage and selection operator; GWAS, genome-wide association; SNPs, single-nucleotide polymorphisms; ADNI, Alzheimer's Disease Neuroimaging Initiative; EDPP, enhanced dual polytope projections; PRS, polygenic risk score; ANNs, artificial neural networks; ReLU, rectified linear unit; SNR, signal-to-noise ratio; MNIST, Mixed National Institute of Standards and Technology; MSE, mean squared error; MCE, mean cross-entropy; TCGA, The Cancer Genome Atlas; NCI, National Cancer Institute; NHGRI, National Human Genome Research Institute; SNV, simple nucleotide variation; GDC, Genomic Data Commons; BRCA, breast cancer; OV, ovary cancer; OS, overall survival; RSF, random survival forest; ROC, receiver operating characteristic; AUC, area under the curve; CI, credible interval; CNV, copy number variation; COSMIC, catalogue of somatic mutations in cancer; lncRNA, long noncoding RNA; fMRI, functional magnetic resonance imaging.

2 Materials and methods

2.1 The construction of auto-encoders

A simple auto-encoder is much similar to the ANNs, which generally contains three layers: an input layer, a hidden layer, and a reconstructed layer (output layer) (15). The hidden layer corresponds to the constructed features, with each neuron node representing one feature. The reconstructed layer and the input layer had the same dimensions, and the objective optimized function for the algorithm was to minimize the difference between the two layers.

Let us recall the traditional auto-encoder model proposed by Bengio et al. (16). As many machine learning methods do, we first normalize the continuous input data by the formula $(x - x_{min})/(x_{max} - x_{min})$. Thus, an auto-encoder with "p" features takes an input vector $\mathbf{x}$ in $[0, 1]^p$. The hidden layer representation $\mathbf{y}$ with "d" dimension is constructed through a

deterministic mapping $\mathbf{y} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b})$, parameterized by $\theta = \{\mathbf{W}, \mathbf{b}\}$. **W** is a "p × d" weight matrix and **b** is a bias vector. Function $s(x)$ is called activation function, which introduces nonlinear properties into the network. Common activation functions include (1) rectified linear unit (ReLU) function and (2) sigmoid function:

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

$$g(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

Equation (1) maps a linear set of input values to an interval ranging from $[0, \infty)$ and equation (2) maps a linear set of input values to an interval in [0, 1]. The value contained in the latent representation **y** for each neuron node is termed the activity value. Then, the resulting hidden layer **y** is mapped back to a "reconstructed" vector **z** in $[0, 1]^p$ in a similar manner, by inputting space $\mathbf{z} = g_{\theta'}(\mathbf{y}) = h(\mathbf{W'y} + \mathbf{b'})$ with $\theta' = \{\mathbf{W'}, \mathbf{b'}\}$. The function $h(x)$ is also an activation function, restoring the latent information to the original information. We could use tied weights if the two activation functions are the same, which means that the transpose of **W** was used for **W'**. The parameters in this neural network are optimized to minimize the average reconstruction loss between the input layer **x** and the reconstructed layer **z**:

$$\theta,\ \theta' = \underset{\theta,\ \theta'}{\arg\min}\ 1/n \sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) \quad (3)$$

where n is the sample size and L is a loss function like squared error loss function $L(\mathbf{x}, \mathbf{z}) = ||\mathbf{x} - \mathbf{z}||^2$. An alternative error loss, cross-entropy loss function, is suggested by the interpretation of **x** and **z** as vectors of bit probabilities:

$$L_H(\mathbf{x},\ \mathbf{z}) = -\sum_{k=1}^{p} [\mathbf{x}_k \log \mathbf{z}_k + (1 - \mathbf{x}_k)\log(1 - \mathbf{z}_k)] \quad (4)$$

Like other feed-forward ANNs, the auto-encoder takes back propagation algorithm and gradient descent algorithm to compute and update target parameters iteratively until reaching an acceptable loss or the given epochs. The specific theory can be referred to the relevant literature (17).

2.2 The LASSO and its selection rules

Given a linear regression with standardized predictors x_{ij} and centered response values y_i for i = 1, 2,…, N (samples) and j = 1, 2,…, p (features), the LASSO solves the $l1$-penalized regression problem for finding $\beta = \{\beta_j\}$ to minimize

$$\sum_{i=1}^{N} (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

where $\lambda \geq 0$ is a tuning parameter.

A main reason for using the LASSO is that the $l1$-penalty tends to set some entries of $\tilde{\beta}$ to 0, and therefore, it performs a kind of variable selection. Furthermore, Tibshirani (18) proposed "strong rules" to discard noise signal in the LASSO-type penalty problems. The results indicated that the LASSO performs well in both low signal-to-noise ratio (SNR) and high sparse regimes by incorporating the "strong rules". However, the predictor matrices from their simulated studies were all generated from Gaussian distribution. Subsequent simulation studies that aimed to improve variable selection algorithm using a LASSO-type penalty still concerned continuous predictors mainly (19–21). Guo et al. considered the power of the LASSO for SNP selection in predicting quantitative traits and proved that the LASSO still has good selection ability for high-dimensional and sparse binary predictors (22). However, when the values of these binary predictors become highly sparse (rare mutation) such as 99.9% of zeros and 0.01% of ones, we observed that the power of the LASSO to select non-zero variables declined. This is briefly illustrated in supplementary file part II and Table S1.

2.3 The property of the auto-encoder to feature selection

We explore the feature extraction capability of the auto-encoder using two visualized image datasets from the Mixed National Institute of Standards and Technology database (MNIST) (23) and fashion MNIST. The MNIST is one of the most widely used benchmark dataset for isolated handwritten digit recognition from 0 to 9. Digits are transformed to 28×28 images, and represented as 784×1 vectors. Each component is a number between 0 and 255, which means the gray levels of each pixel. The number of zeros accounts for about 81%. It has a training set of 60,000 examples, and a test set of 10,000 examples. The fashion MNIST is a substitute for the MNIST dataset and is more complex, consisting of 10 types of wearing images. The number of 0 accounts for about 51%. The above datasets are loaded and accessed through the "Keras" module of TensorFlow. The deep learning framework of the auto-encoder is constructed by the TensorFlow library (2.3.0) of Python (3.7) in the Jupyter Notebook platform (6.3.0).

2.3.1 Handwritten digit recognition

We took the first 1,000 examples of training set as training data and the first 1,000 examples of test set as testing data from the MNIST to study the property of our auto-encoder. First, as mentioned above, we reshaped the 28×28 images to 784×1 vectors and normalized the input data from [0, 255] to [0, 1]. Thus, the dimension of input layer as well as reconstructed layer was 784. We set the hidden layer dimension to 100 (this number is optional). See construction of the auto-encoder in Figure S1.

Activation function $s(x)$ was specified to the ReLU function due to its good property and therefore the activity values in the hidden layer **y** ranging from $[0, \infty)$. The activation function $h(x)$ could be either ReLU function or sigmoid function, corresponding to mean squared error (MSE) loss and mean cross-entropy (MCE) loss. We used the two activation functions respectively and compared the fitting effects.

In terms of configuration training method, we used the "Adam" optimizer from the "Keras" module. The size of each update is controlled by learning rate. To speed up the training, samples were randomly grouped into batches, and the number of samples contained in a batch was termed the batch size, with weight and bias being updated after each batch. Training proceeded through epochs, and samples were re-batched at the beginning of each epoch. Training was stopped after a specified number of epochs (termed epoch size) was reached. We performed a full factorial design over all combinations of the following parameters: a learning rate of 0.001, 0.005, and 0.010; a batch size of 32, 64, and 128; and an epoch size of 50, 100, and 150. After a full factorial parameter sweep, the parameters that we selected were as follows: a learning rate of 0.005, a batch size of 128, and an epoch size of 100, which could achieve fast training speed and smooth loss.

When using the sigmoid function as activation function $h(x)$, the MCE was 0.0683 with a binary accuracy (calculates how often predictions matches labels) of 0.8156 in the training data (see Figure S2A) and 0.0898 MCE with 0.8244 accuracy in the testing data using the model built in training data. We read the first five images of the training data and testing data, as shown in Figures S3A, B. The first row shows the original images, the second row shows the extracted features, and the third row shows that the images were restored accurately with the extracted features. The results show that the model can be used to extract the key features well. Meanwhile, we used the reconstructed data for handwritten digit prediction and found that the probability of predicting the correct classification was close to 1 (see Table S2).

While using the ReLU function as activation function $h(x)$, the MSE was 0.0067 with an accuracy (calculates how often predictions matches labels) of 0.0150 in the training data (see Figure S2B) and 0.0125 MSE with an accuracy 0.0200 in the testing data using the same model. We also read the first five images of the training data and testing data (Figures S3C, D). It shows that the ReLU function performed quite poorer compared to sigmoid function. Because the labels of corresponding output data are normalized data ranging from [0, 1], sigmoid function could work more suitably.

2.3.2 Fashion image recognition

We took the same procedure as *Section 2.3.1* in fashion MNIST data. We selected the first 1,000 examples of training set as training data. The activation function $h(x)$ was directly specified to sigmoid function. We set the same configuration training method except for an epoch size of 200. The MCE was 0.2667 with a binary accuracy of 0.5166 in the training data (see Figure S4A). We read the first six images of the training data, as shown in Figure S5A. We found that the fitting effect was poorer in the fashion MNIST data than in the MNIST data, because the proportion of zeros is lower in the fashion MNIST data (about 51%) than the MNIST data (about 81%).

Inspired by denoising auto-encoders (24), we artificially added some corruption to training data. Specifically, we set values below 0.21 to zeros in the input data, making the proportion of zeros up to about 58.5%. Then, we retrained the model; the MCE was 0.2440 with an accuracy of 0.5924 in the new (corrupted) training data (see Figure S4B). The first six images of the new training data are shown in Figure S5B. The black icon became a little clearer (e.g., the second on the left, the first on the right). Images before and after the corruption are shown in Figure S5C. The first and third images were before the corruption, and the second and fourth images were after the corruption. Our results show that the higher the proportion of 0 and 1, the better the feature extraction effect of the auto-encoder using the sigmoid function.

2.3.3 Auto-encoder feature selection for highly sparse binary predictors

We used the auto-encoder to extract features from the highly sparse binary data. We randomly used simulation data generated from scenario 5 in Table S1. The sample size was 200 with 400 binary predictors. Thus, in the testing auto-coder, the dimension of the input layer as well as the reconstructed layer was 400. We set the hidden layer dimension to 100, i.e., extracting 100 important features. We used the "Adam" optimizer, and the parameters that we selected were as follows: a learning rate of 0.005, a batch size of 32, and an epoch size of 200. The activation function $h(x)$ was set to sigmoid function.

As a result, the MCE was 0.0001 with a binary accuracy of 1.0000 (see Figure S6A). We read the first five "images" of this simulated data, as shown in Figure S6B. The auto-encoder could recover the scattered genetic signals and when there was no genetic signal in the sample, an identical noise signal was given. The extracted 100 signal features were then used in LASSO Cox regression, and 9 features were selected. We calculated Harrell's concordance index (C-index) with 0.670 (standard error, SE = 0.035) and the R^2 was 0.215. If the LASSO Cox regression were applied directly using 400 binary predictors, a total of 65 predictors were selected (of which 5 were real nonzero predictors). The C-index was 0.721 (SE = 0.030) and the R^2 was 0.379. The result obtained using the auto-encoder was much more close to the performance of scenario 1 in Table S1 (average C-index: 0.647, average R^2: 0.244). Due to the selection of more noise predictors, using the LASSO directly had a virtual height of C-index and R^2 that would induce overfitting.

3 Cancer prognosis application

The Cancer Genome Atlas (TCGA) project was started in 2006 by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The database has contained a variety of cancer data from more than 20,000 samples of 33 types of cancer, including transcriptome expression data, genomic variation data, methylation data, and clinical data. As the largest cancer gene database, TCGA has become the first choice for cancer research due to its large sample size, diverse data types and standardized data formats.

We downloaded the latest (in July 2022) single-nucleotide variation (SNV) data and phenotype data of the GDC TCGA Breast Cancer (BRCA) cohort (female) and GDC TCGA Ovary Cancer (OV) cohort from the official website "GDC Data Portal". A total of 977 SNV documents and 1,085 phenotype documents were obtained from BRCA and 480 SNV documents and 597 phenotype documents were obtained from OV. The data type of SNV is masked somatic mutation, read and collated by R package *mafTools*. The overview of SNV in BRCA and OV is shown in Figure S7. We eliminated data with variants that were nonsense mutation. Next, we used the R package *reshape2* to reshape the mutation data to count how many SNV mutations were present in each gene per patient. Zero means wild type, and one means mutated (genotype data with 0/1 values). The interested phenotype in this study was overall survival (OS).

3.1 BRCA data

There were a total of 66,780 SNV items in which 4,910 were nonsense mutation. Many genes had more than one mutation, but we deemed all of them as "mutated" and were labeled "1". A total of 952 BRCA patients with 15,124 genotype data were available. After merging survival data, those with missing survival data were eliminated and 939 subjects were left. Univariable Cox analysis was performed on these 15,124 genotype data as preliminary screening to identify potential contributors, and 1,936 of them with a p-value of less than 0.05 (a rough threshold) were selected for subsequent analysis. We found that if the LASSO Cox regression were applied directly to these 1,936 genotype data, no variables would be selected by LASSO (see Figure S8). This was possible in such a scenario because the proportion of zeros reaches 99.6%. Thus, we thought of using an auto-coder to extract features from these highly sparse binary variables. We also consider random survival forest (RSF) as an alternative to screen the key variables because the random forest method is employed to detect significant SNPs in large-scale GWAS (25).

3.1.1 Feature extraction using an auto-encoder and the development of the prognosis model

Specifically, in our BRCA auto-encoder, the dimension of the input layer as well as the reconstructed layer was 1,936. We set the hidden layer dimension to 100, i.e., extracting 100 important features. Figure 1A shows the construction of the auto-encoder. We used the "Adam" optimizer; the parameters that we selected were as follows: a learning rate of 0.005, a batch size of 32, and an epoch size of 150. The activation function h(x) was set to sigmoid function with MCE loss.

As a result, the MCE was 0.0006 with a binary accuracy of 1.0000 (Figure 1B). We read the first five "images" of these data, as shown in Figure 1C. The auto-encoder could recover the scattered genetic signals well as expected. The extracted 100 signal features were continuous variables (see Table S3 for example) and then thrown into the LASSO Cox regression. Finally, 25 features were selected (see Figure 2). We build a prognosis signature called SNV signature based on these 25 features using the R functions "predict()", "cph()", and "coxph ()" among BRCA patients. The mean C-index of this signature was 0.830 (SE = 0.069), and the mean R^2 was 0.245, which was performed with a fivefold cross-validation process and stepAIC to avoid overfitting.

We used this signature to divide the population into two groups. The optimal cutoff value of the signature was determined using the R package *survminer*. The R package *survival* was used to perform survival analysis between these two groups. The Kaplan–Meier (K-M) curve was used to show difference of survival curves between groups (discrimination). Log-rank test was used to evaluate statistical differences of the survival. The receiver operating characteristic (ROC) curve and its area under the curve (AUC) values were utilized to evaluate the specificity and sensitivity of the signature in a time-dependent manner using the R package *timeROC*. We drew observed survival curves and predicted survival curves to compare the agreement (calibration), by calculating baseline hazard using the R function "basehaz()". We also assessed calibration with calibration plots. A 45°C diagonal line represents perfect calibration, while deviation below or above this line implies overestimation or underestimation of survival.

SNV signature ranging from (−3.564, 6.445) with a mean of 0. Patients were divided into a low-risk group (n = 820) and a high-risk group (n = 119); optimal cutoff value was 1.243 (see Figure 3A). The low-risk group had a much higher survival rate compared to the high-risk group (p < 0.0001). The 8-year survival rate of the low-risk group was over 0.75, whereas that of the high-risk group was almost 0. The time-dependent AUC curve was approximately 0.9 during 8 years (Figure 3B). The 2-, 5-, and 8-year AUC of the signature were 0.928 (95% CI: 0.870–

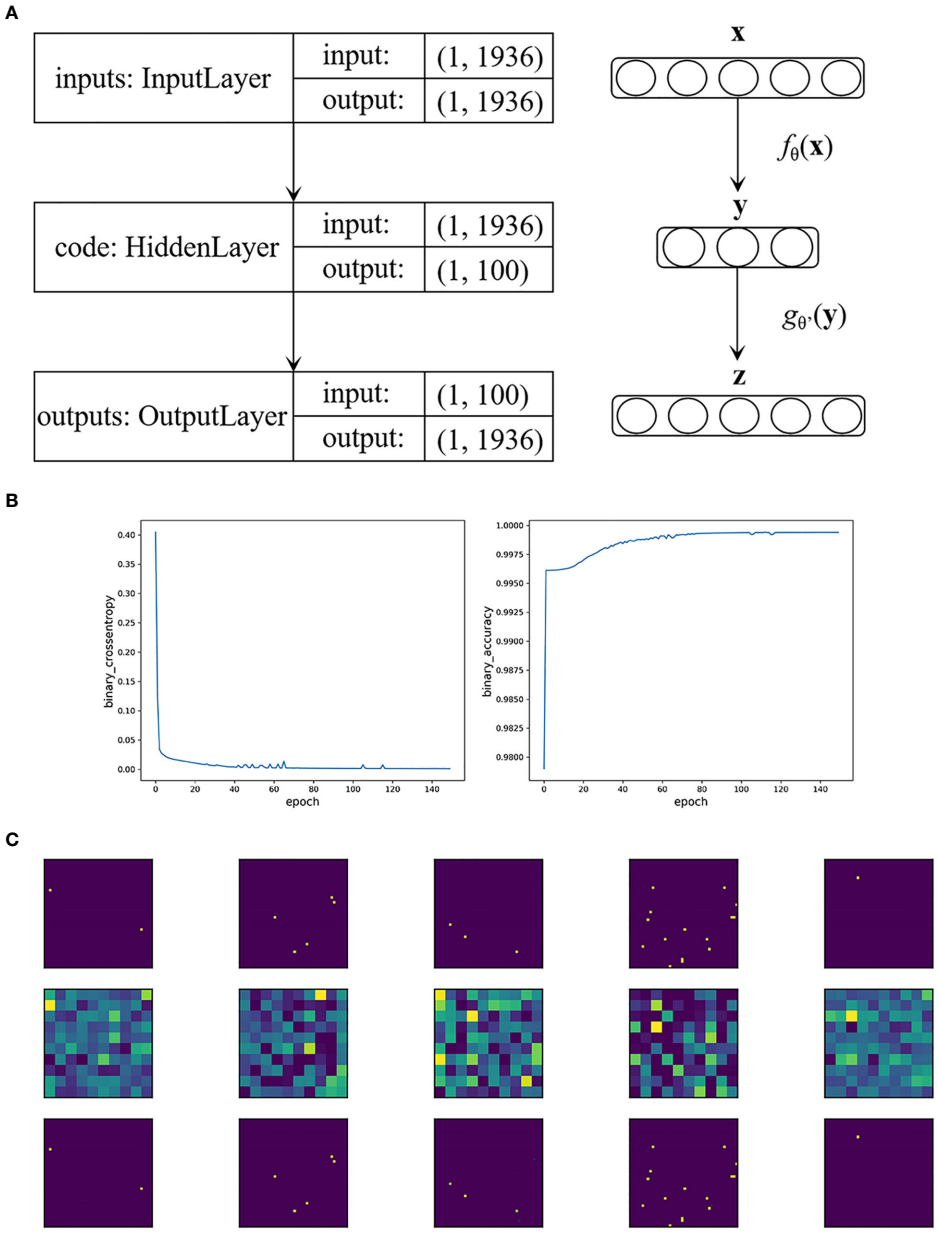


FIGURE 1
(A) The construction of the auto-encoder in BRCA data. **(B)** Loss function value and accuracy of the auto-encoder in BRCA data by the epoch times. **(C)** The first five visualized genetic signal of BRCA data. The first row shows the original images, the second row shows the extracted features, and the third row shows that the images were restored accurately with the extracted features.

0.987), 0.894 (95% CI: 0.840–0.949), and 0.879 (95% CI: 0.821-0.937), respectively. (Figure 3C). The observed survival curves (solid line) and predicted survival curves (dotted line) are shown in Figure 3D. The predicted survival curves were in the credible interval. The signature overestimated survival probability for the low-risk group and underestimated survival probability for the high-risk group. The calibration plot of these two groups shows the same result at 2, 5, and 8 years (Figure 3E).

For a summary of SNVs in both the low-risk group (Figure 4A) and the high-risk group, see Figure 4B. The median of variants per sample in the low-risk group was 30 but 74 in the high-risk group. The rank and distribution of the top 10 mutated genes in the low-risk group was similar to the whole population (Figure S7A). Peculiarly, we plotted the detailed distribution of the top 10 mutated genes in the high-risk group (Figure 4C). Fifty-seven percent of the samples had TP53

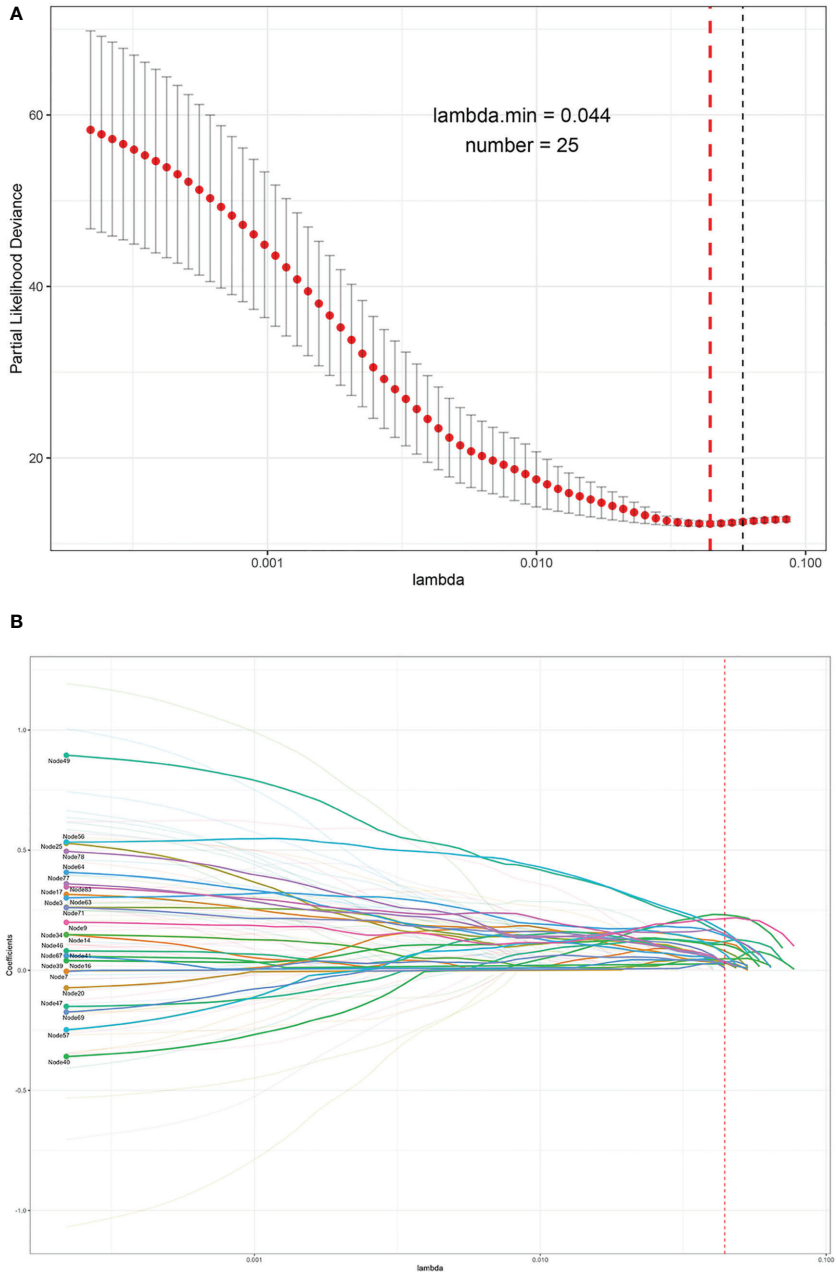


FIGURE 2
The process of the LASSO to select optimal predictors in BRCA data. **(A)** Penalty parameter tuning conducted by 10-fold cross-validation. **(B)** The solution pathway of the 25 features.

mutation in the high-risk group compared to 31% in the low-risk group; 38% of the samples had TTN mutation in the high-risk group compared to 14% in the low-risk group.

3.1.2 RSF for variable screening

RSF is used for prediction and variable selection for right-censored survival and competing risk data (26). A random forest of survival trees is used for ensemble estimation of cumulative hazard function in right-censored settings. Different survival tree splitting rules are used to grow trees. An estimate of C-index is provided for assessing prediction accuracy. Variable importance for single or grouped variables can be used to filter variables and to assess variable predictiveness.

We used the R package *randomSurvivalForest* to build an RSF model and ranked the importance of variables. Number of trees to grow was set to 10,000 in order to ensure that every input row got predicted at least a few times. The result of the model is shown in the Figure S9. Prediction error is measured by the 1 –

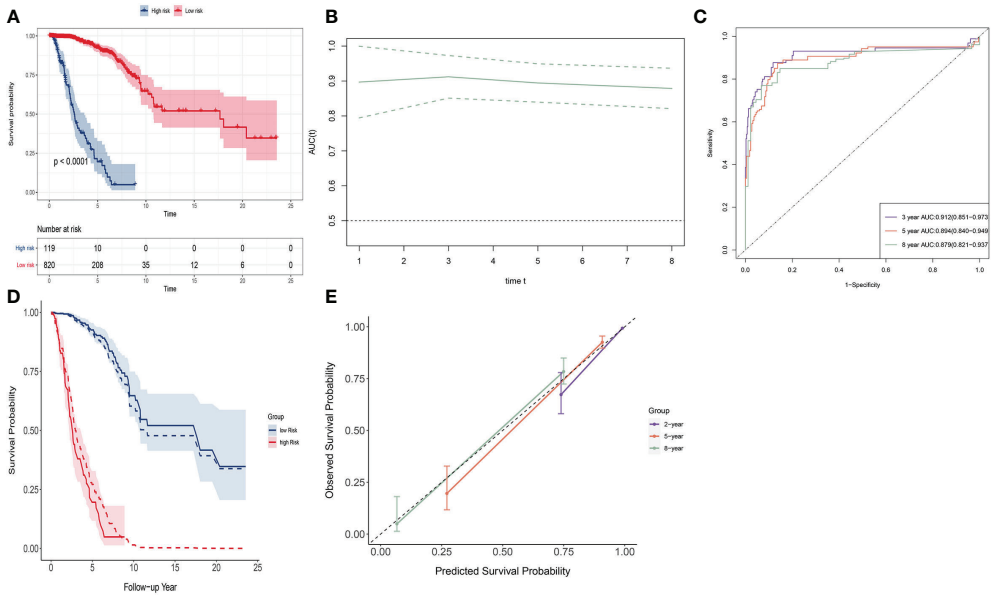


FIGURE 3
Discrimination and calibration of SNV signature in BRCA data. **(A)** The K-M curve of the low-risk group and the high-risk group. **(B)** Time-dependent AUC of SNV signature. **(C)** The 2-, 5-, and 8-year AUC of SNV signature. **(D)** The observed survival curves (solid line) and predicted survival curves (dotted line). **(E)** Calibration plot for 2-, 5-, and 8-year AUC of SNV signature.

C-index. The estimate of prediction error rate of this model was 0.449 (Figure S9A). We selected variables with an importance index greater than 0.3 (21 mutant genes) and plotted them in Figure S9B. However, we selected the 100 most important variables (see Table S4) and threw them into the LASSO Cox regression model. Twenty-three predictors were left (Figure S10). They offered 0.624 (SD = 0.048) of mean C-index and 0.081 of mean R^2 performed with a fivefold cross-validation process and stepAIC. It was not surprising that the C-index and R^2 were much lower using the RSF model when compared to using the auto-encoder (they used a similar number of variables: 25 versus 23) because the RSF model only selected the 100 most important variables and the auto-encoder used whole information.

3.1.3 Genotype and gene expression

We also performed univariable Cox analysis with gene expression data of BRCA. Data category is transcriptome profiling, data type is gene expression quantification, and workflow type is "STAR-Counts". We also selected 1,936 of them with the lowest p-value in univariable Cox analysis. Then, the multivariable LASSO Cox was used to select final predictors. A total of 60 predictors were left (Figure S11). They offered 0.831 (SD = 0.059) of mean C-index and 0.239 of mean R^2 performed with a fivefold cross-validation process and stepAIC. We drew a Venn plot of approximately 1,936 genotypes, 1,936 genes, and 60 predictors (see Figure S12), and found many common genes. Based on an explicit assumption of temporal ordering from genotype, gene expression, and survival outcome, survival mediation analysis of gene expression with multiple genotype exposures is feasible, referring to (27).

3.2 OV data

There were a total of 30,210 SNV items in which 1,650 were nonsense mutation. A total of 406 OV patients with 11,322 genotype data were available. After merging survival data, those with missing survival data were eliminated and 359 subjects were left. Univariable Cox analysis was performed on these 11,322 genotype data, and 1,089 of them with a p-value of less than 0.05 were selected for subsequent analysis. Then, the LASSO Cox regression was applied directly to these data, and a total of 95 predictors were selected by LASSO (see Figure S13A). The mean C-index was 0.707 (SD = 0.032) and the mean R^2 was 0.091 performed with a fivefold cross-validation process and stepAIC. We also used the auto-coder to extract features from the 1,089 binary variables. A total of 19 features were selected from 100 extracted features using the LASSO process (see Figure S13B). The mean C-index of the 19 features was 0.734 (SD = 0.025) and mean R^2 was 0.297 performed with a fivefold cross-validation process and stepAIC.

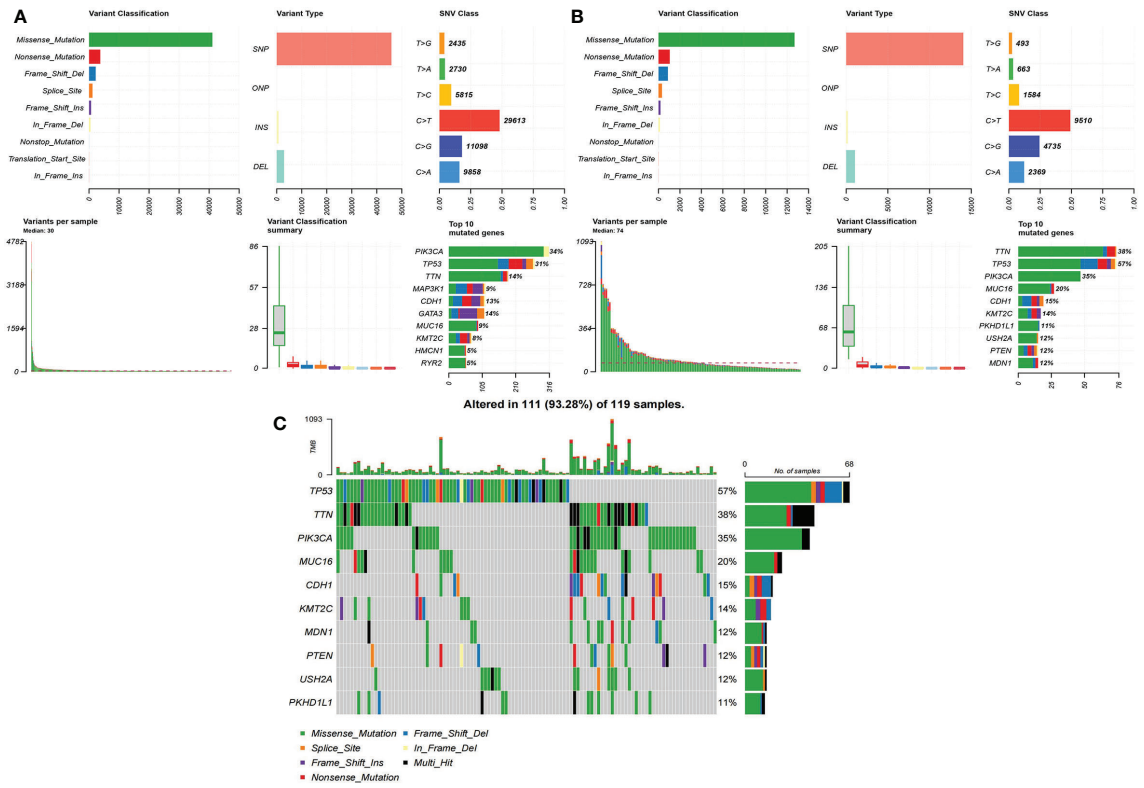


FIGURE 4
The summary of SNVs in two groups in BRCA data. **(A)** Low-risk group. **(B)** High-risk group. **(C)** The detailed distribution of the top 10 mutated genes in the high-risk group.

4 Discussion

The use of transcriptome data to construct cancer prognostic models has become very popular, and its performance in the internal verification is often satisfactory. However, due to different sequencing platforms and sequencing methods, instability of transcriptome data expression, and data standardization problems, extrapolation is still questionable. Trying to get the same desirable results from a random external data is always going to be less than expected.

SNV is a widely studied type of gene mutation (SNP is the most common type), which exists stably in somatic cells and plays a key role in regulating transcriptome expression. Aggregating small effects of SNV is a convincing attempt with promising applications. Our research shows that auto-encoders can extract effective information from dichotomous data well, even in the case of highly sparse variable values. It maps the linear combination of input dichotomous variables to a continuous value space with a lower dimension by neural networks and activation function. These features can retain most of the original information without worrying about overfitting issues, because our goal is to get the original information as possible. In addition, compared to highly sparse binary variables, low-dimensional continuous variables are better utilized. Therefore, we thought of using the auto-encoder to integrate such highly sparse binary SNV data.

Studies have shown that inherited genetic variation is associated with cancer prognosis (28–30). However, few studies have used SNV information to predict cancer prognosis in female patients. A study using multi-omics data [including gene expression data, copy number variation (CNV) data, and SNP] to predict the prognosis of BRCA patients had a 5-year survival AUC of 0.65 through their six-gene signature (31). By contrast, our study shows the power of feature extraction using the deep learning method. Based on the aggregated SNV information, we can greatly improve the ability to predict cancer patients' outcome.

In our study, BRCA patients were stratified into a low-risk group and a high-risk group based on the SNV signature. The high-risk group had higher TP53 and TTN mutation. TP53 is a well-known mutated gene and is a mutant in 30% of all breast cancers. It is clear that the role of TP53 in the management of breast cancer matters (32). Moreover, we searched the existing mutational signatures of BRCA in COSMIC (the catalogue of somatic mutations in cancer, https://cancer.sanger.ac.uk/signatures/) and found that TP53 mutation is validated to be concordant with transcriptome expression (33). TTN-AS1 is a long noncoding RNA (lncRNA) that binds to titin mRNA (TTN). Many studies

have shown that overexpression of TTN-AS1 correlates with poor prognosis in breast cancer and with more advanced pathology (34).

Furthermore, we searched for studies on SNP analysis with the auto-encoder in PubMed (8, 35–37). The most cutting-edge methods take auto-encoders to extract features from SNP data too (35). Specifically, the authors applied a deep canonically correlated sparse auto-encoder to extract key features from SNP data and functional magnetic resonance imaging (fMRI) data and then stacked these features together for classification. Their approach is very interesting and engaging because they addressed the nonlinear dimension reduction and considered the correlation between the above two types of data. The AUC score of their proposed model for the SNP data was 0.984 and that for fMRI data was 0.953, which were the highest AUC scores among all models. The difference of our study is that we have made an interesting experiment on the feature extraction property of auto-encoders. We compared the selection of activation functions in the output layer and found that the sigmoid function was more suitable for feature extraction than the ReLU function. The effect of dichotomous data was better than continuous data. In addition, the data involved in our study were from publicly available databases; thus, all results are reliable and reproducible.

Our study has its limitations. First, a person's entire sequencing genome data are not easy to come by, which makes it difficult to verify the performance of the prediction model externally, but it is hoped to be achieved in the future. Second, although we considered the correlation between covariates within and between groups in our simulation study in supplementary files, we did not incorporate genetic elements such as linkage disequilibrium. Third, due to the randomness of parameter initialization, results of deep neural network training are also random. Therefore, the characteristics obtained from each training time are always different. For example, in the BRCA dataset, each time the auto-encoder was retrained, the obtained features used for the LASSO analysis were different, as well as the C-index. However, the difference was not apparent, only causing the raw C-index to move around an interval, say 0.865 to 0.915 (see Table S5). Therefore, any training result is feasible in a single test. Furthermore, there may be many other scenarios where deep neural networks can be used to extract features and make use of them. This remains to be discovered by the scholars.

5 Conclusion

Integrating minor effects from highly sparse genetic genome data could improve prediction power. We studied the feature extraction property of the auto-encoder and found that it can work well to extract features by transforming highly sparse binary data (e.g., rare mutation) to lower-dimensional continuous data in a non-linear way. We applied this method to two cancer prognosis studies that had genotype data and achieved good predictive performance. This idea may provide something for everyone involved in cancer research, risk reduction, treatment, and patient care.

Data availability statement

We obtained image data information from MNIST (http://yann.lecun.com/exdb/mnist/) and fashion MNIST (https://jobs.zalando.com/en/tech/?gh_src=281f2ef41us). We obtained data information of BRCA and OV from official website "GDC Data Portal" (https://portal.gdc.cancer.gov/repository).

Author contributions

Study conception and design: JS, KL, and ZT. Data collection and cleaning: JS and HL. Real data analysis and interpretation: JS, XY, and JC. Drafting of the manuscript: JS, LB, and YD. All authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China (81773541), funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions at Soochow University and the State Key Laboratory of Radiation Medicine and Protection (GZK1201919) to ZT, the National Natural Science Foundation of China (81872552 and U1967220) to JC, and the National Natural Science Foundation of China (82172441) and Suzhou Key Clinical Diagnosis and Treatment Technology Project (LCZX201925) to KL. The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgments

We acknowledge the contributions of the TCGA cohort study and MNIST team.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.1091767/full#supplementary-material

References

1. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* (2018) 19(5):299–310. doi: 10.1038/nrg.2018.4

2. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings Bioinf* (2018) 19(2):286–302. doi: 10.1093/bib/bbw114

3. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* (2021) 13(1):152. doi: 10.1186/s13073-021-00968-x

4. El Ghaoui L, Viallon V, Rabbani T. Safe feature elimination in sparse supervised learning. *Pacific J Optimization.* (2012) 8(4):667–98.

5. Tibshirani R. Regression shrinkage and selection *via* the lasso. *J R Stat Soc Ser B-Methodological* (1996) 58(1):267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

6. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies *via* penalized logistic regression. *Genet Epidemiol* (2010) 34(8):879–91. doi: 10.1002/gepi.20543

7. Long N, Gianola D, Rosa GJ, Weigel KA. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in holsteins. *J Anim Breed Genet* (2011) 128(4):247–57. doi: 10.1111/j.1439-0388.2011.00917.x

8. Prive F, Aschard H, Blum MGB. Efficient implementation of penalized regression for genetic risk prediction. *Genetics* (2019) 212(1):65–74. doi: 10.1534/genetics.119.302019

9. Yang T, Wang J, Sun Q, Hibar DP, Jahanshad N, Liu L, et al. Detecting genetic risk factors for alzheimer's disease in whole genome sequence data *via* lasso screening. *Proc IEEE Int Symp BioMed Imaging* (2015) 2015:985–9. doi: 10.1109/ISBI.2015.7164036

10. Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput Biol* (2017) 13(1):e1005357. doi: 10.1371/journal.pcbi.1005357

11. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* (2007) 17(10):1520–8. doi: 10.1101/gr.6665407

12. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* (2012) 13(2):135–45. doi: 10.1038/nrg3118

13. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intelligence* (2013) 35(8):1798–828. doi: 10.1109/TPAMI.2013.50

14. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* (2006) 313(5786):504–7. doi: 10.1126/science.1127647

15. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp Biocomput* (2015), 20:132–43. doi: 10.1142/9789814644730_0014

16. Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layerwise training of deep networks. *Adv Neural Inf Process Syst 19* (2007), PP. 153–60.

17. Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol* (2019) 29(7):R231–R6. doi: 10.1016/j.cub.2019.02.034

18. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc Ser B Stat Methodol* (2012) 74(2):245–66. doi: 10.1111/j.1467-9868.2011.01004.x

19. Guo P, Zeng FF, Hu XM, Zhang DM, Zhu SM, Deng Y, et al. Improved variable selection algorithm using a LASSO-type penalty, with an application to assessing hepatitis b infection relevant factors in community residents. *PloS One* (2015) 10(7):e0134151. doi: 10.1371/journal.pone.0134151

20. Wang J, Wonka P, Ye JP. Lasso screening rules *via* dual polytope projection. *J Mach Learn Res* (2015) 16:1063–101. doi: 10.48550/arXiv.1211.3966

21. Jiang Y, He YX, Zhang HP. Variable selection with prior information for generalized linear models *via* the prior LASSO method. *J Am Stat Assoc* (2016) 111(513):355–76. doi: 10.1080/01621459.2015.1008363

22. Guo W, Elston RC, Zhu X. Evaluation of a LASSO regression approach on the unrelated samples of genetic analysis workshop 17. *BMC Proc* (2011) 5 Suppl 9:S12. doi: 10.1186/1753-6561-5-S9-S12

23. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc Ieee* (1998) 86(11):2278–324. doi: 10.1109/5.726791

24. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. ICML '08: Proceedings of the 25th international conference on Machine learning (2008), 1096–103 p. doi: 10.1145/1390156.1390294

25. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, et al. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* (2005) 28(2):171–82. doi: 10.1002/gepi.20041

26. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics* (2014) 15(4):757–73. doi: 10.1093/biostatistics/kxu010

27. Shao ZH, Wang T, Zhang M, Jiang Z, Huang SP, Zeng P. IUSMMT: Survival mediation analysis of gene expression with multiple DNA methylation exposures and its application to cancers of TCGA. *PloS Comput Biol* (2021) 17(8):e1009250. doi: 10.1371/journal.pcbi.1009250

28. Barrdahl M, Canzian F, Lindstrom S, Shui I, Black A, Hoover RN, et al. Association of breast cancer risk loci with breast cancer survival. *Int J Cancer* (2015) 137(12):2837–45. doi: 10.1002/ijc.29446

29. Rafiq S, Tapper W, Collins A, Khan S, Politopoulos I, Gerty S, et al. Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. *Cancer Res* (2013) 73(6):1883–91. doi: 10.1158/0008-5472.CAN-12-3377

30. Lu LG, Katsaros D, Mayne ST, Risch HA, Benedetto C, Canuto EM, et al. Functional study of risk loci of stem cell-associated gene lin-28B and associations with disease survival outcomes in epithelial ovarian cancer. *Carcinogenesis* (2012) 33(11):2119–25. doi: 10.1093/carcin/bgs243

31. Mo WJ, Ding YQ, Zhao S, Zou DH, Ding XW. Identification of a 6-gene signature for the survival prediction of breast cancer patients based on integrated multi-omics data analysis. *PLoS One* (2020) 15(11):e0241924. doi: 10.1371/journal.pone.0241924

32. Shahbandi A, Nguyen HD, Jackson JG. TP53 mutations and outcomes in breast cancer: Reading beyond the headlines. *Trends Cancer* (2020) 6(2):98–110. doi: 10.1016/j.trecan.2020.01.007

33. Smid M, Rodriguez-Gonzalez FG, Sieuwerts AM, Salgado R, Prager-Van der Smissen WJ, Vlugt-Daane MV, et al. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat Commun* (2016) 7:12910. doi: 10.1038/ncomms12910

34. Zheng QX, Wang J, Gu XY, Huang CH, Chen C, Hong M, et al. TTN-AS1 as a potential diagnostic and prognostic biomarker for multiple cancers. *Biomedicine Pharmacother* (2021) 135:111169. doi: 10.1016/j.biopha.2020.111169

35. Li G, Han DP, Wang C, Hu WX, Calhoun VD, Wang YP. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. *Comput Methods Programs Biomed* (2020) 183:105073. doi: 10.1016/j.cmpb.2019.105073

36. Massi MC, Gasperoni F, Ieva F, Paganoni AM, Zunino P, Manzoni A, et al. A deep learning approach validates genetic risk factors for late toxicity after prostate cancer radiotherapy in a REQUITE multi-national cohort. *Front Oncol* (2020) 10. doi: 10.3389/fonc.2020.541281

37. Fergus P, Montanez CC, Abdulaimma B, Lisboa P, Chalmers C, Pineles B. Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women. *Ieee-Acm Trans Comput Biol Bioinf* (2020) 17(2):668–78. doi: 10.1109/TCBB.2018.2868667

frontiers | Frontiers in Oncology

TYPE Methods
PUBLISHED 17 January 2023
DOI 10.3389/fonc.2022.999324

Check for updates

OPEN ACCESS

EDITED BY
Mohammad Kohandel,
University of Waterloo, Canada

REVIEWED BY
Jiaqian Qi,
The First Affiliated Hospital of Soochow University, China
Alexander Kaizer,
University of Colorado Anschutz Medical Campus, United States

*CORRESPONDENCE
Peng Wei
PWei2@mdanderson.org

SPECIALTY SECTION
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

RECEIVED 20 July 2022
ACCEPTED 14 December 2022
PUBLISHED 17 January 2023

CITATION
Zhou J, Jiang X, Xia HA, Hobbs BP and Wei P (2023) Landmark mediation survival analysis using longitudinal surrogate.
Front. Oncol. 12:999324.
doi: 10.3389/fonc.2022.999324

COPYRIGHT
© 2023 Zhou, Jiang, Xia, Hobbs and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Landmark mediation survival analysis using longitudinal surrogate

Jie Zhou[1], Xun Jiang[2], H. Amy Xia[2], Brian P. Hobbs[3] and Peng Wei[4]*

[1]Department of Biostatistics and Pharmacometrics, Neuroscience Global Drug Development, Novartis, East Hanover, NJ, United States, [2]Center for Design and Analysis, Amgen, Thousand Oaks, CA, United States, [3]Department of Population Health, The University of Texas, Austin, TX, United States, [4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

Clinical cancer trials are designed to collect radiographic measurements of each patient's baseline and residual tumor burden at regular intervals over the course of study. For solid tumors, the extent of reduction in tumor size following treatment is used as a measure of a drug's antitumor activity. Statistical estimation of treatment efficacy routinely reduce the longitudinal assessment of tumor burden to a binary outcome describing the presence versus absence of an objective tumor response as defined by RECIST criteria. The objective response rate (ORR) is the predominate method for evaluating an experimental therapy in a single-arm trial. Additionally, ORR is routinely compared against a control therapy in phase III randomized controlled trials. The longitudinal assessments of tumor burden are seldom integrated into a formal statistical model, nor integrated into mediation analysis to characterize the relationships among treatment, residual tumor burden, and survival. This article presents a frameworkfor landmark mediation survival analyses devised to incorporate longitudinal assessment of tumor burden. R^2 effect-size measures are developed to quantify the survival treatment mediation effects using longitudinal predictors. Analyses are demonstrated with applications to two colorectal cancer trials. Survival prediction is compared in the presence versus absence of longitudinal analysis. Simulation studies elucidate settings wherein patterns of tumor burden dynamics require longitudinal analysis.

KEYWORDS

functional principal component analysis, landmark analysis, longitudinal analysis, mediation analysis, oncology, RECIST

1 Introduction

Defined as the time from randomization until death from any cause, overall survival (OS) is considered the most reliable cancer endpoint for randomized controlled trials.

Measurement of OS is unbiased and precise. Prolonging OS is considered the most convincing demonstration that an emerging cancer therapeutic intervention is superior to an existing standard of care. Despite these advantages, a long duration is required to sufficiently follow trial participants to observe OS. Especially in earlier stages, for which recent advances in surgical, radiation, and non-cytotoxic treatments have extended patient survival to an extent that there are now considerable numbers of patients alive 5 years after diagnosis (such as breast cancer and colon cancer). Authors report that the median duration of an industry sponsored phase III oncology trial that completes enrollment is approximately 48 months (1).

This latency period required to acquire sufficiently mature OS data effectively delays the confirmation of OS benefit. For many cancer types, radiographic tumor assessments are used to directly measure components of the disease and trigger treatment decisions in clinical practice. While a variety of tumor response criteria are considered appropriate for regulatory reviews, the revised Response Evaluation Criteria in Solid Tumors (RECIST) (2) is the predominate technique for quantifying changes in tumor burden following treatment for solid tumors. The RECIST criteria is a composite endpoint which characterizes changes in tumor size by the sum of longest diameters (LDSUM). Thought to be a surrogate endpoint for OS, changes in tumor burden are described by four levels of response for target and non-target lesions. For example, for target lesions, takingas reference the smallest LDSUM on study, observing a 20% increase yields the worst result of progressive disease (PD). The best result occurs with the disappearance of all target lesions, or complete response (CR). A 30% decrease from the baseline LDSUM yields a partial response (PR). A patient is considered to have stable disease (SD) if criteria for PD, PR, and CR are not satisfied.

Tumor assessments occur longitudinally over the course of study, typically planned to be acquired at 4-week or 8-week intervals until tumor progression. Planned trial statistical analyses, however, reduce the longitudinal curves to a single response level. At the patient level, the objective response indicates whether a patient experienced PR or CR during the course of study. At the cohort level, the objective response rate (ORR), the proportion of patients that achieved an objective response, is a commonly used primary endpoint for single-arm trials. Objective response is used alongside endpoints describing the duration of response to support applications for accelerated approval to the U.S. Food and Drug Administration.

In oncology settings, complex relationships exist among therapies, tumor response (often used as the basis for evaluation in phase II trials), and survival. Reducing the longitudinal tumor assessment to a single value, however, may not describe the cumulative effect of the patient's tumor burden. Mediation modeling provides a framework for elucidating the mechanisms by which an intervention impacts an endpoint through a third intermediate response variable. The extent to which tumor response representsa reliable surrogate of survival can be measured through the application of mediation models, which decompose the total effect of an intervention into direct effect and indirect effects. In the context of oncologic drug development, the indirect effect defines the extent of survival benefit that is achieved from improving the objective response rate, while direct treatment effects characterize the extent of survival benefit attributable to all other factors. Authors have developed mediation models integrating categorical surrogate endpoints with OS (3, 4). However, to our knowledge a mediation model for OS that leverages the entire tumor assessment trajectory for each patient has not been developed.

Longitudinal data is often analyzed using the linear mixed effects models or generalized estimation equations (5, 6). Functional principal component analysis (FPCA) is another popular method that provides a powerful approach for modelling noisy and irregularly measured longitudinal data. Summary statistics derived from FPC scores yield dimension reduction of the trajectories, while preserving most of the information. Survival models with longitudinal predictors have been developed under the joint modeling framework (7, 8). Specifically, a mixed effects model with normal random effects is commonly assumed for the longitudinal observations. The hazard function or the survival probability is assumed to depend on the true longitudinal outcome at each time point. Naive two-stage approaches (9, 10) were first proposed for estimating the association between the longitudinal and survival outcomes. More advanced estimating procedures based on the Expectation-Maximization (EM) algorithm (11) and Markov chain Monte Carlo (MCMC) (12) were subsequently proposed to reduce bias in parameter estimation. Dynamic predictions of survival probabilities conditional on the available longitudinal data have been developed for joint models (13–15). The application of such joint models is challenging in practice, however. The fitting algorithms are computationally expensive and many underlying parametric distributional assumptions cannot be verified from the data. Furthermore, the extensions to estimate the time-varying effects of different types of longitudinal response variables are not straightforward.

Landmark analysis first proposed in Anderson et al. (16) provides a straightforward approach to approximate the association between longitudinal and OS outcomes at a sequence of landmark times. Landmark models have been developed to estimate the varying effect of a longitudinal predictor on the survival outcome and predict the survival outcomes dynamically (17). Cao et al. (18) calculated the cumulative effect of longitudinal outcome based on FPCs and applied the measure in models with binary disease outcomes. Yan et al. (19) used landmark analyses with FPCs derived based on moving time windows in predicting survival probabilities dynamically. Shi et al. (20) applied the FPCA based landmark analysis in estimating the effect of longitudinal measures of total cholesterol with respect to risk of coronary heart disease.

Though it is not a comprehensive probability model of the longitudinaland the event time processes, landmark analysis circumvents the assumptions and computational burden associated with a joint model.

This article presents a method for quantifying the treatment mediation effects of tumor response on survival using longitudinally observed tumor measurements. R^2 effect-size measures (21, 22) are extended to the settingof longitudinal survival mediation. Different from the traditional product-based and difference-based mediation effect measures, which are based on coefficients estimated in regression models (23), R-squared type measures are derived from the extent of explained variation of survival outcomes by treatment and mediators. Several R^2 measures have been proposed in the literature. Kent and O'Quigley (24) derived an explained risk measure called "explained randomness" using Weibull models with entropy loss function. Korn and Simon (25)Korn and Simon (, 26) discussed explained risk measures for residual variation in survival analysis. To account for censoring in the survival data, authors have incorporated inverse probability censoring weights (27, 28), but the measures depend on the maximum follow-up time and could be sensitive to the values in the right tail of the survival distribution. To avoid any modification or approximation to the metric, Heller (29) proposed a measure called explained relative risk for the proportional hazards model, which is unaffected by non-informative censoring times. Shi et al. (30) compared the different R^2 measures for survival outcomes and suggested two R^2 measures proposed in Kent and O'Quigley (24) and Heller (29) that satisfy the properties conveyed by Royston (31). These methods have not be used to compare tumor burden (TB) trajectories with landmark analyses. We compare mediation analysis for OS using longitudinal measures of tumor response versus conventional approaches using the single-valued objective response. Measures are estimated using patient-level data acquired from two colorectal cancer studies [Goldberg et al. (32); Peeters et al. (33)], which we analyzed before as secondary analyses in different contexts [Hobbs et al. (34); 120 Zhou et al. (3); Zhou et al. (4)].

The remainder of this article is organized as follows. In Section 2.1, we introduce the varying coefficient models for the longitudinal outcomes and different response summary statistics. The landmark analysis models are discussed in Section refsec:osmod for estimating the time-varying effect of longitudinal outcomes and predicting survival probability for new patients. The R^2 measures are introduced in Section 2.3 for measuring the mediation effects of longitudinal tumor measurements. The models comparing different measures are applied to two colorectal cancer studies in Section 2.4. Comprehensive simulation studies are performed in Section 3 to evaluate the predictive performance of summary statistics of longitudinal measures under settings assuming various relationships among treatment, longitudinal and survival outcomes. Finally, we summarize the content and have some discussions in Section 5.

2 Methodology

Let $A \in A$ indicate treatment arm. For simplicity, we present notation assuming only two treatment options $A=(0,1)$ as extensions to three or more are relatively straightforward. X and Z denote covariate vectors for longitudinal and survival models, respectively. X and Z may represents different features or shared common variables. lationships among treatment, TB, and OS.

2.1 Longitudinal tumor burden

Let $Y_{ij}=Y_i(t_{ij})$ denote the observed longitudinal measures of TB for the i th patient at visit time t_{ij} , $i=1,\cdots,N$; $j=1,\cdots,m_i$. We assume Y_{ij} follows a varying coefficient model:

$$Y_i(t) = \eta_0(t) + D_i(t) + \epsilon_i(t)$$

$$= \eta_0(t) + \eta_1(t) \times A_i + \eta_2(t) \times X_i + \epsilon_i(t), \quad (1)$$

Where $\eta_0(t)$ is the overall population mean trajectory of TB and $\eta_1(t)$ and $\eta_2(t)$ are the time-varying treatment and covariate effects on the tumor. The error term $\epsilon_{ij}=\epsilon_i(t_{ij})$ is assumed to follow a Normal distribution $N(0,\sigma^2)$. Note that $D_i(t)$ is the true longitudinal trajectory with the overall mean trajectory removed, and therefore contains the patient level variability of TB.

2.2 Functional principal component analysis

Yao et al. (35) proposed an approach called Principal Components Analysis through Conditional Expectation (PACE) for sparse and irregularly measured longitudinal data. According to the Karhunen–Loeve decomposition, the patient level longitudinal trajectory $D_i(t)$ in Equation (1) can be decomposed as $D_i(t) = \sum_{k=1}^{\infty} \gamma_{ik} p_k(t)$ where $\rho_k(t)$ is the k th eigenfunction satisfying the orthonomal conditions: $\int \rho_k(s)\rho_l(s) ds=0$ and $\int \rho_k(s)^2 ds=1$ for any $k \neq l$ and $k=1,\cdots,\infty$. Parameter $\gamma_{ik}=\int (Y_i(t)-\eta_0(t))\rho_k(t)dt$ is the functional principal component (FPC) score corresponding to the k th eigenfunction.

The variability contained by the functional components decreases as k increases. Usually the trajectories can be well approximated by the first finite number (say K) of components. Therefore we estimate the smoothed trajectories

$$\hat{D}_i(t) = \sum_{k=1}^{K} \hat{\gamma}_{ik} \rho_k(t),$$

where $\hat{\gamma}_{ik}$ is the FPC score estimated through conditional expectation. The estimation of the overall mean function, eigenfunctions, and FPC scores can be achieved in the R package "fdapace".

To make full use of the longitudinal TB information, we consider two approaches to estimation. One approach uses the first K FPC scores that capture at least 90% of the total outcome variability. The FPC scores summarize a patient trajectory's proximity to the patterns described by the eigenfunctions. Restricting estimation to 90% of variability explained captures most information of the longitudinal outcome. The other approach summarizes the cumulative effect of longitudinal outcomes. Let τ_i denote the upper limit of observational time of the longitudinal measures for patient i. We define the integrated smoothed outcome of $D_i(t)$ as $I\hat{D}_i = \int_0^{\tau_i} \hat{D}_i(s)ds$, which can be calculated through numerical integration.

2.3 Time-varying models for overall survival

Let T denote overall survival duration since trial enrollment. To evaluate treatment effect on OS after adjusting for TB, we assume T follows a proportional hazards (PH) model (36)

$$h(t|A_i, R_i, Z_i) = h_0(t) \times \exp\{\beta_1 A_i + \beta_2 Z_i + \alpha R_i\},$$

where $h_0(t)$ is the baseline hazard function. The baseline hazard is positive and can be estimated nonparametrically. R_i denotes the tumor response surrogate measure. For conventional mediation analysis, R_i may be the binary indicator of an objective RECIST tumor response. Leveraging the longitudinal model, one could use the first K FPC scores as R_i or the fully integrated smoothed outcome $I\hat{D}_i$) of TB discussed in Section 4.2. Of note, the classical causal mediation analysis model for survival outcomes assumes no unmeasured confounders for the exposure-outcome (A-T), exposure-mediator (A-R), and mediator-outcome (R-T) relationships VanderWeele (23). As our goal here is prediction rather causal inference, we do not make such assumptions. Nevertheless, adjusting covariates, i.e., potential confounders, in the PH models is expected to improve the prediction accuracy Vandenberghe et al. (37).

The true survival time T is subject to right censoring. Let C_i denote the censoring duration of patient i. For modeling justification, we need to assume that the censoring mechanism is random or non-informative. Therefore, T_i and C_i are independent conditional on the treatment A_i , covariates Z_i, and longitudinal outcome variable R_i . The observed survival outcome for patient i includes an observed survival duration $\tilde{T}_i = min(T_i, C_i)$ and censoring indicator $\delta_i = I(T_i \leq C_i)$.

With complete dataset $O = \{(\tilde{T}_i, \delta_i, A_i, X_i, Z_i, R_i), i = 1, \cdots, N\}$, the overall effects of treatment and covariates can be estimated with the traditional nonparametric Cox PH model. However, this model assumes that the effects of treatment, TB, and covariate on OS are fixed over time. Ignoring the possibility of time-varying effects, this strong assumption is usually difficult to justify. easures. This limitation can be addressed by adopting a landmark analysis approach.

2.3.1 Landmark analysis

Let $0<L_1<L_2<\cdots<L_P<\tau$ denote a sequence of landmark times of interest, where $\tau = max(\tau_1, \cdots, \tau_N)$ is an upper limit of the longitudinal follow-up time of all patients. For each landmark time L_p, only patients in the risk set at L_p are used for estimation. Define $t_i^{(p)} = \{t_{ij} : t_{ij} < L_p, j = 1, \cdots, m_i\}$ to be the observational time points before landmark time L_p . $Y_i^{(p)}$ denotes the corresponding longitudinal outcomes with measurement time in $t_i^{(p)}$. The response summary variable $R_i^{(p)}$ is then derived based on available data $Y_i^{(p)}$ at landmark time L_p. Specifically, a conventional model using binary objective response will allow different patterns for responders after the response is observed, while non-responders will assume constant values at 0 for all time. When is defined based on FPCA, as discussed in Section 4.2, FPCA can be applied according to a moving time window (19).

Initially, FPCA is performed with all available longitudinal data using the maximum time window $(0, \tau)$. The estimated overall mean function $\hat{\eta}_0(t)$, as well as the eigenvalues and eigenvectors corresponding to the first K FPCs are saved. Then at each landmark time L_p , FPC scores are recalculated as $\hat{\gamma}_{ik}^{(p)}$ based on available longitudinal outcomes $Y_i^{(p)}$. The integrated score at landmark timL_p is calculated as $R_i\ I\hat{D}_i^{(p)} = \int_0^{\tau_i^p} \hat{D}_i(s)ds$, where $\tau_i^p = \min(\tau_i, L_p)$.

The landmark data at L_p is defined as $O^{(p)} = \{(\tilde{T}_i, \delta_i, A_i, X_i, Z_i, R_i^{(p)}) : \tilde{T}_i > L_p; i = 1, \cdots, N\}$ and the survival model for $T–L_P$ given $O^{(p)}$ is defined as:

$$h(t|O^{(p)}) = h_0^{(p)}(t) \times \exp\left\{\beta_1(L_p) \cdot A_i + \beta_2(L_p) \cdot Z_i + \alpha(L_p) \cdot R_i^{(p)}\right\}, \quad (2)$$

where $h_0^{(p)}(t)$ is the nonparametric baseline hazard function evaluated at time L_p+t . Trajectories of estimated coefficients $\hat{\beta}_1(\cdot)$, $\hat{\beta}_2(\cdot)$ and $\hat{\alpha}(\cdot)$ represent the changing patterns of the effects of treatment, covariate, and TB measures on overall survival. These model features provide estimates for time-varying coefficients at landmark time points of interest. Moreover, the estimated models can be used for prediction at each landmark time, which will be discussed in the next section.

2.3.2 Prediction for new patients

After estimating the model in (2) for all the landmark times, we can dynamically predict survival probabilities for new patients at each landmark time. Specifically, the conditional survival probability for a new patient N+1 , who survivedlonger than landmark time L_p and had data $O_{N+1}^{(p)} = (A_{N+1}, Z_{N+1}, R_{N+1}^{(p)})$, can be written as

$$\hat{S}(L_p + t|L_p, O_{N+1}^{(p)})$$
$$= Pr\left\{T_{N+1} > L_p + t|T_{N+1} \rangle L_p, A_{N+1}, Z_{N+1}, R_{N+1}^{(p)}\right\}$$
$$= \hat{S}_0^{(p)}(t)^{\exp\left\{\hat{\beta}_1(L_p)\cdot A_{N+1} + \hat{\beta}_2(L_p)\cdot Z_{N+1} + \hat{\alpha}(L_p)\cdot R_{N+1}^{(p)}\right\}}, \quad (3)$$

where $\hat{S}_0^{(p)}(t) = \exp\left\{-\int_0^t h_0^{(p)}(s)ds\right\}$ is the baseline survival function corresponding to landmark time L_p.

The predictive power of various tumor burden summaries, R_i, discussed in Section 4.1 can be evaluated by comparing the resulting area under curves (AUC):

$$AUC_p(u) = Pr\left\{S(L_p + t\left|L_p, O_i^{(p)}\right.) < S(L_p + t\left|L_p, O_j^{(p)}\right.)\left|L_p \langle T_i \le L_p + u, T_j \rangle L_p + u\right.\right\}. \quad (4)$$

The estimated value $\hat{AUC}_p(u)$ can be obtained by plugging the predicted survival probabilities (3) in equation (4). Models with larger values of $\hat{AUC}_p(u)$ facilitate more accurate predictions at landmark time L_p.

Brier score (38) is another measure of prediction performance characterized by the mean squared error for the predicted survival probabilities calculated in equation (3). This measure is useful for model comparison in simulation studies where the true survival probabilities are known Shi et al. (20). However, the Brier score cannot be directly applied in real data analysis with right-censored observations unless modified using the inverse probability of censoring weighting method Graf et al. (27), which can be numerically unstable Seaman and White (39).

2.4 R2 measures for mediation effect

R^2 effect-size measures were originally proposed to assess the variance accounted for in mediation models with uncensored continuous outcomes (21). According to Yang et al. (22) and Shi et al. (30), three differentregression models need to be fit to calculate the R^2 measures for mediation effects: (1) a model with the independent variable of interest, (2) a model with the mediators and (3) a full model with all predictors. Specifically, we fit the following three PH models:

$$h(t|A_i) = h_0^D(t) \times \exp\left\{\beta_1^D A_i\right\}, \quad (5)$$

$$h(t|R_i) = h_0^M(t) \times \exp\left\{\alpha^M R_i\right\}, \quad (6)$$

$$h(t|A_i, R_i) = h_0^F(t) \times \exp\left\{\beta_1^F A_i + \alpha^F R_i\right\}, \quad (7)$$

where the superscripts 'D', 'M' and 'F' denote the "direct", "mediated" and "full" models. These three models are used to distinguish the baseline hazard and regression coefficients. Let $R_{T,D}^2$ denote the total variation of survival time T explained by treatment. This can be derived based on model (5). Similarly, let $R_{T,M}^2$ calculated based on model (6) represent the variation of T explained by the response summary variable, and $R_{T,F}^2$ calculated based on the full model (7) is the the variation of T explained by the treatment and response summary variables conjointly. The R-squared measure of mediated effect is then calculated as

$$R_{med}^2 = R_{T,D}^2 + R_{T,M}^2 - R_{T,F}^2. \quad (8)$$

The proportion of mediated effect with respect to the total treatment effect on survival is defined as the shared over simple (SOS) effect. SOS effect is calculated as $SOS = R_{med}^2/R_{T,D}^2$. Provided a nonzero total effect and non-negative direct and mediated effects (i.e. $0 \le R_{med}^2 \le R_{T,D}^2$ and $R_{T,D}^2 > 0$), the mediation effect of the response summary variable increases as SOS increases from 0 to 1.

To calculate the R^2 measures in Equation (8), Shi et al. (30) compared five available approaches and suggested the R_b^2 proposed by Heller (29) as well as R_w^2 proposed by [15] which is based on the five properties suggested in (31) for evaluating the R^2 measures in survival models. n the manuscript, and results with other approaches can be found in the Supplementary Materials. Let θ denote the vector of regression coefficients in models (5)-(7) and P_i denote the corresponding vector of predictor values for patient i, i=1,⋯,N. The R-squared measure R_b^2 derived by Heller (29), which they referred to as the explained relative risk, is calculated as

$$R_b^2 = \frac{log\left[N^{-1}\sum_i exp\ \left(\hat{\theta}'\widetilde{P}_i\right)\right]}{1.5772 + log\left[N^{-1}\sum_i exp\ \left(\hat{\theta}'\widetilde{P}_i\right)\right] - 1},$$

where $\widetilde{P}_i$ is the covariate value centered around zero and $\hat{\theta}$ is the maximum partial likelihood estimate for θ. Note that 1.5772 is the approximate value for the entropy for the covariate model under the extreme value distribution. The other measure R_w^2 using the approach proposed by [15] was developed from a Weibull model as an approximation to the explained risk based on the product moment correlation coefficient with a standard normal error variance:

$$R_w^2 = \frac{\hat{\theta}'\widehat{\Sigma_P}\hat{\theta}}{1 + \hat{\theta}'\widehat{\Sigma_P}\hat{\theta}},$$

where $\hat{\Sigma_P}$ is the estimated variance covariance matrix of the covariate vector.

Note that the R-squared based approaches require large sample sizes in practice. Based on our numerical studies, the estimated SOS effects using the above approaches are not always restricted to the interval [0,1], especially when the total treatment effects are small while the sample sizes are moderate.

3 Colorectal cancer studies

This section applies the landmark survival mediation analysis to two colorectal cancer studies. The predictive power of various tumor burden measures are compared at different landmark times. Landmark times were selected as the 10th, 20th, …,90th quantiles of event times observed in each study. Estimated time-varying coefficients of treatment and baseline LDSUM were compared among four models: (i) model with no response variable adjusted; (ii) model adjusting binary response; (iii) model adjusting the first three FPC components that explained more than 99% variability in the outcome and (iv) model adjusting the integrated smoothed TB trajectory.

3.1 Goldberg study

The colorectal cancer study reported by Goldberg et al. (32) included 795 patients with metastatic colorectal cancer who had not been treated previously for advanced disease. Secondary analysis of the trial data has been reported Hobbs et al. (34). These patients were enrolled in the study between May 1999 and April 2001 and had a median follow-up of 20.4 months (88 weeks). There were three treatment arms including irinotecan and bolus fluorouracil plus leucovorin (IFL), oxaliplatin and infused fluorouracil plus leucovorin (FOLFOX), or irinotecan and oxaliplatin (IROX). FOLFOX and IROX were two new regimens under investigation while IFL was considered as the standard of care.

Our case study compares the tumor response and overall survival outcomes between patients who received the FOLFOX regime (treatment group) and the standard of care IFL (control group). Analysis was limited to patients with at least three measurements of tumor burden as defined by RECIST LDSUM. As a result, the analysis set included a total of 311 patients (157 in treatment group and 154 in control group). Tumor burden was evaluated at each treatment cycle (every 2 weeks). (This is the description in the study, but we assumed 2 wks for each cycle in data. Do we need to note here)? Longitudinal LDSUM measures are plotted by treatment groups in Figure 1A. The median number of follow-up visits is five. Baseline LDSUM is included as a covariate in models for both longitudinal LDSUM and OS. The baseline LDSUM ranged from 1.5 to 38.5 centimeters (cm) with mean of 10.18 cm in the treatment group compared with 10.11 cm in the control group.

According to RECIST criteria, patients in the treatment group experienced more reductions in tumor burden than patients in the control group: 114 (72.6%) responders (14 CR and 100 PR) among patients receiving FOLFOX versus 93 (60.4%) responders (5 CR and 88 PR) for patients receiving IFL. We performed FPCA on the scaled longitudinal LDSUM. The time scale was map into [0, 1] by dividing the observational time by the maximum follow-up duration observed. Figure 1B, plots the smoothed outcome up to the last observational time. Solid lines depict observed domain for each patient, while dashed lines are used to depict the predicted trajectories for each patient. Functional data regression was fit using equation (1) to the smoothed trajectories yielding the estimated intercept and coefficients for treatment and baseline LDSUM in plotted in Figure 1C. Based the resultant estimated coefficient trajectories, we find that the treatment effect on the longitudinalLDSUM measures fluctuates around zero after adjusting for baseline tumor burden. Patients with larger baseline LDSUM also had larger LDSUM measures but the correlation decreases with time.

The C-index (or AUC) evaluated at median survival time is computed and compared for the four models with complete data (dashed lines) at each landmark time (solid lines) in Figure 1D. With complete data, models (iii) and (iv) based on the FPCs resulted in larger AUC values than model (ii) which described changes in TB as the binary objective response. All the three models containing tumor response information yielded larger AUC than model (i). Models (iii) and (iv) yielded larger AUC values than models (i) and (ii) for landmark time points before 800 days, while the AUC values are comparable beyond 800 days. Figures 1E, F present the estimated coefficients for treatment and baseline LDSUM for four models at eachlandmark time. The point estimates based on the complete dataset are marked as dashed lines in the plot. Patients receiving the FOLFOX regime experienced lower hazard of death compared to patients receiving IFL after adjusting baseline LDSUM measures andresponse based on all four models.

Table 1 reported the estimated R^2 measures for models (ii)-(iv). Note the estimated measures fall below zero occasionally in this data due to the small sample size and mode R^2 rate total effects. All R^2 measures are close to zero, indicating limited mediated effect of the response summary variables. With limited differentiation between treatment and the longitudinal LDSUM, as estimated in Figure 1C, the mediation path from treatment to response to OS is not identified from the data.

3.2 Study of Panitumumab

The other colorectal cancer study was sponsored by Amgen (33). The study enrolled 1186 patients with Metastatic Colorectal Cancer (mCRC). Patients were randomized to receive either standard treatments of FOLFIRI (control group) or FOLFIRI plus Panitumumab (treatment group). The median follow-up time of the patients was 59 weeks. Our case study uses data made available on Project Data Sphere (PDS) (40), which comprised approximately 80% of patient-level data reported in the completed study. Analyses included patients with at least 2 measures of LDSUM during follow-up, yielding 841 patients (417 in control group and 424 in treatment group) for analysis.

10.3389/fonc.2022.999324

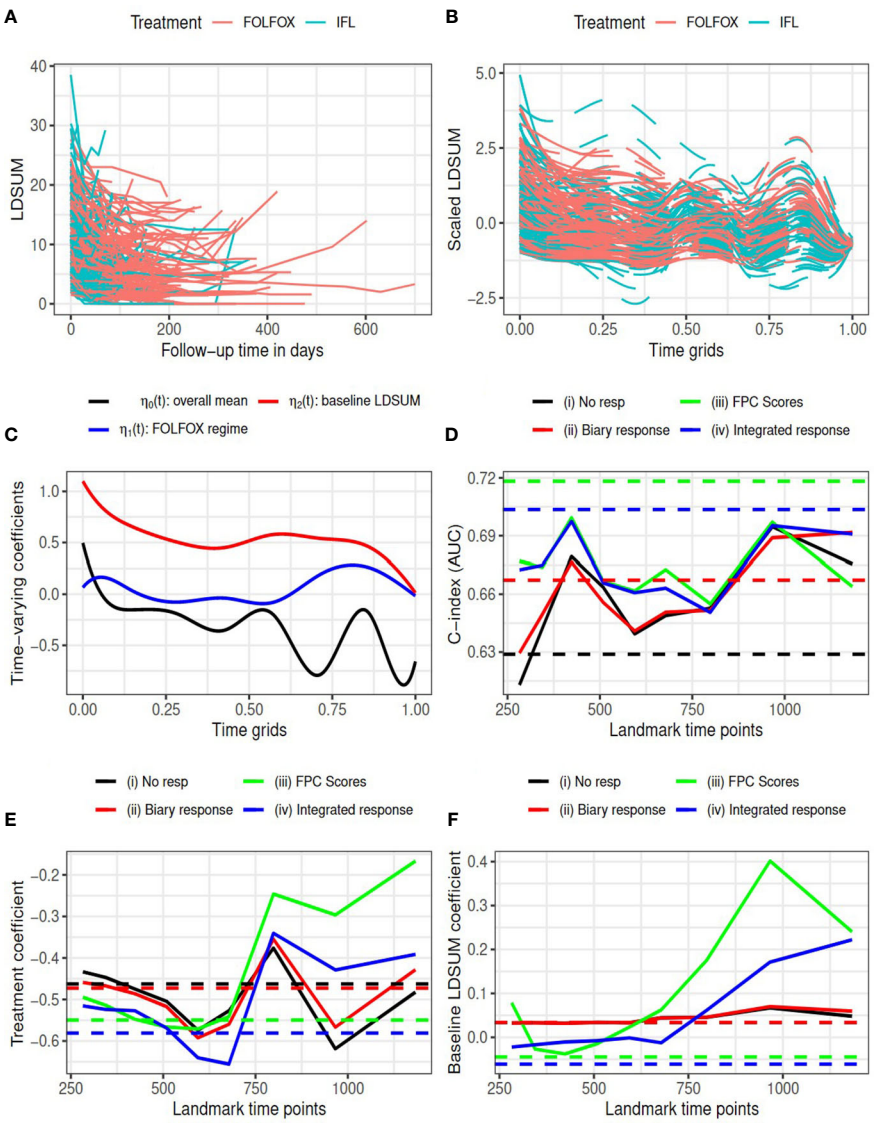


FIGURE 1
Goldberg study: **(A)** observed LDSUM by treatment groups; **(B)** smoothed (solid lines) and predicted (dashed lines) values for scaled LDSUM based on FPCA; **(C)** estimated varying coefficients in functional data regression analysis; **(D)** AUC values with different response measures; (e/f) estimated coefficient for treatment/baseline LDSUM in survival model. Solid lines are based on analysis and dashed lines are based on complete data in **(D–F)**.

TABLE 1 Colorectal studies: estimated R^2 and SOS measures for mediation effects.

		(ii) Binary response		(iii) FPC scores		(iv) Integrated response	
Study	measure	R^2_{med}	SOS	R^2_{med}	SOS	R^2_{med}	SOS
Goldberg	R^2_b	-0.00992	-0.22303	-0.00106	-0.02394	-0.00684	-0.15374
	R^2_w	-0.00993	-0.19275	0.00457	0.08869	-0.00613	-0.11905
Amgen	R^2_b	-0.00001	-0.00447	0.00114	0.68428	0.0009	0.54028
	R^2_w	0.0002	0.10568	0.00069	0.35978	0.00118	0.61047

SOS: the shared over simple effect. It is defined as the ratio of R-squared measures of the mediated and the total treatment effects on OS.

10.3389/fonc.2022.999324

All patients with measurable disease at the baseline central review had their objective tumor response assessed every 8 weeks until progressive disease or death. The longitudinal LDSUM measures are plotted in Figure 2A. The median number of follow-up visits is four. Similar to analysis of the Goldberg study, we include the baseline LDSUM measures as predictors in models for both longitudinal LDSUM and OS. The baseline LDSUM ranged from 20 to 762 millimeter (mm) with an average size of target tumor at baseline of 168.9 mm and 164.1 mm in the treatment and control groups, respectively.

The resultant RECIST objective response rate in the treatment group is around 30%, which is much higher than the response rate in control group (12.5%). Figure 2B presents the predicted smoothed trajectories of longitudinal LDSUM. Estimates of the time-varying intercept and coefficient effects for treatment and baseline LDSUM are shown in Figure 2C. The plots demonstrate that the addition of Panitumumab had the effect of decreasing the LDSUM early following treatment. The effect, however, diminished with further follow-up. Similarly, patients with larger baseline LDSUM maintained larger LDSUM

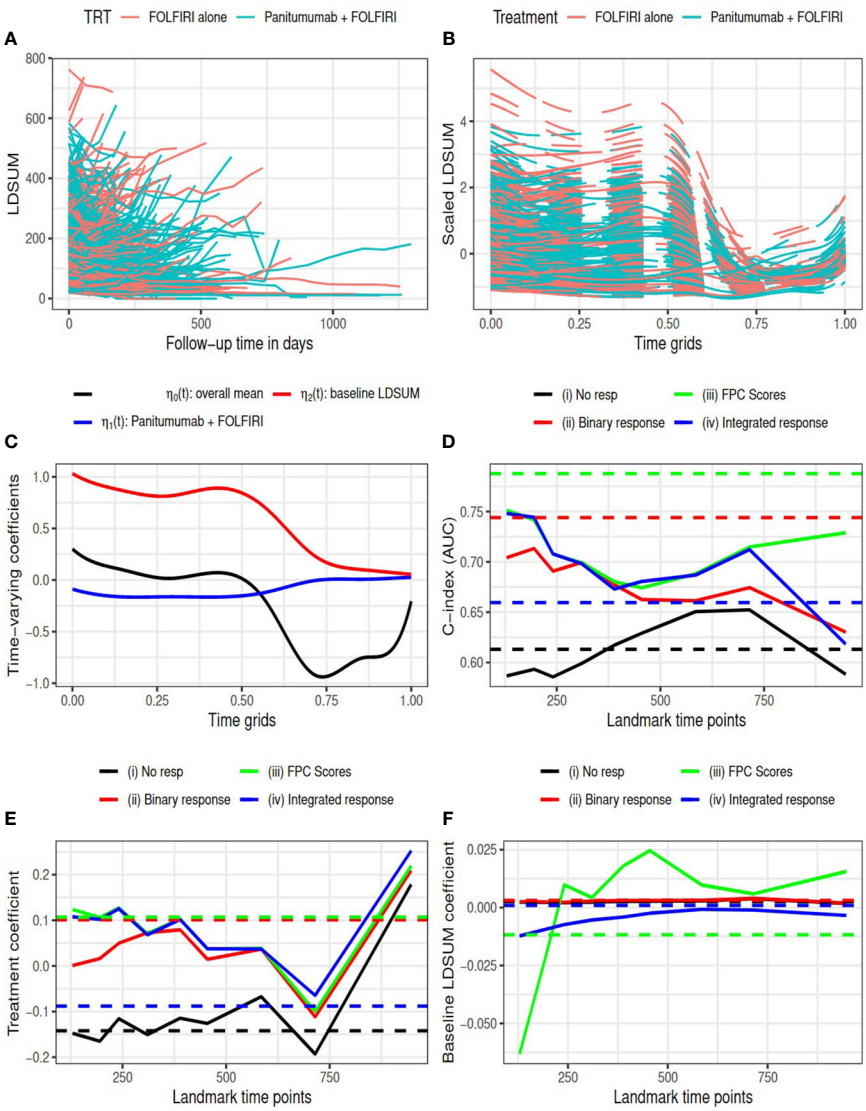


FIGURE 2
Amgen colorectal cancer study: **(A)** observed LDSUM by treatment groups; **(B)** smoothed (solid lines) and predicted (dashed lines) values for scaled LDSUM based on FPCA; **(C)** estimated varying coefficients in functional data regression analysis; **(D)** AUC values with different response measures; (e/f) estimated coefficient for treatment/baseline LDSUM in survival model. Solid lines are based on analysis and dashed lines are based on complete data in **(D–F)**.

measures early on, The baseline effect of LDSUM also decreased to zero with time.

C-index (or AUC) evaluated at median survival time is compared among four models with complete data (dashed lines) and at each landmark time (solid lines) in Figure 2D. With complete data, model (iii) had the largest AUC and model (ii) with best response had larger AUC than model (iv). Models (ii), (iii) and (iv), which contained response information, were found to have larger AUC than model (i) at all the landmark time points.

In Figures 2E, F, the estimated coefficients for treatment and baseline LDSUM based on the four models were compared at each landmark time. The point estimates based on the complete dataset are marked as dashed lines in the plot. The treatment coefficients was negative in model (i) and positive in the other three models adjusted for response variables. The difference here indicated that the protective effect of additional Panitumumab on survival outcome strongly depended on the response information. The coefficients for baseline LDSUM were close to zero and the estimated values from model (iii) had larger variability.

The R^2 measures are listed in Table 1. The estimated *SOS* effects are close to zero for model (ii) and between 0.35 and 0.7 for models (iii) and (iv). This suggests that 35% to 70% of the treatment effect on OS is mediated by the longitudinal TB measures. Additionally, the FPC scores capture this mediation effect more efficiently than the conventional RECIST objective response. on effect during the follow-up between 300 to 700 days.

4 Simulation

4.1 Data generation

This section presents a simulation study devised to compare the four models (i)-(iv) discussed in Section 3. Using estimates from our case studies, we assume that the true trajectory of the longitudinal outcome was $\eta_0(t)+\eta_1(t)A+\eta_2(t)X$ with range for longitudinal follow-up time of $t\in(0,1)$ Treatment A was generated from the Bernoulli distribution with probability of 0.5. Covariate matrix X was generated from the standard Normal distribution. For the time-varying coefficients, we used the estimated overall mean from the Goldberg study for $\eta_0(t)$, and assume different time varying coefficients $\eta_1(t)$ and $\eta_2(t)$ for treatment and covariate X based on a cubic spline functions. Specifically, we defined four cubic B-spline basis function $(B_1(t),\cdots,B_4(t))$ in the range of $t\in(0,1)$ Model coefficient functions were determined by $\eta_j(t) = \sum_{l=1}^{4}\xi_{jl}B_l(t)$.

We generate the true event time T from the survival model

$$h(t|A,Z,D(\tau)) = h_0(t) \times \exp\left\{\beta_1 A + \beta_2 Z + \alpha \int_0^{\tau} D(s)ds\right\},$$

where $D(s)=\eta_1(s)A+\eta_2(s)X$ is the true longitudinal trajectory after removing the overall mean function and $D(\tau)$ includes the history of $D(s)$ up to time $\tau=min\ (T,1)$. The baseline hazard function assumed Weibull distribution with shape 1.682 and scale 1.024, which were estimated from the Goldberg data. Covariate Z was generated from the standard Normal distribution.

Since the longitudinal predictor in the survival model is time-dependent when $T<1$, we generated the event time using the following procedures. First, we set a sequence of grids $0=s_0<s_1<\cdots<s_M=1$, where $s_m-s_{m-1}=0.001$. At each grid s_m, theintegration part in the model can be approximated using the numerical integration. Specifically, we approximate $R(s_m) = \int_0^{s_m} D(s)ds$ by

$$R(s_m) \approx \tilde{R}(s_m) = \sum_{l=1}^{m}(D(s_l) - D(s_{l-1}))(s_l - s_{l-1}),$$

and the cumulative hazard function evaluated at time grid s_m can be approximated as

$$H(s_m) \approx \tilde{H}(s_m) = \exp\ (\beta_1 A + \beta_2 Z)\sum_{l=1}^{m}(s_l - s_{l-1})\left\{h_0(s_l)e^{\alpha\tilde{R}(s_l)} - h_0(s_{l-1})e^{\alpha\tilde{R}(s_{l-1})}\right\}.$$

The survival probability at time s_m is then $S(s_m) \approx \exp\ (-\tilde{H}(s_m))$. For $T\geq 1$, the cumulative hazard function is:

$$H(T) = \tilde{H}(s_M) + \exp\ \left\{\beta_1 A + \beta_2 Z + \alpha\tilde{R}(s_M)\right\}\int_1^T h_0(u)du\,.$$

To generate event time T that follows the desired distribution, we first generate $U\sim Unif(0,1)$, and compare it with the survival probability $S(s_M)$. If $U>S(s_M)$, the event time $T=max\ \{s_m:S(s_m)\geq U\}$, otherwise, T has closed form solution

$$T = H_0^{-1}\left\{H_0(s_M) - \frac{log\ \ (U) + \tilde{H}(s_M)}{exp\ \ (\beta_1 A + \beta_2 Z + \alpha\tilde{R}(s_M))}\right\},$$

Where $H_0(s) = \int_0^s h_0(u)du$ is the cumulative baseline hazard function and H_0^{-1} is its inverse function.

We then generate independent censoring time C from uniform distribution $Unif(0,10)$, and the observed survival outcome as $\tilde{T} = \min\ (T, C)$ and $\delta=I(T\leq C)$. The longitudinal follow-up time $t=(t_1,\cdots,t_m)$ was randomly selected as the time grids on the interval $(0,\tilde{\tau})$, where $\tilde{\tau} = \min\ (1,\tilde{T})$, and the number of total post-baseline visits m was generated from Poisson distribution with mean of 5. The observed longitudinal observations represent the true trajectory $\eta_0(t)+D(t)$ evaluated at visit time t plus error terms generated from $N(0,\sigma^2=0.01)$. Only patients with at least 3 post-baseline observations $(m\geq q3)$ were included in analysis.

4.2 Simulation settings and results

Our simulation study assumed five fundamentally different relationships among the treatment, longitudinal and survival

outcomes. The coefficients for longitudinal and survival models are listed in Table 2 for each setting. In the first two settings 'a1' and 'a2', there is only direct treatment effect on survival outcome with the response effect on OS assumed to be zero (i.e. α=0). The treatment and the longitudinal outcomes are not related in 'a1', while in setting 'a2', treatment has positive effect on the longitudinal outcome. Setting 'b' is another typical case where all the treatment effect on survival is mediated through the longitudinal response outcome. Setting 'c' has both direct and indirect treatment effects on the survival outcomes, while setting 'd' is a null case where treatment, longitudinal response and survival outcomes assume no dependence.

We used sample size N=1000 and 100 replications for each setting. The censoring time was generated from Uniform distribution $U(0,10)$, resulting in censoring proportions between 10% and 40%. Results comparing the four models based on the complete data are summarized in Figure 3. The mean and the 2.5% and 97.5% percentiles of the Brier scores and the C-index (AUC) evaluated at the median survival time are plotted in subfigures (a) and (b), respectively. The FPCA based models (iii) and (iv) have smaller Brier scores and larger AUCs in settings (b) and (c), where there is non-zero indirect treatment effect on survival through longitudinal outcome. The estimated R^2 measure and *SOS* effects using method R_b^2 are plotted in subfigures (c) and (d) with first and third quartiles. Subfigures (e) and (f) describe method R_w^2. The results using the two approaches are similar. *SOS* estimated for models (iii) and (iv) are closer to 1 (the truth), in setting (b) than model (ii). In addition, larger variation of SOS is observed for model (ii) in some settings, especially in the setting (d), where treatment neither impacts the longitudinal LDSUM nor OS. Summary statistics of the estimated coefficients β_1 and β_2 can be found in the Supplementary Materials. The mean estimates of the four models are close to the truth in settings 'a1', 'a2' and 'd', while models (iii) and (iv) have smaller bias than model (ii) in settings (b) and (c).

We selected the landmark time as a sequence of time from 0.2 to 2 with a step of 0.2 on the scaled time grids. Note that the summary variables for the longitudinal outcome were calculated at each landmark time based on available longitudinal data up to that time. The Brier scores and AUCs are reported in Figure 4. The estimated coefficients at landmark time points are summarized in the Supplementary Materials. The conclusions are similar to those based on the complete data. The models (iii) and (iv) have consistently better predictive ability (smaller Brier scores and larger AUCs) than model (i) and (ii) in settings 'b' and 'c'. The four models have similar performance in the other three settings 'a1', 'a2' and 'd', where tumor response is assumed to have no effect on survival.

5 Discussion

Oncology studies routinely acquire measures of tumor burden longitudinally over the course of patient follow-up. This information is predominately dimension reduced to a binary tumor response variable indicating the occurrence of a partial or complete response as defined by the RECIST criteria. This article presented a landmark mediation survival model devised to estimate conjoint effects of treatment and longitudinal tumor assessments. Prediction performance was compared using different characterizations of tumor response following treatment. Conventional binary response based on the RECIST criteria was compared to analysis of the full longitudinal TB assessments using FPC scores as well as the integrated response. R^2 measures were adopted to quantifythe extent of treatment survival mediation effect attributable to longitudinal TB assessments. Implementation was demonstrated with two colorectal cancer studies: the Goldberg study comparing FOLFOX with IFL and the Amgen study on the additional effect of Panitumumab on FOLFIRI. The time-varying effects of treatment and baseline LDSUM were compared with models that leveraged different extent of information from the longitudinal tumor assessments. Prediction performance was compared using AUC. R^2 measures were adopted to quantifying the mediation effect of the longitudinal tumor burden.

We found that the longitudinal models with prediction based on FPC scores tended to yield larger AUCs when compared to models with conventional RECIST objective response. Moreover, it was discovered that the predictive utility of binary tumor response depends on the shape of the underlying longitudinal trajectories. With U-shaped trends for tumor burden following treatment, as observed in the Amgen study, binary objective response (PR or CR based on the RECIST

TABLE 2 Simulation settings and coefficients values.

Settings	Longitudinal model		Survival model		
	ξ_1 (A)	ξ_2 (X)	β_1 (A)	β_2 (Z)	α (R)
a1: direct effect only	(0,0,0,0)	(0.6,0.5,0.5,0.6)	-1	1	0
a2: direct effect only	(0,-0.2,-0.25,-0.7)	(0.6,0.5,0.5,0.6)	-1	1	0
b: indirect effect only	(0,-0.2,-0.25,-0.7)	(0.6,0.5,0.5,0.6)	0	1	10
c: both effects exist	(0,-0.2,-0.25,-0.7)	(0.6,0.5,0.5,0.6)	-1	1	10
d: no effect exists	(0,0,0,0)	(0.6,0.5,0.5,0.6)	0	1	0

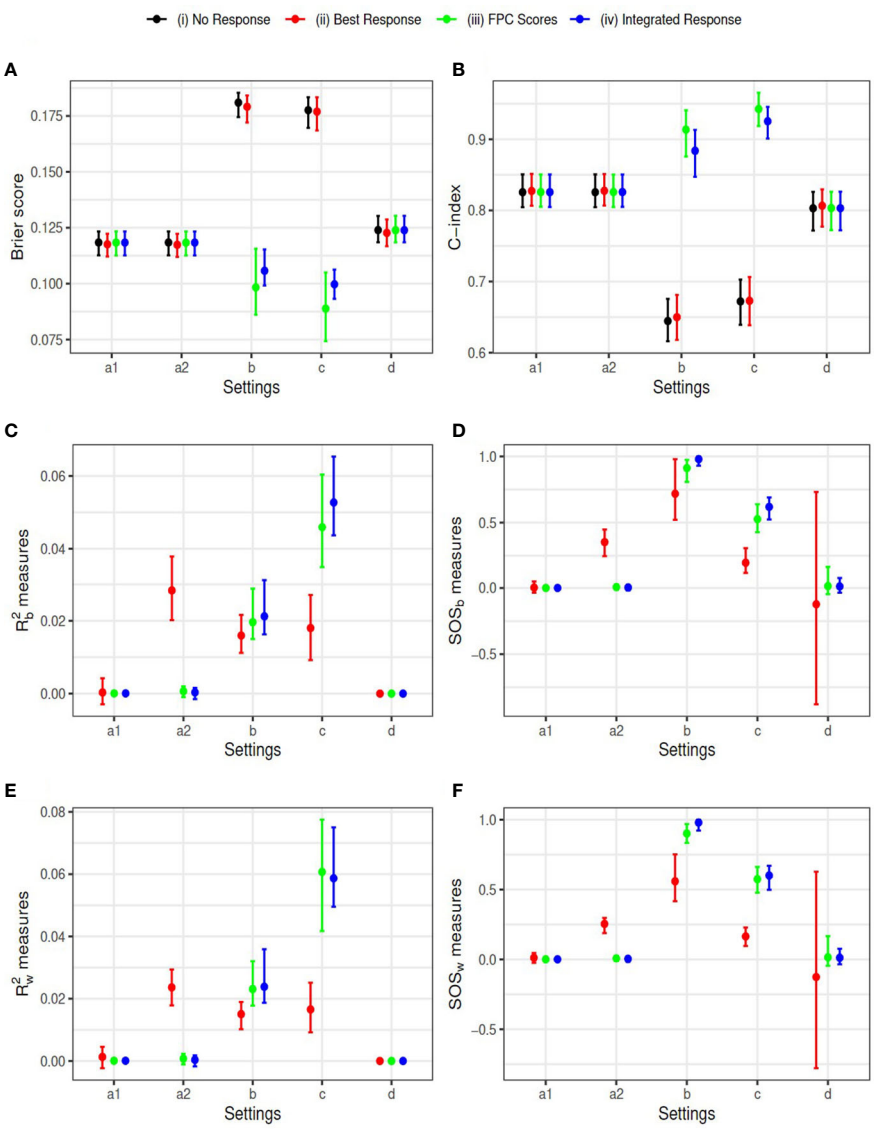


FIGURE 3
Simulation Results: Y-axis corresponds to **(A)** Brier score, **(B)** C-index, **(C)** R^2_b measure, **(D)** SOS_b measure, **(E)** R^2_w measure, and **(F)** SOS_w measure, respectively. Simulation settings include: direct effect only (settings a1 and a2), indirect effect only (setting b), both effects exist (setting c), and neither effect exists (setting d).

criteria) was sufficient to characterize the most pertinent information contributed by the longitudinal data. AUCs obtained from models incorporating tumor response information were much larger than corresponding model absent tumor response.

In the absence of U-shaped trends in tumor burden over time, reducing the longitudinal TB data to a binary response discards important information regarding treatment survival mediation. Models using binary objective response applied to the Goldberg study, for example, yielded AUCs that were very close to those obtained in the model without tumor response. Based on the estimated R^2 and *SOS* effects, the longitudinal TB as defined by LDSUM presented no mediation effect in the Goldberg data. For the Amgen study, however, an estimated 35% to 70% of the treatment effect on OS was mediated through the pattern of longitudinal tumor assessments captured by FPC scores. Simulation demonstrated that FPCA based longitudinal predictors yielded smaller Brier scores and larger AUCs than the binary response model under all settings. Consequently, the complex relationships between treatment, survival, and tumor burden may be better elucidated with the widespread adoption of longitudinal analysis. FPC scores in particular offer a practical

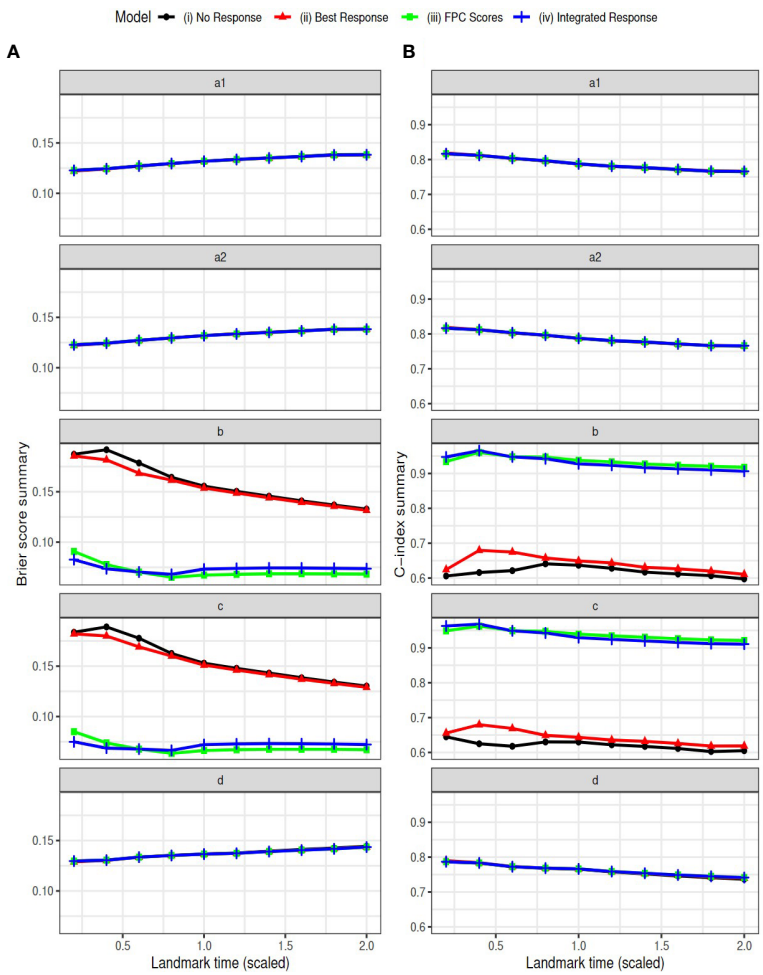


FIGURE 4
Simulation Results: Brier scores (column **A**) and C-index (AUC) (column **B**) in landmark analysis for the following simulation settings: direct effect only (settings a1 and a2), indirect effect only (setting b), both effects exist (setting c), and neither effect exists (setting d).

approach to synthesizing the longitudinal patterns with sufficient flexibility to capture the trends that describe treatment survival mediation.

Several limitations should be noted. The estimation approaches presented were founded on large sample theory which requires caution with the application to small sample data in practice. The interpretation of the mediation effect is uninterruptible when the estimated R^2_{med} is negative or the resulted SOS effects are out of the [0,1] range. This may have resulted in our case study from the additional random variation induced by the joint model and/or the relatively small sample size of the Goldbergcase. Further investigation is required to define the minimal number of OS event one needs to observed before fitting mediation models for OS with longitudinal surrogates.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: www.projectdatasphere.org.

Author contributions

BH and PW contributed to conception and design of the study. JZ organized the database and performed the statistical analysis. JZ and BH wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

The authors are grateful to two reviewers for their helpful comments.

Conflict of interest

JZ is employed by Novartis. XJ and HX are employed by Amgen. BH reports Scientific Advisory for Telperian and Amgen outside the scope of submitted work.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

Supplementary Material is available on the journal website.

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.999324/full#supplementary-material

References

1. Applied Clinical Trials. Oncology trials are taking longer. *Accessed February* (2014) 14:2021.

2. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *Eur J Cancer* (2009) 45:228–47. doi: 10.1016/j.ejca.2008.10.026

3. Zhou J, Jiang X, Amy Xia H, Wei P, Hobbs BP. A survival mediation model with bayesian model averaging. *Stat Methods Med Res* (2021) 30:2413–27. doi: 10.1177/09622802211037069

4. Zhou J, Jiang X, Xia HA, Wei P, Hobbs BP. Predicting outcomes of phase iii oncology trials with bayesian mediation modeling of tumor response. *Stat Med* (2022) 41:751–68. doi: 10.1002/sim.9268

5. Pinheiro JC, Bates DM. Linear mixed-effects models: Basic concepts and examples. In: *Mixed-Effects Models in S and S-PLUS. Statistics and Computing*. New York, NY: Springer (2000). doi: 10.1007/0-387-22747-4_1

6. Hardin JW, Hilbe JM. *Generalized Estimating Equations. 1st edition*. (New York: Chapman and Hall/CRC) (2002). doi: 10.1201/9781420035285

7. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. 1st edition*. (New York: Chapman and Hall/CRC) (2012). doi: 10.1201/b12208

8. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sin* (2004) 14:809–34.

9. Tsiatis AA, Degruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and cd4 counts in patients with aids. *J Am Stat Assoc* (1995) 90:27–37. doi: 10.1080/01621459.1995.10476485

10. Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* (2001) 88:447–58. doi: 10.1093/biomet/88.2.447

11. Zeng D, Cai J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann Stat* (2005) 33:2132–63. doi: 10.1214/009053605000000480

12. Rizopoulos D. The r package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *J of Statist Software* (2016) 72:1–46. doi: 10.18637/jss.v072.i07

13. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* (2011) 67:819–29. doi: 10.1111/j.1541-0420.2010.01546.x

14. Njagi EN, Rizopoulos D, Molenberghs G, Dendale P, Willekens K. A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Stat Model* (2013) 13:179–98. doi: 10.1177/1471082X13478880

15. Andrinopoulou E-R, Rizopoulos D, Geleijnse ML, Lesaffre E, Bogers AJ, Takkenberg JJ. Dynamic prediction of outcome for patients with severe aortic stenosis: Application of joint models for longitudinal and time-to-event data. *BMC Cardiovasc Disord* (2015) 15:28. doi: 10.1186/s12872-015-0035-z

16. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* (1983) 1:710–9. doi: 10.1200/JCO.1983.1.11.710

17. Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian J Stat* (2007) 34:70–85. doi: 10.1111/j.1467-9469.2006.00529.x

18. Cao Y, Rajan SS, Wei P. Mendelian randomization analysis of a time-varying exposure for binary disease outcomes using functional data analysis methods. *Genet Epidemiol* (2016) 40:744–55. doi: 10.1002/gepi.22013

19. Yan F, Lin X, Li R, Huang X. Functional principal components analysis on moving time windows of longitudinal data: Dynamic prediction of times to event. *J R Stat Society: Ser C (Applied Statistics)* (2018) 67:961–78. doi: 10.1111/rssc.12264

20. Shi B, Wei P, Huang X. Functional principal component based landmark analysis for the effects of longitudinal cholesterol profiles on the risk of coronary heart disease. *Stat Med* (2021) 40:650–67. doi: 10.1002/sim.8794

21. Fairchild AJ, MacKinnon DP, Taborga MP, Taylor AB. R 2 effect-size measures for mediation analysis. *Behav Res Methods* (2009) 41:486–98. doi: 10.3758/BRM.41.2.486

22. Yang T, Niu J, Chen H, Wei P. Estimation of total mediation effect for high-dimensional omics mediators. *BMC Bioinformatics* (2021) 22:414. doi: 10.1186/s12859-021-04322-1

23. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology* 22:582–5. doi: 10.1097/EDE.0b013e31821db37e

24. Kent JT, O'Quigley J. Measures of dependence for censored survival data. *Biometrika* (1988) 75:525–34. doi: 10.1093/biomet/75.3.525

25. Korn EL, Simon R. Measures of explained variation for survival data. *Stat Med* (1990) 9:487–503. doi: 10.1002/sim.4780090503

26. Korn EL, Simon R. Explained residual variation, explained risk, and goodness of fit. *Am Statistician* (1991) 45:201–6. doi: 10.2307/2684290

27. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* (1999) 18:2529–45. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

28. O'Quigley J, Xu R. Explained variation in proportional hazards regression. In: *Handbook of statistics in clinical oncology*. (New York: Chapman and Hall/CRC) (2005). p. 367–84.

29. Heller G. A measure of explained risk in the proportional hazards model. *Biostatistics* (2012) 13:315–25. doi: 10.1093/biostatistics/kxr047

30. Shi B, Huang X, Wei P. Comparison of effect size measures for mediation analysis of survival outcomes with application to the framingham heart study. *arXiv preprint arXiv* (2022) 2205:03303. doi: 10.48550/arXiv.2205.03303

31. Royston P. Explained variation for survival models. *Stata J* (2006) 6:83–96. doi: 10.1177/1536867X0600600105

32. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK, et al. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *J Clin Oncol* (2004) 22:23–30. doi: 10.1200/JCO.2004.09.046

33. Peeters M, Price T, Cervantes A, Sobrero A, Ducreux M, Hotko Y, et al. Final results from a randomized phase 3 study of folfiri ± panitumumab for second-line treatment of metastatic colorectal cancer. *Ann Oncol* (2014) 25:107–16. doi: 10.1093/annonc/mdt523

34. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. *Clin Trials* (2013) 10:430–40. doi: 10.1177/1740774513483934

35. Yao F, Müller H-G, Wang J-L. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc* (2005) 100:577–90. doi: 10.1198/016214504000001745

36. Cox DR. Regression models and life-tables. *J R Stat Society: Ser B (Methodological)* (1972) 34:187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x

37. Vandenberghe S, Duchateau L, Slaets L, Bogaerts J, S V. Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Stat Methods Med Res* (2018) 27:3367–85. doi: 10.1177/0962280217702179

38. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather Rev* (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

39. Seaman S, White I. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* (2013) 22:278–95. doi: 10.1177/0962280210395740

40. *Project data sphere*. PDSweb. Available at: www.projectdatasphere.org (Accessed June 29, 2020).

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 19 January 2023
DOI 10.3389/fmolb.2023.1094321

Check for updates

OPEN ACCESS

EDITED BY
George Bebis,
University of Nevada, United States

REVIEWED BY
Jianing Xi,
Guangzhou Medical University, China
Chaowang Lan,
Guilin University of Electronic Technology, China

*CORRESPONDENCE
Sung-Young Shin,
Sungyoung.Shin@monash.edu
Lan K. Nguyen,
lan.k.nguyen@monash.edu

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 10 November 2022
ACCEPTED 06 January 2023
PUBLISHED 19 January 2023

CITATION
Shin S-Y, Centenera MM, Hodgson JT, Nguyen EV, Butler LM, Daly RJ and Nguyen LK (2023), A Boolean-based machine learning framework identifies predictive biomarkers of HSP90-targeted therapy response in prostate cancer. *Front. Mol. Biosci.* 10:1094321. doi: 10.3389/fmolb.2023.1094321

COPYRIGHT
© 2023 Shin, Centenera, Hodgson, Nguyen, Butler, Daly and Nguyen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A Boolean-based machine learning framework identifies predictive biomarkers of HSP90-targeted therapy response in prostate cancer

Sung-Young Shin[1,2]*, Margaret M. Centenera[3,4], Joshua T. Hodgson[3,4], Elizabeth V. Nguyen[1,2], Lisa M. Butler[3,4], Roger J. Daly[1,2] and Lan K. Nguyen[1,2]*

[1]Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC, Australia, [2]Cancer Program, Biomedicine Discovery Institute, Monash University, Clayton, VIC, Australia, [3]South Australian Immunogenomics Cancer Institute and Freemasons Foundation Centre for Men's Health, University of Adelaide, Adelaide, SA, Australia, [4]South Australian Health and Medical Research Institute, Adelaide, SA, Australia

Precision medicine has emerged as an important paradigm in oncology, driven by the significant heterogeneity of individual patients' tumour. A key prerequisite for effective implementation of precision oncology is the development of companion biomarkers that can predict response to anti-cancer therapies and guide patient selection for clinical trials and/or treatment. However, reliable predictive biomarkers are currently lacking for many anti-cancer therapies, hampering their clinical application. Here, we developed a novel machine learning-based framework to derive predictive multi-gene biomarker panels and associated expression signatures that accurately predict cancer drug sensitivity. We demonstrated the power of the approach by applying it to identify response biomarker panels for an Hsp90-based therapy in prostate cancer, using proteomic data profiled from prostate cancer patient-derived explants. Our approach employs a rational feature section strategy to maximise model performance, and innovatively utilizes Boolean algebra methods to derive specific expression signatures of the marker proteins. Given suitable data for model training, the approach is also applicable to other cancer drug agents in different tumour settings.

KEYWORDS

precision oncology, predictive biomarker, machine learning, feature selection, Boolean function minimization, prostate cancer, Hsp90 inhibitor, 17-AAG

Introduction

Precision treatment has become an important treatment modality in oncology, where the molecular makeup of patients' tumour dictates therapeutic decisions. Identifying predictive biomarkers of treatment response that aid stratification of patients is critical for effective deployment of personalized oncology (Barretina et al., 2012; Geeleher et al., 2014; Nalejska et al., 2014). However, for most existing cancer drug agents including those that have been clinically approved (Pernas et al., 2018; Zhong et al., 2021), we currently lack companion predictive biomarkers that can reliably predict treatment response and inform patient selection. Thus,

identification of predictive response biomarkers for cancer therapies represents a broad and unmet clinical need.

A major challenge that complicates the identification of response biomarkers is the multi-factorial determinant of cellular response to drug treatment, which is further accentuated by the extensive tumour heterogeneity between patients (Turajlic et al., 2019; Xi et al., 2019; Lee et al., 2021). Consequently, except for a few notable cases of clinical success (Quintás-Cardama and Cortes, 2009; Dieci et al., 2020), single-gene biomarkers are insufficient for predicting treatment responses and unlikely to be clinically useful (Nguyen et al., 2016). Instead, multi-gene biomarker panels are more likely to capture the complexity underpinning tumour drug response, and deliver better prediction (Zhu et al., 2011; Lima et al., 2019). Biomarker discovery approaches therefore should explicitly model combinations of relevant marker genes/proteins.

Computational methods have been key in the derivation of response biomarkers for cancer therapeutics (Menden et al., 2013; Tabl et al., 2019; Fortino et al., 2020). A simple but commonly used approach is to identify genes (or proteins) that are differentially expressed between treatment-sensitive and -resistant groups using–omics data such as transcriptomic or proteomic data (Chen et al., 2016; Nguyen et al., 2018). However, the degree of differential expression of a gene (based on fold-change and/or *p*-value) is not a good indicator of its predictive power towards treatment responsiveness. Moreover, the lists of DEGs are typically very long, and without further analysis to prioritize and narrow them down, the applicability of DEGs-based approaches remains limited. Recently, more sophisticated approaches such as machine learning (ML) have been applied to the biomarker discovery domain (Parca et al., 2019; Tabl et al., 2019; Fortino et al., 2020; Nguyen et al., 2021).

Yet, the 'curse of dimensionality' widespread in pharmacogenomics data - where the number of molecular features often far exceeds the number of biological samples - necessitates the development of feature selection strategies for ML algorithms (Huang et al., 2018; Nguyen et al., 2021; Ogunleye et al., 2022.). For example, Ballester and others have developed a scheme termed Optimal Model Complexity (OMC) aimed at identifying a smaller subset of informative features from the much larger original feature space, and integrated OMC with various ML algorithms (Bomane et al., 2019; Naulaerts et al., 2020; Nguyen et al., 2021). OMC works by ranking the features using the *p*-values obtained from univariate statistical tests to correlate between each feature and the corresponding labels (e.g., IC50 values of treated drugs), thereby pinpointing the most relevant features prior to model training (Nguyen et al., 2021). OMC-based XGBoost was employed to predict cancer drug response using pharmacogenomic data derived from either cancer cell lines (Yang et al., 2013; Naulaerts et al., 2020), or cancer patient-derived xenografts (Gao et al., 2015; Nguyen et al., 2021). In addition, Bomane et al. has also applied OMC to other ML algorithms, including Random Forest and LightGBM, to predict response to paclitaxel treatment in breast cancer (Bomane et al., 2019). On the other hand, Parca et al. (2019) selected potentially informative molecular genes for predicting cancer drug response by analysing the variance in gene expression profiles using cell lines based pharmacogenomic datasets. To predict cancer patient response to chemotherapeutic drugs, Huang et al. (2018) employed standard recursive feature elimination method to select for most relevant features (gene expression data) and applied it on top of support vector machine algorithm. Other studies utilise knowledge-based approaches to select likely relevant subsets of features: for example, by leveraging the protein-protein interaction network surrounding the drug targets (Kong et al., 2020), or restricting those to genes in the cancer gene census set (Futreal et al., 2004). However, due to lack of relevant patient-derived pharmacogenomic data, most ML studies to date have been performed using panels of cancer cell lines (Barretina et al., 2012; Garnett et al., 2012; Seashore-Ludlow et al., 2015; Iorio et al., 2016), which do not necessarily reflect the heterogeneity and drug sensitivity in human tumours (Borst and Wessels, 2010; Gillet et al., 2013).

In this study, we have developed a generally applicable machine learning framework for identification of multi-gene predictive biomarker panel and associated expression signatures for anti-cancer drugs. The approach comprises two phases (Figure 1). The first is to identify optimal biomarker panels that predict drug response using ML coupled with a new feature selection strategy. The second is to derive expression signatures of the identified biomarkers for different response groups utilizing a new Boolean function minimization-based pipeline. We applied the new approach to identify predictive biomarker panels and expression signatures for 17-AAG, a small-molecule inhibitor targeting heat shock protein 90 (Hsp90), using pharmacoproteomic data obtained from prostate cancer patient-derived explants (PDEs) (Cardillo and Ippoliti, 2006). Blocking Hsp90 is considered as an attractive therapeutic strategy for prostate cancer. This is because Hsp90 is commonly overexpressed in prostate cancer compared to normal prostate cells (Cardillo and Ippoliti, 2006); prostate cancer cells are selectively sensitive to Hsp90-directed agents; and Hsp90 clients include the androgen receptor (Trepel et al., 2010), a major driver of prostate tumorigenesis. However, despite the anti-tumour activity of Hsp90 inhibitors (e.g., 17-AAG) in preclinical models (Solit et al., 2002), the lack of companion predictive biomarkers for rational patient stratification have in part contributed to the poor response rates to these agents seen in clinical trials (Heath et al., 2008).

We derived a 16-protein biomarker panel that achieved 92% response prediction accuracy to 17-AAG. To facilitate clinical translation, we further reduced this to a compact 5-protein panel having 80% prediction accuracy, and identified associated expression signatures. Interrogation of prostate cancer patient data identified almost half of the patients with matching expression signatures, who may benefit from 17-AAG treatment. Overall, this work presents a novel ML framework that aids the discovery of predictive biomarker panels for improved patient selection and treatment of cancer.

Materials and methods

Patient data analysis

Patient data from two prostate cancer patient cohorts were used to interrogate the utility of our derived 5-gene biomarker panel as a potential patient stratification tool. These include the TCGA (Pancancer Atlas, (Hoadley et al., 2018)) cohort containing 494 patients; and another independent prostate cancer patient (PNAS 2019, (Abida et al., 2019)) cohort containing 208 patients. Patient-specific transcriptomic data was downloaded from the cBioPortal for Cancer Genomics database (Cerami et al., 2012; Gao et al., 2013)) for analysis (see Figure 5E, left panels).

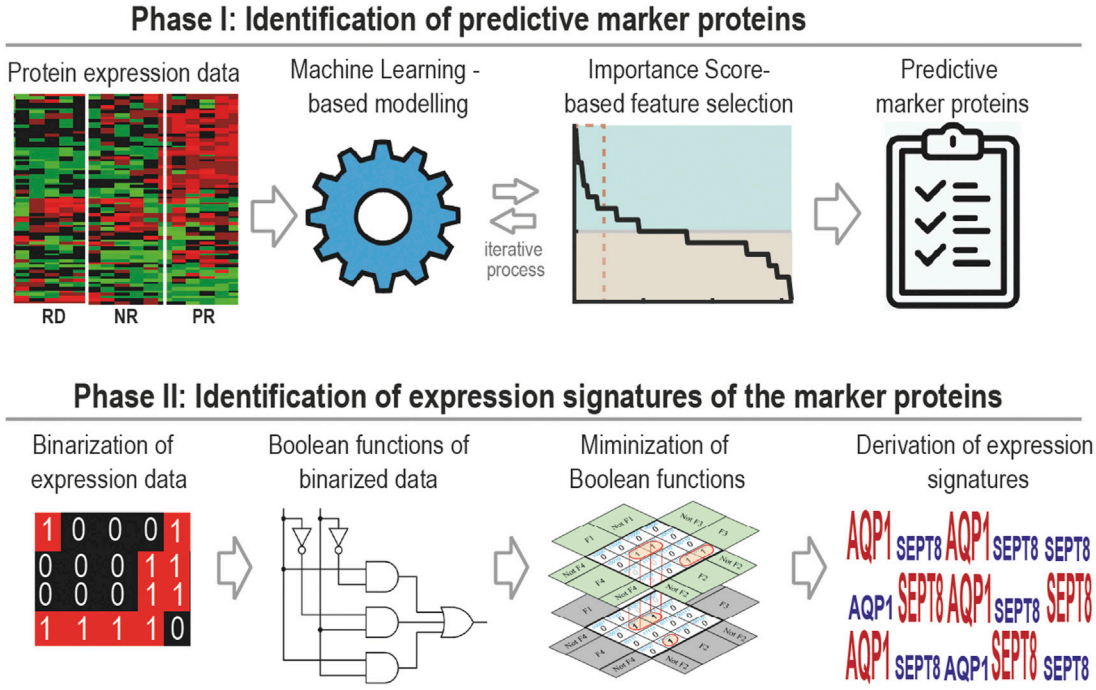


FIGURE 1
A general workflow of our two-phase computational framework, which couples supervised machine learning-based biomarker discovery with Boolean algebra-based signature derivation for the identification of predictive biomarkers.

Patient-derived explant (PDE)

Fresh prostate cancer specimens were obtained with written informed consent through the Australian Prostate Cancer BioResource from men undergoing robotic radical prostatectomy at the Royal Adelaide Hospital and St Andrew's Hospital (Adelaide, South Australia). Tumors from two cohorts of patients were used for this study: a discovery cohort (n = 40, obtained from (Nguyen et al., 2018)), and a test cohort (n = 7). A single 6 mm core of tissue was obtained per patient. A longitudinal section of the entire core was taken for hematoxylin and eosin (H&E) analysis of tumor content. The remaining tissue was dissected into 1 mm3 pieces and cultured in triplicate on a presoaked gelatin sponge (Johnson and Johnson, New Brunswick, NJ) in 24-well plates containing 500 L RPMI 1640 with 10% FBS, 1 antibiotic/antimycotic solution (Sigma, St Louis, MO), 0.01 mg/ml hydrocortisone, 0.01 mg/ml insulin (Sigma) and cultured for 48 h with 17-AAG (500 nM) or DMSO vehicle alone as previously described (Nguyen et al., 2018). Mass spectrometry-based proteomic profiling were performed on the discovery cohort as described in (Nguyen et al., 2018), and these data were used as inputs (features) for our ML models. Treatment response was quantified based on the relative expression of the proliferative marker Ki67, measured post drug treatment by immunohistochemical (IHC) assay (Nguyen et al., 2018).

Immunohistochemical staining

Paraffin-embedded tissues were sectioned (2 mM) on Ultraplus slides prior to H&E staining and IHC detection of Ki67 (Agilent, M7240 antibody; 1:200 dilution, Santa Clara, CA). IHC staining was performed and tissues assessed for tumor content and Ki67 positivity in a blinded fashion as described in (Armstrong et al., 2016).

qRT-PCR

Real-Time Quantitative Reverse Transcription PCR was used to measure the baseline expression of the five genes identified in our 5-gene biomarker panel: AQP1, SEPT8, RBM17, TRIM47, and VPS25 for the testing PDE cohort (Supplementary Table S5). qRT-PCR was also used to measure the baseline and post-treatment expression of *MKi-67*, the gene encoding of the proliferative marker Ki67. Cultured patient derived explants were placed in a Precellys Tissue Homogenizer (Bertin instruments) for 2 cycles at 6500rpm. RNA was extracted from tissue homogenate using miRNeasy mini kit (Qiagen) according to manufactures instructions. RNA (700 ng) was reversed transcribed to cDNA using IScript cDNA synthesis kit (Bio-Rad). QRT-PCR was performed with a 1:10 dilution of cDNA using SYBR green (Bio-Rad) on a CFX 384 real time system (Bio-Rad). Relative gene expression was calculated using the comparative ct method and normalized to internal control genes GAPDH & TUBA1B. Primer sequences used for PCR are given in Supplementary Table S5.

ML implementation

To classify drug response groups, we developed a multi-class Support Vector Machine (SVM) and an artificial neural network

(ANN) classifier using the MATLAB function *fitcecoc* and *patternnet*, respectively. For the multi-class SVM model, we set the "standardized" option to "true," which normalized the predictor data and used the option *linear* as the kernel function of mSVM. For the ANN model, the predictor data was also normalized, and the size of the hidden layers was set to 10. Protein expression data profiled from 40 prostate cancer PDE samples was used for model training and testing. For training and testing, the functions' default settings were used (e.g., scaled conjugate gradient backpropagation algorithm (Møller, 1993), implemented using MATLAB function *trainscg*), with 80%–20% data split ratio. For the implementation of K-Nearest Neighbor, Naïve Bayes, Random Forest, and AdaBoost we used Matlab functions *fitcknn* (Distance = 'Euclidean'), *fitcnb* (Kernel = 'Normal'), *fitrensemble* (Method = 'Bag'), *fitcensemble* (Method = 'AdaBoostM2'), respectively. For Deep Forest, we utilized the Matlab codes developed by (Zhou and Feng, 2017), available at Github at https://github.com/cnzakimuena/casForest.git.

These data were deposited onto the Mass spectrometry Interactive Virtual Environment (www.massive.ucsd.edu) with identifier: MSV000082244 (Nguyen et al., 2018). Model validation was performed using the function *predict* for mSVM and *sim* for ANN, respectively. ROC curves and confusion table were generated using functions *roc* and *confusionchart* in MATLAB. All the relevant codes were deposited on Github at https://github.com/NguyenLabNetworkModeling/GFFS-Biomarker.

Importance score calculation

The importance score (IS) associated with a feature was calculated through performing a systematic 'feature drop-out' analysis. For this, each feature (e.g., DEP) was removed from the feature list, one at a time, and the effect on model prediction performance was assessed. IS measures the difference in prediction accuracy between the 'drop-out' and the original mSVM model, computed as follows:

$$IS(i) = -\frac{PA_i - PA_O}{PA_O} \quad (1)$$

where PA_O and PA_i represent the prediction accuracy of the original mSVM and the "drop-out" model where input feature i is removed from the feature list. Thus, IS > 0, <0, = 0 indicate the dropped-out feature has a positive, negative, or no impact on the model predictive performance, respectively.

Feature selection

Our feature selection strategy GFFS was implemented based on the IS values, as described in the text. To compare different feature selection algorithms, we implemented ReliefF using the function *relief* and MRMR using the function *fscmrmr* in MATLAB. For LASSO regression, we used the function *fitcecoc* and *lasso* as a regularization method. To calculate the importance score of the Boosting and Bagging ensemble models, we used the functions predictorImportance and oobPermutedPredictorImportance in MATLAB. We also implemented RFE and FFE strategies on top of the SVM.

Explainable ML analyses

SHAP and LIME analyses were implemented using the *shapley* and *lime* functions in MATLAB, and *AdaBoostM2* as a 'black-box ensemble model'.

Boolean functions and minimization

A Boolean function is an algebraic expression consisting of n-binary variables, $f(x_1, x_2, \ldots, x_n)$. Boolean functions can be formulated through Sum of Product (SOP) or Product of Sum (POS). In SOP, different product terms of inputs are summed together, where the products are logical AND the sum are OR operators. For example: $x'+xy + yz$' where x, y and z are binary variables and prime (') represent complement of a variable, that is if x = 0 then $x' = 1$. On the other hand, in POS products of different summation terms of inputs are taken, e.g. $(x')\cdot(x + y)\cdot(y + z')$. Boolean functions can be simplified using Boolean laws and theorems (Hanf, 1975; Whitesitt, 2012). The process of simplifying the algebraic expression of a Boolean function is called 'minimization'. To minimize the Boolean function, we employed the Quinine-MacCluskey algorithm (Jain et al., 2008) implemented in MATLAB (http://www.tu-harburg.de/~rtsap/#Programs).

Statistical and bioinformatic analysis

Statistical t-tests were performed using GraphPad Prism 9 and Matlab R2022b. For the KEGG pathway and the GO function analysis we utilized Enrichr web application (Kuleshov et al., 2016), which can be accessed at https://maayanlab.cloud/Enrichr/.

Pharmacoproteomic data from prostate cancer PDEs for machine learning

To recapitulate the *in vivo* response of prostate cancer to therapies, we have previously developed an *ex vivo* culturing model of prostate cancer tissue that retains the structure and stromal-epithelial interactions of the tumor microenvironment and provides the level of disease heterogeneity seen in patients (Centenera et al., 2012). Using this system, we established in a previous study 40 prostate cancer patient-derived explants (PDEs) and subjected them to either vehicle (DMSO) or 17-AAG (500 nM) treatment for 48 h (Figure 2A, see also Materials and Methods) (Nguyen et al., 2018). Treatment response was quantified based on the relative expression of the proliferative marker Ki67, measured post drug treatment by immunohistochemical assay (Nguyen et al., 2018). In addition, we performed mass spectrometry-based proteomic profiling and HRM-DIA data analysis on the corresponding 40 PDEs, which identified the expression of 3,766 quantifiable proteins prior to 17-AAG treatment (Nguyen et al., 2018). These datasets will be used in this study to develop companion biomarkers that accurately predict response to 17-AAG treatment.

To label the data, the PDE samples were classified into three distinct response groups based on changes in Ki67 positivity upon treatment with 17-AAG (Nguyen et al., 2018). These are depicted in Figure 2B: (i) RD (responders) group containing PDEs having > two-

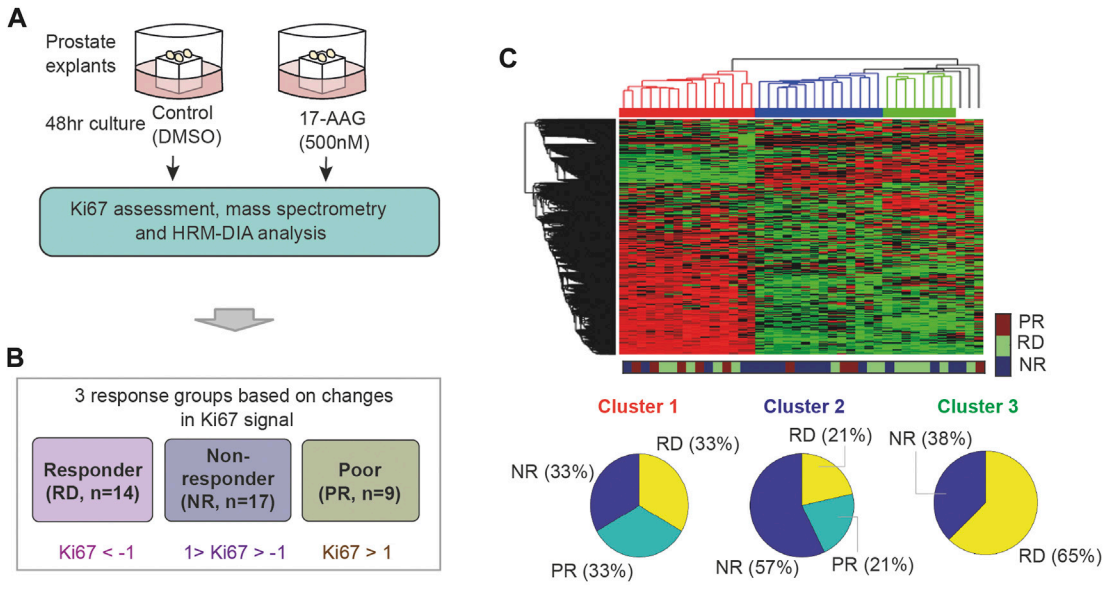


FIGURE 2
Unsupervised clustering of the PDEs' response to 17-AAG treatment. **(A)** Pharmacoproteomic data obtained from 40 prostate cancer PDEs (Nguyen et al., 2018). **(B)** Classification of the PDEs into three response groups based on Ki67 positivity post 17-AAG treatment: RD, PR, and NR with indicated fold-change in Ki67 positivity (in log2 scale). **(C)** Unsupervised hierarchical clustering of the PDEs using the proteomic data, which failed to reasonably predict response to 17-AAG.

fold decrease in Ki67 positivity; (ii) PR (poor responders) group containing PDEs with < two-fold increase in Ki67 positivity; and (iii) NR (non-responders) group with Ki67 positivity in between. As a result, 14 PDEs were classified as RD, 17 as NR and 9 as PR (Figure 2B; Supplementary Table S2). Of note, the PDE proteomic data has 0.16% missing (undetectable) values and they were imputed with random values generated from a uniform distribution between 0 and 1 (1 is the minimal machine-detectable protein amount) (Wei et al., 2018). Together, the PDE data consists of protein expression levels of 3,766 proteins serving as 'input features' and Ki67-based response classification serving as 'labelled outputs' for development of ML models.

Results

Supervised ML using differentially expressed proteins (DEPs) sub-optimally predicts 17-AAG response

Using expression levels of all the 3,766 proteins as inputs, we first tested whether unsupervised hierarchical clustering could predict the PDE response to 17-AAG treatment. While this identified three distinct clusters, they poorly reflected the labelled response groups (Figure 2C). Each of the three clusters comprises a good mixture of RD, NR and PR samples, suggesting that unsupervised clustering could not reasonably predict response to 17-AAG.

Next, to examine if supervised ML methods would improve the response prediction, we developed a multi-class support vector machine (mSVM) model using the protein expression as inputs and the labelled drug responses (RD, PR, NR) as outputs (Figure 3A). The dataset was randomly divided into a training (80%, 32 PDEs) and a test set (20%, 8 PDEs). To avoid biases in data splitting and mitigate model overfitting, we held out the test set and trained the model with the training set. This training and test process were repeated 50 times to obtain reliable and robust performance evaluation (). We found that the model displayed an average prediction accuracy of ~39% (Supplementary Figure S1). This poor performance is somewhat expected because the number of input variables/features (3,766) greatly exceeds the number of samples (40), a phenomenon known as 'curse of dimensionality' in ML (Hughes, 1968). By this principle, the prediction power of a ML classifier typically improves as the number of the features gradually increases, but after an threshold (i.e., optimal) number of features, adding more starts to diminish the model performance (Hughes, 1968). This is because the high dimensionality of the input data causes every observation to appear equidistant from the others, preventing meaningful clustering (Hughes, 1968). Moreover, irrelevant or partially relevant features can negatively impact model performance (John et al., 1994).

In order to circumvent the curse of dimensionality, we carried out a feature selection strategy with the goal to rationally reduce the number of non-relevant features (Cai et al., 2018; Gopika and Meena Kowshalya, 2018). As differentially expressed proteins (DEPs) often provide a good starting point for identifying potential biomarkers (Chen et al., 2016; Nguyen et al., 2018), we first performed differential expression analysis between the three response groups using analysis of variance (ANOVA) tests, and obtained a total of 157 DEPs (p-value <0.05, Supplementary Table S2). Unsupervised hierarchical clustering using these DEPs still failed to appropriately cluster the PDE samples (Supplementary Table S2), confirming the suboptimal performance of this approach. Next, we retrained the mSVM model using the 157 DEPs as input features and found that it displayed an average prediction accuracy of 77% (Figure 3B).

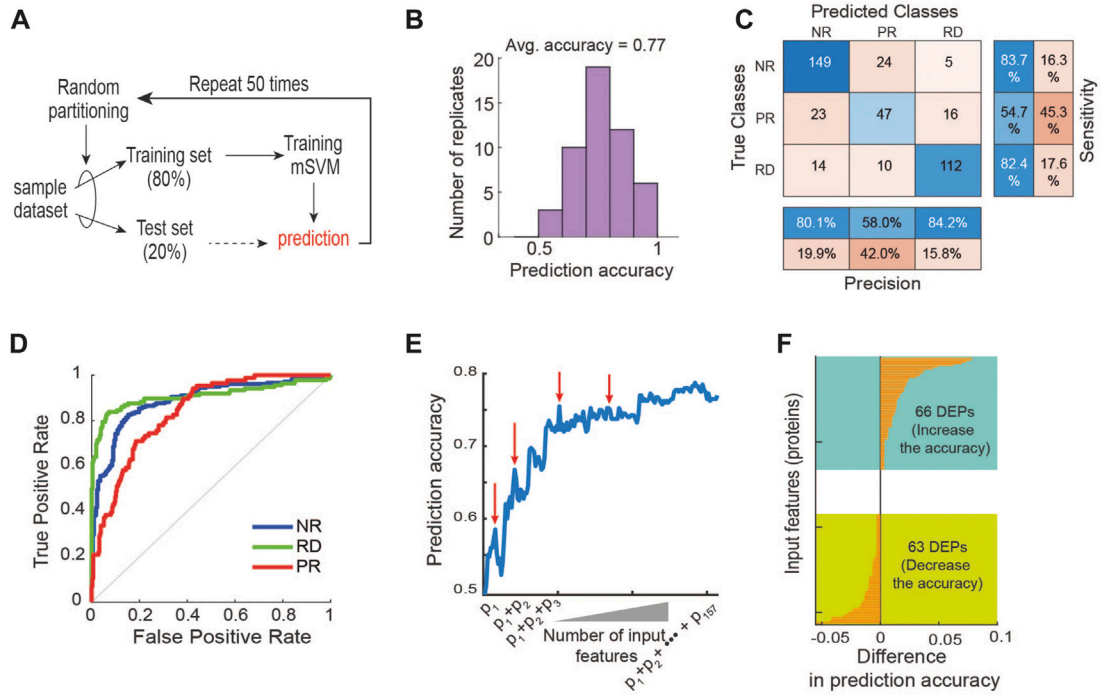


FIGURE 3
Machine learning-based prediction of response to 17-AAG. **(A)** A machine learning pipeline utilizing the repeated holdout method for training and testing with an mSVM model (see Materials and Methods). **(B)** Distribution of prediction accuracy performance across 50 model replicate runs. **(C)** Performance of the mSVM classifier using all 157 DEPs as input features, summarized in a confusion matrix. **(D)** Receiver operating characteristic (ROC) curves of the mSVM classifier for the RD, NR and PR response groups. **(E)** Impact of input feature space modulation on model prediction accuracy. Adding specific input features may worsen model performance, indicated by the red arrows. **(F)** Difference in the prediction accuracy when a new input feature is added to the training data.

Examining the confusion matrix results further showed that while the precision and sensitivity for the RD and NR groups are around and above 80%, they are below 55% for the PR group (Figure 3C), demonstrating the mSVM model did not perform well against the PR group. Consistently, analysis of the receiver operating characteristic (ROC) curves confirms that model performance against the PR group was inferior compared to the other groups (Figure 3D).

In addition to mSVM, for comparison purposes we also performed similar analyses using an array of common ML methods, including artificial neural network (ANN), K-Nearest Neighbor (KNN) (Min-Ling and Zhi-Hua, 2005; Zhang, 2016), Naive Bayes (Yousef et al., 2007), Decision Tree (Navada et al., 2011), AdaBoost (Feng et al., 2020), Random Forest and Deep Forest (Liu et al., 2012; Zhou and Feng, 2018; Su et al., 2019). The results show that mSVM was the best-performing algorithm, followed by ANN and KNN (Supplementary Figure S3). Like mSVM, the ANN model performed relatively poorly in predicting the PR group (Supplementary Figure S4). Together, these results suggest that although supervised ML approaches perform better than unsupervised hierarchical clustering, using all the DEPs as features may be inadequate for optimizing predictive power. This may be due to the noise exhibited by certain DEPs that bear no relevance in predicting response to 17-AAG, which interferes with the predictive signals from the relevant features, thereby lowering the model's overall predictive performance (Blum and Langley, 1997).

To interrogate how modulation of the input feature space may influence performance of the mSVM, we systematically increased the number of features by adding the DEPs one by one to the training set and re-evaluated the model prediction accuracy. Figure 3E shows an overall upward trajectory of prediction accuracy as the number of feature increases. However, there were specific DEPs whose addition to the feature space actually worsened the model's predictive power, evidenced by drops in the trajectory (indicated by red arrows, Figure 3E). Specifically, 66 of the 157 DEPs contributed positively to the model performance while 63 contributed negatively, and some had negligible effects on performance (Figure 3F). These results support the idea that irrelevant features can negatively impact the model's ability to predict drug response, and thus rational selection of informative features is key in improving predictive performance.

A novel ML framework maximises prediction accuracy through rational feature selection

To select the most relevant features from the DEPs, we first performed a systematic feature drop-out analysis. One at a time, each DEP was removed from the feature space and the effect on performance of the mSVM was assessed, as compared to the original model using all the 157 DEPs as features (workflow in Figure 4A). If removal of a protein attenuates/improves the model prediction accuracy, then the protein is deemed to have a positive/negative impact on drug response prediction. We quantified these effects by defining an 'importance score' (IS) as in Eq. 1 that computes the difference in prediction accuracy between the drop-out and original

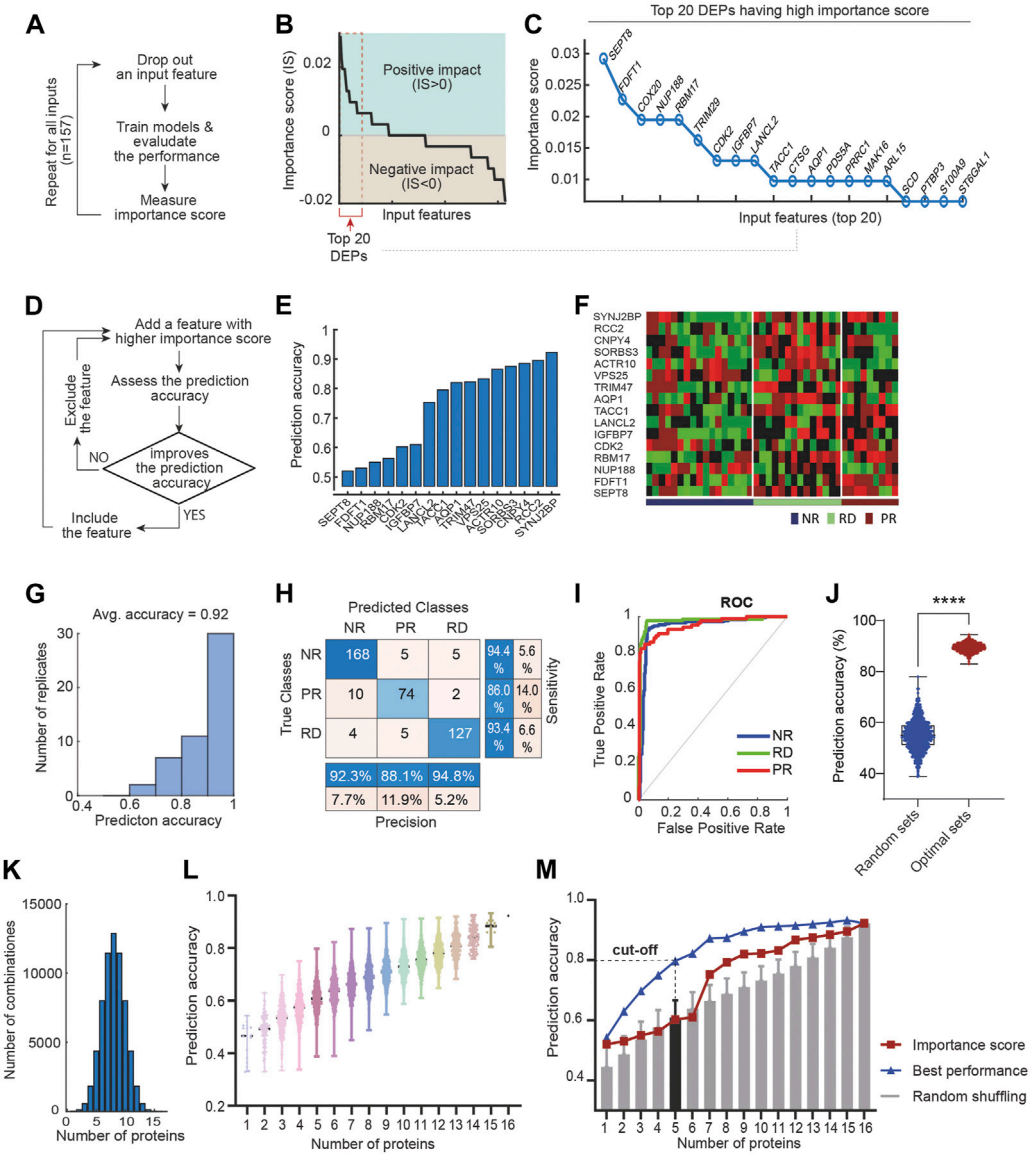


FIGURE 4
Rationalized feature selection optimizes ML model's prediction accuracy. **(A)** A schematic of the drop-out analysis that enables calculation of the importance score (IS) for each feature. **(B)** Importance score values of all the DEPs, sorted in a descending order. IS > 0, <0, and = 0 indicates proteins having negative-impact, positive-impact and no-impact on drug response prediction, respectively. **(C)** Top 20 of 157 DEPs having the highest importance score values. **(D)** A schematic workflow of our IS-based feature selection strategy. **(E)** Gradual incorporation of 16 proteins that ultimately leads to optimal prediction accuracy of drug response. **(F)** Heatmap displaying the expression data of the identified 16 marker proteins across the discovery PDE cohort. **(G)** Distribution of prediction accuracy performance across 50 model replicate runs. **(H)** Performance of the optimal mSVM classifier using the 16 DEPs as input features, summarized in a confusion matrix. **(I)** ROC curves of the mSVM classifier for the RD, NR and PR response groups. **(J)** Comparison of performance between the optimal mSVM model and models using 16 randomly selected features (from 157 DEPs); **** p-value <0.0001 (unpaired t-test, n = 1,000). **(K)** Tally of possible combinations of proteins with increasing size ranging from 1 to 16. **(L)** Prediction accuracy significantly varies depending on the combination specification of input features. **(M)** Performance comparison between the optimal model and those with randomly selected features (shuffled from the 16 identified biomarkers), displayed for increasing panel size.

mSVM models. Thus, IS > 0, <0, and = 0 indicates proteins having positive-impact, negative-impact and no-impact on drug response prediction, respectively (Materials and Methods). Figure 4B displays a sorted list of the 157 DEPs according to the respective IS values. Interestingly, a large fraction (48%) of the DEPs had a negative impact on drug response prediction (Figure 4B), suggesting inclusion of these in the feature space may diminish the model performance. In contrast, more than half of the DEPs had a positive impact on the drug response prediction (Figure 4B), with the top 20 proteins shown in Figure 4C.

We reasoned that the positive-impact DEPs would represent good candidate features for maximizing model prediction. We next introduced a new algorithm, termed greedy forward feature selection (GFFS), which aimed to select the optimal combination of features from the pool of positive-impact DEPs. A schematic of the algorithm is given in Figure 4D. First, we trained the mSVM using the positive-impact DEP having the highest IS (i.e., SEPT8; Figure 4E) as the single input variable, employing a similar training/validation data splitting scheme as in Section 3.2. Unsurprisingly, this single-feature

model achieved ~50% accuracy (Figure 4E), much worse compared to the model using all 157 DEPs. Next, we retrained the model by adding the second most influential DEP (i.e. FDFT1 having the second highest IS) to the feature space and re-evaluated the model performance. Because the new model had a better overall prediction accuracy, FDFT1 was kept as an input feature (Figure 4E). This process was repeated by gradually adding the next most important DEP to the feature space: if the new DEP improves prediction accuracy then it is kept; however, if it attenuates (or does not affect) accuracy, the protein is skipped and we move to the next positive-impact DEP. This was done until all the positive-impact DEPs were considered and the model performance did not further increase (Figure 4D). As a result, we determined an optimal feature space containing 16 DEPs, depicted in Figure 4E. The corresponding mSVM model achieved an overall prediction accuracy of 92% (Figures 4E–G), which was significantly superior to the initial model using all the 157 DEPs (77%, Figure 3B). This was further confirmed by examining the confusion matrix (Figure 4H) and the ROC curves (Figure 4I), indicating significantly improved prediction of 17-AGG response within each of the response groups.

A key attribute of our ML-based algorithm is the rationalized selection of features guided by prior IS-based ranking. To determine if this was critical in enhancing prediction accuracy, we assessed the performance of mSVM models using randomly selected features instead, and replicated this 1,000 times. The result shows that the model with feature selection consistently and significantly outperformed the random-feature models (Figure 4J), suggesting our IS-based feature selection strategy was key in boosting predictive power.

Next, we comparatively evaluated the performance of our IS-based GFFS approach with a range of available feature selection algorithms, including filter (ReliefF; minimum redundancy maximum relevance (MRMR)) (Ding and Peng, 2005; Stief et al., 2018), wrapper [recursive feature elimination (RFE); forward feature selection (FFS)] (Aha and Bankert, 1996; Tang et al., 2007; Marcano-Cedeño et al., 2010; Zhang et al., 2013) and embedded (boosting; bagging; least absolute shrinkage and selection operator (LASSO)) methods (Vasquez et al., 2016; Alsahaf et al., 2022) (Supplementary Figure S5A). GFFS showed significantly better predictive accuracy than all of the tested methods except for RFE, with which GFFS had comparable performance (Supplementary Figure S5A). Interestingly, the top three performers were GFFS, RFE, and FFS, highlighting the importance of rational feature selection in this context. We note that the maximal performance of FFS and RFE was achieved with 41 and 28 features, respectively (Supplementary Figures S5C, D), which were higher than GFFS, but at the cost of much larger number of features. Importantly, among of the top three methods, GFFS's running time scales linearly and was significantly better than RFE and FFS (Supplementary Figure S5B). Thus overall, GFFS-based feature selection achieved a strong and balanced performance in terms of predictive accuracy and computational cost.

Identification of a compact biomarker panel for 17-AAG treatment response

There is a general trade-off between the size of a biomarker panel and its practical applicability. A panel having more relevant proteins tends to deliver enhanced prediction, but this comes at a cost of having to detect more readouts from patients–a non-trivial task for poorly characterized biomarkers. In order to facilitate translation of the predictive biomarkers for 17-AAG-based therapy, here we aim to derive a more compact-size panel from the 16 identified marker proteins while maintaining high predictive performance. To this end, we considered all possible ways to combine the marker proteins into panels with increasing size, ranging from 1 to 16 (Figure 4K). As such, there are 16 possible panels with size 1; 120 panels with size 2; 4368 panels with size 5; and so on. We then evaluated the predictive performance of the mSVM model using each panel as input features. The results, displayed in Figure 4L, show that for each panel size the prediction accuracy varied significantly depending on the specific composition of the feature proteins (Supplementary Table S3). For instance, among 4368 5-protein panels, the one comprising AQP1, SEPT8, RBM17, TRIM47, and VPS25 exhibits the highest prediction accuracy of 80% (Figure 4L–M). Interestingly, this panel significantly outperformed the 5-protein panel derived from ranked IS score (accuracy 60%, Figure 4M), and panels derived from random shuffling (accuracy 61%, Figure 4M). Moreover, this 5-protein panel also outclassed the model using all the 157 DEPs (accuracy 77%, Figure 3B). Taken together, given its small size yet excellent predictive power, we concluded [AQP1, SEPT8, RBM17, TRIM47, VPS25] as a novel, practical biomarker panel for predicting response to 17-AAG treatment in prostate cancer.

Machine learning models have traditionally been treated as "black boxes". As ML applications become more widespread, it is important to better interpret ML-based predictions and decision-making processes. Largely, the model interpretability (or explainability) methods can be categorized in two types: (i) global and (ii) local approaches (Ribeiro et al., 2016; Lundberg and Lee, 2017; Linardatos et al., 2021). Global explainability approaches explain the model's behavior as a whole (across whole samples). For example, which features in the model contribute to the model's prediction performance and how important they are. In Figure 4C, we have analysed the importance of individual features through performing a systematic "feature drop-out" analysis, which exactly corresponds to a global explainability method (Guidotti et al., 2019). On the other hands, local explainability approaches explain why and how the model make a particular decision for a particular sample (Guidotti et al., 2019). Among these, LIME (Local Interpretable Model-agnostic Explanations; (Ribeiro et al., 2016; Lundberg and Lee, 2017)] and SHAP (SHapley Additive exPlanations, (Lundberg and Lee, 2017; Linardatos et al., 2020)] have emerged as state-of-the-art approaches. For example, Gardiner et al. have recently applied SHAP to infer important features associated with drug responses (5-ASA, Prednisolone, BIRB796) for patients having inflammatory bowel diseases (Gardiner et al., 2022).

Thus, to examine the relative contribution of each feature (protein) to the prediction of drug response (RD, NR, and PR classification), we implemented SHAP (Lundberg and Lee, 2017; Linardatos et al., 2021) and LIME (Ribeiro et al., 2016; Lundberg and Lee, 2017) analyses. TRIM47, RBM17 and AQP1 were found to positively contribute to model prediction of the RD class; while the VPS25 and SEPT8 contribute negatively instead to the model prediction or not strong enough (Supplementary Figure S7A); The SHAP results were consistent with the importance score of LIME. On the other hand, RBM17 and AQP1 both have a positive impact on NR and PR classes but the contribution of TRIM47 is less significant for

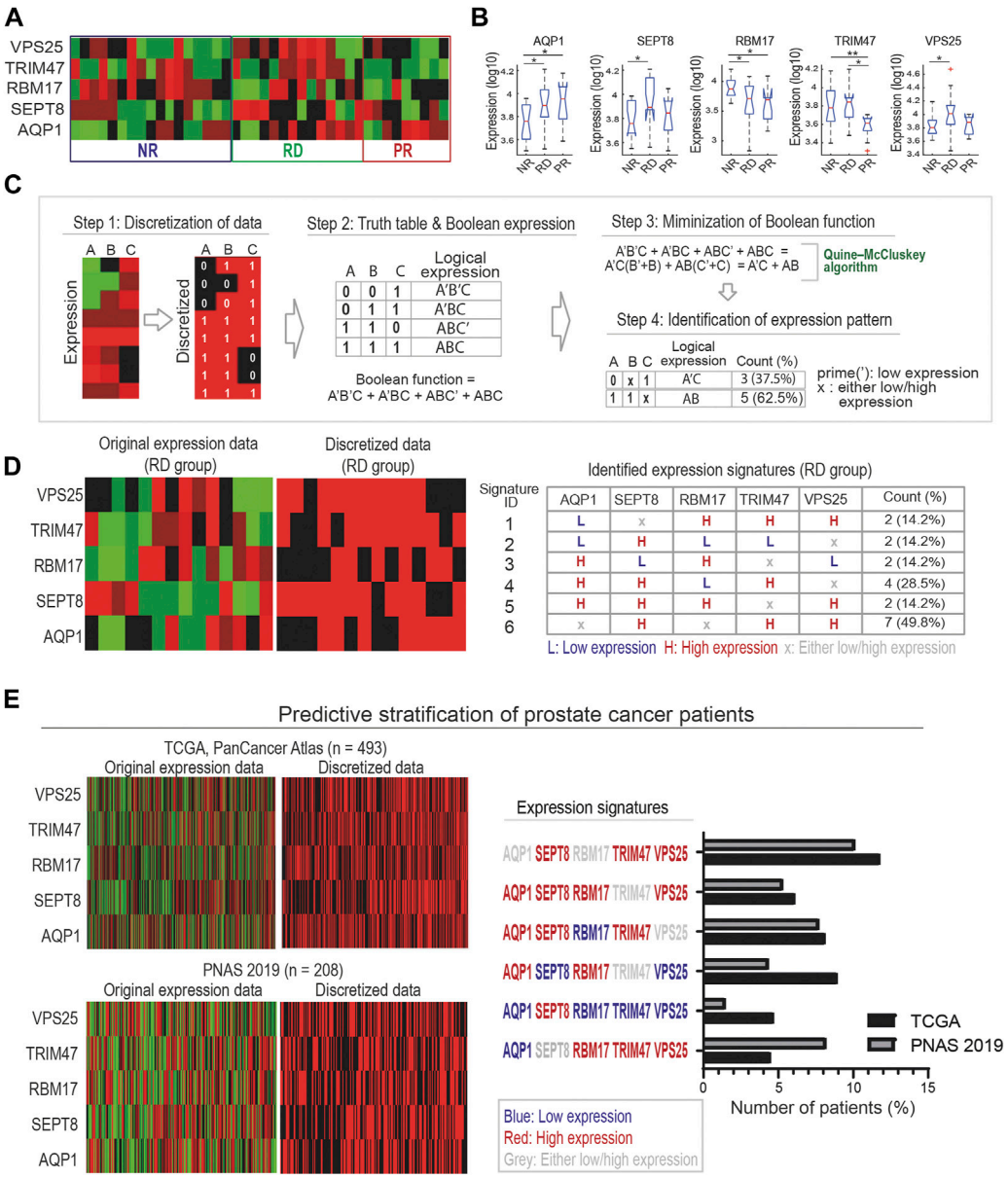


FIGURE 5
A Boolean algebra-based pipeline for derivation of biomarker expression signatures. **(A)** A heatmap displaying protein expression levels of the five marker proteins in our identified compact panel across the discovery PDE cohort. **(B)** Traditional statistical analyses using *t*-test and boxplot to compare expression levels of individual marker proteins between the response groups (* indicates *p*-value <0.05, ** <0.01 (unpaired *t*-test), a red sign indicates outlier data). **(C)** A multi-step Boolean algebra-based pipeline designed to identify combinatorial expression signatures of the biomarkers for each response group. Step 1: Discretization of protein expression levels into binary values. Step 2: Generation of truth table for binarized expression levels that are then transformed into Boolean expressions. Step 3: Minimization of Boolean functions using Quine-McCluskey algorithm, which converts it into simpler, more compact forms. Step 4: Identification of expression signatures of biomarkers. **(D)** Left: the original expression levels and corresponding binarized values of the five markers, shown for the RD group (left panels). Right: List of six identified expression signatures (ID 1–6) of the marker proteins, shown for the RD group. Similar data for the NR and PR groups is shown in Supplementary Figure S6. **(E)** Predictive stratification of prostate cancer patients, using two patient cohorts from the cBioPortal (Materials and Methods). Left: the original gene expression levels and corresponding binarized values of the five markers, shown for all the patients in each cohort. Right: number of patients identified with matching RD-specific expression signatures shown in Figure 5D.

NR. The correlation analysis of features (proteins) with the Shapley values revealed that the AQP1 expression has a negative impact on the NR prediction but a positive impact on PR (Supplementary Table S4). VPS25 and SEPT8 showed a strong correlation with RD and NR although they did not contribute to the model prediction of target variables. Overall, these analyses helped enhance the interpretability of our ML model predictions.

A Boolean algebra-based pipeline to derive biomarker expression signatures

Once the biomarkers have been identified, it is important to define specific expression signatures of these markers that could then be utilized for patient stratification. For this, analyses including *t*-test and boxplot are often employed to deduce the differential expression

patterns of the marker proteins across the response groups. For example, the 17-AAG responsive PDEs (RD group) displayed significantly higher VPS25 expression, while those in the PR group have significantly lower TRIM47 expression compared to the other groups (Figures 5A, B). While useful, these approaches do not consider the expression heterogeneity within each response group (evidenced in Figure 5A) and possible hidden interlinks between the markers. Thus, derivation of biomarker signatures that encapsulate the response group-specific heterogeneity and possible functional links between the markers is important.

Here, we propose a new pipeline to identify combinatorial expression signatures for biomarkers characterizing individual response group utilizing methods from Boolean algebra. The pipeline consists of 4 steps and is illustrated in Figure 5C for example proteins A, B, and C. Step 1 discretizes the continuous expression data into binary values where 1 and 0 indicate high and low expression, respectively. This is done by normalizing the protein expression data to its median value across the samples: normalized value >1 or <1 will be converted to 1 or 0, respectively.

In step 2, all combinatorial binary expression patterns of the proteins are identified and summarized in a 'truth' table, which are then converted into logical expressions of the proteins (Figure 5C). Then, the logical expression of the individual patterns are summed together in a Sum-of-Products (SOP) form using the Boolean operator (+) (Materials and Methods) (Huntington, 1933). In step 3, the summed logical expression is reduced to a minimal form without losing information using a Boolean function minimization algorithm, the Quine-McCluskey algorithm (Jain et al., 2008). Finally, in step 4 the resulting reduced logical expression is converted back into binary expression patterns of the biomarkers. In the example in Figure 5C, we started with four expression patterns involving 3 proteins (A′B′C + A′BC + ABC' + ABC) that were simplified into two patterns (A′C + AB) involving only 2 proteins (Hanf, 1975; Whitesitt, 2012) (see Figure 5C). Here, the prime (') sign indicates the respective protein should be low, and high otherwise.

Next, we applied the new pipeline to our previously identified 5-protein biomarker panel (Figure 5A). As a result, we identified six, five and four specific expression signatures of the biomarkers for the RD, NR and PR response groups, respectively (Figure 5D; Supplementary Figure S6). As an example, Figure 5D displays the six expression signatures for the RD group. Among these, signature ID 6, characterized by concomitant high expression of SEPT8, TRIM47, VPS25 while the expression of AQP1 and RBM17 could be either high or low, represents the most common signature among the RD-group PDEs (~49.8%). The next most common signature, signature ID 4, is however characterized by high expression of AQP1, SEPT8, TRIM47 coupled with low expression of RBM17, while VPS25 expression could be high or low (Figure 5D). The biomarker signatures identified for the NR and PR groups are given in Supplementary Figure S6. In summary, our new Boolean logics-based pipeline has allowed us to identify specific expression signatures of the biomarkers that could be utilized to stratify patients for 17-AAG response.

Validation of the biomarker signatures using independent PDE and patient cohorts

To demonstrate the utility of our derived biomarker signatures as a tool for patient stratification, we interrogated whether there are patients with matching 17-AAG-responsive signatures using publicly available prostate cancer patient datasets. To this end, two prostate cancer patient cohorts were obtained from the cBioPortal for Cancer Genomics database for analysis (see Materials and Methods). Patient-specific gene expression data of the 5 proteins in our biomarker panel were binarized as in step 1 of our pipeline (Figure 5E, left panels). Comparing the expression patterns of these proteins in the patients with the six identified signatures for the RD group showed that in both cohorts, a substantial fraction of the patients displays matching expression signatures (Figure 5E, right). Consistent with our PDE-based prediction, signature ID 6, the most frequent signature of the RD group (Figure 5D), was actually found in more patients than any other RD-specific signatures. Together, these findings support the utility of the derived biomarker signatures in identifying subsets of patients with specific drug response behaviour.

To further validate the predictive power of our identified biomarker signatures, we generated an independent cohort of prostate cancer patient derived explants (n = 7). Tissues were collected, cultured and analysed as previously described for our discovery cohort (Nguyen et al., 2018). These PDEs were treated with DMSO and 500 nM 17-AAG for 48 h. Treatment response to 17-AAG was assessed based on changes in Ki-67 positivity compared to vehicle treatment, detected using immunohistochemical staining and using similar cut-offs as done in the discovery cohort (Figure 6A, Materials and Methods). Baseline expression levels of the five biomarker genes AQP1, SEPT8, RBM17, TRIM47 and VPS25 were measured using qRT-PCR for each PDE under DMSO control (Figure 6B). Then, for each PDE we predicted the drug response using the expression signatures of the biomarkers identified using the Boolean optimization-based pipeline for the different response classes (Figure 5D; Supplementary Figure S6). Following the pipeline, for each PDE we first binarized the gene expression levels of the biomarkers into low or high expression, as shown in Figure 6C. The biomarker expression patterns for each PDE were then mapped to the identified signatures for the three response groups. For example, the expression pattern for PDE X34393R matches with Signature ID 5 of the RD group, which correctly predicted this PDE to be responsive to 17-AAG. On the other hand, PDE X34380R matches with Signature ID4 of the NR group, thus correctly predicting this PDE to be non-responsive to the drug. Overall, cross-validating model predictions with measured drug response, our identified signatures correctly predicted response classification for the responsive and non-responsive PDEs, but did not correctly predict the poor-responders, achieving an overall >71% accuracy on this independent dataset. Despite the small size of the validation cohort, this independent validation analysis has provided a proof-of-concept demonstrating the potential of our predictive pipeline. We envisage as more similar data become available in the future, further validation will be done to strengthen the validity of the identified biomarkers.

Discussion

Precision oncology embraces cancer treatment strategies that are based on the distinct molecular characteristics of a tumour. However, lack of predictive companion biomarkers that help forecast patient-specific treatment response remains a barrier to widespread adoption of this paradigm (Mateo et al., 2022; Pich et al., 2022). In this study, we have developed a novel computational framework that couples

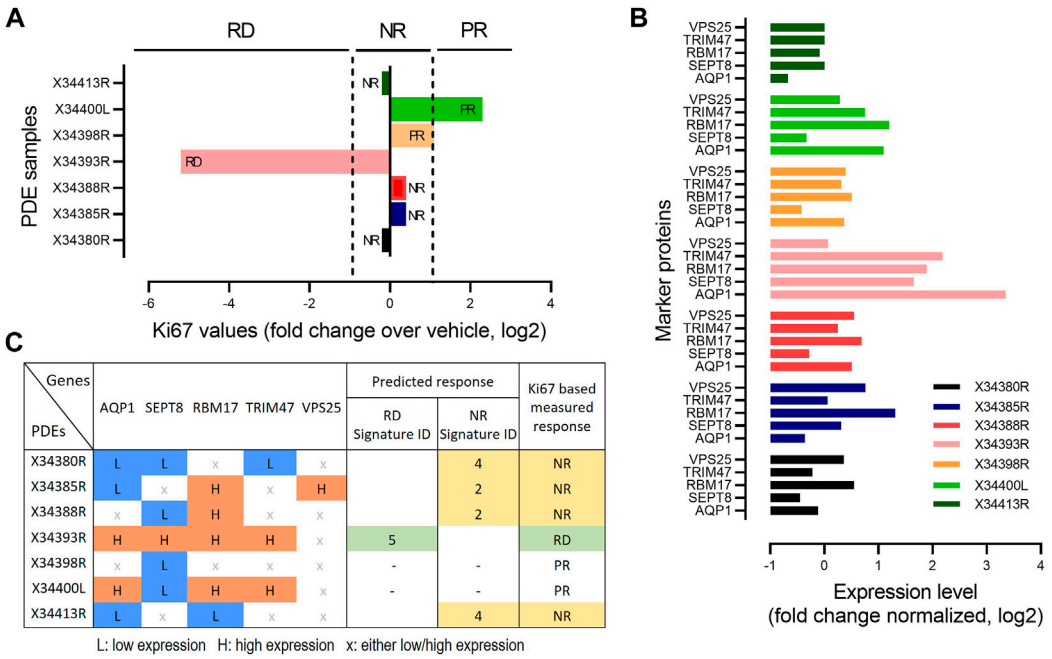


FIGURE 6
Validation of biomarker signatures. **(A)** Drug responsiveness of the seven PDEs to 17-AAG treatment based on fold change in *MKi-67* expression levels by the drug treatment relative to DMSO, measured by IHC. PDEs having > two-fold decrease in Ki-67 levels in response to the drug treatment compared to vehicle were defined as RD; having < two-fold increase as PR); otherwise, NR as in Figure 2B. **(B)** mRNA expression levels of the 5 genes in our identified compact biomarker panels measured by qRT-PCR for each PDE sample. The mRNA expression was normalized to GAPDH and TUBA1B. **(C)** Prediction of drug response based on the identified biomarker expression signatures, compared against Ki67-based response classification. The expression levels of the marker genes were binarized based on the median values by applying 30% and 70% quantile cut-offs: 0 indicates low expression (<30% quantile) and 1 indicates high expression (>70% quantile).

supervised machine learning-based biomarker discovery with Boolean algebra-based signature derivation, in order to identify predictive multi-gene biomarker signatures for cancer therapies (Figure 1). We demonstrated the utility of the approach by applying it to the HSP90 inhibitor 17-AAG in the context of prostate cancer. The approach is however broadly applicable, and given suitable data, can be deployed for different drugs in various tumour types.

The new framework possesses two salient distinguishing properties. First, it rationalizes the most predictive input features based on an importance score that measures how each feature influences the model's predictive performance. Only features that contribute positively to the classification accuracy are retained in the feature space. These are then ranked by their IS values, and increasingly combined one-by-one to identify the optimal combinatorial panel of features that delivers the maximal predictive accuracy. Second, once the biomarkers have been identified, our framework innovatively utilizes Boolean algebra and function minimization [Quine-McCluskey algorithm (Quine, 1955; McCluskey, 1956)] techniques to deduce common expression patterns of the response-specific biomarkers. Because Quine-McCluskey algorithm enables the minimal form of a Boolean function to be reached, our framework helps derive the most compact response group-specific biomarker expression patterns. These easily-interpretable patterns thus constitute biomarker signatures that ultimately allows predictive selection of patient subgroups having a particular drug response, an ability invaluable for precision clinical trials and treatment.

Boolean function minimization algorithms aim to identify the core logics of the underlying phenomenon and are routinely used in engineering fields, such as to design digital logic circuits (Huntington, 1933; Jain et al., 2008). We have previously applied Boolean function minimization to identify core combinatorial feedback loop structures that generate switch-like behavior of E-cadherin (Shin et al., 2010). To our best knowledge, the current study represents the first attempt to apply Boolean function minimization to the problem of biomarker signature derivation. Nevertheless, Boolean logics-based approaches have been used to predict drug response. For example, the LOBICO (Logic Optimization for Binary Input to Continuous Output) modelling framework was developed to explain drug response in cancer cell lines based on binary mutation data of 60 selected genes (Knijnenburg et al., 2016). Using integer linear programming, LOBICO aims to identify the logic combinations of mutations that best explain the response of cancer cell lines to cancer drug agents. In a similar vein, MOCA (Multivariate Organization of Combinatorial Alterations) has been applied to predict drug response by inferring logic combinations of genomic input features (Masica and Karchin, 2013). Overall, our framework represents a novel effort in repurposing Boolean function minimization techniques for derivation of drug-response biomarker signatures.

The results in this study emphasize the importance of rational feature selection in optimizing drug response prediction accuracy by machine learning classifiers. While our GFFS approach is similar to FFS in the sense that it starts with no feature, it differs in two key aspects. Firstly, it pre-determines the relative importance of the input features by calculating the IS beforehand *via* comparison of model performance, and it does this only once (Figures 4A–C). This is

opposed to the classical implementation of FFS, where the relative importance ranking of the remaining features is repeatedly evaluated at each iteration. Secondly, GFFS goes through the pre-ranked features, adds each feature and keeps the feature only if it improves the model performance; otherwise, the feature is dropped and the algorithm moves on to the next feature in the ranked list. Again, this differs from the classical FFS, where addition of a new best feature (n+1 features) at the current round may reduce the overall model performance as compared to the optimal model at the previous round (n features), as shown in Fig. S5C. This difference stems from the fact that in FFS-based implementation, a specific number of features is typically specified prior to model running whereas GFFS does not require such specification. Another consequential difference is that the model performance increases monotonically as more features are added with our approach (Figure 4E), while for FFS the model performance could exhibit a drop as new features are added (Supplementary Figure S5C). Importantly, because of the pre-determined feature ranking, GFFS is computationally much more efficient than FFS and RF. In Big-O notation, our algorithm has O (2n) complexity compared to O (n (n+1)/2) complexity displayed by RFE/FFS (see Supplementary Figure S5B). This superiority in computational cost makes GFFS highly scalable compared to other feature selection techniques, particularly when the number of input features to be assessed is in the range of thousands to tens of thousands.

It is worthy to note that in our two-phase framework, the machine learning coupled GFFS feature selection (phase 1) is integrated with but can work independently from the Boolean logics-based biomarker signature identification (phase 2). As such, in principle the Boolean logics-based biomarker signature identification part can be plugged into any other feature section ML approaches (e.g., using FFS or RFE) and serves as a downstream analysis. This plug-and-play flexibility provides another strength of the framework.

Due to its ability to stabilize client oncogenic proteins and thereby maintain the survival of cancer cells, HSP90 presents an attractive therapeutic target and has been explored in a variety of cancers including prostate, breast, and colon cancer (Caldas-Lopes et al., 2009; Wang et al., 2016; Nguyen et al., 2018). Although limited, several studies have attempted to identify predictive markers for HSP90-based therapy. For example, Nguyen et al. (Nguyen et al., 2018) has identified PCBP3, an RNA binding protein important in post-transcriptional control of gene expression, as a potential predictive biomarker for 17-AAG response in prostate cancer. In colorectal cancer, high expression of the UDP glucuronosyltransferase 1A (UGT1A) gene was found to correlate with poor sensitivities to the HSP90 inhibitor ganetespib, and its related compound NVP-AUY922, suggesting UGT1A levels in tumour tissues may be a suitable predictive biomarker for ganetespib treatment (Landmann et al., 2014). Interestingly, gene expression levels of UGT1A did not show correlation with 17-AAG response, implying different classes of HSP90 inhibitors may have different predictive biomarkers (Landmann et al., 2014). In addition, in acute lymphoblastic leukemia (ALL), patients with high levels of phosphorylated Src were more sensitive to the Hsp90 inhibitor NVP-BEP800 compared to those with low phosphorylated Src (Mshaik et al., 2021), suggesting Src phosphorylation may serve as a predictive biomarker. Moreover, since Hsp90 inhibition regulates Akt phosphorylation and Bcl-xL, expression levels of these effector proteins may be suitable predictive of response to Hsp90 inhibition in triple negative breast cancer (Caldas-Lopes et al., 2009). Similarly, as Hsp90 inhibition downregulates c-Myc expression and upregulates the expression of tumour repressor proteins such as p53 and pRB, which inhibits the G1/S transition (Yamaki et al., 2011), expression levels of cell cycle regulatory proteins such as pRB, E2F, cyclin–cyclin-dependent kinase (CDK) complexes could inform predictive biomarkers in specific tumour contexts. However, there are several limitations associated with current studies of predictive biomarkers for Hsp90 inhibitors, including: (i) their derivation was largely based on correlation analyses; (ii) the biomarkers are mostly single-gene markers and so unlikely to be clinically robust; (iii) and lack of patient-derived data. Together, these factors may explain the fact that so far, no companion predictive biomarkers of Hsp90-based therapy are employed for clinical practice.

In this study, we have aimed to alleviate these limitations through utilization of patient-derived data from unique explant models; implementation of predictive ML modelling rather than association analyses; and derivation of multi-gene rather than single-gene biomarkers. As a result, we have identified a highly-predictive biomarker panel (92% accuracy) consisting of 16 proteins. Its superior performance to individual DEPs and to using all the 157 DEPs points to the importance of selectively combining the relevant input features in optimizing drug-response prediction. The result also highlights the need to venture beyond the contemporary single-marker paradigm. Reassuringly, the identified panel contains proteins that have been implicated in prostate tumorigenesis and drug resistance, including CDK2, IGFBP7, TRIM47, and RBM17. For example, CDK2 was identified as a therapeutic target in prostate cancer (Yin et al., 2018). Its activation is significantly associated with disease recurrence, and its inhibition reduces invasion of prostate cancer cell lines (Yin et al., 2018). Moreover, CDK2 mediates androgen-dependent inhibition of AR+, castration-resistant prostate cancer cell proliferation (Kokontis et al., 2014). IGFBP7, a member of the insulin growth factor binding protein family, is involved in a variety of cancers including prostate cancer (Sullivan et al., 2012; Jin et al., 2020). Aberrant promoter hypermethylation of IGFBP7 and consequential gene silencing were found in prostate cancer cell lines (Sullivan et al., 2012). On the other hand, the tripartite motif (TRIM) protein TRIM47 is significantly increased in prostate cancer compared to normal tissues (Fujimura et al., 2016). In addition, SPF45, a splicing factor, is overexpressed in select tumours including prostate cancer, and it confers resistance to multiple anti-cancer drugs (Sampath et al., 2003; Perry et al., 2005). Overall, these evidences support the validity of our predictive multi-protein biomarker panel.

Translation of predictive biomarkers into clinical usage depends strongly on the ability to develop assays for detection of these markers in patient samples. We therefore reasoned that compact biomarker panels displaying strong predictive power are optimal for clinical application. With this in mind, we reduced the panel from 16 to 5 proteins, which achieved excellent prediction accuracy (80%). The 5-protein panel includes VPS25 (Vacuolar Protein Sorting 25 Homolog), TRIM47 (Tripartite motif 47), RBM17 (RNA Binding Motif Protein 17), SEPT8 (Septin-8) and AQP1 (Aquaporin 1). In addition to TRIM47's involvement in prostate cancer mentioned above, RBM17 is frequently overexpressed in a variety of carcinomas, including prostate cancer (Sampath et al., 2003). Importantly, RBM17 confers resistance to doxorubicin and vincristine, two chemotherapeutic drugs commonly used in cancer treatment (Perry

et al., 2005). Septins are GTP-binding proteins that are evolutionarily and structurally related to the *RAS* oncogenes (Abbey et al., 2019). Septin's expression levels are altered in hormonally regulated cancers such as prostate, breast, ovarian and endometrial cancers (Dolat et al., 2014; Angelis and Spiliotis, 2016). AQP1 is known to be upregulated by hypoxia that leads to increased cell water permeability, motility, and migration in neuroblastoma, lung and prostate cancer cells (Mobasheri et al., 2005; Hwang et al., 2012; Wei and Dong, 2015; Huo et al., 2021). Further, AQP1 is involved in microvascular alteration during prostate tumour angiogenesis (Mobasheri et al., 2005); and it promotes sensitivity of anthracycline chemotherapy in breast cancer (Chong et al., 2021). Taken together, these studies provide evidence linking the identified marker proteins to prostate cancer, supporting to the validity of the simplified panel. Further understanding of the roles of these proteins in prostate cancer tumorigenesis, and how they mechanistically modulate 17-AAG sensitivity are important areas of future research.

In addition, we conducted KEGG pathway and GO function analysis using both the 5-protein and 16-protein biomarker panels. As shown in the Supplementary Figure S8, the 5-proteins panel is mainly related to the proximal tubule bicarbonate reclamation (Dubose, 1990) and renin secretion (Kurtz, 2012) in the KEGG pathway analysis. Proximal tubule bicarbonate reclamation is a process by which the proximal tubules in the kidney reclaim bicarbonate ions from the filtrate in the renal tubules (Rector et al., 1998). This process helps to maintain electrolyte balance in the body by reabsorbing bicarbonate ions and preventing their excretion in urine. Renin secretion is important in cancer development as it regulates the production of angiotensin II, which has been shown to stimulate cancer cell growth and proliferation (Sobczuk et al., 2017). The 16-protein panel was found to be mainly related to steroid biosynthesis, a process by which the body produces steroid hormones. Abnormal steroid hormone production, which can be influenced by abnormalities in steroid biosynthesis pathways, has been linked to prostate cancer (Wilding, 1992; Mostaghel, 2013).

Our GO function analysis identified that both the 5- and 16-protein biomarker panels are mainly related to polyol transmembrane transporter activity (GO:0015166) and intracellular cGMP-activated cation channel activity (GO:0005223). Polyol transmembrane transporter activity involves the transport of small sugar molecules, such as glucose, across cell membranes. Dysregulation of this activity has been implicated in cancer development, as it can contribute to increased cellular proliferation and survival (Jones and Morris, 2016). Intracellular cGMP-activated cation channels are proteins activated by the signaling molecule cGMP, which allow ions to enter cells (Biel and Michalakis, 2009). Dysregulation of these channels has been linked to the development of various types of cancer, including breast, prostate, and ovarian cancer (Di Iorio et al., 2021).

Importantly, in an effort to validate the identified 5-protein biomarker signatures, we have generated an independent validation cohort of PDEs, and predicted their responses to 17-AAG treatment based on the PDE-specific expression levels of the five marker proteins. Overall, our framework correctly predicted the response for the responsive and non-responsive PDEs, but did not correctly predict the poor-responders, achieving >71% accuracy on this independent dataset. A limitation of the current validation is pertained to the small size of the validation cohort, due primarily to the challenge in accessing a large number of suitable patient samples and establishing the corresponding PDEs. This, however, is a general issue in biomarker studies utilizing pharmacogenomic data derived from cancer patients (Huang et al., 2018; Parca et al., 2019; Nguyen et al., 2021). We envisage as additional PDEs are generated in the future, the data will provide a more robust validation of our pipeline.

In summary, we have developed a new computational framework based on machine learning that aids the identification of multi-gene predictive biomarkers for targeted cancer drugs. While we have demonstrated its power focusing on prostate cancer as a proof-of-concept, the framework has broad applicability and can be applied to other drugs and cancer types in future studies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: TCGA, cancer_study_identifier: prad_tcga_pan_can_atlas_2018, PNAS cancer_study_identifier: prad_su2c_2019.

Author contributions

Conceptualization, S-YS and LN; methodology, S-YS, MC, JH, EN, LB, RD, and LN; writing—original draft preparation and revision, S-YS and LN; supervision, S-YS and LN; funding acquisition, LN. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by a Victorian Cancer Agency Mid-Career Research Fellowship (MCRF18026) and an Investigator Initiated Research Scheme grant from National Breast Cancer Foundation and Love Your Sister (IIRS-20–094) that were awarded to LN.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1094321/full#supplementary-material

References

Abbey, M., Gaestel, M., and Menon, M. B. (2019). Septins: Active GTPases or just GTP-binding proteins? *Cytoskelet. Hob.* 76, 55–62. doi:10.1002/cm.21451

Abida, W., Cyrta, J., Heller, G., Prandi, D., Armenia, J., Coleman, I., et al. (2019). Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* 116, 11428–11436. doi:10.1073/pnas.1902651116

Aha, D. W., and Bankert, R. L. (1996). "A comparative evaluation of sequential feature selection algorithms," in *Learning from data: Artificial intelligence and statistics V*. Editors D. Fisher and H.-J. Lenz (New York, NY: Springer), 199–206.

Alsahaf, A., Petkov, N., Shenoy, V., and Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Syst. Appl.* 187, 115895. doi:10.1016/j.eswa.2021.115895

Angelis, D., and Spiliotis, E. T. (2016). Septin mutations in human cancers. *Front. Cell Dev. Biol.* 4, 122. doi:10.3389/fcell.2016.00122

Armstrong, H. K., Koay, Y. C., Irani, S., Das, R., Nassar, Z. D., Selth, L. A., et al. (2016). A novel class of Hsp90 C-terminal modulators have pre-clinical efficacy in prostate tumor cells without induction of a heat shock response. *Prostate* 76, 1546–1559. doi:10.1002/pros.23239

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi:10.1038/nature11003

Biel, M., and Michalakis, S. (2009). Cyclic nucleotide-gated channels. *Handb. Exp. Pharmacol.* 191, 111–136. doi:10.1007/978-3-540-68964-5_7

Blum, A. L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.* 97, 245–271. doi:10.1016/s0004-3702(97)00063-5

Bomane, A., Gonçalves, A., and Ballester, P. J. (2019). Paclitaxel response can Be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. *Front. Genet.* 10, 1041. doi:10.3389/fgene.2019.01041

Borst, P., and Wessels, L. (2010). Do predictive signatures really predict response to cancer chemotherapy? *Cell Cycle* 9, 4836–4840. doi:10.4161/cc.9.24.14326

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing* 300, 70–79. doi:10.1016/j.neucom.2017.11.077

Caldas-Lopes, E., Cerchietti, L., Ahn, J. H., Clement, C. C., Robles, A. I., Rodina, A., et al. (2009). Hsp90 inhibitor PU-H71, a multimodal inhibitor of malignancy, induces complete responses in triple-negative breast cancer models. *Proc. Natl. Acad. Sci.* 106, 8368–8373. doi:10.1073/pnas.0903392106

Cardillo, M. R., and Ippoliti, F. (2006). IL-6, IL-10 and HSP-90 expression in tissue microarrays from human prostate cancer assessed by computer-assisted image analysis. *Anticancer Res.* 26, 3409–3416.

Centenera, M. M., Gillis, J. L., Hanson, A. R., Jindal, S., Taylor, R. A., Risbridger, G. P., et al. (2012). Australian Prostate CancerEvidence for efficacy of new Hsp90 inhibitors revealed by *ex vivo* culture of human prostate tumors. *Clin. Cancer Res.* 18, 3562–3570. doi:10.1158/1078-0432.CCR-12-0782

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi:10.1158/2159-8290.CD-12-0095

Chen, C., Zhang, L. G., Liu, J., Han, H., Chen, N., Yao, A. L., et al. (2016). Bioinformatics analysis of differentially expressed proteins in prostate cancer based on proteomics data. *Onco Targets Ther.* 9, 1545–1557. doi:10.2147/OTT.S98807

Chong, W., Zhang, H., Guo, Z., Yang, L., Shao, Y., Liu, X., et al. (2021). Aquaporin 1 promotes sensitivity of anthracycline chemotherapy in breast cancer by inhibiting β-catenin degradation to enhance TopoIIα activity. *Cell Death Differ.* 28, 382–400. doi:10.1038/s41418-020-00607-9

Di Iorio, P., Ronci, M., Giuliani, P., Caciagli, F., Ciccarelli, R., Caruso, V., et al. (2021). Pros and cons of pharmacological manipulation of cGMP-PDEs in the prevention and treatment of breast cancer. *Int. J. Mol. Sci.* 23, 262. doi:10.3390/ijms23010262

Dieci, M. V., Miglietta, F., Griguolo, G., and Guarneri, V. (2020). Biomarkers for HER2-positive metastatic breast cancer: Beyond hormone receptors. *Cancer Treat. Rev.* 88, 102064. doi:10.1016/j.ctrv.2020.102064

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform Comput. Biol.* 3, 185–205. doi:10.1142/s0219720005001004

Dolat, L., Hu, Q., and Spiliotis, E. T. (2014). Septin functions in organ system physiology and pathology. *Biol. Chem.* 395, 123–141. doi:10.1515/hsz-2013-0233

Dubose, T. D. (1990). Reclamation of filtered bicarbonate. *Kidney Int.* 38, 584–589. doi:10.1038/ki.1990.246

Feng, D.-C., Liu, Z.-T., Wang, X.-D., Chen, Y., Chang, J.-Q., Wei, D.-F., et al. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Constr. Build. Mater.* 230, 117000. doi:10.1016/j.conbuildmat.2019.117000

Fortino, V., Wisgrill, L., Werner, P., Suomela, S., Linder, N., Jalonen, E., et al. (2020). Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc. Natl. Acad. Sci.* 117, 33474–33485. doi:10.1073/pnas.2009192117

Fujimura, T., Inoue, S., Urano, T., Takayama, K., Yamada, Y., Ikeda, K., et al. (2016). Increased expression of tripartite motif (TRIM) 47 is a negative prognostic predictor in human prostate cancer. *Clin. Genitourin. Cancer* 14, 298–303. doi:10.1016/j.clgc.2016.01.011

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi:10.1038/nrc1299

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* 6, pl1. doi:10.1126/scisignal.2004088

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* 21, 1318–1325. doi:10.1038/nm.3954

Gardiner, L. J., Carrieri, A. P., Bingham, K., Macluskie, G., Bunton, D., Mcneil, M., et al. (2022). Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS One* 17, e0263248. doi:10.1371/journal.pone.0263248

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575. doi:10.1038/nature11005

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines. *Genome Biol.* 15, R47. doi:10.1186/gb-2014-15-3-r47

Gillet, J. P., Varma, S., and Gottesman, M. M. (2013). The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.* 105, 452–458. doi:10.1093/jnci/djt007

Gopika, N., and Meena Kowshalya, A., M.E. (2018). "Correlation based feature selection algorithm for machine learning," in 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, October 15–16, 2018, 692–695.

Guidotti, R., Monreale, A., Ruggieri, S., Turin, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *Acm Comput. Surv.* 51, 1–42. doi:10.1145/3236009

Hanf, W. (1975). The Boolean algebra of logic. *Bull. Am. Math. Soc.* 81, 587–589. doi:10.1090/s0002-9904-1975-13747-5

Heath, E. I., Hillman, D. W., Vaishampayan, U., Sheng, S., Sarkar, F., Harper, F., et al. (2008). A phase II trial of 17-allylamino-17-demethoxygeldanamycin in patients with hormone-refractory metastatic prostate cancer. *Clin. Cancer Res.* 14, 138–141. doi:10.3816/cgc.2005.n.024

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-Origin patterns dominate the molecular classification of 10, 000 tumors from 33 types of cancer. *Cell* 173, 291–304. doi:10.1016/j.cell.2018.03.022

Huang, C., Clayton, E. A., Matyunina, L. V., Mcdonald, L. D., Benigno, B. B., Vannberg, F., et al. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Sci. Rep.* 8, 16444. doi:10.1038/s41598-018-34753-5

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* 14, 55–63. doi:10.1109/tit.1968.1054102

Huntington, E. V. (1933). New sets of independent postulates for the algebra of logic, with special reference to Whitehead and Russell's principia mathematica. *Trans. Am. Math. Soc.* 35, 274–304. doi:10.1090/s0002-9947-1933-1501684-x

Huo, Z., Lomora, M., Kym, U., Palivan, C., Holland-Cunz, S. G., and Gros, S. J. (2021). AQP1 is up-regulated by hypoxia and leads to increased cell water permeability, motility, and migration in neuroblastoma. *Front. Cell Dev. Biol.* 9, 605272. doi:10.3389/fcell.2021.605272

Hwang, I., Jung, S.-I., Hwang, E.-C., Song, S. H., Lee, H.-S., Kim, S.-O., et al. (2012). Expression and localization of aquaporins in benign prostate hyperplasia and prostate cancer. *Chonnam Med. J.* 48, 174–178. doi:10.4068/cmj.2012.48.3.174

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* 166, 740–754. doi:10.1016/j.cell.2016.06.017

Jain, T. K., Kushwaha, D. S., and Misra, A. K. (2008). "Optimization of the quine-McCluskey method for the minimization of the boolean expressions," in Fourth International Conference on Autonomic and Autonomous Systems (ICAS'08), 165–168.

Jin, L., Shen, F., Weinfeld, M., and Sergi, C. (2020). Insulin growth factor binding protein 7 (IGFBP7)-Related cancer and IGFBP3 and IGFBP7 crosstalk. *Front. Oncol.* 10, 727. doi:10.3389/fonc.2020.00727

John, G. H., Kohavi, R., and Pfleger, K. (1994). "Irrelevant features and the subset selection problem," in *Machine learning proceedings 1994*. Editors W. W. Cohen and H. Hirsh (San Francisco (CA): Morgan Kaufmann), 121–129.

Jones, R. S., and Morris, M. E. (2016). Monocarboxylate transporters: Therapeutic targets and prognostic factors in disease. *Clin. Pharmacol. Ther.* 100, 454–463. doi:10.1002/cpt.418

Knijnenburg, T. A., Klau, G. W., Iorio, F., Garnett, M. J., Mcdermott, U., Shmulevich, I., et al. (2016). Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci. Rep.* 6, 36812. doi:10.1038/srep36812

Kokontis, J. M., Lin, H. P., Jiang, S. S., Lin, C. Y., Fukuchi, J., Hiipakka, R. A., et al. (2014). Androgen suppresses the proliferation of androgen receptor-positive castration-resistant prostate cancer cells via inhibition of Cdk2, CyclinA, and Skp2. *PLoS One* 9, e109170. doi:10.1371/journal.pone.0109170

Kong, J., Lee, H., Kim, D., Han, S. K., Ha, D., Shin, K., et al. (2020). Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nat. Commun.* 11, 5485. doi:10.1038/s41467-020-19313-8

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi:10.1093/nar/gkw377

Kurtz, A. (2012). Control of renin synthesis and secretion. *Am. J. Hypertens.* 25, 839–847. doi:10.1038/ajh.2011.246

Landmann, H., Proia, D. A., He, S., Ogawa, L. S., Kramer, F., Beißbarth, T., et al. (2014). UDP glucuronosyltransferase 1A expression levels determine the response of colorectal cancer cells to the heat shock protein 90 inhibitor ganetespib. *Cell Death Dis.* 5, e1411. doi:10.1038/cddis.2014.378

Lee, D., Park, Y., and Kim, S. (2021). Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. *Brief. Bioinform* 22, bbaa188. doi:10.1093/bib/bbaa188

Lima, A. R., Pinto, J., Azevedo, A. I., Barros-Silva, D., Jerónimo, C., Henrique, R., et al. (2019). Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine. *Br. J. Cancer* 121, 857–868. doi:10.1038/s41416-019-0585-4

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy (Basel)* 23, 18. doi:10.3390/e23010018

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy* 23, 18. doi:10.3390/e23010018

Liu, Y., Wang, Y., and Zhang, J. (2012). "New machine learning algorithm: Random forest," in *Information computing and applications*. Editors B. Liu, M. Ma, and J. Chang (Springer Berlin Heidelberg), 246–252.

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA: Curran Associates Inc).

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., and Andina, D. (2010). "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, 2845–2850.

Masica, D. L., and Karchin, R. (2013). Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer Res.* 73, 1699–1708. doi:10.1158/0008-5472.CAN-12-3122

Mateo, J., Steuten, L., Aftimos, P., André, F., Davies, M., Garralda, E., et al. (2022). Delivering precision oncology to patients with cancer. *Nat. Med.* 28, 658–665. doi:10.1038/s41591-022-01717-2

Mccluskey, E. J. (1956). Minimization of boolean functions. *Bell Syst. Tech. J.* 35, 1417–1444. doi:10.1002/j.1538-7305.1956.tb03835.x

Menden, M. P., Iorio, F., Garnett, M., Mcdermott, U., Benes, C. H., Ballester, P. J., et al. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 8, e61318. doi:10.1371/journal.pone.0061318

Min-Ling, Z., and Zhi-Hua, Z. (2005). "A k-nearest neighbor based algorithm for multi-label classification," in 2005 IEEE International Conference on Granular Computing, 718–721.

Mobasheri, A., Airley, R., Hewitt, S. M., and Marples, D. (2005). Heterogeneous expression of the aquaporin 1 (AQP1) water channel in tumors of the prostate, breast, ovary, colon and lung: a study using high density multiple human tumor tissue microarrays. *Int. J. Oncol.* 26, 1149–1158. doi:10.3892/ijo.26.5.1149

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw.* 6, 525–533. doi:10.1016/s0893-6080(05)80056-5

Mostaghel, E. A. (2013). Steroid hormone synthetic pathways in prostate cancer. *Transl. Androl. Urol.* 2, 212–227. doi:10.3978/j.issn.2223-4683.2013.09.16

Mshaik, R., Simonet, J., Georgievski, A., Jamal, L., Bechoua, S., Ballerini, P., et al. (2021). HSP90 inhibitor NVP-BEP800 affects stability of SRC kinases and growth of T-cell and B-cell acute lymphoblastic leukemias. *Blood Cancer J.* 11, 61. doi:10.1038/s41408-021-00450-2

Nalejska, E., Maczynska, E., and Lewandowska, M. A. (2014). Prognostic and predictive biomarkers: tools in personalized oncology. *Mol. Diagn Ther.* 18, 273–284. doi:10.1007/s40291-013-0077-9

Naulaerts, S., Menden, M. P., and Ballester, P. J. (2020). Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. *Biomolecules* 10, 963. doi:10.3390/biom10060963

Navada, A., Ansari, A. N., Patil, S., and Sonkamble, B. A. (2011). "Overview of use of decision tree algorithms in machine learning," in 2011 IEEE Control and System Graduate Research Colloquium, 37–42.

Nguyen, L., Dang, C. C., and Ballester, P. J. (2016). Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *F1000Res* 5. doi:10.12688/f1000research.10529.2

Nguyen, E. V., Centenera, M. M., Moldovan, M., Das, R., Irani, S., Vincent, A. D., et al. (2018). Identification of novel response and predictive biomarkers to Hsp90 inhibitors through proteomic profiling of patient-derived prostate tumor explants. *Mol. Cell Proteomics* 17, 1470–1486. doi:10.1074/mcp.RA118.000633

Nguyen, L. C., Naulaerts, S., Bruna, A., Ghislat, G., and Ballester, P. J. (2021). Predicting cancer drug response *in vivo* by learning an optimal feature selection of tumour molecular profiles. *Biomedicines* 9, 1319. doi:10.3390/biomedicines9101319

Ogunleye, A. Z., Piyawajanusorn, C., Gonçalves, A., Ghislat, G., and Ballester, P. J. (2022). Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microRNA profiles. *Adv. Sci.* 9, e2201501. doi:10.1002/advs.202201501

Parca, L., Pepe, G., Pietrosanto, M., Galvan, G., Galli, L., Palmeri, A., et al. (2019). Modeling cancer drug response through drug-specific informative genes. *Sci. Rep.* 9, 15222. doi:10.1038/s41598-019-50720-0

Pernas, S., Tolaney, S. M., Winer, E. P., and Goel, S. (2018). CDK4/6 inhibition in breast cancer: current practice and future directions. *Ther. Adv. Med. Oncol.* 10, 1758835918786451. doi:10.1177/1758835918786451

Perry, W. L., 3rd, Shepard, R. L., Sampath, J., Yaden, B., Chin, W. W., Iversen, P. W., et al. (2005). Human splicing factor SPF45 (RBM17) confers broad multidrug resistance to anticancer drugs when overexpressed--a phenotype partially reversed by selective estrogen receptor modulators. *Cancer Res.* 65, 6593–6600. doi:10.1158/0008-5472.CAN-03-3675

Pich, O., Bailey, C., Watkins, T. B. K., Zaccaria, S., Jamal-Hanjani, M., and Swanton, C. (2022). The translational challenges of precision oncology. *Cancer Cell* 40, 458–478. doi:10.1016/j.ccell.2022.04.002

Quine, W. V. (1955). A way to simplify truth functions. *Am. Math. Mon.* 62, 627–631. doi:10.1080/00029890.1955.11988710

Quintás-Cardama, A., and Cortes, J. (2009). Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood* 113, 1619–1630. doi:10.1182/blood-2008-03-144790

Rector, F. C., Carter, N. W., and Seldin, D. W. (1998). The mechanism of bicarbonate reabsorption in the proximal and distal tubules of the kidney. *J. Am. Soc. Nephrol.* 9, 1134–1145; discussion 1134-1139. doi:10.1681/ASN.V961134

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA: Association for Computing Machinery).

Sampath, J., Long, P. R., Shepard, R. L., Xia, X., Devanarayan, V., Sandusky, G. E., et al. (2003). Human SPF45, a splicing factor, has limited expression in normal tissues, is overexpressed in many tumors, and can confer a multidrug-resistant phenotype to cells. *Am. J. Pathol.* 163, 1781–1790. doi:10.1016/S0002-9440(10)63538-9

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., et al. (2015). Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* 5, 1210–1223. doi:10.1158/2159-8290.CD-15-0235

Shin, S. Y., Rath, O., Zebisch, A., Choo, S. M., Kolch, W., and Cho, K. H. (2010). Functional roles of multiple feedback loops in extracellular signal-regulated kinase and Wnt signaling pathways that regulate epithelial-mesenchymal transition. *Cancer Res.* 70, 6715–6724. doi:10.1158/0008-5472.CAN-10-1377

Sobczuk, P., Szczylik, C., Porta, C., and Czarnecka, A. M. (2017). Renin angiotensin system deregulation as renal cancer risk factor. *Oncol. Lett.* 14, 5059–5068. doi:10.3892/ol.2017.6826

Solit, D. B., Zheng, F. F., Drobnjak, M., Münster, P. N., Higgins, B., Verbel, D., et al. (2002). 17-Allylamino-17-demethoxygeldanamycin induces the degradation of androgen receptor and HER-2/neu and inhibits the growth of prostate cancer xenografts. *Clin. Cancer Res.* 8, 986–993.

Stief, A., Ottewill, J. R., and Baranowski, J. (2018). "Relief F-based feature ranking and feature selection for monitoring induction motors," in 2018 23rd International Conference on Methods & Models in Automation & Robotics (MMAR), 171–176.

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Sullivan, L., Murphy, T. M., Barrett, C., Loftus, B., Thornhill, J., Lawler, M., et al. (2012). IGFBP7 promoter methylation and gene expression analysis in prostate cancer. *J. Urol.* 188, 1354–1360. doi:10.1016/j.juro.2012.06.002

Tabl, A. A., Alkhateeb, A., Elmaraghy, W., Rueda, L., and Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Front. Genet.* 10, 256. doi:10.3389/fgene.2019.00256

Tang, Y., Zhang, Y. Q., and Huang, Z. (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4, 365–381. doi:10.1109/TCBB.2007.70224

Trepel, J., Mollapour, M., Giaccone, G., and Neckers, L. (2010). Targeting the dynamic HSP90 complex in cancer. *Nat. Rev. Cancer* 10, 537–549. doi:10.1038/nrc2887

Turajlic, S., Sottoriva, A., Graham, T., and Swanton, C. (2019). Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* 20, 404–416. doi:10.1038/s41576-019-0114-6

Vasquez, M. M., Hu, C., Roe, D. J., Chen, Z., Halonen, M., and Guerra, S. (2016). Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. *BMC Med. Res. Methodol.* 16, 154. doi:10.1186/s12874-016-0254-8

Wang, C. Y., Guo, S. T., Wang, J. Y., Yan, X. G., Farrelly, M., Zhang, Y. Y., et al. (2016). Reactivation of ERK and Akt confers resistance of mutant BRAF colon cancer cells to the HSP90 inhibitor AUY922. *Oncotarget* 7, 49597–49610. doi:10.18632/oncotarget.10414

Wei, X., and Dong, J. (2015). Aquaporin 1 promotes the proliferation and migration of lung cancer cell *in vitro*. *Oncol. Rep.* 34, 1440–1448. doi:10.3892/or.2015.4107

Wei, R., Wang, J., Jia, E., Chen, T., Ni, Y., and Jia, W. (2018). GSimp: A gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput. Biol.* 14, e1005973. doi:10.1371/journal.pcbi.1005973

Whitesitt, J. E. (2012). *Boolean algebra and its applications*. Mineola, New York: Dover Publications.

Wilding, G. (1992). The importance of steroid hormones in prostate cancer. *Cancer Surv.* 14, 113–130.

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2019). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi:10.1093/bioinformatics/btz793

Yamaki, H., Nakajima, M., Shimotohno, K. W., and Tanaka, N. (2011). Molecular basis for the actions of Hsp90 inhibitors and cancer therapy. *J. Antibiot. (Tokyo)* 64, 635–644. doi:10.1038/ja.2011.60

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi:10.1093/nar/gks1111

Yin, X., Yu, J., Zhou, Y., Wang, C., Jiao, Z., Qian, Z., et al. (2018). Identification of CDK2 as a novel target in treatment of prostate cancer. *Future Oncol.* 14, 709–718. doi:10.2217/fon-2017-0561

Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C., and Showe, M. K. (2007). Naïve Bayes for microRNA target predictions—Machine learning for microRNA targets. *Bioinformatics* 23, 2987–2992. doi:10.1093/bioinformatics/btm484

Zhang, F., Kaufman, H. L., Deng, Y., and Drabier, R. (2013). Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood. *BMC Med. Genomics* 6, S4. doi:10.1186/1755-8794-6-S1-S4

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* 4, 218. doi:10.21037/atm.2016.03.37

Zhong, L., Li, Y., Xiong, L., Wang, W., Wu, M., Yuan, T., et al. (2021). Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct. Target Ther.* 6, 201. doi:10.1038/s41392-021-00572-w

Zhou, Z. H., and Feng, J. (2017). "Deep forest: Towards an alternative to deep neural networks," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 3553–3559.

Zhou, Z.-H., and Feng, J. (2018). Deep forest. *Natl. Sci. Rev.* 6, 74–86. doi:10.1093/nsr/nwy108

Zhu, C. S., Pinsky, P. F., Cramer, D. W., Ransohoff, D. F., Hartge, P., Pfeiffer, R. M., et al. (2011). A framework for evaluating biomarkers for early detection: validation of biomarker panels for ovarian cancer. *Cancer Prev. Res. Phila. Pa.* 4, 375–383. doi:10.1158/1940-6207.CAPR-10-0193

frontiers | Frontiers in Medicine

TYPE Review
PUBLISHED 15 February 2023
DOI 10.3389/fmed.2023.1086097

Check for updates

OPEN ACCESS

EDITED BY
Mamoru Kato,
National Cancer Centre, Japan

REVIEWED BY
Amin Emad,
McGill University, Canada
Duc-Hau Le,
Thuyloi University, Vietnam

*CORRESPONDENCE
Alexander Partin
✉ apartin@anl.gov

SPECIALTY SECTION
This article was submitted to Precision Medicine, a section of the journal Frontiers in Medicine

RECEIVED 01 November 2022
ACCEPTED 23 January 2023
PUBLISHED 15 February 2023

CITATION
Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Overbeek J and Stevens RL (2023) Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. *Front. Med.* 10:1086097. doi: 10.3389/fmed.2023.1086097

COPYRIGHT
© 2023 Partin, Brettin, Zhu, Narykov, Clyde, Overbeek and Stevens. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Deep learning methods for drug response prediction in cancer: Predominant and emerging trends

Alexander Partin[1]*, Thomas S. Brettin[1], Yitan Zhu[1], Oleksandr Narykov[1], Austin Clyde[1], Jamie Overbeek[1] and Rick L. Stevens[1,2]

[1]Division of Data Science and Learning, Argonne National Laboratory, Lemont, IL, United States, [2]Department of Computer Science, The University of Chicago, Chicago, IL, United States

Cancer claims millions of lives yearly worldwide. While many therapies have been made available in recent years, by in large cancer remains unsolved. Exploiting computational predictive models to study and treat cancer holds great promise in improving drug development and personalized design of treatment plans, ultimately suppressing tumors, alleviating suffering, and prolonging lives of patients. A wave of recent papers demonstrates promising results in predicting cancer response to drug treatments while utilizing deep learning methods. These papers investigate diverse data representations, neural network architectures, learning methodologies, and evaluations schemes. However, deciphering promising predominant and emerging trends is difficult due to the variety of explored methods and lack of standardized framework for comparing drug response prediction models. To obtain a comprehensive landscape of deep learning methods, we conducted an extensive search and analysis of deep learning models that predict the response to single drug treatments. A total of 61 deep learning-based models have been curated, and summary plots were generated. Based on the analysis, observable patterns and prevalence of methods have been revealed. This review allows to better understand the current state of the field and identify major challenges and promising solution paths.

KEYWORDS

deep learning, drug sensitivity, multiomics, neural networks, precision medicine, precision oncology, personalized medicine, drug response prediction

1. Introduction

Cancer treatment response prediction is a problem of great importance for both clinical and pharmacological research communities. Many believe it will pave the way to devising more efficient treatment protocols for individual patients and provide insights into designing novel drugs that efficiently suppress disease. Currently, however, cancer treatment remains extremely challenging, often resulting in inconsistent outcomes. For example, tumor heterogeneity contributes to differential treatment responses in patients with the same tumor type (1, 2). Nevertheless, conventional tumor type-dependent anticancer treatments such as chemotherapy often lead to suboptimal results and substantial side effects, and therefore, are notoriously regarded as one-size-fits-all therapies (3, 4). Alternatively, targeted therapies and certain immunotherapies are prescribed based on known biomarkers, observable within individual patients (5). Cancer biomarkers refer to abnormalities in omics data (genomic, transcriptomic, etc.) which can be predictive of treatment response (3). Biomarker-driven treatment plans, either standalone or in combination with chemotherapies, are the mainstream of nowadays personalized (or precision) oncology. Discovery of biomarkers and their subsequent utilization

in clinical settings are attributed to advances in tumor profiling technologies and high-throughput drug screenings (3, 4).

An alternative direction to leverage large-scale screenings and high-dimensional omics data in the cancer research community is to build analytical models designed to predict the response of tumors to drug treatments. Typically, such models use tumor and drug information without explicitly specifying biomarkers (6). These models, often referred to as drug response prediction (DRP) models, can be used to prioritize treatments, explore drug repurposing, and reaffirm existing biomarkers. Artificial intelligence (AI) is the core methodology in designing DRP models, demonstrating encouraging results in retrospective evaluation analyzes with pre-clinical and clinical datasets. Many DRP models use classical machine learning (ML) and deep learning (DL), i.e., multi-layer neural networks (NNs). While DL is generally considered a subset of ML, we differentiate between the two terms, where ML is referred here to learning algorithms that do not involve the use of NNs.

Papers in this field are being published constantly, exploiting learning algorithms for DRP. To cope with increasing rate of publications, a recent special issue in Briefings in Bioinformatics was dedicated to DRP in cancer models (7). Collectively, at least 18 review-like papers have been published since 2020 in an attempt to summarize progress and challenges in the field, as well as provide discussions on promising research directions. Each paper aims to review the field from a unique perspective but certain topics substantially overlap as shown in Table 1 which lists some of the major topics and associated references. Common topics include data resources for constructing training and test datasets, prevalent data representations, ML and DL approaches for modeling drug response, and methods for evaluating the predictive performance of models.

Following the revival of artificial neural networks (NNs) more than a decade ago (29), DL methods have become a promising research direction across a variety of scientific and engineering disciplines (30–33). This trend is also observed in cancer research, including the prediction of tumor response to treatments, as shown in Figure 1. In 2013, Menden et al. demonstrated that a single hidden layer NN predicts drug response with comparable performance to random forest (RF) (34). The authors used genomic and response data from the Genomics of Drug Sensitivity in Cancer (GDSC) project which was published in 2012 (35). Despite the accelerated popularity of DL and the availability of substantial screening and omics data, it was not until the emergence of open-source frameworks dedicated for building and training NNs (36–38) that DL has become an integral part of DRP research (note the gap between 2013 and 2018 in Figure 1). Owing to the abundance of omics and screening data, the availability of DL frameworks and the unmet need for precision oncology, we have been witnessing a growth of scientific publications exploring NN architectures for DRP.

Dozens of DL-based models have been published, exploring diverse feature representations, NN architectures, learning schemes, and evaluation methods. However, only three out of the many existing reviews explicitly focus on modeling DRP with DL (10, 11, 18). These reviews present a categorized summary of methods, citing and discussing a limited number of examples from each category. As a result, only a selective overview of methods is provided and limited aspects of published DRP models are considered, exhibiting a primary limitation of these reviews. In addition, the scope of existing reviews does not require a comprehensive search for models which could reveal predominant and emerging trends. Considering the current rate of publications, the diversity of approaches, and the limitations of existing reviews, a comprehensive review is necessary and timely.

This review is the first one to conduct a comprehensive search to include all relevant papers that utilize NNs for DRP. As of the time of completing this paper, a total of 61 peer-reviewed publications have been identified. Only models predicting response to single-drug treatments are included in the current review, excluding models that make predictions to combination therapies. We identified three major components involved in developing

TABLE 1 Categorization of topics covered in existing review papers on drug response prediction (DRP).

Category	Topic	References
Data resources[a]	Characterization of cancer tumors	(8–18)
	Characterization of drugs	(15, 17, 19–21)
	Monotherapy screenings	(9, 10, 12–16, 18, 19)
	Drug combination screenings	(6, 9–11, 19, 22)
	Complimentary resources	(6, 8, 10, 11, 19, 20, 22, 23)
Data representation methods[b]	Representation of cancer tumors	(8, 9, 12, 18, 24)
	Representation of drugs	(17, 21, 24)
	Representation of treatment response	(8, 9, 14, 18, 24)
	Feature selection and extraction	(6, 8, 13, 16, 25)
Prediction models[c]	DL models for monotherapy	(6, 10, 14, 16–19, 21, 23, 26)
	DL models for drug combination therapies	(6, 10, 11, 16, 19, 21, 22)
	ML models for monotherapy	(6, 8, 12–16, 18, 21, 24, 27)
	ML models for drug combination therapies	(6, 16, 21, 22, 28)
Assessment and evaluation[d]	Experimental validation of prediction models	(14, 24–28)
	Evaluation methods of model performance	(8, 10, 12, 14, 18, 22, 26)

[a]Characterization of cancer tumors: resources of cancer omics data; Characterization of drugs: tools for generating drug feature representations; Monotherapy screenings: resources of single-drug response data; Drug combination screenings: resources of drug-combination response data; Complimentary resources relevant to DRP: additional data resources and computational methods relevant to the DRP problem such as drug-target interactions, pathway information, etc.

[b]Representation of cancer tumors: gene expressions, copy-number variations, and other representations of tumors; Representation of drugs: drug molecular descriptors, fingerprints, molecular graph structures, etc.; Representation of treatment response: drug efficacy metrics such as IC50, AUC, sensitive/resistant, and others; Feature selection and extraction: methods for feature selection and dimensionality reduction.

[c]Models for monotherapy: discussion of models predicting single-drug response; Models for drug-combination therapies: discussion of models predicting drug-combination response.

[d]Experimental validation of prediction models: assessment studies that conduct experiments to assess performance of DRP models; Evaluation methods of model performance: discussion of methods for evaluating performance of DRP models.

A

Table S1. A list of drug response prediction models that utilize neural networks. The data were collected until August of 2022, considering only peer-reviewed publications since 2013.

Paper	Model	Framework	Methods	Cancer features	Drug features	Response cell line	Response PDX	Response patient	Evaluation scheme
Hostallero et al. 2022	BiG-DRP	PyTorch	GNN	GE	DD	IC50	NA	IC50-to-Bin	Cancer-blind, Mixed-set
Yan et al. 2022	DGSDRP	PyTorch	1D-CNN, GNN, bRNN	Mu	MG	IC50	NA	NA	Cancer-blind, Drug-blind, Mixed-set
Xia et al. 2022	UnoMT	TF1	RC, AE, MTL	GE	DD, FP	AUC	NA	NA	Cross-dataset, Mixed-set
Wang et al. 2022	MvMo	PyTorch	MVL	GE, Methyl, Mu	MG	IC50	NA	NA	Cancer-blind, Drug-blind, Mixed-set
Su et al. 2022	SRDFM	TF1 (w/o Keras)	DeepFM	GE	FP	Rank	NA	NA	Cross-dataset, Drug-blind
Pu et al. 2022	CancerOmics Net	PyTorch	Att, GNN	GE	KIP	GR	NA	NA	Cancer-blind
Prasse et al. 2022	Conv NN	TF2	1D-CNN, TL	GE	SMILES	IC50	CTR	NA	Cancer-blind, Cross-dataset, Drug-blind
Prasse et al. 2022	None	TF2 (w/o Keras)	1D-CNN, Att, RC	GE	SMILES	Rank	NA	NA	Cancer-blind
Peng et al. 2022	MOFGCN	PyTorch	GNN	CNV, GE, Mu	FP	Bin-enc-IC50	NA	NA	Cancer-blind, Drug-blind, Mixed-set
Nguyen et al. 2022	GraOmicDRP	PyTorch	1D-CNN, GNN	CNV, GE, Methyl, Mu	MG	IC50, Bin-enc-IC50	NA	NA	Cancer-blind, Drug-blind, Mixed-set
Ma et al. 2022	DualGCN	TF1	GNN	CNV, GE	MG	IC50	NA	IC50-to-Bin	Cross-dataset, Mixed-set
Liu et al. 2022	PPORank	NA	RNN, RL	GE, Methyl	NA	Rank	NA	Rank	Cancer-blind, Cross-dataset
Jiang et al. 2022	DeepTTA	PyTorch	Tran	GE	FP	IC50-to-Bin	NA	NA	Cancer-blind, Drug-blind, Mixed-set

B

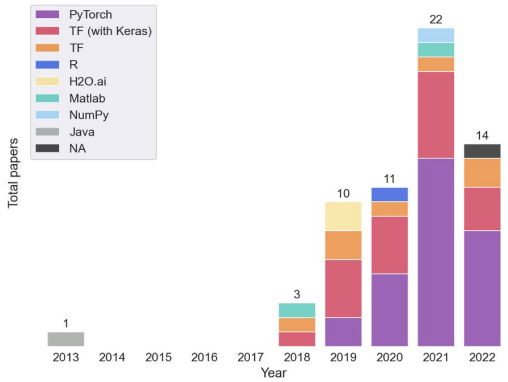


FIGURE 1
(A) A snapshot of Supplementary Table 1 that lists peer-reviewed papers proposing monotherapy drug response prediction (DRP) models (the full list can be found in the Supplementary material). The papers have been curated to identify various properties (shown in table columns), such as deep learning (DL) methods, feature types, evaluation methods, etc., as discussed in detail in this paper. Multiple plots in this paper have been generated using the data from Supplementary Table 1 [**(B)**, Figures 4–7C, D]. **(B)** Distribution of papers by year that use DL methods for DRP. The neural network models are designed using popular computational frameworks which include both proprietary and open-source software. The bar plots are color-coded by the different computational frameworks. TensorFlow/Keras and PyTorch are the most popular frameworks based on this plot. Data were collected until August 2022, considering only peer-reviewed publications.

DRP models, including data preparation, model development, and performance analysis. These components have been used to guide the curation of papers with special focus on representation methods of drugs, cancers, and measures of treatment response, DL related methods including NN modules and learning schemes, and methods for evaluating the prediction performance. This information is summarized in Supplementary Table 1. Summary plots have been generated, revealing the prevalence of methods used in these papers. Observing the prevalence of methods will assist in revealing approaches that have been investigated in multiple studies as well as emerging methods which are rather underexplored for DRP. We believe this review would serve as a valuable reference for new and experienced researchers in this field.

Section 2 formulates the DRP as a DL problem. Sections 3–5 provide a detailed review of the drug, cancer, and response representations used in papers listed in Supplementary Table 1. Existing design choices are summarized in Sections 6.1, 6.2 in terms of fundamental NN building blocks and learning schemes, respectively. Section 7 compiles existing approaches for evaluating model performance. In Section 8, we discuss the current state of DRP field, determine primary challenges, and propose further directions.

2. Deep learning-based drug response prediction workflow

A DRP model can be represented by $r = f(d, c)$, where f is the analytical model designed to predict the response r of cancer c to the treatment by drug d. The function f is implemented with a NN architecture in which the weights are learned through backpropagation. This formulation is for pan-cancer and multi-drug[1] prediction model where both cancer and drug representations are needed to predict response. A special case is drug-specific models designed to make predictions for a drug or drug family [e.g., drugs with the same mechanism of action (MoA)] (39). These models learn from cancer features only and can be formulated as $r = f_D(c)$. Another type of models is multi-task learning models which take only cancer representations as inputs and generate multiple outputs where each output produces predictions for a specific drug. As compared to drug-specific models, the multi-task formulation enables learning from larger amounts of drug response data while exploiting common characteristics among drugs, thereby allowing to further improve generalization of the entire model (further discussed in Section 6.2.3).

The general workflow for developing DRP models is not much different from developing supervised models for other applications (Figure 2). The challenges come in the specific details typical to predicting treatment response. The process can be divided into three components: (1) data preparation, (2) model development, and (3) performance analysis (10, 12). The choice of methods associated with each one of these components can have a significant impact on the overall workflow complexity and the potential application of the final model.

2.1. Data preparation

Data preparation is generally the initial step in designing a prediction model, requiring expertise in bioinformatics and statistical methods. During this step, heterogeneous data types are aggregated from multiple data sources, preprocessed, split into training and test sets, and structured to conform to an API of a DL framework. The generated drug response dataset with N samples, denoted by $S = \{d, c, r\}_{i=1}^{N}$, includes representations for drug (d), cancer (c), and response (r). Prediction generalization is expected to improve with a larger number of training samples as demonstrated with multiple cell line datasets (40–42), normally creating preference for larger datasets when developing DL models. Recent research focusing on data-centric approaches suggests that efficient data representations and proper choice of a training set are at least as important as the dataset size for improving predictions,

1 Sometimes referred as pan-drug.

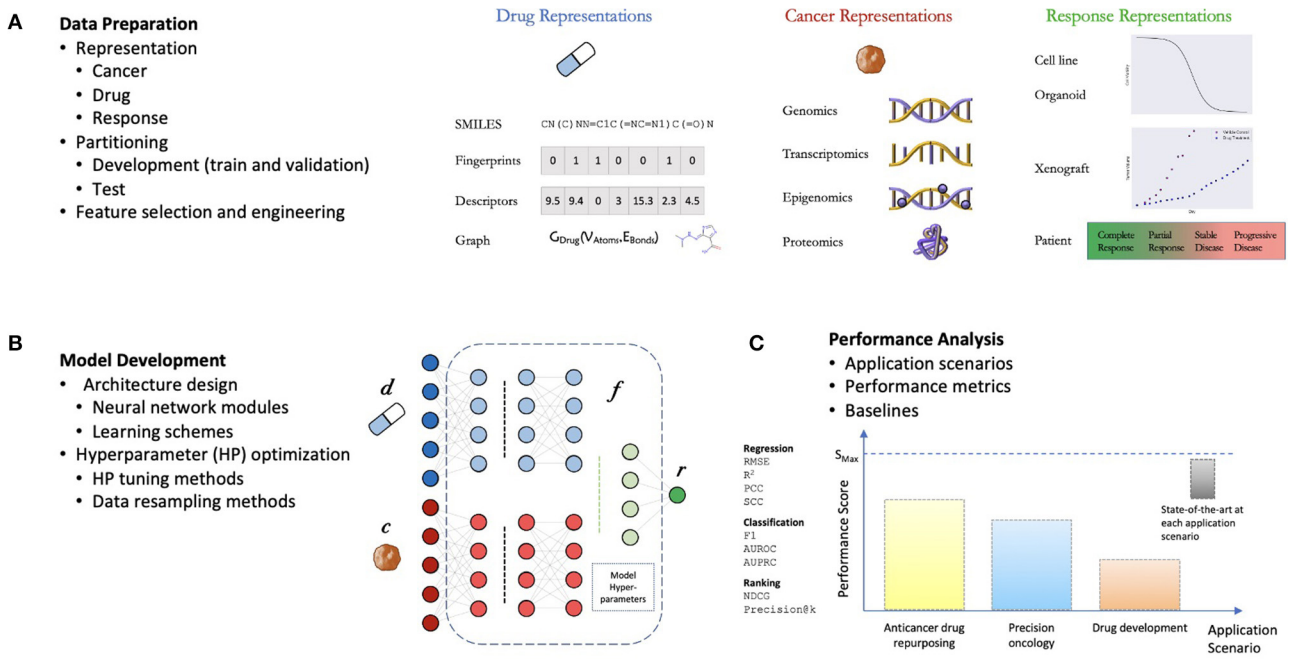


FIGURE 2
General components of a drug response prediction (DRP) workflow. **(A)** Data preparation: requires generating representations of features and treatment response, partition the dataset into development set (for training and hyperparameter (HP) tuning) and test set (for performance analysis), and any additional preprocessing such feature selection/engineering. **(B)** Model development: the process of generating a deep learning model which involves the design of a neural network (NN) architecture (choice of NN modules and learning schemes) and model training including HP optimization. **(C)** Performance analysis: assessment of prediction generalization and other metrics allowing to evaluate the utility of the DRP model for different applications in oncology such as personalized recommendation of treatments, drug repurposing, and drug development. The performance is benchmarked against one or more baseline models which should ultimately be chosen from available state-of-the-art models for the investigated application.

and further emphasize the importance of the data preprocessing step (43).

2.2. Model development

Model development refers to NN architecture design and optimization of model hyperparameters (HPs). To design NNs, developers often resort to common heuristics which rely on intuition, experimentation, and adoption of architectures from related fields. This process involves choosing the basic NN modules, the architecture, and learning schemes. Diversity of data representations for cancers (9) and drugs (21) and potential utilization of DRP models in several pre-clinical and clinical settings have led researchers to explore a wide range of DL methods.

2.3. Performance analysis

A desirable outcome of a model development workflow is a robust model that produces accurate predictions across cancers and drugs as evaluated by appropriate performance metrics. DRP models can be used in various scenarios such as personalized recommendation of treatments, exploration of drug repurposing, and assisting in development of new drugs. Therefore, both performance metrics and appropriate evaluation schemes (e.g., design of training and test sets) are critical for proper evaluation of prediction performance. The abundance of DRP papers in recent years (Figure 1) and lack of benchmark datasets (10), strongly suggest that a rigorous assessment of model performance is required where state-of-the-art baseline models serve as a point of reference.

2.4. Source and target domains

A vital characteristic that affects the entire workflow is the source domain data used to develop the DRP model, and the target domain data representing the biological domain on which the model is expected to operate. Data suitable for modeling DRP (Sections 3–5) come from multiple biological domains, such as cell lines, organoids, xenografts, and patients (9). Most models utilize data from a single domain (usually cell lines, Section 3.1, due to the abundance of response data). Certain models exploit data from a mix of domains with the goal to improve predictions in a target domain which suffers from insufficient data. Since a primary goal is to make DRP models useful in improving patient care, the biological domains can be further categorized into clinical (human patient data) and pre-clinical (cell lines, etc.). Considering this distinction, models can exploit data in different ways: train and test on preclinical (44–46), train and test on clinical, train on preclinical and test on clinical (47–49), and models that leverage both preclinical and clinical data during the training process and then test on clinical (50–54). Models that use data from mixed domains are generally more challenging to develop because they require extra steps in data preparation,

advanced modeling techniques, and robust performance evaluation analysis (Section 6.2.2).

3. Representations of treatment response

Drug screening platforms enable testing the sensitivity of cancer samples in controlled lab environments, ultimately producing data for predictive modeling (Figure 3). A major objective is a discovery of potential anticancer treatments through the screening of compound libraries against diverse cancer panels. Systematic drug screening platforms have been established for *in vitro* cancer models such as cell lines (55–57) and organoids (58), and *in vivo* models such as patient-derived xenografts (PDXs) (59). An overview of preclinical cancer models and the corresponding methods for sensitivity profiling is available in a recent review article (9). Drug sensitivity data can be transformed into continuous or categorical variables, representing the treatment response. Accordingly, prediction models have been designed to solve regression, classification, and ranking problems. This section reviews methods for representing treatment response and the corresponding prediction tasks.

3.1. Cell lines

Cell line studies constitute the most abundant resource of response data. High-throughput drug screenings with cell lines is performed with a compounds library, where each cell-drug combination is screened at multiple drug concentrations. The response of *in vitro* cells at each concentration is assessed *via* a cell viability assay which quantifies the surviving (viable) cells after treatment vs. the untreated control cells. Performing the experiments over a range of concentrations results in a vector of non-negative continuous dose-response values for each cell-drug pair. Those data points are summarized *via* a dose-response curve obtained by fitting a four parameters logistic Hill equation.

3.1.1. Continuous measures of response

The cell line dose-dependent responses lack a direct translation into the space of *in vivo* cancer models. A common approach is to extract from the dose-response curve a single-value summary statistic which represents the response for a cell-drug pair. Several methods exist allowing to calculate continuous response values which serve as the prediction target with supervised regression models. The two most common metrics are IC50 and AUC, with IC50 being substantially more prevalent (Figure 4).

Observing such predominance of IC50, an immediate question is whether IC50 exhibits substantial benefits over other measures. The half-maximal inhibitory concentration, i.e., IC50, is the concentration at which the drug reaches half of its maximal inhibition power on the fitted dose-response curve. Alternatively, measures such as AUC (area under the dose-response curve), AAC (area above the dose-response curve or activity area), and DSS (drug sensitivity score), are obtained by aggregating the cell viability values across a range of concentrations of the dose-response curve, providing what is considered a more global measure of response. Arguments in favor of these global measures are available, discussing the benefits of AUC (41, 60) and DSS (61) as opposed to IC50. In addition, empirical analyzes suggest better prediction generalization in the case of using AAC values as opposed to the alternative of using IC50 (14).

3.1.2. Categorical measures of response

Despite the prevalence of IC50 and AUC, there are arguments suggesting that continuous measures of response (blue in Figure 4) lack a straightforward interpretation in the context of decision-making purposes. Alternatively, a categorical output representing a discrete level of response such as *sensitive* vs. *resistant* is more comprehensible for humans, and thus, better supports actionable outcomes. Two primary approaches were used to produce categorical responses with DRP models.

In the more common approach, continuous responses were first categorized usually using one of the available methods such as waterfall algorithm (55, 62), LOBICO (56, 63), or a histogram-based method (64, 65). DL classifiers were trained on the transformed values to predict class probability, which is subsequently translated into a discrete response label (orange in Figure 4). The predicted probability can also be utilized as a quantitative measure of prediction uncertainty, an essential aspect in decision-making. Most models were trained to predict a binary response, representing that cancer is either sensitive or resistant to treatment. A multi-class classifier with three classes corresponding to low, intermediate, and high responsiveness, was also explored (66). The second approach is to train regression models to predict one of the continuous responses and then categorize it into two or more classes (green in Figure 4). Only a few papers explored this approach, including binary (67–70) and multi-class labels (71). This approach naturally allows assessing model performance using both regression and classification metrics, possibly offering a more robust generalization analysis.

3.1.3. Ranking

In addition to producing continuous or categorical predictions, models can be trained to learn a ranking function with drug response data. In the context of clinical precision oncology, the clinician might be interested in obtaining a ranked list of the top-k drugs that are likely to be most beneficial for the patient. In our search, we obtained three ranking models, all of which targeting personalized treatment recommendation. By framing the problem as a ranking task, the authors proposed models that learn to produce a ranked list of drugs per cell-line that are most likely to inhibit cell viability. Prasse et al. (72) transformed IC50 into drug relevance scores and derived a differentiable optimization criteria to solve the ranking problem which can be combined with different NN architectures. By combining their ranking learning method with the PaccMann architecture (73) and a fully-connected NN (FC-NN), they significantly improved the ranking performance as compared to baseline models. SRDFM used a deep factorization machine (DeepFM) with pairwise ranking approach which generates relative rankings of drugs rather than exact relevance scores (74). PPORank generates drug rankings using deep reinforcement learning which enables to sequentially improve the model as more data becomes available. SRDFM and PPORank rankings outperform non-DL models.

10.3389/fmed.2023.1086097

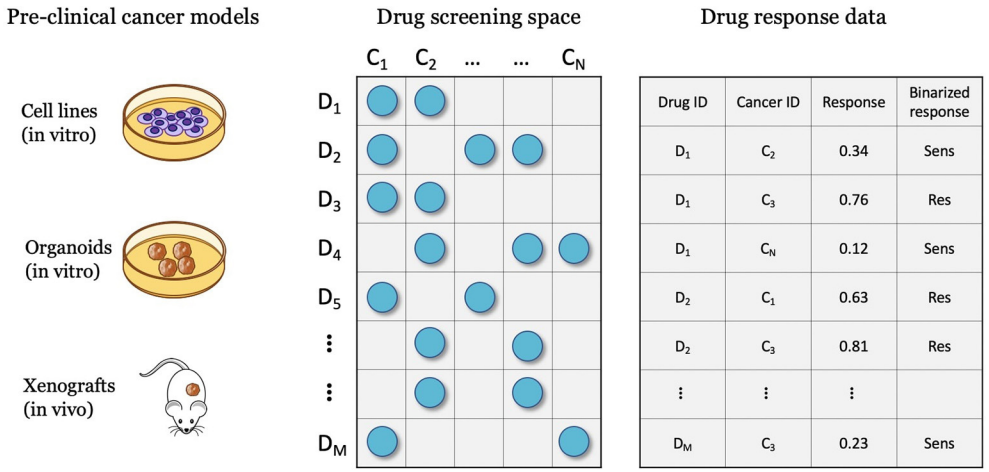


FIGURE 3
Drug screening experiments are performed with various cancer models such as cell lines, organoids, and xenografts, where cancer samples are screened against a library of drug compounds. The screening data is transformed into a drug response dataset that can be used for developing drug response prediction models, including regression, classification, and ranking.

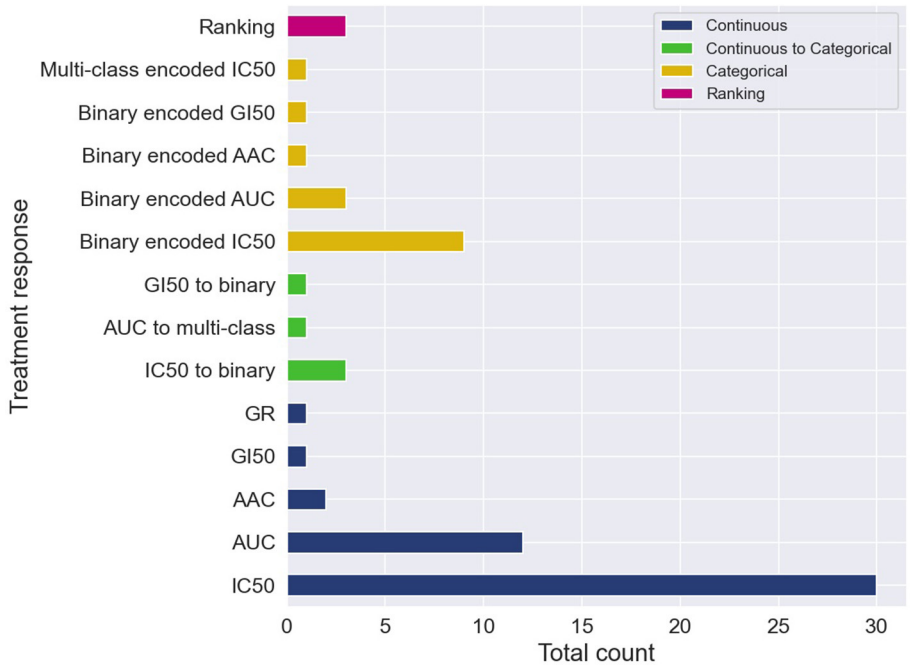


FIGURE 4
Cell-line drug response data is usually represented with continuous or categorical values. Drug response prediction (DRP) models use the different drug response representations to train regression, classification and ranking models. The histogram illustrates the prevalence of the difference representations and learning tasks. Certain papers exploit several representations of response and learning tasks, and therefore, these papers contribute more than one item to the histogram. The label *categorical* means that the continuous response was first categorized and then a DRP classifier was trained, while *continuous to categorical* means that a DRP regressor was trained and then the predicted response was categorized (in both cases, classification metrics were used for performance analysis).

3.2. Patient derived xenografts and patient tumors

Whereas, cell line data serve as the primary resource for training DL models, several papers proposed methods for predicting response in PDX and patient tumors. PDX is a contemporary cancer model that was developed to better emulate human cancer in medium-scale drug screenings, providing a controlled environment for studying the disease and systematically testing treatments in pre-clinical settings. While systematic drug screenings with patients is practically impossible (6), public data containing treatment response in individual patients are available.

Tumor response in humans is obtained *via* a standardized evaluation framework called Response Evaluation Criteria in Solid Tumors (RECIST) (75). RECIST involves non-invasive imaging followed by evaluation of change in tumor size. Four categories used for grading tumor change include Complete Response (CR), Partial Response (PR), Progressive Disease (PD) and Stable Disease (SD). Tumor change in PDXs is evaluated by monitoring tumor volume over time using more traditional methods (e.g., calipers) as opposed to using RECIST, primarily due to cost. The main data resource for modeling drug response with PDXs, provides continuous response values calculated based on tumor volume (59). To create a counterpart to RECIST, they also set cutoff criteria allowing to transform the continuous response values into the four categories.

Because discrete response labels are available in patient and PDX datasets, DRP models are primarily designed as classifiers, where the four-label RECIST categories are transformed into sensitive (CR and PR) and resistant (PD and SD) labels. Small sample size, however, is a major challenge in developing DL models with these datasets. This issue is generally addressed by incorporating abundant cell line data and transfer learning schemes to improve predictions in PDX and patients (discussed in Section 6.2.2).

4. Cancer representations

As personalized oncology treatments rely on omics biomarkers, progress in high-throughput tissue profiling plays an important role in existing and future cancer therapies. Pharmaco-omic studies conduct multiomic profiling of cancer models and drug screening experiments (9). Advances in sequencing technologies allowed to substantially scale the experimental studies by increasing the throughput rate and reducing the cost of profiling. The algorithms for processing the raw data evolve as well, thereby providing more reliable representations of the underlying cancer biology and allowing projects to update their repositories with refined versions of existing data while utilizing improved processing techniques (e.g., DepMap portal[2]).

Dedicated bioinformatics pipelines for transforming raw data depend on the omic type and the profiling technology, where common steps include alignment, quantification, normalization, and quality control. Profiling at each omic level produces high-dimensional representation of cancer that can be used as features in the downstream modeling of drug response. Existing prediction models utilize features that were obtained primarily at four omic levels including genomic [mutation, copy-number variation (CNV)], transcriptomic (gene expression microarrays, RNA-Seq), epigenomic (methylation), and proteomic [Reverse Phase Protein Arrays (RPPA)] (9). Figure 5A shows the prevalence of these representations in DRP models. Many models use these omic features directly as inputs to NNs while others optimize representations *via* additional preprocessing with the goal of increasing predictive power (76, 77).

In 2014, the NCI-DREAM challenge was a community effort that assessed performance of various classical ML models in predicting drug response in breast cancer cell-lines (27). It was reported that for most models participated in the challenge, gene expression microarrays provided greater predictive power than other data types. Results of this challenge seem to set the tone for future research in this field as gene expression was used in approximately 90% of the models either alone or in combination with other feature types, including mutation, CNV, methylation, and RPPA. Mutation and CNV are also common but rarely used without gene expression (with only 12% of such models).

Learning from multiple omic types have shown to improve generalization, and as consequence, integration of multiomics have been a recent trend (Figure 5C). A straightforward approach is to concatenate the multiomic profiles to form a feature vector and pass it as an input to a NN model, a method commonly referred to as *early integration*. It has been shown, however, that *late integration* significantly improves predictions where the different omic profiles are passed through separate subnetworks before the integration (39, 40, 78). Yet, caveats related to data availability hinder the pertinence of multiomics as opposed to using single-omics. Not having all omics available for all samples is a common occurrence in pharmaco-omic projects. A common approach to address this issue is filtering the dataset to retain a subset of cancer samples that contain all the required omic types as well as drug response data. The multiomics data is closely related to multimodal learning which refers to prediction models that learn jointly from multiple data modalities (i.e., representations). It was demonstrated that leveraging multimodal data generally improves predictions.

5. Representations of drug compounds

Contemporary cancer treatment often involves administering therapeutic drugs to patients to inhibit or stop growth of cancer cells, destroy tumors, or boost cancer-related immune system. Drug molecules are 3-D chemical structures consisting of atoms and bonds with complex atomic interactions. Developing numerical representation of molecules is an active research area in several related disciplines, including *in silico* design and discovery of anticancer drugs.

Ultimately, drug representations should be able to encode essential physical and chemical properties of molecules in compact formats. Alternatively, when using these representations as drug features for DRP, the model should be able to properly ingest these features and extract information predictive of treatment response. Therefore, the inherent information encoded with numerical drug representations and the model capability to learn from that specific representation are closely related and mutually important for producing efficient DRP models.

Neural networks provide substantially more flexibility than classical ML in learning from unstructured data such as strings, images, and graphs. This flexibility facilitates a noticeable trend of exploring various drug representations as features in DL-based DRP models (21). Primary types of drug representations include SMILES, fingerprints, descriptors, and graph-based structures (Figure 5B).

5.1. SMILES

Perhaps the most common format for querying, handling and storing molecules when working with modern chemoinformatics software tools and databases is a linear notation format called SMILES (simplified molecular-input line-entry system). With this

2 Cancer Dependency Map Portal. Last accessed: January (2023). https://depmap.org/portal.

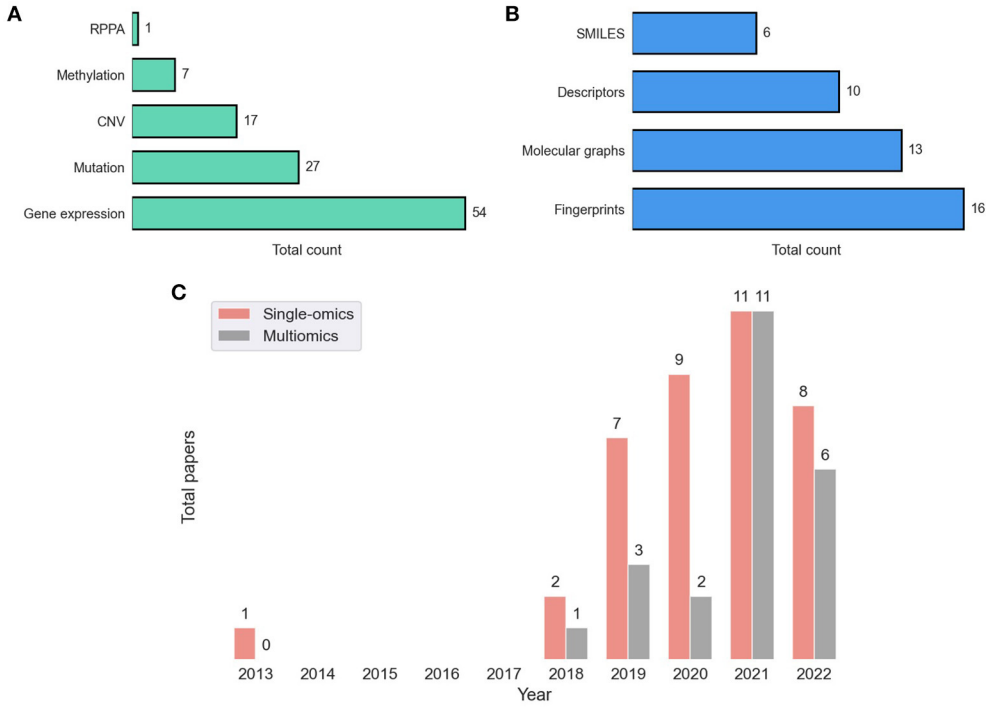


FIGURE 5
Feature representations in drug response prediction (DRP) models. Various representations can be used to represent cancers and drugs in DRP models, as described in Sections 4, 5, respectively. Each DRP model usually exploits one or multiple feature types (Supplementary Table 1 lists the feature types that each model have used). **(A)** The prevalence of omics (cancer) feature representations in DRP models. **(B)** The prevalence of drug feature representations in DRP models. **(C)** The distribution of papers that utilized single-omics and multiomics features in DRP models. We used Supplementary Table 1 to generate these figures, where we considered only peer-reviewed publications, collected until August 2022. CNV, copy number variations; RPPA, reverse phase protein arrays.

data structure, each molecule is represented with a string of symbolic characters generated by a graph traversal algorithm (21, 79). Although the use of SMILES in DRP is quite limited as compared to other applications (e.g., molecule property prediction, QSAR), this format provides several benefits.

Following common text preprocessing steps (e.g., tokenization, one-hot encoding), SMILES can be naturally used with common sequence-aware modules such as RNN (72, 73, 80) and 1-D CNN (81). Moreover, each molecule can be represented with different strings depending on the initial conditions when applying the SMILES generating algorithm (i.e., starting node of graph traversal). The resulting strings are referred to as randomized or enumerated SMILES and have been reported to improve generalization when utilized for drug augmentation (82, 83). SMILES are less prevalent in DRP models but remain a popular notation for describing molecules because it often serves as an intermediate step for generating other representations such as fingerprints, descriptors, and graph structures.

5.2. Descriptors and fingerprints

Fingerprints (FPs) and descriptors are the two most common feature types for representing drugs in DRP papers (Figure 5B). Unlike with SMILES, where the string length varies for different molecules, we can specify the same number of features for all drugs in a dataset with either FPs or descriptors. The consistent feature dimensionality across drugs makes descriptors and FPs easy to use with NNs and classical ML. With FPs, a drug is a binary vector where each value encodes presence or absence of a molecular substructure with a common vector size of 512, 1,024, or 2,048. Multiple algorithms for generating FPs are available, implemented by several chemoinformatics packages. For example, Morgan FPs refers to Extended Connectivity Fingerprints (ECFPs) generated *via* the Morgan algorithm. ECFP is a class of circular FPs where atom neighborhoods are numerically encoded with binary values. The open-source package RDKit provides an API for generating these FPs which are often used for DRP.[3] Descriptors are a vector of continuous and discrete values representing various physical and chemical properties, usually containing hundreds or a few thousands of variables. Both open-source and proprietary software tools are available for generating descriptors, where PaDEL (84), Mordred (85), and Dragon (86) being particularly used for DRP. A systematic assessment of various features and HPs suggests that DRP with DL exhibits no significant difference when utilizing ECFP, Mordred, or Dragon representations as drug features (87).

3 RDKit: Open-source cheminformatics. Last accessed: January (2023). https://www.rdkit.org.

5.3. Graph structures

Graphs are powerful representations where complex systems can be represented using nodes and edges. Recent advancements in GNNs have opened promising research directions allowing efficient learning of predictive representations from graph data (88–90), and contributing to the widespread interest in GNN among AI researchers and application domain experts (91, 92). Graph-based representations have emerged as a new trend in computational drug development and discovery (93, 94). In graph notation, each molecule is described with a unique graph of nodes and edges, where each atom (i.e., node) and each bond (i.e., edge) can be represented with multiple features characterizing physical and chemical properties. Resources and examples for constructing graphs are well documented in popular chemoinformatics toolkits, as well as software packages dedicated for GNNs (95–97).

By borrowing methodologies from related fields such as drug property prediction and drug design, developers of DRP models exploit graph molecular structures for representing drugs combined with GNN-based architectures. We observe concordant results among those papers investigating molecular graphs with GNNs, reporting superior performance of their models compared to baselines that use non-graph representations. Based on these recent results and the assumption that graphical description is arguably a more natural way to represent drugs as compared to the aforementioned alternatives, molecular graphs combined with GNNs should be a default modeling choice. However, only few papers actually report results with rigorous ablation analysis with extensive HP tuning comparing molecular graphs with other representations such as SMILES, FPs, and descriptors for the application of DRP.

6. Deep learning methods for drug response prediction

We review existing approaches for modeling DRP with DL in terms of two perspectives: (1) NN modules that are used as building blocks for constructing DRP architectures, (2) learning schemes that are used to train models with the goal to improve prediction generalization. The diversity of methods is highlighted in Figure 6.

6.1. Neural network modules

Constructing a NN architecture requires choosing appropriate components, their properties, and the way these components are organized and connected (i.e., network topology). These components are often available as modules in DL frameworks and range from simple dense layers to more advanced structures such as attention. Certain structures, however, are not generally available as is and should be constructed manually using the available modules (e.g., residual connections). While there is an absence of rigorous methodologies for designing network topology, the choice of certain modules can be driven by the characteristics of input data.

6.1.1. Dense layers

Menden et al. (34) is a pioneering work where cell line and drug features were used to train an FC-NN with a single hidden layer to predict IC50. Recently, FC-NN models with multiple dense hidden layers have been proposed to predict response in cell lines (71) and humans (67). Several other models with only dense layers were explored, combining advanced NN attributes. MOLI utilizes triplet loss function and late integration of multiomic inputs to generalize across cancer models by training with cell lines and predicting in PDXs and patients (39). DrugOrchestra is multi-task model that jointly learns to predict drug response, drug target, and side effects (98). RefDNN explores data preprocessing techniques to generate more predictive cell line and drug representations (99). For cell line representation, multiple ElasticNets produce a vector representing drug resistance of a cell line to a set of reference drugs, while for drug representation, a structure similarity profile is computed for each input drug with respect to the set of reference drugs. Another model with late integration of multiomic data is proposed for predicting drug response and survival outcomes which also uses neighborhood component analysis for feature selection of multiomic data (69). PathDSP utilizes multi-modal data preprocessed *via* pathway enrichment analysis and integrated into an FC-NN for DRP (45). All the aforementioned models demonstrate promising prediction performance while utilizing dense layers only in their architectures.

6.1.2. Convolutional layers

Due to state-of-the-art performance in various applications (100) and fewer learning parameters required as compared to dense layers, 1-D CNNs are a popular choice in DRP models, particularly for processing omics data. DRP models utilizing CNNs have been published every year since 2018.

DeepIC50 is a multi-class classifier exploiting an early integration of cell line mutations and drug features (descriptors and FPs) passed to a three-layer CNN and followed by an FC-NN prediction module (66). tCNNs exploits late integration of two parallel CNN subnetworks with three layers each (101). Binary features of cell lines (mutations and CNV) and drugs (one-hot encoded SMILES) are propagated through the respective subnetworks, concatenated, and passed through an dense layer for the prediction of IC50. GraphDRP is very similar to tCNNs in terms of CNN utilization, with the primary difference being the utility of molecular graphs and GNN layers for learning drug representations. CDRScan is an ensemble of five different architectures with varying design choices (late and early integration, shallow and deep NNs, with and without dense layers) which all contain CNNs and learn to predict IC50 from mutation and FP data (102).

DeepCDR takes the design further by exploiting late integration of drug and multiomic features, including genomic mutation, gene expression, and DNA methylation, where only the mutations are passed through a CNN layers (103). Following DeepCDR, two similar models were published, GraTransDRP (104) and GraOmicDRP (105) which both use late integration of multiomics and molecular graphs for drug representation. While DeepCDR uses CNNs only for genomic features, all three subnetworks for multiomic inputs consist of CNNs in GraTransDRP and GraOmicDRP. SWNet is another model utilizing late integration of cell lines and drugs with three CNN layers in the cell-line subnetwork and one such layer in the prediction subnetwork (106).

We may gather that 1-D CNNs carry substantial benefits in the context of DRP, considering their prevalence (Figure 6). Based

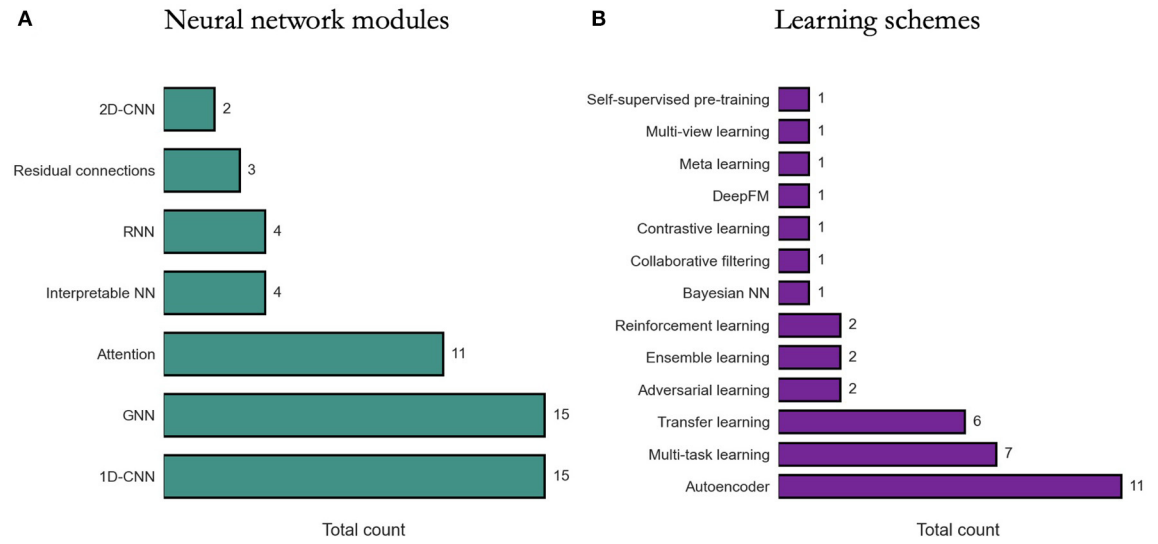


FIGURE 6
Prevalence of deep learning methods in drug response prediction (DRP) papers. Various methods have been used to build DRP models which can be categorized into neural network (NN) modules (Section 6.1) and learning schemes (Section 6.2). The prevalence of NN modules and learning schemes across papers (from Supplementary Table 1) is shown, respectively, in **(A, B)**. 1D-CNN and 2D-CNN, one- and two-dimensional convolutional NN; GNN, graph NN; DeepFM, deep factorization machine.

on our search, however, only two papers explicitly report their empirical findings assessing CNNs vs. alternative learning modules. Interestingly, both papers suggest that CNNs underperform dense layers in their respective architectures. Zhao et al. (71) reports that in a twelve-layer model predicting response from gene expression, dense layers outperform CNNs or RNNs. Manica et al. (73) present extensive analyzes exploring various architectures for encoding drug representations directly from SMILES, including CNNs, bidirectional RNNs (bRNNs), and various attention modules. Their results suggest that a combination of convolutional and attentions modules produces the most predictive model while a CNN-only encoding subnetwork performs the worst as compared to the explored variants.

With the objective to leverage 2-D CNNs, two algorithms were recently published for converting tabular data into images and utilizing CNNs for predicting drug response. Both algorithms, REFINED (76) and IGTD (77), convert gene expressions and drug descriptors into images as a preprocessing step, and then learn to predict drug response with late integration of 2-D CNN subnetworks. The papers report superior performance of proposed algorithms as compared to various baseline models.

6.1.3. Attention mechanism

Presenting PaccMann (80), Oskooei et al. were the first to report the use of attention-based NN for DRP, exploiting late feature integration with attention mechanisms incorporated both in the cell line and drug subnetworks. On the cell line path, gene expressions are encoded with self-attention producing a gene attention (GA) vector. On drug path, SMILES embeddings are combined with GA *via* contextual-attention, where the GA vector serves as the context. This design which was further described in Manica et al. (73) improves generalization as compared to baselines as well as facilitates interpretability *via* attention weights.

Attention-based designs have been further explored in recent years. CADRE is a collaborative filtering model with contextual-attention and pre-trained gene embeddings (gene2vec), designed to recommend treatments based on cell line gene expressions (107). Ablation analyzes against simpler models suggest that CADRE's attributes yield improved generalization. Moreover, the authors show how attention weights could be used to identify biomarker genes, partially addressing the notorious black-box property associated with many DL models. In AGMI, attentions are used for aggregating heterogeneous feature types, including raw multiomics as well as engineered features *via* protein-protein interaction (PPI), gene pathways, and gene correlations with PCC (Pearson correlation coefficient), contributing to improved generalization. Overall, papers report the predictive benefits of attentions (106–111), while a few papers also explore attention weights for interpretability (107, 110).

Transformer is an attention-based architecture that have recently been explored in DRP models for encoding drug representations (70, 104). In GraTransDRP (104), the authors propose to extend an existing model, GraOmicDRP (105), by modifying the drug subnetwork to include graph attention network (GAT), graph isomorphism network (GIN), and a graph transformer which learns from graph data. In DeepTTA (70), a transformer module encodes drug information represented as text data [ESPF substructures (112)]. Both models report significant improvement in generalization thanks to transformer modules. Transformers have shown immense success first in language and later vision applications, and expected to gain further attention from the DRP community.

6.1.4. Graph neural network layers

Designed to efficiently learn from graph data, GNNs have been extremely popular in applications where information can be represented as graphs. With at least 15 GNN-based DRP models published since 2020 (Supplementary Table 1), it is apparent that

the use of GNNs with graph data becomes an emerging alternative to some of the more traditional approaches discussed above. In DRP papers, a graph, $G = (V, E)$, is often characterized by a set of nodes V (a.k.a. vertices), a set of edges E (a.k.a. links), an adjacency matrix W which contains values representing associations between the nodes, and an attribute matrix A containing feature vectors representing node attributes (113). A primary challenge that researchers are facing when designing GNN-based models lies in how to transform components of the DRP problem into graphs. Luckily, accumulated knowledge and methodologies allow viewing biological and chemical systems as networks.

Due to substantial progress in applying GNNs to drug discovery and development (94, 95), it was relatively straightforward to adopt similar techniques to DRP. Papers propose models that convert drug SMILES into graphs and utilize modern GNN layers to encode latent drug representations (49, 103–106, 114). The common approach is constructing a unique graph per drug with nodes and edges representing atoms and bonds, respectively, where node features are the properties of each atom. The subnetworks consist of several (usually 3–5) GNN layers such as GCN and GIN.

Another approach is constructing graphs using biological information of cancer samples, where genes are the graph nodes and gene relationships are the edges. There is slightly more diversity of approaches in this space as compared to drugs partially due to the use multiomic data which allow constructing graphs with heterogeneous node attributes. The multiomics can be utilized to encode gene relationships and attributes using one or more data modalities, including correlations between genes (109, 111, 115, 116), known protein interactions (i.e., PPI) (109, 111, 117, 118) using STRING database (119), and relationships based on known gene pathways (109) using GSEA dataset (120). Recently, novel approaches have been explored such as heterogeneous graphs where both cell lines and drugs are encoded as graph nodes (108, 116, 118, 121), and a model that utilizes diverse data types for building graphs, including differential gene expressions, disease-gene association scores and kinase inhibitor profiling (111). Despite the diversity of the different approaches, most papers report superior performance as compared baseline models.

6.2. Learning schemes

Various learning schemes have been proposed to improve drug response prediction in cancer models. These techniques can be used in different configurations, usually regardless of the fundamental NN building blocks that are used to construct the architecture.

6.2.1. Autoencoders

When considering the large feature space of cancer and drug representations (Sections 4, 5) and the size of drug response datasets, the number of input features outnumbers the number of response samples which potentially leads to model overfitting (6). Learning predictive representations with high-dimensional data manifests a primary strength of multi-layer NNs. A prominent example are autoencoders (AEs) which are NNs that are trained to compress and then reconstruct data in an end-to-end unsupervised learning fashion. Given that an AE can reliably reconstruct the input data, the compressed latent representation is characterized by reduced data redundancy which improves the feature to sample ratio and can be subsequently used as inputs in downstream drug response prediction task.

The primary application of AEs in DRP models is dimensionality reduction. The compressed representation is used as input to a DRP model which is usually a NN or in some cases a classical ML model. DeepDSC reduces gene expressions from about 20,000 genes down to 500 and concatenates the downsampled data with drug FPs as inputs to a FC-NN (122). VAEN exploits variational AEs (VAEs) to learn a low-dimensional representation of gene expressions which are fed into ElasticNet (48). DEERS compresses cancer features (gene expression, mutation, and tissue type) and drugs features (kinase inhibition profiles) into 10 dimensions which are concatenated for a FC-NN (44). Dr.VAE learns pre- and post-treatment embeddings of gene expressions and leverages those embeddings for DRP (123). AutoBorutaRF combines AEs with a feature selection method followed by Boruta algorithm and RF (124). Ding et al. (125) utilized learned representations of multiple hidden layers of the encoder as input features, rather than using just the latent representation.

AEs have also been utilized with external datasets as part of model pretraining. DeepDR uses gene expressions and mutations from the TCGA database to train separate AEs for each omic type (126). The pre-trained encoders are extracted from the individual AEs, concatenated, and passed as inputs to a FC-NN. The combined model was trained to predict drug response in cell lines. Another model uses separate AEs to encode cancer and drug features (127). A GeneVAE encodes gene expressions from CCLE dataset, and a Junction-Tree VAE (JT-VAE) (128) encodes drug molecular graphs from the ZINC database. Similar to DeepDR, the pre-trained encoders are concatenated and combined for DRP.

Dimensionality reduction and model pre-training are common applications of AEs. Yet, AEs have also been utilized in less conventional ways. Xia et al. (41) proposed UnoMT, a multi-task learning (MTL) model that predicts drug response in cell lines *via* late-integration of multiomic features and drug representations. In addition to the primary DRP task, the model predicts several auxiliary tasks, where one of them is a decoder that reconstructs gene expressions. TUGDA exploits advanced learning schemes such as domain adaption and MTL to improve generalization in datasets with limited sample size (51). Each prediction task in the MTL corresponds to a different drug. With the goal to mitigate the influence of unreliable tasks and reduce the risk of negative transfer, a regularization AE takes the model output and reconstructs an intermediate hidden layer.

6.2.2. Transfer learning

Transfer learning refers to learning schemes designed to improve generalization performance in a target domain T by transferring knowledge acquired in a source domain S (129). The most common scenario is to transfer learned representations from a source domain with large amounts of data into a target domain which suffers from insufficient data. Transfer learning with DL have shown remarkable success in various applications (130) and have recently been applied to DRP (14, 42, 50, 51, 78, 131).

Zhu et al. proposed an ensemble transfer learning (ETL) that extends the classical transfer learning by combining predictions from multiple models, each trained on a cell-line dataset and fine-tuned on a different cell-line dataset. Considering three application scenarios,

a NN with late-integration of gene expression and drug descriptors outperformed baselines in most cases, while the LightGBM showed superior performance in certain experimental settings. AITL was proposed to improve generalization across biological domains (50). The model was trained using gene expressions and drug responses from source and target domain data, where abundant cell line data serve as source and the relatively scarce data from either PDXs or patients serves as target. A primary component for accomplishing knowledge transfer is a feature extractor subnetwork that learns shared representations using source and target features, which are passed to multi-task subnetwork for response prediction of source and target samples. Ma et al. (131) proposed TCRP, a two-phase learning framework with meta-learning as a pre-training step followed by a few-shot learning for context transfer. With the goal to obtain transferable knowledge, a meta-learning is applied iteratively by training a simple NN with different subsets of cell lines. The same NN was trained to predict drug response in both phases but with different data. With a few samples from the target context, the pre-trained NN was further trained for a single iteration, demonstrating performance with patient-derived cell line (PDTC) and PDXs.

ETL and AITL can be categorized as inductive transfer learning, where labels from both the source and target data are used to improve generalization on target data (130). However, abundance of pre-clinical and clinical data are not labeled, and therefore, remain unused with classical transfer learning. Velodrome is a semi-supervised model that exploits labeled cell lines and unlabeled patient data to improve out-of-distribution (OOD) generalization on datasets that remained unused during training. Generalization has been demonstrated with PDX, patient, and cell lines from non-solid tissue types. TUGDA is an MTL and domain adaptation model which relies on cell line data and advanced learning schemes to improve predictions in data-limited cancer models (51). The model takes gene expressions and learns a latent shared representation that is propagated to predict drug response in a multi-task fashion with each task corresponding to a different drug. To enable domain adaptation from cell lines to other cancer models such as PDXs and patients, adversarial learning is used where a discriminator is branched from the shared layer to classify the type of cancer model. The discriminator is employed in supervised and unsupervised steps, with source- and target domain samples, respectively.

6.2.3. Multi-task learning

As transfer learning and multi-task learning (MTL) are related approaches (129, 130), several models have integrated both methods in their DL-based DRP models (41, 50, 51). However, utilizing MTL without transfer learning is also common. DrugOrchestra is an original work where multiple heterogeneous data sources were curated to develop a multi-task model for simultaneously learning to predict drug response, drug target, and side effects while utilizing advanced weigh sharing approaches (98). SWnet proposed a gene weight layer which scales the contributions of mutations when combined with gene expressions as a multiomic input (106). In its simplest form, the weight layer is trained for the entire dataset. The authors also explored a weight matrix where each vector is determined for a single drug. Considering N drugs in the dataset, the model was trained in an MTL fashion with N prediction tasks, one task for each drug, outperforming the single weight layer configuration. Zhu et al. (118) used MTL in a pre-training step to accumulate chemical knowledge by training a GNN to predict more than a thousand biochemical properties of molecules, a strategy adopted from another work (132). The pre-trained model was integrated into the complete DRP model called TGSA.

7. Model evaluation and comparison

Papers proposing DRP models explore advanced approaches for data representation, NN architecture design, and learning schemes. To motivate the increasing complexity of methods, it is essential to demonstrate their utility for cancer medicine and advantages against existing approaches (i.e., baseline models). Model performance in these papers is primarily assessed *via* prediction generalization, i.e., prediction accuracy on a test set of unseen data samples, with several common metrics used for regression, classification, and ranking. The choice of the test set is critical as it may illustrate the potential utility of the model for a given application in cancer medicine and significantly affect the observed performance. To make a comparison between proposed and existing models compelling, consistent evaluations should be utilized in terms of drug response datasets, training and test sets, evaluation metrics, and optimization efforts (e.g., HP tuning). This section presents common evaluation approaches used for prediction generalization in terms of drug screening datasets (single or multiple datasets), data splitting strategies (train and test), and baselines models (Figure 7).

7.1. Single-dataset evaluation

A single drug screening study is most commonly used to develop and assess performance of DRP models. Regardless of specific cancer models, the screening data space can be characterized by the involved drugs and cancers. The 2-D matrix in Figure 3 illustrates such a space composed of a finite number of known cancer cases and drugs, where the marked coordinates symbolize that treatment responses are available for these combinations. Considering this 2-D space, four data splitting strategies are commonly used, where each can imitate a different application scenario of drug response prediction (Figure 7A).

7.1.1. Known cancer and drugs (mixed-set)

To ensure a valid assessment of prediction generalization, training and test sets should contain unique cancer-drug pairs without overlap. The overlap may occur, however, on cancer or drug (but not both) in which case a model encounters drug or cancer features during testing that also showed up during the training phase. For example, consider two drug response samples $r_{11} = \{d_1, c_1\}$ included in the training set and $r_{12} = \{d_1, c_2\}$ in the test set. While the pairs are unique, the drug features of d_1 show up during both training and testing. The cancers or drugs that show up in test and training set are deemed as *known* or *seen*. This splitting strategy is also known as mixed-set, where drugs and cancers are mixed in training and test sets. Randomly splitting cancer-drug pairs naturally results in a mixed-set analysis, where drug and cancer features can show up in both training and test sets. Due to the overlap on both feature dimensions, this analysis naturally leads to the highest performance, and it is also the most common and straightforward to implement.

10.3389/fmed.2023.1086097

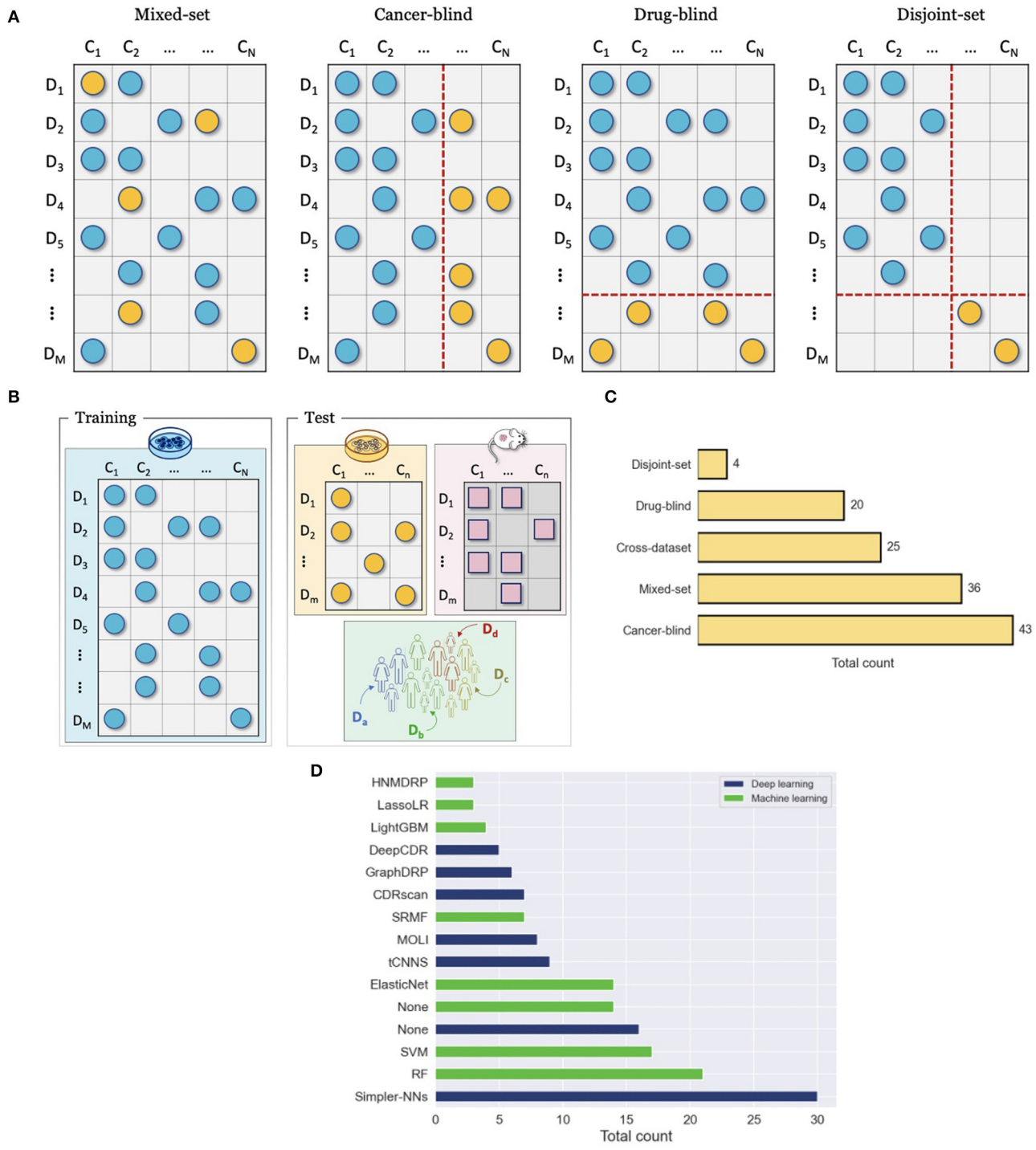


FIGURE 7
Evaluation and comparison of drug response prediction (DRP) models. **(A)** Data splitting strategies for evaluating performance with a single drug screening study include mixed-set, cancer-blind, drug-blind, and disjoint-set. **(B)** Cross-dataset evaluation where drug response data that is used for training and evaluation come from different studies. **(C)** Prevalence of common evaluation methods across studies (from Supplementary Table 1). The methods include mixed-set, cancer-blind, drug-blind, disjoint-set, and cross-dataset, as described in Sections 7.2, 7.1. **(D)** Histogram of the top fifteen most popular baseline models that were used to benchmark prediction performance of DRP models. The baselines are color-coded as either deep learning (DL) or classical machine learning (ML). The label Simpler-NNs refers to simpler versions of proposed models, where ablation analysis was usually conducted. The label None refers to cases where this type of baseline (i.e., ML or DL) was not used in performance analysis at all (e.g., the blue bar for None shows that 16 of the papers have not used DL-based baselines in their analysis).

Because of the simplicity of this strategy, almost all papers include it in their analysis. An application of prediction models developed with the mixed-set validation is repurposing of known anticancer drugs for unexplored cancer conditions that have not been treated with these drugs. When considering the 2-D space of known cancer cases and drugs in a given dataset, not all cancer cases and drugs

have been screened against each other. If a DRP model exhibits high generalization, it can be deployed as a virtual screening tool which helps in identifying highly sensitive cancer-drug pairs and thereby assisting in the design of experiments or treatments.

7.1.2. Unknown cancers and known drugs (cancer-blind)

When a dataset is split such that an overlap occurs only on drugs but not on cancer cases, it is said that drugs are known but the cancer cases are unknown with respect to the test set. This splitting strategy is also sometimes referred to as cancer-disjoint, cancer-blind, unseen-cancer, and a special case called leave-n-cancers-out where a certain number of cancers is excluded from the training set. Splitting a drug response dataset this way requires extra work which is, therefore, less common than random splitting but still prevalent, especially in recent years. The performance is usually worse than in the mixed-set analysis due to increased complexity in generalizing across the omics feature space of cancers. Alternatively, this type of analysis is perhaps the most adequate in simulating the utility of DRP models for the design of personalized cancer treatment. When considering a potential workflow of utilizing a DRP model for personalized cancer treatment, the patient tumor data is not expected to be included in the training set that was used for model development. However, the drug treatments would be chosen from a collection of known compounds (approved or investigational) that may have been screened against other cancer cases, the drug features of which were used to train the prediction model during model development. Models that exhibit high generalization in this scenario are considered more suitable for designing personalized cancer treatment.

7.1.3. Known cancers and unknown drugs (drug-blind)

Analogously to cancer-blind analysis, the drug-blind refers to data partitioning where an overlap between the training and test sets may occur on cancers but not on drugs. This data splitting strategy is also sometimes referred to as drug-disjoint, unseen-drugs, and a special case called leave-n-drugs-out where a certain number of drugs is excluded from the training set. Interestingly, the generalization in drug-blind analysis as evaluated by common performance metrics is significantly worse than in mixed-set and cancer-blind analysis and, in some cases, models completely fail to make effective predictions. The worse performance can be attributed to the immense chemical space of drug compounds which imposes a challenge on predictive models of learning generalizable feature embeddings with just a few hundreds of drugs that were screened in typical drug screening studies. In addition, it has been shown empirically that drug diversity contributes to a majority of response variation (41, 133), which can explain the performance drop in drug-blind analysis. Models that excel in generalizing for unknown drugs can be useful for repurposing of non-cancer therapies to cancer indications and development of novel drugs for cancer treatment. Existing drugs approved for non-cancer diseases can be repurposed for cancer treatment, thereby decreasing expenses and speeding up the time to market. A DRP model, exhibiting high performance in drug-blind scenario, can be utilized for *in silico* drug screening across cancer types and libraries of approved compounds.

7.1.4. Unknown cancers and drugs (disjoint-set)

By extending the drug-blind and cancer-blind analysis to both dimensions, generalization analysis can be performed where both drugs and cancers remain disjoint between training and test sets. This analysis exhibits the worst generalization performance. It is primarily used to assess the capacity of models to generalize in this challenging scenario where the application of such model in clinical or preclinical setting is not essentially obvious. Models exhibiting significantly better performance as compared to baselines could be interesting cases for further exploration of the model architecture to scenarios more relevant in pre-clinical and clinical settings.

7.2. Cross-dataset evaluation

Another validation scheme is to assess generalization across datasets, where response data for model development and model testing are derived from different drug screening studies (Figure 7B). The source dataset (D_S), used for model development, and the target dataset (D_T), used for evaluation, can be of the same or different cancer models. Several selected works have been summarized in Sharifi-Noghabi et al. (14), showing that cell lines constitute the primary cancer model for D_S while cell lines, PDX, and patient response data are all common options for D_T. Cross-dataset analysis is less prevalent as compared to single-dataset analysis. Primary challenges are related to data preparation and arise when standardizing metadata across datasets, including annotations of drugs, cancer samples, and omics features, as well as utilizing consistent data preprocessing steps for feature normalization (e.g., addressing batch effect) and computation of response metrics.

Inconsistencies in drug screening data across cell line studies, partially due to different experimental setups, are an acknowledged reality which impedes naive integration of data from multiple sources into a single dataset (62, 134). A large-scale empirical study focused on assessing generalization of DL and ML models across five cell line datasets suggests that generalization may depend on multiple attributes of D_S and D_T datasets, including the number of unique drugs and cell lines in D_S and D_T, the drug and cell line overlap between D_S and D_T, and the sensitivity assays used for measuring drug response. Although this type of analysis requires adequate knowledge and substantial effort in data preparation and model development, demonstrating the ability to make accurate predictions across datasets or cancer models can be essential for a DRP model to become a favorable candidate for preclinical and clinical studies.

7.3. Baselines

Proposing novel DRP models could be seen unnecessary unless it leads to better performance as compared to existing models. In predictive modeling, baseline models usually refer to models that serve as a point of reference. Figure 7D shows the number of times each model was used as a baseline across papers, where ML-based and DL-based models are color-coded to distinguish between the two types. The label *None* refers to cases where this type of baseline (i.e., ML or DL) was not used in performance analysis at all (e.g., the blue bar for *None* shows that 16 of the papers have not used DL-based baselines in their analysis).

The most popular choice for a baseline is to use a *simpler* version of a proposed model, where simplicity is defined in the context of a given model. For example, in the case of MOLI which advocates for multi-omics late integration with triplet loss function, simpler versions include combinations of single- instead multi-omics, early instead of late integration, and binary cross-entropy loss function instead of triplet loss function (39). Using simpler versions of a model as baselines can be both straightforward and insightful. In many cases, modifying a model to construct a baseline involves minor changes of little complexity where model attributes are replaced with simpler alternatives followed by ablation analysis demonstrating the necessity of the proposed attributes. However, the actual baselines and ablation experiments that are chosen to benchmark models are a critical aspect that affects credibility. Some DL models explore alternative data representations, such as images (76, 77) or graphs (105, 118). In these cases, ablation studies with common vector-based representations (e.g., FPS, descriptors) should be considered. Certain models incorporate structural biological information such as pathways into NN architectures with the goal to produce interpretable DRP models. With these models, a baseline may need to incorporate random pathway information to serve as a reference point to any claimed pathway interpretability, as discussed in Li et al. (135).

The second most popular baseline model is random forest (RF) presumably due to its availability, simplicity, and predictive power. Boosting algorithms such XGBoost and LightGBM are less prevalent although poses similar characteristics as RF yet faster and often more predictive. Tree-based models and boosting algorithms have shown tremendous success in various prediction tasks and data science competitions. HP tuning is often an essential step to squeeze the maximum performance from these algorithms. When these models are used as baselines in DRP papers, very little (if any) HP tuning is performed. When the focus of proposed models is on prediction performance, extensive HP tuning should be performed with the baselines the reported.

Another common approach to demonstrate model performance is to compare it with other community models. The main problem with this approach is that many papers extract reported performance scores from original publications without actually implementing the models or using the same dataset. Model performance is likely to depend on feature types and specific samples allocated for training and testing. Thus, this approach makes sense in cases where benchmark datasets have been established with clearly designated training and testing samples. However, this is not yet the case with DRP tasks. Reproducing results from papers is challenging, time-consuming, and often contradicts the reported scores (26). Alternatively, there are papers that diligently reproduce models and use these as baselines. So long as benchmark datasets have not been established, reproducing models and benchmarking with consistent datasets should be considered the standard if one decides to compare a model with another published community model.

Oftentimes, simple analysis can lead to useful insights. For example, in the same study that analyzed baselines for interpretable pathway-based architectures (135), the authors also explored a *naïve* baseline which simply reports the average of drug response values across all samples. While very easy to implement, this baseline performed surprisingly well. This simple and model-agnostic baseline can be very informative not only with respect to a specific model, but more generally for the entire prediction task. Despite the extra amount of work which may require the design of appropriate baselines, rigorous and comprehensive benchmark analysis is likely to establish fidelity and attract the community.

8. Discussion

8.1. Practices

Development of DRP models is a comprehensive task, success of which depends on multiple factors, such as prepossessing and representation of data, model training, and performance evaluation. Good design choices related to each component of the DRP workflow (Section 2) can positively contribute to model performance, while bad practices can result in adverse effects, leading to poor performance. Here, we provide a short summary of practices that can guide researchers in developing DRP models.

Feature scaling is a common preprocessing step, usually applied to both cancer and drug features. However, when the dataset contains omics data from multiple sources or batches, active measures should be considered for mitigating systematic differences between the batches, which otherwise are likely to bias the downstream ML analysis (51, 136, 137). Possible measures include batch-correction (51) and architectures that are specifically designed to address discrepancies in the input feature space (50, 52). Notice that models can also benefit from standardizing data in the output space. For example, in order to combine response data from multiple sources, the drug response AUC values shall be calculated based on the same dose range (41). Still in the context of *in vitro* drug response, there is a consensus that the global measures of response such AUC and AAC are more robust and produce more predictive models as compared to models trained with IC50 (14, 41, 61).

Due to the diversity of NN architectures and learning schemes (Sections 6.1, 6.2), it is challenging to decipher best practices for DRP model development. Yet, some consensus seems to exist regarding several design principles. Late feature integration exhibits better generalization as compared to early integration approaches (39, 40, 78). Transfer learning and domain adaption significantly improve generalization across datasets and biological cancer models (50, 51). Similar to other application domains, DL for DRP usually requires extensive tuning of HPs to boost performance and integration of early stopping during the training process to avoid model over-fitting. Further comprehensive and systematic studies are required to produce unbiased conclusions regarding DL design methodologies for DRP.

Cross-validation with a DRP model can be performed through different data partitioning schemes, each simulating a different application of the model (Section 7). For example, if authors target the application of precision oncology, then cancer-blind analysis should be considered for model evaluation (Section 7.1.2). Alternatively, if mixed-set analysis is used (Section 7.1.1), the model will produce inflated scores, exhibiting overoptimistic results in the context of precision oncology. Similarly, drug-blind analysis should be used with models targeting drug development applications

which commonly leads to significantly lower performance scores as compared to mixed-set and cancer-blind analysis (Section 7.1.3). Whenever possible and regardless of the chosen evaluation scheme, extensive cross-validation should be performed with multiple data splits (e.g., k-fold cross-validation repeated several times) that would generate enough data points for statistical significance analysis and comparison with baseline models (i.e., mean and standard deviation).

The choice of adequate baseline models is critical for objectively assessing the true capabilities of proposed methods, where ablation analysis plays a central role. Claims made about the predictive power of promising data representations such as molecular graphs (Section 5.3) should be benchmarked against common alternatives such as descriptors and FPs (Section 5.2). Another reasonable baseline is a model trained on one-hot encoded labels of drugs, which utilize only drug identities while ignoring the features (133, 138). Similar one-hot encoding strategies can be applied to cancer features (138) as well as ablation analysis exploring various omics cancer representations (105). Novel NN modules that aim to better leverage DRP data should be benchmarked against alternative and generally less complex modules (e.g., attention modules vs. dense layer). For models integrating cancer pathway data, randomly generated pathways can be used to evaluate the gain obtained by leveraging the pathway information (135). Another baseline is to take the average of sample response values. While very easy to implement, this naïve baseline often exhibits surprisingly good results and can put the model performance into the appropriate perspective, allowing to control for inflated performance scores (135).

Finally, there are two more practices that could be highly beneficial for increasing visibility, but are often neglected. First, many papers finalize their performance analysis by demonstrating performance scores using various validation schemes. Instead, it would be highly beneficial to discuss how the proposed DRP model could be integrated in a larger patient care or drug development workflow and what are the primary challenges preventing these type of models from being deployed in real-world applications. Use case scenario could be highly relevant in this context. The second practice relates to software implementation. Many papers are accompanied by software and data that were used to build their models. However, the code repositories often miss certain pieces of information required to reproduce the paper results. Instead, it is recommended for papers to publish well-documented repositories with reproduce code, including data preprocessing scripts and installation instructions of the computational environment. This practice will improve the usability of models, their adoption as baselines, and visibility within the community.

8.2. Challenges

In recent years, AI approaches have taken a notable place in health care (139). Their promise of utilizing large data arrays to improve treatment protocols, aid drug development efforts, and increase diagnostic precision is attractive to the clinical research community. However, navigating a complex landscape of factors affecting clinical outcomes is a challenging task. We can see it in the number of emerging papers on the DRP problem (140). The research community put in monumental efforts to develop methods that leverage multiomics and drug data, explore the applicability of the existing methods, and bridge the gap between results on ubiquitously used cell lines and patient-derived models. However, this area is facing multiple challenges related to the lack of unified evaluation framework, generalizability, interpretability as well as challenges relating to computational representations of omic data, biological response, and drug representations.

8.2.1. Unified framework for model evaluation and comparison

Currently, there is no standard or accepted framework for evaluation or comparison of cancer drug response models. Model development and performance comparison against baselines is frequently accomplished with different compositions of datasets, inconsistent training and testing splits, and diverse scoring metrics, using varying protocols for hyperparameter optimization or none at all. Yet, the majority of papers report to outperform current state-of-the-art (SOTA) in the task of drug response prediction. In ML, SOTA refers to the best model for a specific task as evaluated by concrete performance criteria on a benchmark dataset of predetermined test set samples. Thousands of benchmark datasets and prediction tasks are publicly available for different applications ranging from vision and language to drug discovery and tumor segmentation. Whereas, most papers follow the same general model development workflow (Figure 2), benchmark datasets and agreed-upon evaluation criteria have yet been established and adopted in the community. It makes identifying the most promising research areas challenging and impedes the research community from making directed efforts to breach them. Creating an ImageNet moment for DRP *via* established benchmark datasets, consistent test sets, robust evaluation criteria, and a platform for publishing and monitoring SOTA models should be on the critical path for our community.

8.2.2. Generalizability

Despite the breadth of methods, the majority of papers have focused on improving predictions in cell lines, often demonstrating only a marginal improvement in prediction generalization. While cell lines remain a primary biological media to study cancer and conduct drug screenings, the potential utility of prediction models for improving patient care is not immediately evident, and several questions naturally arise. How well models trained with cell lines would generalize to xenografts or patients? How much one could rely on any given model prediction in a decision-making process? What are the potential clinical application for these models? Addressing questions important from the user perspective and suggesting use case scenarios have not been the driving mechanism behind the majority of published models. Yet, certain trends aiming to respond to these challenges emerge: transfer learning that utilize abundant cell line datasets to improve predictions in PDX and patients, uncertainty quantification allowing to estimate the confidence for each model prediction (51), and ranking learning models principally suitable for personalized treatment recommendation (72, 74, 141). Demonstrating the utility of DRP models, integrated

in a larger patient care workflow, could provide the needed user-centric view and navigate the community toward developing application-aware models.

8.2.3. Interpretability

Despite the demonstrated success of the DL on expansive datasets, delegating the decision-making process on patient well-being to the black box is widely contested (142). Thus, providing not only precise but also salient results is a task of paramount importance for the clinical community. To address this challenge, model-agnostic methods such as Integrated Gradients (143) and SHAP (144), designed to explain model predictions, have been applied post-training to DRP models (42, 45, 145), where the explanation is provided in the form of most important features attributing to model predictions. However, it has been shown that these methods can lead to highly misleading information in applications more comprehensible to humans such as image classification, and therefore, attempting to explain black boxes with *post-hoc* methods is perceived as dangerous in high-stakes decision-making domains (146). Instead of explaining predictions, the alternative is to design interpretable models that are understandable by domain experts or provide insights into the decision-making process (147, 148). Unfortunately, no single definition exists, and identifying a model as interpretable is considered domain-specific (146, 147). Several papers referring to their DRP models as interpretable integrate domain structural knowledge into the model form (145, 149–151). Currently, no clear definition of an interpretable DRP model yet exists, nor the extent to which it might improve cancer treatment is known. Therefore, it is challenging to evaluate whether existing efforts in this direction are in line with the views of the clinical community at large on this important matter. Addressing the interpretability issue will perhaps require better framing of what interpretable DRP models are and how much performance drop could be tolerated, if any, for improved interpretability. As discussed in Section 7.3, baseline models that can assess the quality and robustness of model interpretability can serve to bolster the claim regarding the benefits and usability of interpretable models (135). Collaboration with cancer biologists and clinical oncologists would be essential to advance this direction. And, as discussed in Section 7.3, baseline models which compare the quality and robustness of interpretability can serve to bolster claims regarding the improvement or introduction of model interpretability.

8.2.4. Variability and high dimensionality associated with omics data

Omics values differ substantially depending on the underlying biological model (e.g., cell line, organoid, xenograft), experimental protocols and selected platforms (e.g., Illumina, Nanopore), technical variations (batch effect), and computational pipeline for processing raw data. Identifying the potential sources of variation and taking directed measures to mitigate the undesired differences is critical, especially in cross-domain generalization scenarios where the similarity between the source and target domains is a fundamental assumption (50–52, 67). A major conclusion of the NCI-DREAM challenge was that gene expression modality provides most of the predictive power for cell line drug sensitivity prediction with additional improvement when combined with other omics types (27). This collaborative effort contributed to the adoption of gene expression (GE; e.g., microarrays or RNA-Seq) or combining them with other omics in DRP models. Integrating multiomics into the learning process further exacerbates the already problematic feature size of single omics data. A few papers indeed demonstrate significant performance boost with multiomics (45, 49, 108) but the majority report only marginal improvement (104, 108, 118, 126, 141, 152). As opposed to DREAM participators which utilized agreed-upon datasets and scoring metrics, the DRP models in Supplementary Table 1 substantially differ among them as discussed earlier, largely contributing to discrepancies and mixed conclusions. Analysis across multiple models and omics types is required to evaluate the predictive capabilities of individual data types and their subsets and make unbiased and coherent conclusions. Such analysis should incorporate recent trends which encode biological information such as protein-protein interactions (PPI), gene correlations, and pathway information (45, 49, 109, 111).

8.2.5. Representation of biological treatment response

In vitro response data (e.g., cell line and organoid) is usually derived from a series of inherently noisy cell viability experiments and subsequent curve fitting. However, these curves do not always have a good fit, and usage of a single point on the fitted curve may result in a substantial information loss. While the use of IC50 as a prediction variable has been the most prevalent in regression models (Figure 4), recent assessment studies could shift this trend toward the use of global and generally more robust measures such as AUC and AAC (14, 41). With the objective of utilizing prediction models in decision-making situations, another option is converting continuous into discrete values. This is particularly common in cross-domain generalization from *in vitro* to *in vivo*, where *in vivo* response is generally encoded as discrete values. The third option is generating a rank list of items. This has been applied to a personalized treatment recommendation given cancer information. Surprisingly, despite the success of recommendation algorithms and the direct relevance to precision oncology, only few methods have been explored (72, 74, 141).

8.2.6. Vastness of chemical space and its representations

Reviewed papers consistently report the response prediction accuracy drop for compounds, previously unseen by the model. Inability to reliably overcome this limitation severely limits the practical value of DRP models to virtual drug screening or drug design applications. Representing drug molecules with graphs and using GNN for learning response prediction has recently inspired many DRP architectures. The use of graph structures is motivated by the proposition that molecular graphs better capture intrinsic chemical properties of molecules (49, 108, 153). Whereas, this is presumably expected to produce a better prediction of drug response, we have not found a comprehensive study that could assert this hypothesis. In fact, in a related field of molecular property prediction, a comparison with multiple datasets and prediction tasks suggests that on average models that use FPs or descriptors outperform graph-based models (154). We have found only one study comparing side-by-side the added value of molecular graphs

against SMILES, reporting <0.5% improvement as evaluated by Pearson correlation coefficient (155). As in the case of cancer representation, further studies are required across various models and datasets to assess the predictive capabilities of molecular graphs and other drug representations to DRP. Several underexplored but interesting directions include SMILES combined with transformer-based models, kinase inhibition profiles representing kinase inhibitor therapies (44), and 2D-CNNs that learn from FPs or descriptors transformed into images (77).

This list of challenges is in no way conclusive, as we refrained from studying even more complicated scenarios emerging in polypharmacy, even though the treatment of patients in the clinic often involves combination therapy. Meanwhile, most of the prediction models to date focus on single-drug treatments. This highlights a significant disconnect between the modeling community and the current patient standard of care. However, even this simplified problem raises many obstacles, and making progress in overcoming them would significantly benefit the scientific community because they are relevant to a much wider domain of AI-driven biological research. Each of the highlighted problems has multiple potential solutions, and we hope to see subsequent progress in the near future, allowing AI to deliver on its grand promise.

Author contributions

Writing the original draft: AP. Collection of prediction models and other references: AP, TB, YZ, and AC. Review and editing: AP, TB, YZ, ON, AC, JO, and RS. All authors have read and agreed to the published version of the manuscript.

Funding

This study has been funded in whole or in part with Federal funding by the NCI-DOE Collaboration established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health, Cancer Moonshot Task Order No. 75N91019F00134 and under Frederick National Laboratory for Cancer Research Contract 75N91019D00024. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract 89233218CNA000001, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1086097/full#supplementary-material

References

1. Yancovitz M, Litterman A, Yoon J, Ng E, Shapiro RL, Berman RS, et al. Intra- and inter-tumor heterogeneity of BRAFV600EMutations in primary and metastatic melanoma. *PLoS ONE*. (2012) 7:e0029336. doi: 10.1371/journal.pone.0029336

2. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*. (2013) 108:479–85. doi: 10.1038/bjc.2012.581

3. Collins DC, Sundar R, Lim JSJ, Yap TA. Towards precision medicine in the clinic: from biomarker discovery to novel therapeutics. *Trends Pharmacol Sci*. (2017) 38:25–40. doi: 10.1016/j.tips.2016.10.012

4. Hartl D, de Luca V, Kostikova A, Laramie J, Kennedy S, Ferrero E, et al. Translational precision medicine: an industry perspective. *J Transl Med*. (2021) 19:245. doi: 10.1186/s12967-021-02910-6

5. Prasad V, Kaestner V, Mailankody S. Cancer drugs approved based on biomarkers and not tumor Type-FDA approval of pembrolizumab for mismatch repair-deficient solid cancers. *JAMA Oncol*. (2018) 4:157–8. doi: 10.1001/jamaoncol.2017.4182

6. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol*. (2020) 4:19. doi: 10.1038/s41698-020-0122-1

7. Ballester PJ, Stevens R, Haibe-Kains B, Huang RS, Aittokallio T. Artificial intelligence for drug response prediction in disease models. *Brief Bioinform*. (2021) 23:bbab450. doi: 10.1093/bib/bbab450

8. De Niz C, Rahman R, Zhao X, Pal R. Algorithms for drug sensitivity prediction. *Algorithms*. (2016) 9:77. doi: 10.3390/a9040077

9. Piyawajanusorn C, Nguyen LC, Ghislat G, Ballester PJ. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. *Brief Bioinform*. (2021) 22:bbab312. doi: 10.1093/bib/bbab312

10. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform*. (2021) 22:360–79. doi: 10.1093/bib/bbz171

11. Kumar V, Dogra N. A comprehensive review on deep synergistic drug prediction techniques for cancer. *Arch Comput Methods Eng*. (2022) 29:1443–61. doi: 10.1007/s11831-021-09617-3

12. Azuaje F. Computational models for predicting drug responses in cancer research. *Brief Bioinform*. (2017) 18:820–9. doi: 10.1093/bib/bbw065

13. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev*. (2019) 11:31–9. doi: 10.1007/s12551-018-0446-z

14. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, Hon C, Mammoliti A, Nair SK, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief Bioinform*. (2021) 22:bbab294. doi: 10.1093/bib/bbab294

15. Güvenç Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Brief Bioinform*. (2021) 22:346–59. doi: 10.1093/bib/bbz153

16. Rafique R, Islam SMR, Kazi JU. Machine learning in the prediction of cancer therapy. *Comput Struct Biotechnol J*. (2021) 19:4003–17. doi: 10.1016/j.csbj.2021.07.003

17. Zadorozhny K, Nuzhna L. Deep generative models for drug design and response. *arXiv:210906469 [cs*] (2021). doi: 10.48550/arXiv.2109.06469

18. Firoozbakht F, Yousefi B, Schwikowski B. An overview of machine learning methods for monotherapy drug response prediction. *Brief Bioinform*. (2022) 23:bbab408. doi: 10.1093/bib/bbab408

19. Chiu YC, Chen HIH, Gorthi A, Mostavi M, Zheng S, Huang Y, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. *Brief Bioinform*. (2020) 21:2066–83. doi: 10.1093/bib/bbz144

20. Tanoli Z, Vähä-Koskela M, Aittokallio T. Artificial intelligence, machine learning, and drug repurposing in cancer. *Expert Opin Drug Disc*. (2021) 16:977–89. doi: 10.1080/17460441.2021.1883585

21. An X, Chen X, Yi D, Li H, Guan Y. Representation of molecules for drug response prediction. *Brief Bioinform*. (2021) 23:bbab393. doi: 10.1093/bib/bbab393

22. Wu L, Wen Y, Leng D, Zhang Q, Dai C, Wang Z, et al. Machine learning methods, databases and tools for drug combination prediction. *Brief Bioinform*. (2022) 23:bbab355. doi: 10.1093/bib/bbab355

23. Wu Z, Lawrence PJ, Ma A, Zhu J, Xu D, Ma Q. Single-cell techniques and deep learning in predicting drug response. *Trends Pharmacol Sci*. (2020) 41:1050–65. doi: 10.1016/j.tips.2020.10.004

24. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform*. (2021) 22:232–46. doi: 10.1093/bib/bbz164

25. Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for drug sensitivity prediction. *Sci Rep*. (2020) 10:9377. doi: 10.1038/s41598-020-65927-9

26. Chen Y, Zhang L. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? *Brief Bioinform*. (2021) 23:bbab378. doi: 10.1093/bib/bbab378

27. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. (2014) 32:1202–12. doi: 10.1038/nbt.2877

28. Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*. (2019) 10:2674. doi: 10.1038/s41467-019-09799-2

29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. (2015) 521:436–44. doi: 10.1038/nature14539

30. Stevens R, Taylor V, Nichols J, Maccabe AB, Yelick K, Brown D. AI for science: report on the department of energy (DOE) town halls on artificial intelligence (AI) for science. In: *Argonne National Lab*. (*ANL*), Argonne, IL (2020).

31. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. *ArXiv:1906.05433 [cs, stat*] (2019). doi: 10.48550/arXiv.1906.05433

32. Cohen R, Zhang Y, Solomon O, Toberman D, Taieb L, van Sloun RJ, et al. Deep convolutional robust PCA with application to ultrasound imaging. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE (2019). p. 3212–6.

33. Utyamishev D, Partin-Vaisband I. Multiterminal pathfinding in practical VLSI systems with deep neural networks. *ACM Trans Design Autom Electron Syst*. (2022) 2022:3564930. doi: 10.1145/3564930

34. Menden MP, Iorio F, Garnett M, Mcdermott U, Benes CH. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE*. (2013) 8:61318. doi: 10.1371/journal.pone.0061318

35. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. (2013) 41:955–61. doi: 10.1093/nar/gks1111

36. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous systems. *arXiv:1603.04467 [cs.DC*] (2015). doi: 10.48550/arXiv.1603.04467

37. Chollet F, others. *Keras*. (2015). Available online at: https://keras.io

38. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Vancouver, BC: Curran Associates, Inc. (2019). p. 8024–35.

39. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*. (2019) 35:501–9. doi: 10.1093/bioinformatics/btz318

40. Partin A, Brettin T, Evrard YA, Zhu Y, Yoo H, Xia F, et al. Learning curves for drug response prediction in cancer cell lines. *BMC Bioinform*. (2021) 22:252. doi: 10.1186/s12859-021-04163-y

41. Xia F, Allen J, Balaprakash P, Brettin T, Garcia-Cardona C, Clyde A, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Brief Bioinform*. (2022) 23:bbab356. doi: 10.1093/bib/bbab356

42. Prasse P, Iversen P, Lienhard M, Thedinga K, Herwig R, Scheffer T. Pre-Training on *in vitro* and fine-tuning on patient-derived data improves deep neural networks for anti-cancer drug-sensitivity prediction. *Cancers*. (2022) 14:3950. doi: 10.3390/cancers14163950

43. Mazumder M, Banbury C, Yao X, Karlaš B, Rojas WG, Diamos S, et al. DataPerf: benchmarks for data-centric AI development. *arXiv:2207.10062 [cs.LG*] (2022). doi: 10.48550/arXiv.2207.10062

44. Koras K, Kizling E, Juraeva D, Staub E, Szczurek E. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *Sci Rep*. (2021) 11:15993. doi: 10.1038/s41598-021-94564-z

45. Tang YC, Gottlieb A. Explainable drug sensitivity prediction through cancer pathway enrichment. *Sci Rep*. (2021) 11:3128. doi: 10.1038/s41598-021-82612-7

46. Partin A, Brettin T, Zhu Y, Dolezal JM, Kochanny S, Pearson AT, et al. Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. *ArXiv:2204.11678 [q-bio*] (2022). doi: 10.48550/arXiv.2204.11678

47. Hostallero DE, Wei L, Wang L, Cairns J, Emad A. A deep learning framework for prediction of clinical drug response of cancer patients and identification of drug sensitivity biomarkers using preclinical samples. *biorxiv*. (2021). doi: 10.1101/2021.07.06.451273

48. Jia P, Hu R, Pei G, Dai Y, Wang YY, Zhao Z. Deep generative neural network for accurate drug response imputation. *Nat Commun*. (2021) 12:1740. doi: 10.1038/s41467-021-21997-5

49. Ma T, Liu Q, Li H, Zhou M, Jiang R, Zhang X. DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinform*. (2022) 23:129. doi: 10.1186/s12859-022-04664-4

50. Sharifi-Noghabi H, Peng S, Zolotareva O, Collins CC, Ester M. AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics*. (2020) 36:i380–8. doi: 10.1093/bioinformatics/btaa442

51. Peres da Silva R, Suphavilai C, Nagarajan N. TUGDA: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from *in vitro* to *in vivo* settings. *Bioinformatics*. (2021) 37:i76-i83. doi: 10.1093/bioinformatics/btab299

52. Sharifi-Noghabi H, Harjandi PA, Zolotareva O, Collins CC, Ester M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nat Mach Intell*. (2021) 3:962–72. doi: 10.1038/s42256-021-00408-w

53. Mourragui S, Loog M, van de Wiel MA, Reinders MJT, Wessels LFA. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics*. (2019) 35:i510–9. doi: 10.1093/bioinformatics/btz372

54. Mourragui SMC, Loog M, Vis DJ, Moore K, Manjon AG, van de Wiel MA, et al. Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proc Natl Acad Sci USA*. (2021) 118:2118. doi: 10.1073/pnas.2106682118

55. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. (2012) 483:603–7. doi: 10.1038/nature11003

56. Iorio F, Knijnenburg TA, Vis DJ, Saez-Rodriguez J, Mcdermott U, Garnett MJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017

57. Seashore-Ludlow B, Rees MG, Cheah JH, Coko M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov*. (2015) 5:1210–23. doi: 10.1158/2159-8290.CD-15-0235

58. Larsen BM, Kannan M, Langer LF, Leibowitz BD, Bentaieb A, Cancino A, et al. A pan-cancer organoid platform for precision medicine. *Cell Rep*. (2021) 36:109429. doi: 10.1016/j.celrep.2021.109429

59. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med*. (2015) 21:1318–25. doi: 10.1038/nm.3954

60. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput*. (2013) 2014: 63–74. doi: 10.1142/9789814583220_0007

61. Yadav B, Pemovska T, Szwajda A, Kulesskiy E, Kontro M, Karjalainen R, et al. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci Rep*. (2014) 4:5193. doi: 10.1038/srep05193

62. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, et al. Inconsistency in large pharmacogenomic studies. *Nature*. (2013) 504:389–93. doi: 10.1038/nature12831

63. Knijnenburg TA, Klau GW, Iorio F, Garnett MJ, McDermott U, Shmulevich I, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci Rep*. (2016) 6:36812. doi: 10.1038/srep36812

64. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA*. (2001) 98:10787–92. doi: 10.1073/pnas.191368598

65. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*. (2015) 15:489. doi: 10.1186/s12885-015-1492-6

66. Joo M, Park A, Kim K, Son WJ, Lee HS, Lim G, et al. A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. *Int J Mol Sci*. (2019) 20:6276. doi: 10.3390/ijms20246276

67. Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep*. (2019) 29:3367–73.e4. doi: 10.1016/j.celrep.2019.11.017

68. Daoud S, Mdhaffar A, Jmaiel M, Freisleben B. Q-Rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy. *IEEE J Biomed Health Inform*. (2020) 24:3154–61. doi: 10.1109/JBHI.2020.3004663

69. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics*. (2021) 22:214. doi: 10.1186/s12864-021-07524-2

70. Jiang L, Jiang C, Yu X, Fu R, Jin S, Liu X. DeepTTA: a transformer-based model for predicting cancer drug response. *Brief Bioinform*. (2022) 23:bbac100. doi: 10.1093/bib/bbac100

71. Zhao Z, Li K, Toumazou C, Kalofonou M. A computational model for anti-cancer drug sensitivity prediction. In: *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Nara: IEEE (2019). p. 1–4.

72. Prasse P, Iversen P, Lienhard M, Thedinga K, Bauer C, Herwig R, et al. Matching anticancer compounds and tumor cell lines by neural networks with ranking loss. *NAR Genomics Bioinform*. (2022) 4:lqab128. doi: 10.1093/nargab/lqab128

73. Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Rodríguez Martínez M. Toward explainable anticancer compound sensitivity prediction *via* multimodal attention-based convolutional encoders. *Mol Pharm*. (2019) 16:4797–806. doi: 10.1021/acs.molpharmaceut.9b00520

74. Su R, Huang Y, Zhang Dg, Xiao G, Wei L. SRDFM: siamese response deep factorization machine to improve anti-cancer drug recommendation. *Brief Bioinform*. (2022) 23:bbab534. doi: 10.1093/bib/bbab534

75. Schwartz LH, Litière S, Vries Ed, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1–Update and clarification: From the RECIST committee. *Eur J Cancer*. (2016) 62:132–7. doi: 10.1016/j.ejca.2016.03.081

76. Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat Commun*. (2020) 11:4391. doi: 10.1038/s41467-020-18197-y

77. Zhu Y, Brettin T, Xia F, Partin A, Shukla M, Yoo H, et al. Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep*. (2021) 11:11325. doi: 10.1038/s41598-021-90923-y

78. Zhu Y, Brettin T, Evrard YA, Partin A, Xia F, Shukla M, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep*. (2020) 10:1–11. doi: 10.1038/s41598-020-74921-0

79. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J Cheminform*. (2020) 12:56. doi: 10.1186/s13321-020-00460-5

80. Oskooei A, Born J, Manica M, Subramanian V, Sáez-Rodríguez J, Rodríguez Martínez M. PaccMann: prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. In: *NIPS*. Montreal, QC (2018).

81. Liu H, Zhao Y, Zhang L, Chen X. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol Therapy Nucleic Acids*. (2018) 13:303–11. doi: 10.1016/j.omtn.2018.09.011

82. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv:1703.07076 [cs.LG]* (2017). doi: 10.48550/arXiv.1703.07076

83. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, et al. Randomized SMILES strings improve the quality of molecular generative models. *J Cheminform*. (2019) 11:71. doi: 10.1186/s13321-019-0393-0

84. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. (2010) 32:1466–74. doi: 10.1002/jcc.21707

85. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform*. (2018) 10:4. doi: 10.1186/s13321-018-0258-y

86. Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: an easy approach to molecular descriptor calculations. *MATCH Commun Math Comput Chem*. (2006) 56:237–48.

87. Clyde A, Brettin T, Partin A, Shaulik M, Yoo H, Evrard Y, et al. A systematic approach to featurization for cancer drug sensitivity predictions with deep learning. *arXiv:200500095 [cs, q-bio]*. (2020). doi: 10.48550/arXiv.2005.00095

88. Kipf TN, Welling M. Semi-Supervised classification with graph convolutional networks. In: *International Conference on Learning Representations*. Toulon (2017).

89. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: *International Conference on Learning Representations*. Vancouver, BC (2018).

90. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *arXiv:170602216 [cs, stat]*. (2018). doi: 10.48550/arXiv.1706.02216

91. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. (2018) 34:i457–66. doi: 10.1093/bioinformatics/bty294

92. Utyamishev D, Partin-Vaisband I. Knowledge graph embedding and visualization for pre-silicon detection of hardware Trojans. In: *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. Austin, TX: IEEE (2022). p. 1–5.

93. Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F. Graph convolutional networks for computational drug development and discovery. *Brief Bioinform*. (2020) 21:919–35. doi: 10.1093/bib/bbz042

94. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. *Brief Bioinform*. (2021) 23:bbab159. doi: 10.1093/bib/bbab159

95. Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. *Deep Learning for the Life Sciences*. O'Reilly Media (2019). Available online at: https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837

96. Grattarola D, Alippi C. Graph neural networks in tensorflow and Keras with Spektral. *arXiv:2006.12138 [cs.LG]*. (2020). doi: 10.1109/MCI.2020.3039072

97. Li M, Zhou J, Hu J, Fan W, Zhang Y, Gu Y, et al. DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega*. (2021) 6:27233–8. doi: 10.1021/acsomega.1c04017

98. Jiang Y, Rensi S, Wang S, Altman RB. DrugOrchestra: Jointly predicting drug response, targets, and side effects *via* deep multi-task learning. *biorxiv*. (2020). doi: 10.1101/2020.11.17.385757

99. Choi J, Park S, Ahn J. RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci Rep*. (2020) 10:1861. doi: 10.1038/s41598-020-58821-x

100. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: a survey. *Mech Syst Signal Process*. (2021) 151:107398. doi: 10.1016/j.ymssp.2020.107398

101. Liu P, Li H, Li S, Leung KS. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinform*. (2019) 20:408. doi: 10.1186/s12859-019-2910-6

102. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep*. (2018) 8:1–11. doi: 10.1038/s41598-018-27214-6

103. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*. (2020) 36:i911–8. doi: 10.1093/bioinformatics/btaa822

104. Chu T, Nguyen TT, Hai BD, Nguyen QH, Nguyen T. Graph transformer for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform*. (2022) 1–9. doi: 10.1109/TCBB.2022.3206888

105. Nguyen GTT, Vu HD, Le DH. Integrating molecular graph data of drugs and multiple -Omic data of cell lines for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform*. (2022) 19:710–7. doi: 10.1109/TCBB.2021.3096960

106. Zuo Z, Wang P, Chen X, Tian L, Ge H, Qian D. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC Bioinform*. (2021) 22:434. doi: 10.1186/s12859-021-04352-9

107. Tao Y, Ren S, Ding MQ, Schwartz R, Lu X. Predicting drug sensitivity of cancer cell lines *via* collaborative filtering with contextual attention. In: *Proceedings of the 5th Machine Learning for Healthcare Conference*. PMLR (2020). p. 660–84.

108. Liu X, Song C, Huang F, Fu H, Xiao W, Zhang W. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief Bioinform*. (2021) 23:bbab457. doi: 10.1093/bib/bbab457

109. Ruiwei F, Yufeng X, Minshan L, Danny C, Ji C, Jian W. AGMI: attention-guided multi-omics integration for drug response prediction with graph neural networks. *arXiv:211208366 [cs, q-bio]*. (2021). doi: 10.48550/arXiv.2112.08366

110. Jin I, Nam H. HiDRA: hierarchical network for drug response prediction with attention. *J Chem Inf Model*. (2021) 61:3858–67. doi: 10.1021/acs.jcim.1c00706

111. Pu L, Singha M, Ramanujam J, Brylinski M. CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling. *Oncotarget*. (2022) 13:695–706. doi: 10.18632/oncotarget.28234

112. Huang K, Xiao C. Explainable substructure partition fingerprint for protein, drug, and more. In: *NeurIPS*. Vancouver, BC (2019).

113. Yi HC, You ZH, Huang DS, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. *Brief Bioinform*. (2022) 23:bbab340. doi: 10.1093/bib/bbab340

114. Nguyen TT, Nguyen GTT, Nguyen T, Le DH. Graph convolutional networks for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform*. (2021) 19:146–54. doi: 10.1101/2020.04.07.030908

115. Ahmed KT, Park S, Jiang Q, Yeu Y, Hwang T, Zhang W. Network-based drug sensitivity prediction. *BMC Med Genomics*. (2020) 13:193. doi: 10.1186/s12920-020-00829-3

116. Peng W, Chen T, Dai W. Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J Biomed Health Inform*. (2022) 26:1384–93. doi: 10.1109/JBHI.2021.3102186

117. Kim S, Bae S, Piao Y, Jo K. Graph convolutional network for drug response prediction using gene expression Data. *Mathematics*. (2021) 9:772. doi: 10.3390/math9070772

118. Zhu Y, Ouyang Z, Chen W, Feng R, Chen DZ, Cao J, et al. TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics*. (2021) 38:461–8. doi: 10.1093/bioinformatics/btab650

119. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. (2021) 49:D605–12. doi: 10.1093/nar/gkaa1074

120. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102

121. Hostallero DE, Li Y, Emad A. Looking at the BiG picture: incorporating bipartite graphs in drug response prediction. *Bioinformatics*. (2022) 38:3609–20. doi: 10.1093/bioinformatics/btac383

122. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, et al. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans Comput Biol Bioinform*. (2021) 18:575–82. doi: 10.1109/TCBB.2019.2919581

123. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: improving drug response prediction *via* modeling of drug perturbation effects. *Bioinformatics*. (2019) 35:3743–51. doi: 10.1093/bioinformatics/btz158

124. Xu X, Gu H, Wang Y, Wang J, Qin P. Autoencoder based feature selection method for classification of anticancer drug response. *Front Genet*. (2019) 10:233. doi: 10.3389/fgene.2019.00233

125. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol Cancer Res*. (2018) 16:269–78. doi: 10.1158/1541-7786.MCR-17-0378

126. Chiu YC, Chen HIH, Zhang T, Zhang S, Gorthi A, Wang LJ, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics*. (2019) 12:18. doi: 10.1186/s12920-018-0460-9

127. Dong H, Xie J, Jing Z, Ren D. Variational autoencoder for anti-cancer drug response prediction. *arXiv:2008.09763 [cs.LG]* (2021). doi: 10.48550/arXiv.2008.09763

128. Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR (2018). p. 2323–32.

129. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. (2010) 22:1345–59. doi: 10.1109/TKDE.2009.191

130. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. *Proc IEEE*. (2021) 109:43–76. doi: 10.1109/JPROC.2020.3004555

131. Ma J, Fong SH, Luo Y, Bakkenist CJ, Shen JP, Mourragui S, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer*. (2021) 2:233–44. doi: 10.1038/s43018-020-00169-2

132. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks. In: *International Conference on Learning Representations*. Addis Ababa (2020).

133. Zhu Y, Brettin T, Evrard YA, Xia F, Partin A, Shukla M, et al. Enhanced co-expression extrapolation (coxen) gene selection method for building anti-cancer drug response prediction models. *Genes*. (2020) 11:1070. doi: 10.3390/genes11091070

134. Gupta A, Gautam P, Wennerberg K, Aittokallio T. A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening. *Commun Biol*. (2020) 3:1–12. doi: 10.1038/s42003-020-0765-z

135. Li Y, Hostallero DE, Emad A. Interpretable deep learning architectures for improving drug response prediction: myth or reality? *biorxiv*. (2022). doi: 10.1101/2022.10.03.510614

136. Dincer AB, Janizek JD, Lee SI. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*. (2020) 36:i573–82. doi: 10.1093/bioinformatics/btaa796

137. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*. (2017) 33:2539–46. doi: 10.1093/bioinformatics/btx196

138. Xia F, Shukla M, Brettin T, Garcia-Cardona C, Cohn J, Allen JE, et al. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinform*. (2018) 19:486. doi: 10.1186/s12859-018-2509-3

139. Bohr A, Memarzadeh K. Chapter 2 - the rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare*. Academic Press (2020). p. 25–60. Available online at: https://www.sciencedirect.com/science/article/pii/B9780128184387000022

140. Clyde A. AI for science and global citizens. *Patterns*. (2022) 3:100446. doi: 10.1016/j.patter.2022.100446

141. Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. *Stat Med*. (2022) 41:4034–56. doi: 10.1002/sim.9491

142. Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg*. (2022) 9:862322. doi: 10.3389/fsurg.2022.862322

143. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34 th International Conference on Machine Learning*. Sydney, NSW (2017).

144. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Neural Information Processing Systems*. Long Beach, CA (2017).

145. Snow O, Sharifi-Noghabi H, Lu J, Zolotareva O, Lee M, Ester M. Interpretable drug response prediction using a knowledge-based neural network. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21*. New York, NY: Association for Computing Machinery (2021). p. 3558–68.

146. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x

147. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci USA*. (2019) 116:22071–80. doi: 10.1073/pnas.1900654116

148. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

149. Deng L, Cai Y, Zhang W, Yang W, Gao B, Liu H. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. *J Chem Inf Model*. (2020) 60:4497–505. doi: 10.1021/acs.jcim.0c00331

150. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*. (2020) 38:672–84. doi: 10.1016/j.ccell.2020.09.014

151. Zhang H, Chen Y, Li F. Predicting anticancer drug response with deep learning constrained by signaling pathways. *Front Bioinform*. (2021) 1:639349. doi: 10.3389/fbinf.2021.639349

152. Park S, Soh J, Lee H. Super.FELT: supervised feature extraction learning using triplet loss for drug response prediction with multi-omics data. *BMC Bioinform*. (2021) 22:269. doi: 10.1186/s12859-021-04146-z

153. Hosseini R, Simini F, Clyde A, Ramanathan A. Deep surrogate docking: accelerating automated drug discovery with graph neural networks. In: *Workshop on AI for Science: Progress and Promises*. New Orleans, LA (2022).

154. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Cheminform*. (2021) 13:12. doi: 10.1186/s13321-020-00479-8

155. Yan X, Liu Y, Zhang W. Deep graph and sequence representation learning for drug response prediction. In: *Artificial Neural Networks and Machine Learning-ICANN*. Bristol: Springer (2022). p. 97–108.

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/fimmu.2023.1036562

Check for updates

OPEN ACCESS

EDITED BY
George Bebis,
University of Nevada, United States

REVIEWED BY
Jingtao Qiu,
Stanford University, United States
Li Zhang,
University of California, San Francisco, United States

*CORRESPONDENCE
Hui Jiang
hjiang03@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cancer Immunity and Immunotherapy, a section of the journal Frontiers in Immunology

RECEIVED 04 September 2022
ACCEPTED 15 February 2023
PUBLISHED 01 March 2023

CITATION
Mou L, Pu Z, Luo Y, Quan R, So Y and Jiang H (2023) Construction of a lipid metabolism-related risk model for hepatocellular carcinoma by single cell and machine learning analysis.
Front. Immunol. 14:1036562.
doi: 10.3389/fimmu.2023.1036562

COPYRIGHT
© 2023 Mou, Pu, Luo, Quan, So and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Construction of a lipid metabolism-related risk model for hepatocellular carcinoma by single cell and machine learning analysis

Lisha Mou[1,2†], Zuhui Pu[1†], Yongxiang Luo[3], Ryan Quan[2], Yunhu So[2] and Hui Jiang[3*]

[1]Imaging Department, Shenzhen Institute of Translational Medicine, The First Affiliated Hospital of Shenzhen University, Shenzhen Second People's Hospital, Shenzhen, China, [2]MetaLife Center, Shenzhen Institute of Translational Medicine, The First Affiliated Hospital of Shenzhen University, Shenzhen Second People's Hospital, Shenzhen, China, [3]Department of General Surgery, The First People's Hospital of Qinzhou/The Tenth Affiliated Hospital of Guangxi Medical University, Qinzhou, Guangxi, China

One of the most common cancers is hepatocellular carcinoma (HCC). Numerous studies have shown the relationship between abnormal lipid metabolism-related genes (LMRGs) and malignancies. In most studies, the single LMRG was studied and has limited clinical application value. This study aims to develop a novel LMRG prognostic model for HCC patients and to study its utility for predictive, preventive, and personalized medicine. We used the single-cell RNA sequencing (scRNA-seq) dataset and TCGA dataset of HCC samples and discovered differentially expressed LMRGs between primary and metastatic HCC patients. By using the least absolute selection and shrinkage operator (LASSO) regression machine learning algorithm, we constructed a risk prognosis model with six LMRGs (*AKR1C1*, *CYP27A1*, *CYP2C9*, *GLB1*, *HMGCS2*, and *PLPP1*). The risk prognosis model was further validated in an external cohort of ICGC. We also constructed a nomogram that could accurately predict overall survival in HCC patients based on cancer status and LMRGs. Further investigation of the association between the LMRG model and somatic tumor mutational burden (TMB), tumor immune infiltration, and biological function was performed. We found that the most frequent somatic mutations in the LMRG high-risk group were *CTNNB1*, *TTN*, *TP53*, *ALB*, *MUC16*, and *PCLO*. Moreover, naïve CD8+ T cells, common myeloid progenitors, endothelial cells, granulocyte-monocyte progenitors, hematopoietic stem cells, M2 macrophages, and plasmacytoid dendritic cells were significantly correlated with the LMRG high-risk group. Finally, gene set enrichment analysis showed that RNA degradation, spliceosome, and lysosome pathways were associated with the LMRG high-risk group. For the first time, we used scRNA-seq and bulk RNA-seq to construct an LMRG-related risk score model, which may provide insights into more effective treatment strategies for predictive, preventive, and personalized medicine of HCC patients.

KEYWORDS

hepatocellular carcinoma, scRNA-seq, lipid metabolism, prediction model, immune microenvironment, TMB, GSEA, Machine learning

1 Introduction

According to recent epidemiological data, 906,000 new cases of liver cancer were diagnosed globally in 2020, making it the sixth most prevalent cancer worldwide (1). Hepatocellular carcinoma (HCC) is a common type of liver cancer (2). The disease is caused by a number of risk factors, including HBV/HCV infection, nonalcoholic steatohepatitis (NASH), alcoholism, and smoking. HCC treatment, including surgery, chemotherapy, and radiation therapy, has significantly enhanced survival and reduced cancer cell proliferation in patients with the disease (3). Early HCC is treatable with tumor resection and liver transplantation; however, many patients are not diagnosed until the late stages (4). As HCC is highly heterogeneous, predictive, preventive, and personalized medicine can improve therapy outcomes. Therefore, it is essential to uncover the mechanisms that drive the progression of HCC, and effective biomarkers must be identified as soon as possible to provide individualized treatment for HCC patients.

Recent studies have revealed that alterations in lipid metabolism are significant metabolic indicators of cancer cells in general (5). Changes in lipid metabolism, for instance, can occur in tumor cells and the tumor microenvironment, which influences the development, proliferation, invasion, and metastasis of cancer cells (6). A previous study showed that TAR DNA-Binding Protein 43 (TDP-43) can suppress apoptosis by facilitating lipid metabolism in HCC (7). In HCC, ovarian cancer, lung adenocarcinoma, pancreatic cancer, renal cell carcinoma, and diffuse glioma, lipid metabolism-related genes (LMRGs) show excellent predictive values (8). Consequently, targeting lipid metabolism has been viewed as a potential way of treating tumors. To date, several prognostic models have examined the value of genes associated with ferroptosis, epithelial-mesenchymal transition, and immunity in HCC, whereas little is known about how LMRGs contribute to HCC and whether LMRGs are correlated with HCC patient prognosis (9–11).

Multiple gene signatures for predicting the prognosis of HCC patients have been created in prior research based on bulk RNA sequencing; however, these signatures have not been used in clinical settings. RNA signals from several cells within a sample are combined during bulk RNA sequencing to reflect the sample's average RNA content. As a result, cell type predominance has a large impact. However, there are certain genetic traits linked to HCC that may differently favor its development. Therefore, uncommon or diverse cell populations cannot be studied by bulk RNA sequencing. Single-cell RNA sequencing (scRNA-seq), in contrast to bulk RNA sequencing analysis, enables the investigation of transcriptional activity inside a single cell and allows the detection of tiny, clinically important tumor subpopulations (12).

Machine learning (ML) research has increased quickly because it provides a practical means to analyze huge and complicated datasets. In practice, a variety of ML algorithms are used (including random survival forest, support vector machine, gradient boosting, Bayesian, and deep learning). Moreover, these machine learning algorithms have been applied to the clinical management and prevention of HCC (13), including the discovery of biomarkers for early diagnosis (14), the development of prediction signatures for HCC recurrence (15, 16), and the production of single-cell atlases of HCC cell heterogeneity in response to immunotherapy (17).

In this study, we aimed to identify a prognostic biomarker that predicts the overall survival of HCC patients by scRNA-seq and bulk RNA-seq. We identified an LMRG signature in a training HCC cohort and further validated it in an external cohort. A novel nomogram incorporating clinical features and an LMRG signature was also constructed. The results demonstrated that this LMRG signature could help in the early diagnosis of patients with HCC, which also plays essential roles in the prognostication process and could be a viable therapeutic target for HCC patients.

Abbreviations: HCC, Hepatocellular carcinoma; DEGs, differentially expressed genes; LMRG, Lipid Metabolism-related genes; TMB, tumor mutational burden; GSEA, Gene set enrichment analysis; OS, overall survival; RS, risk score.

2 Methods

2.1 Data collection and preparation

TCGA and ICGC provide data on gene expression, prognosis, and clinicopathology for hepatocellular carcinoma (HCC) (18). Single-cell RNA sequencing (scRNA-seq) data from ten HCC patients were downloaded from the Gene Expression Omnibus (19). In addition, a total of 260 HCC samples from the ICGC cohorts (20) with clinical data and 232 with gene expression data were used as independent validation sets. Moreover, lipid metabolism-related genes (LMRGs) were obtained from Reactome.

2.2 ScRNA-seq data processing

The transcript count matrix were analyzed with the Seurat package v4.1.0 in R, as mentioned previously. Subsequent analysis was performed for the three specific tumor sites at the primary tumor, portal vein tumor thrombus, and metastatic lymph node in the 10 HCC patients. The resulting matrix was used to select the top 2000 highly expressed and variable genes. These selected genes were then used to compute the independent component (IC). RunUMAP were used to perform expression profile analysis. An absolute value of ($|\log FC|$) > 0.5 and an adjusted P value (adj P) < 0.05 were used as the cutoff values for differentially expressed genes (DEGs).

2.3 Machine learning model construction

We created models utilizing ML algorithms to forecast the prognosis of HCC patients using the TCGA-LIHC dataset. The ML approach, the least absolute shrinkage and selection operator (LASSO) regression model, was chosen based on the distribution of the outcome variable. The LASSO regression approach automates the selection of variables by shrinkage and the deletion of nonsignificant variables by setting them to zero to achieve L1 regularization to maximize prediction accuracy (21). By minimizing

the sum of squares, LASSO adjusts a shrinkage penalty lambda (λ) or tuning hyperparameter to the regression coefficients. As lambda values increase, the model becomes more biased, and further coefficients may be eliminated or set to zero. Using the optimum lambda value, the parameter estimates for the prediction model were calculated. We fitted additional models using a negative binomial depending on the number of zeros in the outcome variable. To determine which factors had the highest predictive value for determining the prognosis of HCC, we assessed the variable relevance rankings. The variables with the highest predictive value were identified among those chosen by the model with the best lambda value. The LMRG signature was then determined using the risk score (RS) formula based on the findings of the multivariate COX regression: RS = Σ (βi * Expi). The accuracy of the risk score model was evaluated by ROC and Kaplan–Meier survival analyses. The Kaplan–Meier survival curve combined with the log-rank test was employed to evaluate the survival differences between the LMRG high- and low-risk groups. The model was further validated in an external dataset of ICGC.

2.4 Nomogram construction

Univariate and multivariate regression analyses were used to analyze the independent clinical factors. Prior to nomogram construction, the LMRG signature and the clinical characteristics were integrated. The predictive accuracy of the prognostic model was evaluated by the time-dependent ROC curve. We evaluated the performance of the established nomogram on the basis of ROC curves, and decision curve analysis for overall survival at 1, 3, and 5 years.

2.5 Histological data analysis of HCC

The protein expression of AKR1C1, CYP27A1, CYP2C9, GLB1, HMGCS2, and PLPP1 in HCC patients was analyzed using histological data from the Human Protein Atlas (HPA) database (http://www.proteinatl.as.org/).

2.6 Western blot

HCC samples from all patients were collected with written informed consent with approval from the institutional research ethics committee of the First People's Hospital of Qinzhou (Approval number, 2021-15). The patients provided their written informed consent to participate in this study. Total proteins were extracted by lysis buffer radioimmunoprecipitation assay (RIPA; Beyotime, USA) with 1% phenylmethylsulfonyl fluoride (PMSF). Polyvinylidene fluoride (PVDF) membranes containing proteins separated on sodium dodecylsulfate–polyacrylamide gel electrophoresis (SDS-PAGE) gels were blocked with 5% nonfat milk and incubated with different primary antibodies, including β-ACTIN (#4970S, Cell Signal Technology), CYP27A1 (#67045-1-Ig, Proteintech), and GLB1 (#TA505544, ORIGENE), overnight at 4°C, and chemiluminescence was used for detection. Protein quantification was performed using ImageJ software.

2.7 Tumor mutational burden

The DNA somatic mutation dataset from TCGA-LIHC was used to determine whether LMRG signatures were associated with TMB. In accordance with our previous description, HCC patients were divided into LMRG-high and LMRG-low risk groups (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). The somatic mutations in the two LMRG groups were further visualized using the "maftools" R package.

2.8 Gene set enrichment analysis

To further understand the relationship of LMRGs and biological processes, we used GSEA software based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) gene set (KEGG C2, MSigDB database v7.5.1) to assess possible differences in biological functions between the LMRG high- and low-risk groups (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). To measure the significance of genomic enrichment, we used the *P* value of normalized enrichment scores and the *Q*-value of FDR.

2.9 Immune infiltration

To analyze the association between LMRG signatures and tumor-infiltrating immune cells, xCell analysis was used to estimate the fraction of the 64 subtype immune cells in each TCGA-LIHC sample (22). Subgroup analysis of signature immune cells for both LMRG high- and low-risk patients was carried out. An illustration of the results is shown using a heatmap and violin plot.

2.10 Statistical analysis

All analyses were performed with R version 4.0.5 and its appropriate packages without special instructions. A P value < 0.05 was set as statistically significant for all the analyses.

3 Results

3.1 Data profiling of the GSE149614 cohort by scRNA-seq

The workflow of this study is shown in Figure 1. After downloading the GSE149614 cohort from the GEO database, the data were profiled using scRNA-seq to determine the differentially expressed genes (DEGs) between primary and metastatic tumor tissues in HCC (19). UMAP algorithms were implemented for nonlinear dimensionality reduction, and samples were clustered, as shown in Figures 2A–C. The samples included 53 clusters

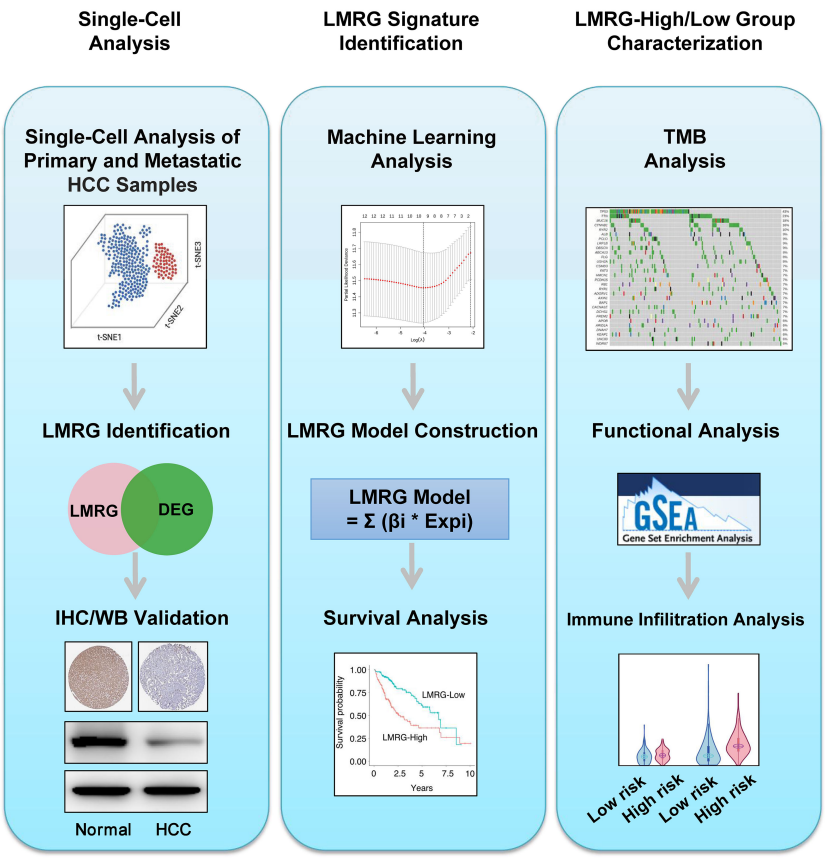


FIGURE 1
Workflow of this study. Step 1: Single-cell analysis. Step 2: LMRG signature identification. Step 3: LMRG-high and -low risk group characterization.

(Figure 2A) and six major cell types (Figure 2B). Among the six major cell types, hepatocytes are the most abundant cells. Myeloid cells, T/NK cells, B cells, fibroblast cells, and endothelial cells are less abundant cells. We also showed the distribution of cells in primary HCC and metastatic HCC (Figure 2C).

Then, DEGs related to HCC development were identified. To identify lipid metabolism-related genes (LMRGs) in HCC development, the DEGs were intersected with the LMRGs in the Reactome Pathway Database (https://reactome.org/), and 28 genes were obtained (Figure 2D and Supplementary Table 1). The 28 genes were defined as the screened LMRGs. Figure 2E shows the expression of 28 genes in different cell types. Most of the 28 genes were highly expressed in hepatocytes. However, *G0S2* and *ALOX5AP* were highly expressed in myeloid cells, *PLPP1* was highly expressed in endothelial cells, and *FABP4* was highly expressed in fibroblasts and endothelial cells.

3.2 Identification and validation of prognostic biomarkers for the risk model

To construct a predictive risk model, 28 previously screened LMRGs were subjected to univariate Cox regression analysis and the LASSO method. Supplementary Table 2 details the univariate Cox analysis of the 28 LMRGs. The LASSO model in Supplementary Figures 1A, B implies that eight genes (*AKR1C1*, *APOA1*, *CYP27A1*, *CYP2C9*, *GC*, *GLB1*, *HMGCS2*, and *PLPP1*) would be assessed in multivariate Cox analysis. As a result, *AKR1C1*, *CYP27A1*, *CYP2C9*, *GLB1*, *HMGCS2*, and *PLPP1* were selected for the LMRG signature (Supplementary Table 2), and the risk score was risk score (RS) = 0.00225 × *AKR1C1* - 0.00200 × *CYP27A1* - 0.00182 × *CYP2C9* + 0.02063 × *GLB1* - 0.00074 × *HMGCS2* - 0.01185 × *PLPP1*. Based on Kaplan-Meier analysis, we calculated survival probabilities for HCC patients with high and low expression of each gene (*AKR1C1*, *CYP27A1*, *CYP2C9*, *GLB1*, *HMGCS2*, and *PLPP1*). The results showed that high expression of *CYP27A1, CYP2C9, HMGCS2*, and *PLPP1* was correlated with better overall survival (OS) outcomes (Figure 3A). The expression of these genes in single-cell data (GSE149614) is shown in Figures 3B. Moreover, to examine the protein levels of AKR1C1, CYP27A1, CYP2C9, GLB1, HMGCS2, and PLPP1 between HCC and normal samples, we used the histological data in the HPA database and identified that the protein expression of AKR1C1 and GLB1 was significantly higher in HCC than in normal patients, while the protein expression of CYP27A1 and HMGCS2 was significantly higher in normal samples than in HCC samples (Figure 4A). Previous studies showed that AKR1C1 was upregulated in HCC and HMGCS2 was downregulated in HCC (23–25), which was consistent with our results. Since AKR1C1 and HMGCS2 were already reported in previous studies, we further

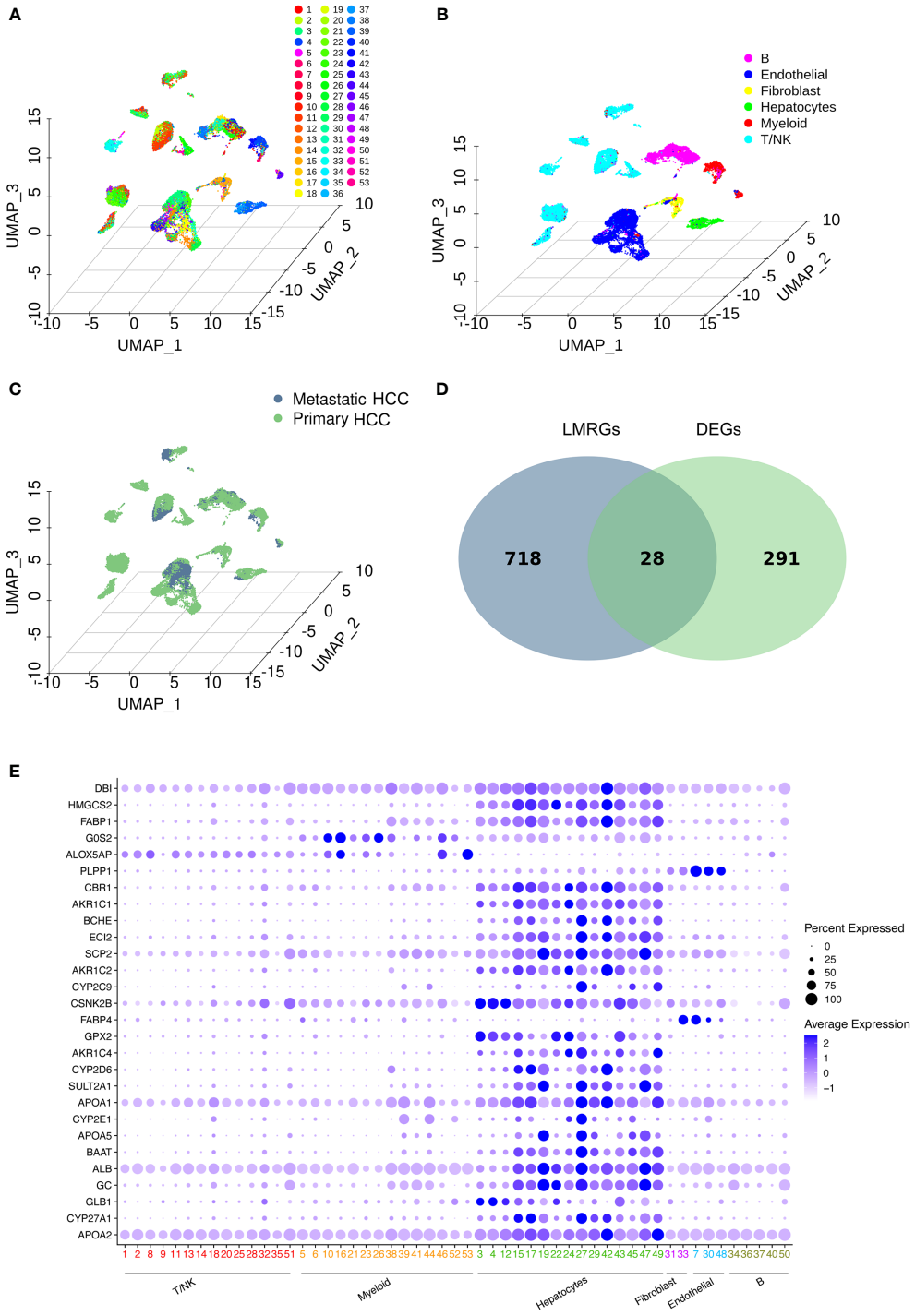


FIGURE 2
Data profiling of the GSE149614 cohort by scRNA-seq. **(A-C)** Samples from primary and metastatic hepatocellular carcinoma (HCC) tissue clustered by the UMAP algorithm. The dimension reduction showed the results of **(A)** 53 clusters, **(B)** six major cell types, and **(C)** primary and metastatic HCC. **(D)** Identification of lipid metabolism-related genes (LMRGs) in HCC development. A total of 746 LMRGs were assessed in the Reactome Pathway Database (https://reactome.org/). A total of 319 differentially expressed genes (DEGs) between primary and metastatic HCC samples were used to intersect with the 746 LMRGs. Finally, 28 screened LMRGs were identified. **(E)** Dot plot showing the expression of 28 LMRGs in different cell types.

validated the expression of GLB1 and CYP27A1 by Western blotting in HCC samples and adjacent normal samples. The results were consistent with the IHC results, as shown in Figure 4B. Subjects in the training cohort (TCGA-LIHC) and the testing cohort (ICGC-LIRI-JP) were divided into LMRG high- and low-risk groups by the median RS (Supplementary Table 3). The median RS for TCGA-LIHC was 1.001, and the median RS for ICGC-LIRI-JP was 0.976. Moreover, the areas under the ROC curves (AUCs) were evaluated, resulting in finding the AUCs of OS. The values for the training cohort were 0.745 (1 year), 0.696 (3

10.3389/fimmu.2023.1036562

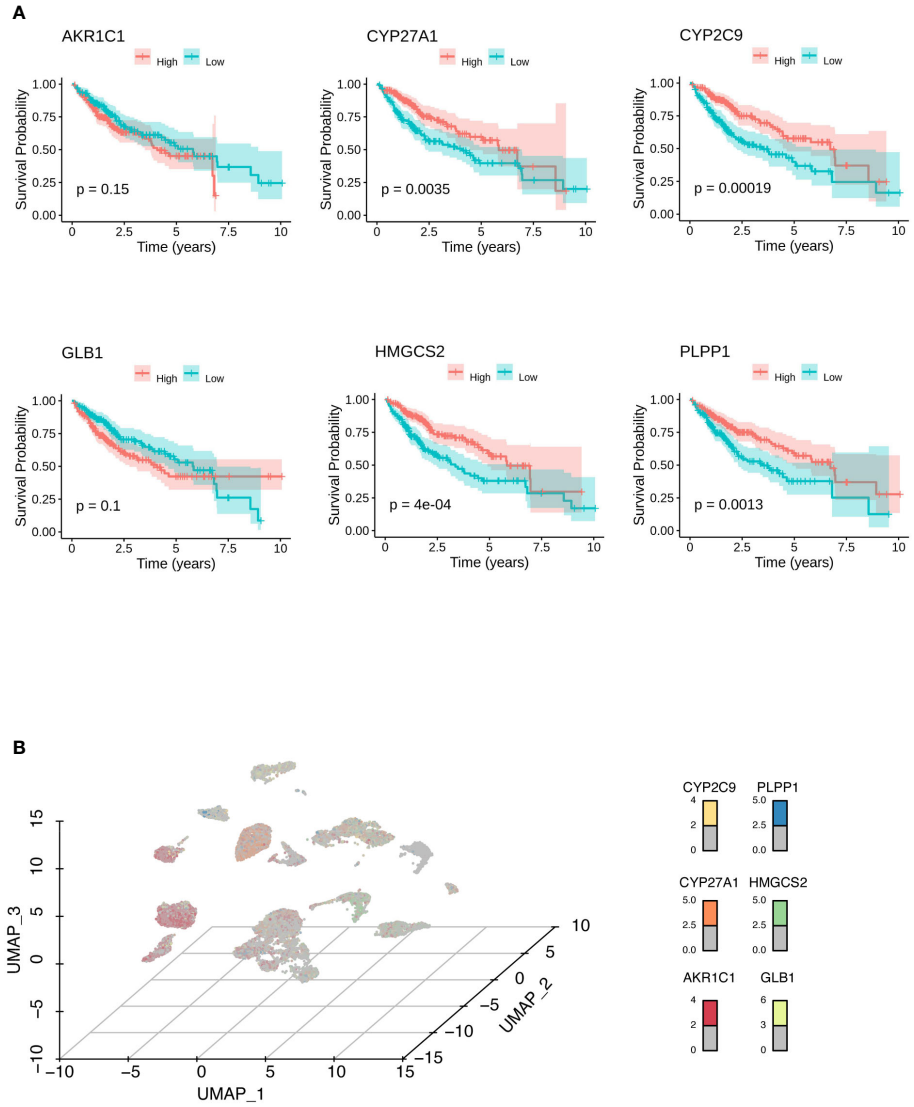


FIGURE 3
Identification of biomarkers for HCC. **(A)** Kaplan-Meier analysis showed the survival probabilities of HCC patients with high and low expression of the hub LMRG gene (*AKR1C1*, *CYP27A1*, *CYP2C9*, *GLB1*, *HMGCS2*, and *PLPP1*). **(B)** UMAP results showed the expression of the hub LMRG genes (*AKR1C1*, *CYP27A1*, *CYP2C9*, *GLB1*, *HMGCS2*, and *PLPP1*) in single-cell data.

years), and 0.702 (5 years), while the values for the testing cohort were 0.792 (1 year), 0.79 (3 years), and 0.761 (5 years) (Supplementary Figures 2A, B). The survival probabilities of HCC patients in both the training and testing cohorts were estimated by Kaplan-Meier analysis. Figures 4C, D show that poorer overall survival outcomes were observed in the LMRG high-risk groups than in the LMRG low-risk groups.

3.3 Construction and validation of the nomogram

On the basis of the training cohort, univariate and multivariate Cox regression models were utilized to screen significantly correlated clinical parameters for prognosis (Supplementary Figures 3A, B). As a result, cancer status and RS were screened (Supplementary Figures 3A, B). Then, a nomogram involving cancer status and RS was developed, and the survival of HCC patients at 1, 3, and 5 years was predicted by summarizing all points of the clinical parameters (Figure 5A). High AUC values (0.704, 0.743, and 0.792 for 1-, 3-, and 5-year survival, respectively) implied that the nomogram performed well in predicting OS (Figure 5B). To estimate the prediction power of the nomogram constructed, a decision curve analysis was performed. The results revealed that the nomogram could provide better benefits to HCC patients than the risk model constructed by genes for 5-year OS prediction (Figure 5C).

3.4 Mutation analysis in the training cohort

Somatic mutations were evaluated to analyze the tumor mutation burden and RS. The mutational landscape was

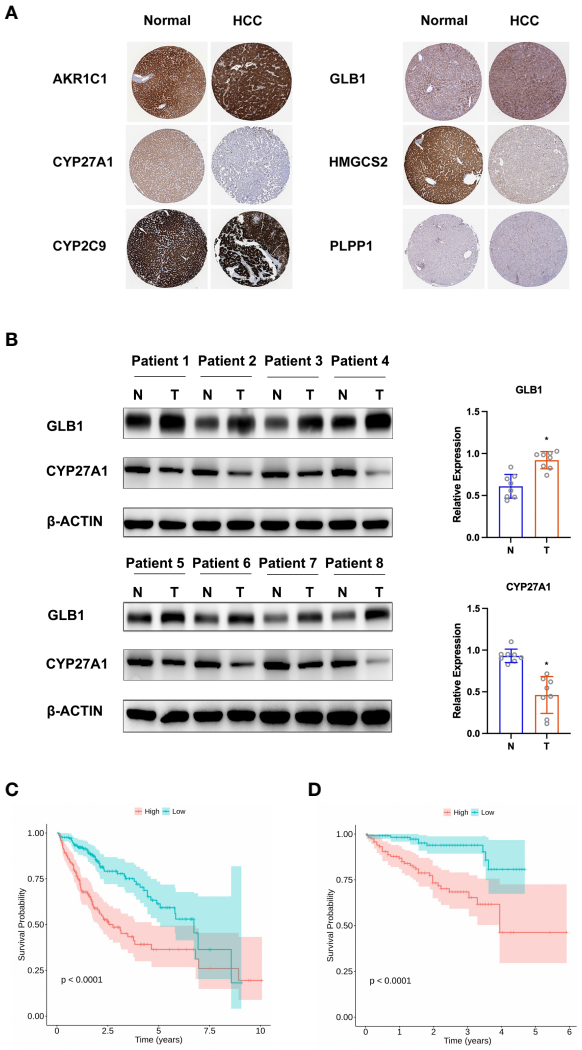


FIGURE 4
Validation of hub LMRG genes. **(A)** Histological data of AKR1C1, CYP27A1, CYP2C9, GLB1, HMGCS2, and PLPP1 from the Human Protein Atlas. **(B)** Western blotting analysis of GLB1 and CYP27A1. **(C)** Kaplan-Meier curves of the risk model in the training cohort. **(D)** Kaplan-Meier curves of the risk model in the testing cohort. Training cohort: TCGA-LIHC; testing cohort: ICGC-LIRI-JP. *p<0.05.

constructed, indicating that the LMRG high-risk group (92.64%) had more frequent mutation events than the low-risk group (86.14%) (Figure 6A). In addition, three significantly mutated genes, *TP53* (LMRG high-risk: 43%, low-risk: 17%), *CTNNB1* (LMRG high-risk: 16%, low-risk: 33%), and *TTN* (LMRG high-risk: 23%, low-risk: 21%), were identified in the two groups (Figure 6B. Accordingly, it was revealed that a mutation event was a risk factor in HCC.

3.5 Gene set enrichment analysis in the training cohort

Next, gene set enrichment analysis provided the pathways that were enriched in the LMRG high- and low-risk groups of the training cohorts. Pathways such as RNA degradation, spliceosome, epithelial cell signaling in Helicobacter pylori infection, lysosome, oocyte meiosis, and progesterone-mediated oocyte maturation were upregulated in the LMRG high-risk group (Figure 7A). On the other hand, pathways such as fatty acid metabolism, drug metabolism-cytochrome P450, glycine, serine and threonine metabolism, retinol metabolism, valine, leucine and isoleucine degradation, and tryptophan metabolism were upregulated in the LMRG low-risk group (Figure 7B). Moreover, other enriched pathways are illustrated in Supplementary Table 4.

3.6 Analysis of immune microenvironment characteristics in the training cohort

A training cohort was implemented in the xCell algorithm for immune infiltration estimation (Supplementary Table 5). The results of the immune infiltration in the TCGA-LIHC cohort were shown in Figure 8A. Remarkably, high levels of naïve CD8+ T cells, common myeloid progenitors, endothelial cells, granulocyte-monocyte progenitors, hematopoietic stem cells, M2

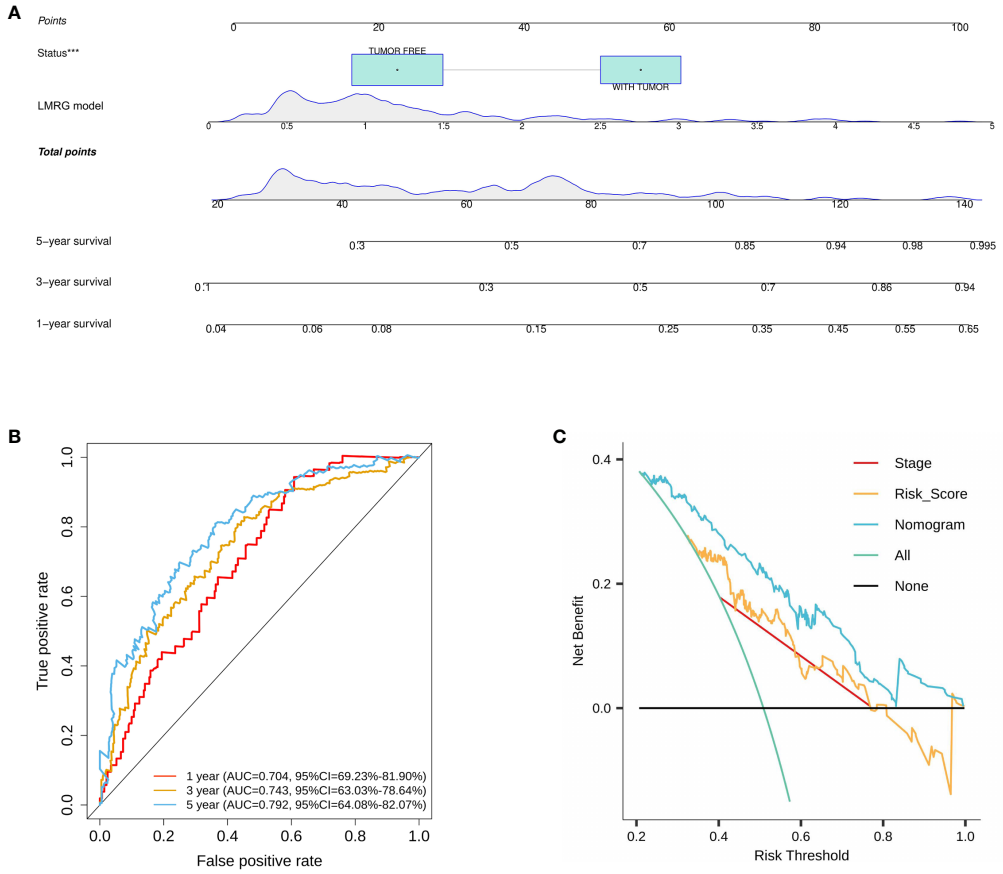


FIGURE 5
Construction and validation of a nomogram for HCC. **(A)** A nomogram was constructed based on cancer status and six LMRGs. **(B)** ROC curves of the nomogram in the training cohort. **(C)** Decision curve analysis of the nomogram in the TCGA cohort at 5 years. ***p<0.001.

macrophages, and plasmacytoid dendritic cells were observed in the LMRG high-risk group. However, high levels of activated myeloid dendritic cells, B cells, memory CD4+ T cells, class-switched memory B cells, common lymphoid progenitors, myeloid dendritic cells, M1 macrophages, mast cells, monocytes, NKT cells, and Th2 CD4+ T cells were observed in the LMRG low-risk group (Figure 8B).

4 Discussion

Hepatocellular carcinoma (HCC), one of the most common and aggressive tumors, has been linked to a high rate of morbidity for patients. Although several environmental or genetic risk factors linked to hepatocellular carcinoma (HCC) have been identified, the molecular processes causing HCC occurrence are still unknown. The proliferation and spread of tumors are facilitated by abnormal lipid metabolism. Research on the mechanism of lipid metabolism might therefore aid in the development of novel, targeted treatments to control or remove these refractory tumor cells, which might lead to the development of new medicines for HCC. A great number of studies have found a link between abnormal lipid metabolism and the onset and progression of malignancies (26). As a result, a lipid metabolism-related gene (LMRG) signature for predicting the survival of HCC patients is needed. Large-scale bulk and single-cell sequencing of tumor samples is now possible because of recent breakthroughs in sequencing technology. Moreover, the direct examination of genetic cell-to-cell variety is made possible by machine learning technology. For the first time, we were able to create a 6-LMRG signature of HCC in this study by scRNA-Seq. *AKR1C1, CYP27A1, CYP2C9, GLB1, HMGCS2*, and *PLPP1*, all six LMRGs, have been implicated in the genesis and progression of cancer. Upregulated *AKR1C1* expression was found in *HPV16*-positive oropharyngeal squamous cell carcinoma with viral integration, and it was linked with a poor prognosis in both HPV-positive and HPV-negative tumors (27). Furthermore, as one of the vitamin D pathway genes, *CYP27A1* has some impact on prostate cancer chemoprevention based on vitamin D metabolism and has the ability to predict the prognosis of prostate cancer patients (28, 29). A variation in the *CYP2C9* gene has been linked to an increased risk of colorectal cancer and adenoma (30). *GLB1* is a lysosomal exoglycosidase that catabolizes glycoconjugates and has been linked to cancer cell senescence (31). *HMGCS2* has been linked to oncogenic activity

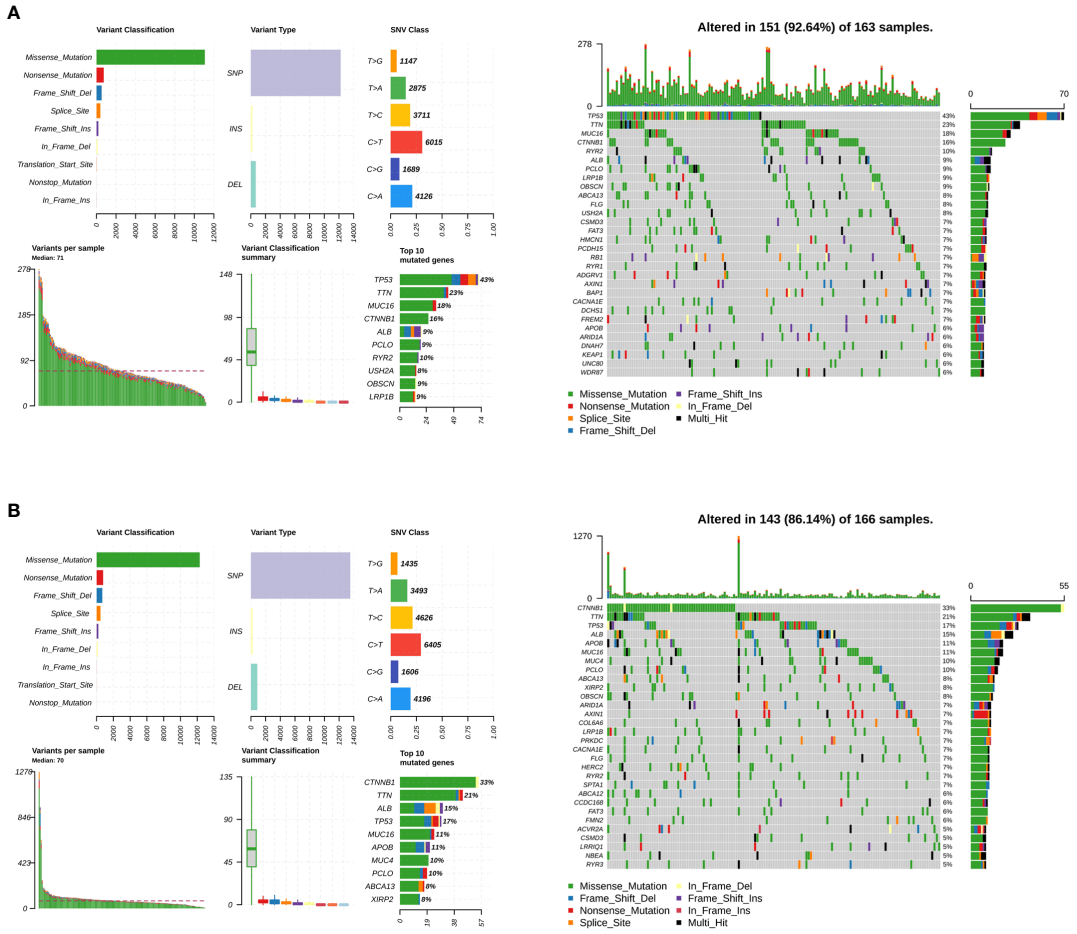


FIGURE 6
The tumor mutation burden characteristics of patients in the LMRG high- and low-risk groups. **(A)** Mutational landscape in the TCGA cohort of the LMRG high-risk groups. **(B)** Mutational landscape in the TCGA cohort of the low-risk groups.

in a variety of human tumors (32, 33). *HMGCS2* was identified as a differential hub gene of lipid metabolism in the pancancer immune microenvironment. Lower levels of *PLPP1* mRNA expression in tumor tissues than in surrounding normal tissues are linked to a worse prognosis (34).

Using a multivariate Cox regression analysis approach, we combined the signatures of several genes. The nomogram model, composed of the tumor status and the risk score derived from the LMRG signature, can visually predict the one-, three- and five-year overall survival outcomes for individual HCC patients. The final six genes demonstrated high accuracy in both the validation set and the overall prognosis for samples. In both TCGA and ICGC data, the LMRG low-risk score group showed worse results than the LMRG high-risk score group. In our present study, the most frequent somatic mutations in the LMRG high-risk group were *CTNNB1, TTN, TP53, ALB, MUC16*, and *PCLO*. Previous studies have shown that *TP53*, *MUC16*, and *TTN* mutations are common in many types of cancer, including gastric cancer and pancreatic and bladder cancers, and are associated with poor prognoses (35–38). In our study, the immune microenvironments of the LMRG high- and low-risk groups were analyzed by the xCell algorithm. Here, naïve $CD8^+$ T cells, common myeloid progenitors, endothelial cells, granulocyte-monocyte progenitors, hematopoietic stem cells, M2 macrophages, and plasmacytoid dendritic cells were significantly correlated with the LMRG high-risk group, which was first revealed in an HCC study.

The results of the gene set enrichment analysis showed that fatty acid metabolism and lysosome pathways, which involve lipid metabolism-related genes, changed between the LMRG high-risk group and the LMRG low-risk group. The fatty acid metabolism pathway participates in energy production, membrane synthesis, and signal transduction in tumor initiation and progression. Cancer cells rely on fatty acids as cellular building blocks for membrane formation, energy storage, and the production of signaling molecules (39). Lysosome pathways were associated with the LMRG high-risk group in our study. Lysosome pathways play an important role in autophagy. Autophagy, which is an evolutionarily conserved cellular degradation process that delivers cellular components to lysosomes, plays a critical role in cellular homeostasis through the degradation of lipids (40). Dysfunction or dysregulation of autophagy has been proven to be associated with HCC (41). The lysosome is a metabolic signaling hub that integrates

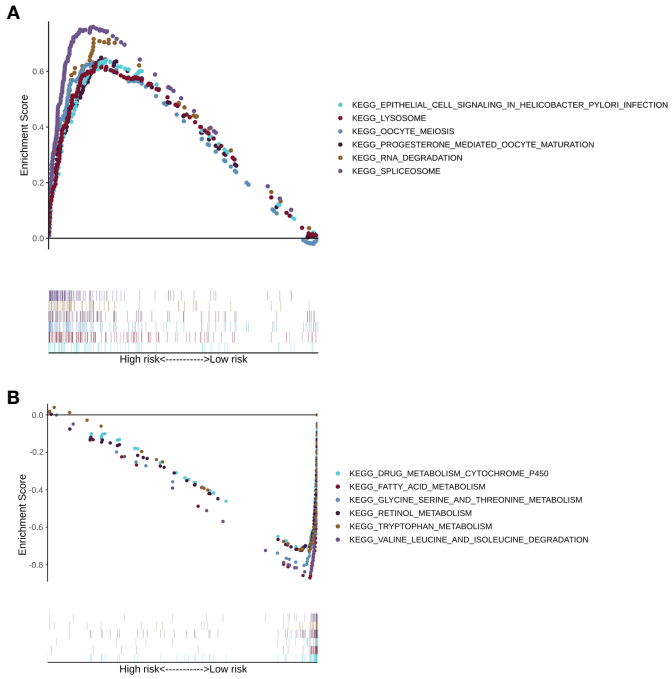


FIGURE 7
The KEGG pathway enrichment analysis of patients in the LMRG high- and low-risk groups. **(A)** Enriched pathways in the LMRG high-risk group. **(B)** Enriched pathways in the LMRG low-risk group.

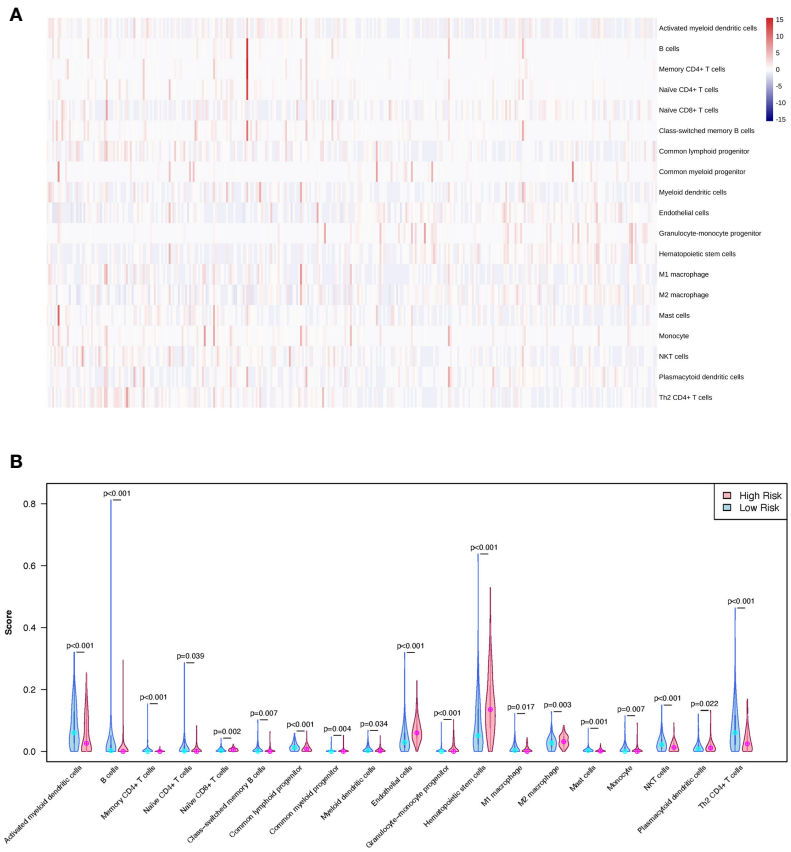


FIGURE 8
The different immune infiltration of patients in the LMRG high- and low-risk groups were identified. **(A)** The profile of immune infiltration in the TCGA-LIHC cohort showed by heatmap. **(B)** The violin plot shows the significantly different immune cells between the two risk groups in the TCGA cohort.

different environmental signals to regulate core anabolic and catabolic pathways critical in the maintenance of cellular homeostasis (42). As a result, the changes in fatty acid metabolism and lysosome pathways between the LMRG high-risk group and the LMRG low-risk group affect the overall survival outcomes of HCC patients through cellular homeostasis.

The findings of our study highlight the significant role of multiomics studies in basic research as well as translational and applied research within the field of personalized medicine for HCC. In particular, biomarkers based on LMRGs are essential for a reliable and effective evaluation of HCC prognosis and diagnosis. Moreover, clarifying molecular mechanisms through LMRGs is essential for the discovery of effective targets to treat HCC personalized treatment. HCC personalized medicine will benefit greatly from the prognosis-related LMRGs, and risk models identified in this study.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the institutional research ethics committee of the First People's Hospital of Qinzhou. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LM and ZP initiated the study. RQ performed the analysis. RQ and YS wrote and revised the manuscript. YL collected the patient samples. LM, ZP, and HJ designed the study and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Guangxi Key Clinical Specialty Construction Project (General Surgery, 2022), and Scientific Research and Technology Development Plan Project of Qinzhou (No. 20230301).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1036562/full#supplementary-material

SUPPLEMENTARY FIGURE 1
LASSO regression analysis of the LMRGs. **(A)** LASSO coefficient profiles of eight prognostic LMRGs. **(B)** Partial likelihood deviance curves of eight prognostic LMRGs.

SUPPLEMENTARY FIGURE 2
ROC curves of the LMRG model. **(A)** ROC curves of the risk model in the training cohort. **(B)** ROC curves of the risk model in the testing cohort. Training cohort: TCGA-LIHC; testing cohort: ICGC-LIRI-JP.

SUPPLEMENTARY FIGURE 3
Cox regression analysis of clinical parameters. **(A)** Univariate analysis of clinical parameters. **(B)** Multivariate analysis of clinical parameters.

References

1. Jafari N, Dolatkhah R. Molecular epidemiology of liver cancer: Liver cancer incidence and mortality pattern worldwide. *Int J Life Sci Res Arch* (2021) 1:018–23. doi: 10.53771/ijlsra.2021.1.1.0046

2. Gomaa AI, Khan SA, Toledano MB, Waked I, Taylor-Robinson SD. Hepatocellular carcinoma: Epidemiology, risk factors and pathogenesis. *World J Gastroenterol* (2008) 14:4300–8. doi: 10.3748/wjg.14.4300

3. Vivarelli M, Montalti R, Risaliti A. Multimodal treatment of hepatocellular carcinoma on cirrhosis: an update. *World J Gastroenterol* (2013) 19:7316–26. doi: 10.3748/wjg.v19.i42.7316

4. Ghavimi S, Apfel T, Azimi H, Persaud A, Pyrsopoulos NT. Management and treatment of hepatocellular carcinoma with immunotherapy: A review of current and future options. *J Clin Transl Hepatol* (2020) 8:168–76. doi: 10.14218/JCTH.2020.00001

5. Zheng M, Mullikin H, Hester A, Czogalla B, Heidegger H, Vilsmaier T, et al. Development and validation of a novel 11-gene prognostic model for serous ovarian carcinomas based on lipid metabolism expression profile. *Int J Mol Sci* (2020) 21:E9169. doi: 10.3390/ijms21239169

6. Anderson NM, Simon MC. The tumor microenvironment. *Curr Biol* (2020) 30: R921–5. doi: 10.1016/j.cub.2020.06.081

7. Liu B-W, Wang X-Y, Cao J-L, Chen L-L, Wang Y-L, Zhao B-Q, et al. TDP-43 upregulates lipid metabolism modulator ABHD2 to suppress apoptosis in hepatocellular carcinoma. *Commun Biol* (2022) 5:816. doi: 10.1038/s42003-022-03788-w

8. Qian H, Lei T, Hu Y, Lei P. Expression of lipid-metabolism genes is correlated with immune microenvironment and predicts prognosis in osteosarcoma. *Front Cell Dev Biol* (2021) 9:673827. doi: 10.3389/fcell.2021.673827

9. Wan S, Lei Y, Li M, Wu B. A prognostic model for hepatocellular carcinoma patients based on signature ferroptosis-related genes. *Hepatol Int* (2022) 16:112–24. doi: 10.1007/s12072-021-10248-w

10. Zhu G, Xia H, Tang Q, Bi F. An epithelial-mesenchymal transition-related 5-gene signature predicting the prognosis of hepatocellular carcinoma patients. *Cancer Cell Int* (2021) 21:166. doi: 10.1186/s12935-021-01864-5

11. Dai Y, Qiang W, Lin K, Gui Y, Lan X, Wang D. An immune-related gene signature for predicting survival and immunotherapy efficacy in hepatocellular carcinoma. *Cancer Immunol Immunother* (2021) 70:967–79. doi: 10.1007/s00262-020-02743-0

12. Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat Rev Genet* (2021) 22:3–18. doi: 10.1038/s41576-020-0265-5

13. Calderaro J, Seraphin TP, Luedde T, Simon TG. Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. *J Hepatol* (2022) 76:1348–61. doi: 10.1016/j.jhep.2022.01.014

14. Johnson P, Zhou Q, Dao DY, Lo YMD. Circulating biomarkers in the diagnosis and management of hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* (2022) 19 (10):670–81. doi: 10.1038/s41575-022-00620-y

15. Lee I-C, Huang J-Y, Chen T-C, Yen C-H, Chiu N-C, Hwang H-E, et al. Evolutionary learning-derived clinical-radiomic models for predicting early recurrence of hepatocellular carcinoma after resection. *Liver Cancer* (2021) 10:572–82. doi: 10.1159/000518728

16. Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: Results from a multicenter study. *J Clin Oncol* (2017) 35:734–42. doi: 10.1200/JCO.2016.68.2153

17. Ma L, Wang L, Khatib SA, Chang C-W, Heinrich S, Dominguez DA, et al. Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *J Hepatol* (2021) 75:1397–408. doi: 10.1016/j.jhep.2021.06.028

18. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* (2015) 19:A68–77. doi: 10.5114/wo.2014.47136

19. Lu Y, Yang A, Quan C, Pan Y, Zhang H, Li Y, et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat Commun* (2022) 13:4594. doi: 10.1038/s41467-022-32283-3

20. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The international cancer genome consortium data portal. *Nat Biotechnol* (2019) 37:367–9. doi: 10.1038/s41587-019-0055-9

21. Misra S, Li H, He J. Chapter 5 - robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods. In: Misra S, Li H, He J, editors. *Machine learning for subsurface characterization*. Houston, TX: Gulf Professional Publishing (2020). p. 129–55. doi: 10.1016/B978-0-12-817736-5.00005-3

22. Aran D, Hu Z, Butte AJ. XCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* (2017) 18:220. doi: 10.1186/s13059-017-1349-1

23. Zhao S-F, Wang S-G, Zhao Z-Y, Li W-L. AKR1C1-3, notably AKR1C3, are distinct biomarkers for liver cancer diagnosis and prognosis: Database mining in malignancies. *Oncol Lett* (2019) 18:4515–22. doi: 10.3892/ol.2019.10802

24. Li J, Li M-H, Wang T-T, Liu X-N, Zhu X-T, Dai Y-Z, et al. SLC38A4 functions as a tumour suppressor in hepatocellular carcinoma through modulating wnt/β-catenin/MYC/HMGCS2 axis. *Br J Cancer* (2021) 125:865–76. doi: 10.1038/s41416-021-01490-y

25. Wang Y-H, Liu C-L, Chiu W-C, Twu Y-C, Liao Y-J. HMGCS2 mediates ketone production and regulates the proliferation and metastasis of hepatocellular carcinoma. *Cancers (Basel)* (2019) 11:1876. doi: 10.3390/cancers11121876

26. Long J, Zhang C-J, Zhu N, Du K, Yin Y-F, Tan X, et al. Lipid metabolism and carcinogenesis, cancer development. *Am J Cancer Res* (2018) 8:778–91.

27. Huebbers CU, Verhees F, Poluschkin L, Olthof NC, Kolligs J, Siefer OG, et al. Upregulation of AKR1C1 and AKR1C3 expression in OPSCC with integrated HPV16 and HPV-negative tumors is an indicator of poor prognosis. *Int J Cancer* (2019) 144:2465–77. doi: 10.1002/ijc.31954

28. Maksymchuk OV, Kashuba VI. Altered expression of cytochrome P450 enzymes involved in metabolism of androgens and vitamin d in the prostate as a risk factor for prostate cancer. *Pharmacol Rep* (2020) 72:1161–72. doi: 10.1007/s43440-020-00133-y

29. Alfaqih MA, Nelson ER, Liu W, Safi R, Jasper JS, Macias E, et al. CYP27A1 loss dysregulates cholesterol homeostasis in prostate cancer. *Cancer Res* (2017) 77:1662–73. doi: 10.1158/0008-5472.CAN-16-2738

30. Barry EL, Poole EM, Baron JA, Makar KW, Mott LA, Sandler RS, et al. CYP2C9 variants increase risk of colorectal adenoma recurrence and modify associations with smoking but not aspirin treatment. *Cancer Causes Control* (2013) 24:47–54. doi: 10.1007/s10552-012-0088-6

31. Vidya B, Palaniswamy M, Angayarkanni J, Ayub Nawaz K, Thandeeswaran M, Krishna Chaithanya K, et al. Purification and characterization of β-galactosidase from newly isolated aspergillus terreus (KUBCF1306) and evaluating its efficacy on breast cancer cell line (MCF-7). *Bioorg Chem* (2020) 94:103442. doi: 10.1016/j.bioorg.2019.103442

32. Su S-G, Yang M, Zhang M-F, Peng Q-Z, Li M-Y, Liu L-P, et al. MiR-107-mediated decrease of HMGCS2 indicates poor outcomes and promotes cell migration in hepatocellular carcinoma. *Int J Biochem Cell Biol* (2017) 91:53–9. doi: 10.1016/j.biocel.2017.08.016

33. Hao Y, Li D, Xu Y, Ouyang J, Wang Y, Zhang Y, et al. Investigation of lipid metabolism dysregulation and the effects on immune microenvironments in pan-cancer using multiple omics data. *BMC Bioinf* (2019) 20:195. doi: 10.1186/s12859-019-2734-4

34. Qian D, Liu H, Zhao L, Wang X, Luo S, Moorman PG, et al. Novel genetic variants in genes of the fc gamma receptor-mediated phagocytosis pathway predict non-small cell lung cancer survival. *Transl Lung Cancer Res* (2020) 9:575–86. doi: 10.21037/tlcr-19-318

35. Ling S, Shan Q, Zhan Q, Ye Q, Liu P, Xu S, et al. USP22 promotes hypoxia-induced hepatocellular carcinoma stemness by a HIF1α/USP22 positive feedback loop upon TP53 inactivation. *Gut* (2020) 69:1322–34. doi: 10.1136/gutjnl-2019-319616

36. Li X, Pasche B, Zhang W, Chen K. Association of MUC16 mutation with tumor mutation load and outcomes in patients with gastric cancer. *JAMA Oncol* (2018) 4:1691–8. doi: 10.1001/jamaoncol.2018.2805

37. Oh J-H, Jang SJ, Kim J, Sohn I, Lee J-Y, Cho EJ, et al. Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *NPJ Genom Med* (2020) 5:33. doi: 10.1038/s41525-019-0107-6

38. Cornella H, Alsinet C, Sayols S, Zhang Z, Hao K, Cabellos L, et al. Unique genomic profile of fibrolamellar hepatocellular carcinoma. *Gastroenterology* (2015) 148:806–818.e10. doi: 10.1053/j.gastro.2014.12.028

39. Currie E, Schulze A, Zechner R, Walther TC, Farese RV. Cellular fatty acid metabolism and cancer. *Cell Metab* (2013) 18:153–61. doi: 10.1016/j.cmet.2013.05.017

40. Liu K, Czaja MJ. Regulation of lipid stores and metabolism by lipophagy. *Cell Death Differ* (2013) 20:3–11. doi: 10.1038/cdd.2012.63

41. Chao X, Qian H, Wang S, Fulte S, Ding W-X. Autophagy and liver cancer. *Clin Mol Hepatol* (2020) 26:606–17. doi: 10.3350/cmh.2020.0169

42. Lamming DW, Bar-Peled L. Lysosome: The metabolic signaling hub. *Traffic* (2019) 20:27–38. doi: 10.1111/tra.12617

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 07 March 2023
DOI 10.3389/fmed.2023.1058919

Check for updates

OPEN ACCESS

EDITED BY
Ken Chen,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Carl-Magnus Svensson,
Leibniz Institute for Natural Product Research and Infection Biology, Germany
Shaoheng Liang,
Rice University, United States

*CORRESPONDENCE
Alexander Partin
✉ apartin@anl.gov

SPECIALTY SECTION
This article was submitted to Precision Medicine, a section of the journal Frontiers in Medicine

RECEIVED 30 September 2022
ACCEPTED 10 February 2023
PUBLISHED 07 March 2023

CITATION
Partin A, Brettin T, Zhu Y, Dolezal JM, Kochanny S, Pearson AT, Shukla M, Evrard YA, Doroshow JH and Stevens RL (2023) Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. *Front. Med.* 10:1058919. doi: 10.3389/fmed.2023.1058919

COPYRIGHT
© 2023 Partin, Brettin, Zhu, Dolezal, Kochanny, Pearson, Shukla, Evrard, Doroshow and Stevens. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images

Alexander Partin[1]*, Thomas Brettin[1], Yitan Zhu[1], James M. Dolezal[2], Sara Kochanny[2], Alexander T. Pearson[2], Maulik Shukla[1], Yvonne A. Evrard[3], James H. Doroshow[4] and Rick L. Stevens[1,5]

[1]Division of Data Science and Learning, Argonne National Laboratory, Lemont, IL, United States, [2]Section of Hematology/Oncology, Department of Medicine, University of Chicago Medical Center, Chicago, IL, United States, [3]Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD, United States, [4]Division of Cancer Therapeutics and Diagnosis, National Cancer Institute, Bethesda, MD, United States, [5]Department of Computer Science, The University of Chicago, Chicago, IL, United States

Patient-derived xenografts (PDXs) are an appealing platform for preclinical drug studies. A primary challenge in modeling drug response prediction (DRP) with PDXs and neural networks (NNs) is the limited number of drug response samples. We investigate multimodal neural network (MM-Net) and data augmentation for DRP in PDXs. The MM-Net learns to predict response using drug descriptors, gene expressions (GE), and histology whole-slide images (WSIs). We explore whether combining WSIs with GE improves predictions as compared with models that use GE alone. We propose two data augmentation methods which allow us training multimodal and unimodal NNs without changing architectures with a single larger dataset: 1) combine single-drug and drug-pair treatments by homogenizing drug representations, and 2) augment drug-pairs which doubles the sample size of all drug-pair samples. Unimodal NNs which use GE are compared to assess the contribution of data augmentation. The NN that uses the original and the augmented drug-pair treatments as well as single-drug treatments outperforms NNs that ignore either the augmented drug-pairs or the single-drug treatments. In assessing the multimodal learning based on the MCC metric, MM-Net outperforms all the baselines. Our results show that data augmentation and integration of histology images with GE can improve prediction performance of drug response in PDXs.

KEYWORDS

drug response prediction, histology whole-slide images, gene expression, multimodal deep learning, preclinical drug studies, data augmentation, patient-derived xenograft (PDX)

1. Introduction

With recent advancements in applications of artificial intelligence in medicine and biology, predictive modeling has gradually become one of the primary directions in cancer research for analytically predicting the response of tumors to anticancer treatments (1, 2). In particular, conventional machine learning (ML) and deep learning (DL) methods have been widely investigated for building computational drug response prediction models for cancer cell lines with large datasets of omics profiles (3). The complex heterogeneities of cancer that occur within and between tumors present a major obstacle to successful discovery of robust biomarkers and therapies (4, 5). Patient-derived tumor xenografts (PDXs) are a contemporary biological model that is created by grafting cancerous tissue, obtained from human tumor specimens, into immunodeficient mice. The *in vivo* environment of PDXs helps preserve tumor heterogeneity as compared to *in vitro* cell lines, and therefore, is presumed to better mimic the response of human patients with certain cancer types. PDXs continue to gain reputation for studying cancer and investigating drug response in preclinical drug studies (6–8).

Predicting the response of tumors to drug treatments with accurate and robust computational models provides a modern approach for identifying top candidates for preclinical drug screening experiments or personalized cancer treatments. A variety of ML and DL approaches have been explored with high-throughput drug screens and cell lines (9, 10). Alternatively, our literature search retrieved only two publications that have used only PDX data to train prediction models for drug response (11, 12). Both studies used the Novartis PDX data (NIBR PDXE), which were generated using a 1 x 1 x 1 experimental design (13, 14), where each drug was tested against each patient PDX model using only one entumored mouse per model. Nguyen et al. (11) used an optimal model complexity (OMC) strategy with random forests to build drug response models for 26 treatment-cancer type combinations. They considered three genomic feature types in their analyses, including gene expressions (GE), copy-number alterations (CNAs), and single-nucleotide variants (SNVs). While considering a single feature type at a time, they used OMC to determine an optimal subset of features to obtain the best performing model for each treatment-cancer type pair. They showed that for the majority of cases, models developed with OMC outperform models that used all the available features. In another study, Kim et al. proposed PDXGEM, a pipeline that identifies biomarkers predictive of drug response in PDX and then uses the identified markers to train prediction models (12). To identify predictive genes based on GE and drug response, the pipeline utilizes a strategy similar to co-expression extrapolation (COXEN) (15, 16), and consequently selects the genes whose co-expression patterns are best preserved between PDXs and patient tumors. They trained prediction models using random forests for six treatment-tumor type combinations and then predicted response in patients.

A primary challenge in modeling drug response with PDXs is the limited availability of drug response data. The sample size of PDXs is usually orders of magnitude smaller than the analogous cell line datasets. It has been shown that increasing the amount of training samples improves generalization performance of supervised learning models in vision and text applications (17, 18), as well as drug response models in cell lines (19, 20). Collecting PDX response data, either through experiments or integration of multiple datasets, carries considerable technical and financial challenges. Alternatively, instead of directly increasing the sample size, the volume of data can be expanded by representing each sample with multiple feature types. Multimodal architectures that integrate genomic and histology images have been shown to improve prognosis prediction of patients with cancer as compared with unimodal architectures that learn only from a single data modality (i.e., feature type) (21–23). Another possible direction to address the limited sample size is data augmentation. Augmentation techniques have been extensively explored with image and text data, but not much with drug response. While augmenting images has become a common practice, tabular datasets such as omics profiles lack standardized augmentation methods.

In this study, we investigate two approaches for predicting drug response in PDX, including multimodal learning and data augmentation. We explore a multimodal neural network (MM-Net) that learns to predict drug response in PDXs using GE and histology whole-slide images (WSIs), two feature types representing cancer tissue, and molecular descriptors that represent drugs. The multimodal architecture is designed to take four feature sets as inputs: (1) GE, (2) histology images, and (3,4) molecular descriptors of a pair of drugs. We benchmark the prediction performance of MM-Net against three baselines: (1) NN trained with drug descriptors and GE, (2) NN trained with drug descriptors and WSIs, and (3) LightGBM model (24) trained with drug descriptors and GE. With multimodal learning, our goal is to explore whether the integration of histology images with GE improves the prediction performance as compared with models that use GE features alone. For data augmentation, we homogenize the drug representation of single-drug and drug-pair treatments in order to combine them into a single dataset. Moreover, we introduce an augmentation method that doubles the sample size of all drug-pair treatments. The proposed augmentations allow us to combine single-drug and drug-pair treatments to train MM-Net and the baselines without changing the architectures and the dataset. We explore the contribution of augmented data for improving the prediction of drug response.

This paper provides unique contributions compared with existing works that train drug response models with PDX data (11, 12). First, we build general drug response models for PDXs across multiple cancer types and drug treatments. Alternatively, prediction models in Kim et al. (12) and Nguyen et al. (11) are built for specific combinations of cancer type and drug treatment. Thus, our study targets a more challenging task. Second, we utilize PDX histology images with multimodal architecture which has not yet been studied for drug response prediction in PDX. Our study presents a framework for integrating image data with genomic measurements and drug chemical structure for predicting treatment effect. Third, we combine multiple treatments into a single dataset by homogenizing single-drug and drug-pair treatments and utilize drug features for training models. This provides an advantage over existing works which built prediction models for unique combinations of drugs and cancer

types, and therefore, disregard drug features when model training. Furthermore, we propose an augmentation method for drug-pairs that doubles the sample size of the drug-pair treatments in the dataset. Fourth, existing studies built prediction models using the PDXE drug screening data which were generated using a 1 x 1 x 1 experimental design, i.e., one mouse per model per treatment. In contrast, we utilized the PDMR dataset where treatment response is measured by comparing a group of treated mice to a group of untreated mice. The group approach allows assessing the variability of response across mice and might be considered as more reliable in capturing tumor heterogeneity.

2. Materials and methods

2.1. Data

2.1.1. Experimental design of drug efficacy in PDX

We used unpublished PDX drug response data from the NCI Patient-Derived Models Repository (PDMR; http://pdmr.cancer.gov). The NCI PDMR performs histopathology assessment, whole-exome sequencing and RNA-Seq analysis of a subset of tumors from each PDX model to establish baseline histology and omic characterization for each model. To date, over six hundred unique PDX models have been characterized and data made available through the public website. Baseline pathology and omic characterization from 487 models were used for this analysis. The efficacy of drug treatments in PDMR is measured through controlled group experiments. Figure 1 illustrates the process of obtaining primary tumor specimen from a patient, engrafting tumor tissue into PDX models, performing baseline characterization, expanding tumor tissue over multiple passages within a lineage, and then using the expanded tumors in drug treatment experiments. A total of 96 PDX models from 89 unique patients were used for the experiments. The tumors are grown subcutaneously in NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) host mice and staged to an approximate tumor weight of 200 mm^3 for the drug studies. The control group is treated with a vehicle only (i.e., a solution that delivers drugs to the treated animals). The preclinical dataset includes twelve single-drug and 36 drug-pair treatment arms (the drugs are still in preclinical and/or clinical investigations and their names and properties are expected to be released in the future). Median tumor volume over time for each vehicle or treatment group is used for response assessment. The GE profiles and WSIs were aggregated and preprocessed for the downstream ML and DL analysis.

2.1.2. Drug response in PDX

The growth of tumor volume over time represents the response of PDX tumors to drug treatments, as shown in Figure 2. There are several methods available in the literature for encoding drug response in PDXs which include both continuous and categorical types, but no consensus currently exists regarding which type or actual representation is better (6, 25). Continuous metrics include percent change in tumor volume, area under the tumor growth curve, best tumor response, best average response, and more. In this study we chose to use the binary representation of response which aligns with other drug response prediction studies that are mentioned earlier (11, 12), and the RECIST criteria, an existing standard for encoding response in patients (26). The group approach intends to capture the variability of PDX drug response across mice of the same lineage (25). Median tumor volume per treatment group is assessed relative to the control group to create a binary variable representing response. Specifically, for each drug treatment experiment, a single experienced preclinical study analyst assessed the curves of median tumor volume over time for each treatment arm, and assigned label "1" for response (regression of at least 30% from staging for more than one consecutive time point at any point during the study) and label "0" for non-response. In essence, a modified RECIST score for regression vs. no regression was used to label the response. Thus, a single best response value was assigned for each treatment arm.

2.1.3. Data generation

Three feature types were used for model training, including drug descriptors, GE, and histology image tiles.

Gene expressions. Gene expressions have been considered to provide more predictive power than other omics data types for drug response prediction (DRP) (27), and therefore, are often used to represent cancer in DRP models, either standalone or in a combination with other multiomics (28). However, the high dimensionality of gene expressions and the relatively small sample size can lead to overfitting (2). To address this issue, several gene selection methods have been utilized, including filtering genes based on variability across samples (29–31) or using gene subsets such as LINCS (10, 32–36) or COSMIC (35, 37–40) that are known to be associated with cancer and/or treatment response. We are not aware of any systematic analysis that studied which filtering method better addresses overfitting and improves prediction generalization. In this study, we filtered the RNA-seq data by selecting 942 landmark genes discovered by the Library of Integrated Network-Based Cellular Signatures (LINCS) project. The LINCS genes have been shown to comprehensively characterize and infer the gene expression variation of more than 80% of the whole transcriptome (32). We used TPM (transcripts per kilobase million) expression values of these genes, which were transformed by $log2x+1$, where x is the TPM value of a gene. The *log* transformed TPM values of each gene were then standardized to have a zero mean and a unit standard deviation across all gene expression profiles.

Drug descriptors. We used the Dragon software package (version 7.0) to calculate numerical descriptors of drug molecular structure. The software calculates various types of molecular descriptors, such as atom types, estimations of molecular properties, topological and geometrical descriptors, functional groups and fragment counts, and drug-like indices. A total of 1,993 descriptors were used for the analysis after removing descriptors with missing values. We standardized the descriptor values across drugs to have a zero mean and a unit standard deviation.

Histology images. During PDX model expansion, entumored mice were sacrificed between 1,000 and 2,000 mm^3 for collection of tumors for representative model characterization including histopathological examination. Hematoxylin and Eosin (H&E)

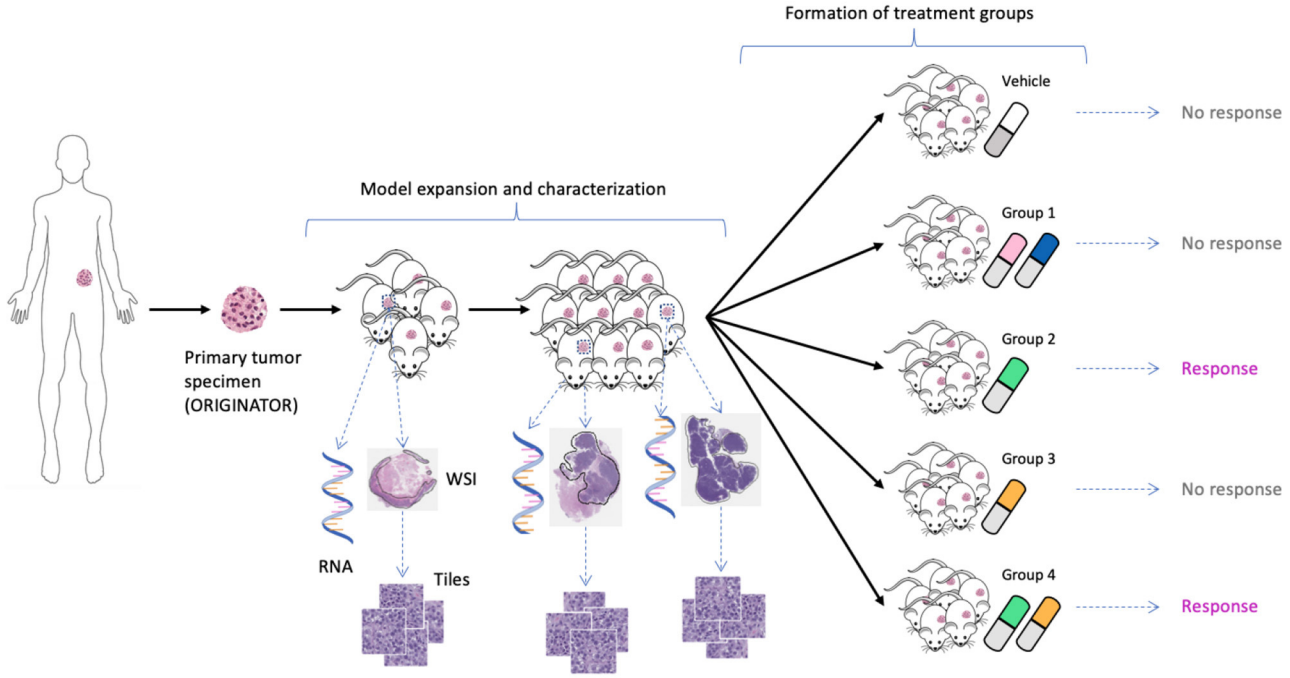


FIGURE 1
Expansion of tumor tissue from the source specimen (ORIGINATOR) to mice across multiple passages. Mice originated from the same specimen are divided into a control group and multiple treatment groups. Tumors from certain mice were histologically and molecularly profiled, resulting in whole-slide images and omics profiles.

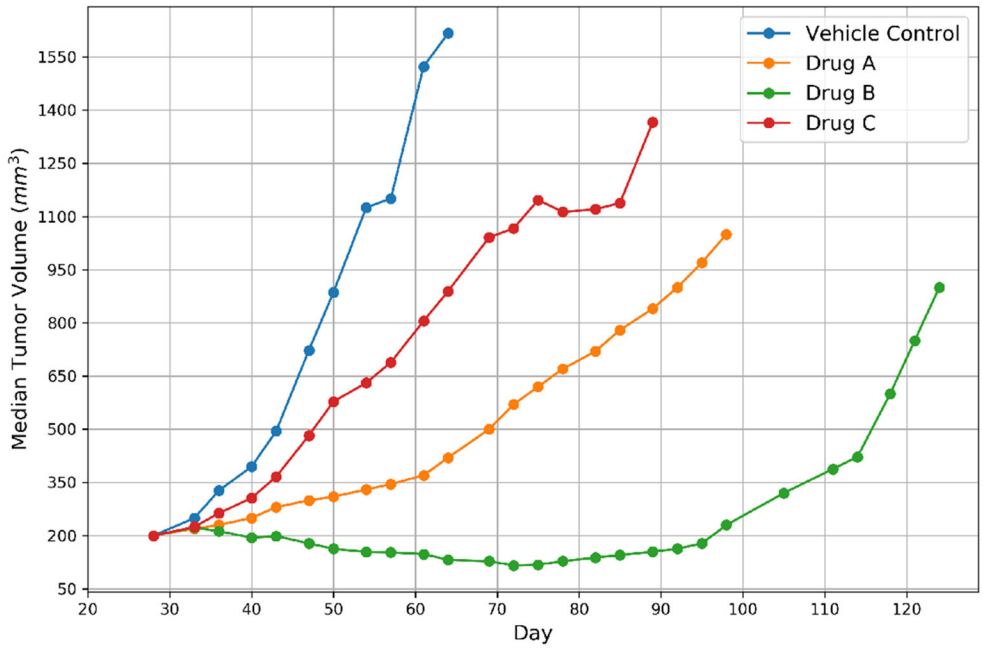


FIGURE 2
Representative tumor growth curves for vehicle control and three drugs. The label of response was assigned to Drug B because the tumor achieved regression and non-response was assigned to the remaining drugs.

stained pathology slides were digitized into WSIs using an Aperio AT2 digital whole slide scanner (Leica Biosystems) at 20x magnification. A board-certified pathologist from Frederick National Laboratory of Cancer Research reviewed the slides to ensure the PDX models were consistent with the original patient diagnosis. Tumor regions of interest (ROIs) were annotated within

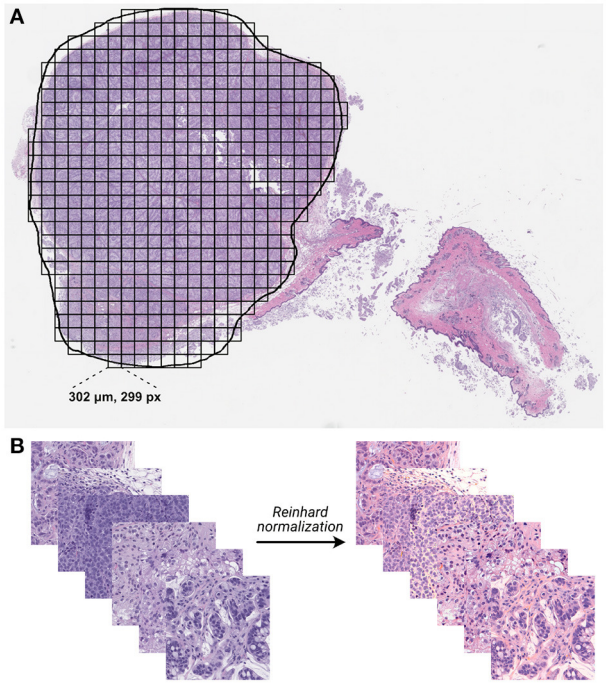


FIGURE 3
Whole-slide histology image processing. **(A)** Whole-slide images were annotated with region of interest (ROI) outlines, and image tiles were extracted from within ROIs in a grid-wise fashion. **(B)** Extracted non-background tiles underwent digital stain normalization using the Reinhard method (44).

the image slides using QuPath (41) by a single University of Chicago pathologist.

Whole slide images were processed into individual tiles using the Slideflow software package (42, 43), as shown in Figure 3. Image tiles were extracted from within annotated ROIs in a grid pattern at 302 μm by 302 μm with no overlap, then downsampled to 299 pixels by 299 pixels, resulting in an effective optical magnification of 10x. Background tiles were removed with grayscale filtering, where each tile is converted to the HSV color space and removed if more than 60% of its pixels have a hue value of less than 0.05. Image tiles then underwent digital stain normalization using the Reinhard method (44) and were subsequently standardized to give each image a mean of zero with a variance of one.

2.1.4. Constructing PDX drug response dataset

In constructing the drug response dataset, we populated samples from each group experiment with the corresponding response label. Each sample that was molecularly and histologically profiled consists of three feature types and a binary response value. The feature types include drug descriptors, GE, and histology tiles, as illustrated in Figure 4. Table 1 lists the summary statistics of the dataset.

The PDMR preclinical dataset contains experiments of single-drug treatments and drug pairs. In order to include both single-drug and drug-pair treatments in the dataset, and ensure consistent dimensionality of drug features, we homogenized single-drug treatments by duplicating drug descriptors to form a pseudo drug-pair that includes two identical drug feature vectors. In this case, the samples of single-drug and drug-pair treatments will follow the same input dimensionality for all ML models. Moreover, because switching the position of drug features in drug-pair treatments should not change the drug response, we augmented all drug-pair samples by switching the position of the two drugs while keeping the drug response value unchanged. Such data augmentation doubles the number of drug-pair samples in the dataset.

Following the integration of group samples into the dataset and the augmentation of drug-pair treatments, the drug response dataset contains 6,962 samples. The total number of treatment groups in the dataset is 959 with 917 non-response and 42 response groups. The dataset contains three feature types (modalities): two vectors of drug descriptors (a vector for each drug), GE profile, and histology tiles. Each sample consists of a unique combination of drug descriptors and GE profiles. However, each such sample contains multiple image tiles from a corresponding histology slide. Concretely, each sample consists of a GE profile, vector of descriptors for drug 1 and drug 2, and multiple image tiles as shown in Figure 4. We store the data in TFRecords (TensorFlow file format) which enables efficient data prefetching and loading, and therefore, considerably decreases the training and inference time.

2.1.5. Data splits

Data leakage can lead to overly optimistic predictions (45). Two primary characteristics of our dataset may lead to leakage if random splitting is used to generate training, validation, and test sets. First, a drug response label is assigned to all the samples in the entire treatment group. To prevent leakage, we make sure that samples from the same treatment group always appear together in one of the training, validation, or test sets. Second, the augmented drug-pair samples represent, in reality, the same experiment, and therefore, are also placed together when generating the splits. With this strategy, we generated 100 data splits for the analysis (10-fold cross-validation repeated ten times with different random seeds), where the tissue features (GE profiles and WSIs) of the same treatment group are kept together and not shared across training, validation, and test sets of each data split.

2.2. Prediction models

We explore the performance of MM-Net, shown in Figure 5, in predicting the drug response in PDXs. The model takes preprocessed feature sets as inputs, including drug descriptors, GE, and histology tiles, and passes them through subnetworks of layers. The encoded features from the subnetworks are merged *via* a concatenation layer and propagated to the output for predicting a binary drug response.

Since the dataset is highly redundant in terms of GE and drug features as shown in Figure 4 and Table 1 (there are 48 unique drug treatments and 487 unique expression sets), we use a single layer of trainable weights to encode these features with the goal to mitigate overfitting. The image tiles are passed through a subnetwork of

10.3389/fmed.2023.1058919

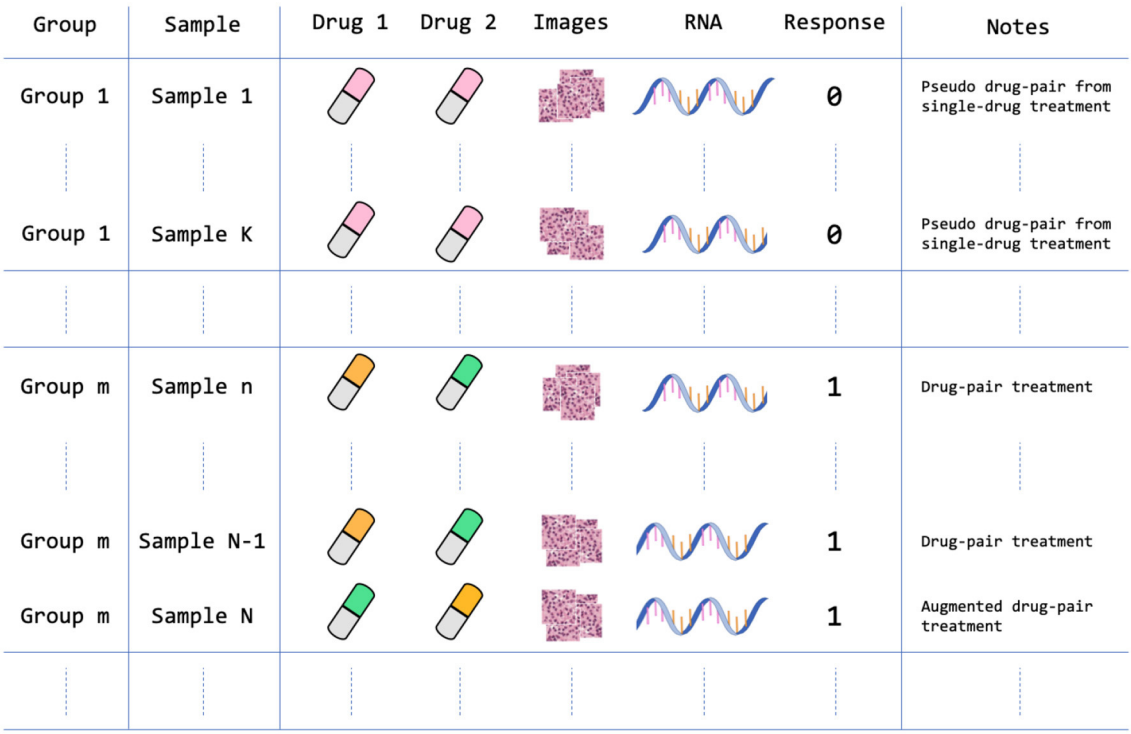


FIGURE 4
Data arrangement of the PDX drug response dataset. The dataset contains 959 treatment groups after homogenizing and augmenting the drug experiments as described in Section 2.1.4. For example, Sample 1 is a single-drug treatment that is structured as a pseudo drug-pair treatment where Drug 1 and Drug 2 features are the same feature vectors; Sample N is an augmented version of Sample N-1, in which the positions of drug feature vectors are switched. Note that each sample contains multiple histology image tiles that were extracted from a large WSI.

TABLE 1 Summary of the PDX drug response dataset used for building prediction models.

Patients	89
Primary tumor specimens	96
Single-drug treatments	12
Drug-pair treatments (excludes augmented samples)	36
Treatment groups	959
Gene expression profiles	487
Histology whole-slide images (WSIs)	487
Histology image tiles (extracted from WSIs)	177,468
Single-drug response values	2,556
Drug-pair response values (includes augmented values)	4,406
Drug response values	6,962

convolutional layers of the Xception model (46) with weights pre-trained on ImageNet (47). The output from the convolutional neural network (CNN) is passed through a series of dense layers before being concatenated with the encoded GE and drug descriptor representations.

We compare the performance of MM-Net with three unimodal baselines that use either GE or WSI as tumor features: (1) UME-Net, NN that uses GE, (2) UMH-Net, NN that uses histology tiles, and (3) LGBM, LightGBM that uses GE. Note that all models use drug descriptors.

2.3. Training and evaluation

We used a randomized search to obtain a set of hyperparameters (HPs) for UME-Net, including optimizer, learning rate, and layer dimensions that encode GE and drug descriptors. The values of these HPs were also used for MM-Net. A few remaining HPs that are unique to MM-Net were determined in a separate search using the MM-Net architecture. Another randomized search was performed to obtain the HPs for LGBM such as the number of leaves in the decision tree and the number of trees.

To mitigate overfitting, we used the early stopping mechanism in TensorFlow and LightBGM where model trainings terminate automatically if the predictions on a validation set have not been improved for a predefined number of training iterations. The early stopping parameter was set to 10 epochs for all NNs and 100 boosting rounds for the LGBM. All NNs were trained for 400 training epochs which triggered early stopping and ensured model convergence. To further address overfitting, we applied standard image augmentation methods such as rotation and horizontal flipping to histology tiles during the training of MM-Net and UMH-Net.

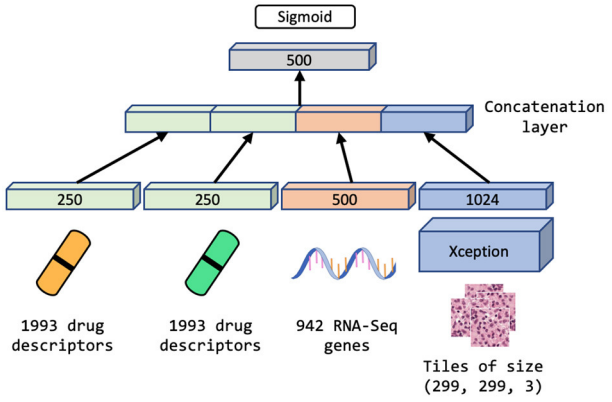


FIGURE 5
Multimodal neural network (MM-Net) learns from drug descriptors, gene expressions, and tiles generated from whole-slide images, to predict drug response in PDXs.

Since the dataset is highly imbalanced in terms of drug response distribution, we used a weighted loss function that penalizes more heavily incorrect predictions of the response samples as opposed to the non-response samples. For training MM-Net, we used only 10% of the image tiles that were available in each WSI, because our preliminary experiments revealed that the prediction performance does not improve if additional tiles are used. The 10% of the tiles have been drawn at random from each WSI.

For model evaluation, we used three performance metrics for binary classification tasks, including Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). We calculated the metrics based on a test set of each one of the 100 splits. To compute each metric for a given test set, we aggregated all the sample predictions in the test set. In the case of baseline models where each tumor sample is represented by a GE vector, the prediction model generates in a single probability value for each sample. However, in the case of MM-Net where each tumor sample is represented by multiple image tiles in addition to the GE profile, the prediction model generates a single probability value for each image tile which results in multiple predicted probability values for a single sample. To conform with the output of the baseline models, the tile predictions from MM-Net were aggregated *via* mean to provide a single probability value for each sample. Note that while only 10% of the available tiles were used for training MM-Net, all tiles in the test set were used to compute predictions and subsequently obtain the performance metrics.

3. Results

A total of six prediction models were analyzed, as summarized in Table 2. The models differ in terms of the feature sets and the samples that were used for training and validation (binary columns in Table 2). All models were evaluated across the same 100 data splits. Figure 6 shows the performance metrics, including MCC, AUPRC, and AUROC where each data point is a metric value calculated for a given split. The average score of each model was aggregated *via* mean across the splits for each metric (listed in Table 2).

In constructing the drug response dataset, we used two approaches to increase the number of response values, as described in Section 2.1.4. We analyzed the effect of these two methods on the prediction performance by comparing three unimodal NNs that were trained with GE and drug descriptors on subsets of the dataset: (1) UME-Net, trained with the full dataset that includes the original and the augmented drug-pair treatments as well as single-drug treatments, (2) UME-Net$_{pairs}$, trained with only drug-pair samples which include the original and the augmented samples, and (3) UME-Net$_{org}$, trained with the original single-drug and drug-pair samples that exclude the augmented drug-pair samples. Figure 6 shows the prediction performance of using the different training subsets across the data splits where each data point is a metric value calculated for a given split. While the distribution of scores across the splits is quite substantial, removing either subset of samples (i.e., single-drug samples or augmented drug-pairs) results in a significant decline in prediction performance, as demonstrated by two statistical tests, including paired t-test and Wilcoxon signed-rank test ($p < 0.05$). In other words, augmentation methods lead to a significant improvement in performance of the NNs when trained with GE and descriptors. Hence, we used the full set of the available training samples for the analyses of multimodal learning.

The MM-Net architecture, shown in Figure 5, was compared against three baseline models, including UME-Net, UMH-Net, and LGBM. The performance metrics represent the ability of the models to generalize to a test set of unseen observations. Statistical tests (paired t-test and Wilcoxon signed-rank test) were performed to assess statistical difference between the models across the 100 data splits. Based on the aggregated MCC score, MM-Net statistically significantly outperforms all the baselines ($p < 0.05$ for both tests) except for UME-Net. When considering the AUPRC, MM-Net outperforms UMH-Net but there is no significant difference when comparing MM-Net with UME-Net or LGBM. No statistical significance was observed when comparing MM-Net with the other models. All the performance scores and statistical tests are provided in Supplementary Table S1.

As compared to DL models that are trained with cell line data, all models in Table 2 generally exhibit a relatively lower performance. For example, the average AUPRC is around 0.27 in Table 2 (all precision-recall curves are provided in the Supplementary material) but models trained on cell lines can exhibit AUPRC of 0.7 and above (48). Yet, we can observe a large spread of scores for all models and metrics (Figure 6). This indicates that for certain data splits, the models exhibit very high generalization performance, while for other splits, the models almost entirely fail to learn a meaningful mapping function for predicting drug response. In practical scenarios, where the goal is to design a highly generalizable model, a careful analysis should be conducted to determine the training and validation sets that adequately represent the test set. Subsequently, the threshold of the classifier can be determined depending on the error rate that the stakeholders can tolerate which will ultimately depend on the specific application that the model was designed for (e.g., precision oncology, drug development, etc.). In our analysis, however, the objective was to conduct large-scale trainings across multiple data splits and examine the overall capacity of MM-Net in predicting

TABLE 2 Performance metrics including MCC, AUPRC, and AUROC are listed for drug response prediction models (UME-Net, UME-Net$_{org}$, UME-Net$_{pairs}$, UMH-Net, MM-Net, and LGBM).

Model	WSI	GE	Single-drug	Drug-pair	Augmented drug-pairs	MCC	AUPRC	AUROC
UME-Net	–	v	v	v	v	0.2958	**0.2996**	0.8047
UME-Net$_{org}$	–	v	v	v	–	0.2391	0.2610	0.7766
UME-Net$_{pairs}$	–	v	–	v	v	0.2039	0.2355	0.7423
UMH-Net	v	–	v	v	v	0.2124	0.2303	0.7977
MM-Net	v	v	v	v	v	**0.3102**	0.2974	0.7978
LGBM	–	v	v	v	v	0.2594	0.2784	**0.8065**

As summarized in the binary columns (with v's), the differences between the models are in the feature sets (GE, WSI, or both) and the samples. To compute the average score for each metric and model, the predictions were aggregated *via* mean across the data splits.
Best scores per metric are shown in bold.

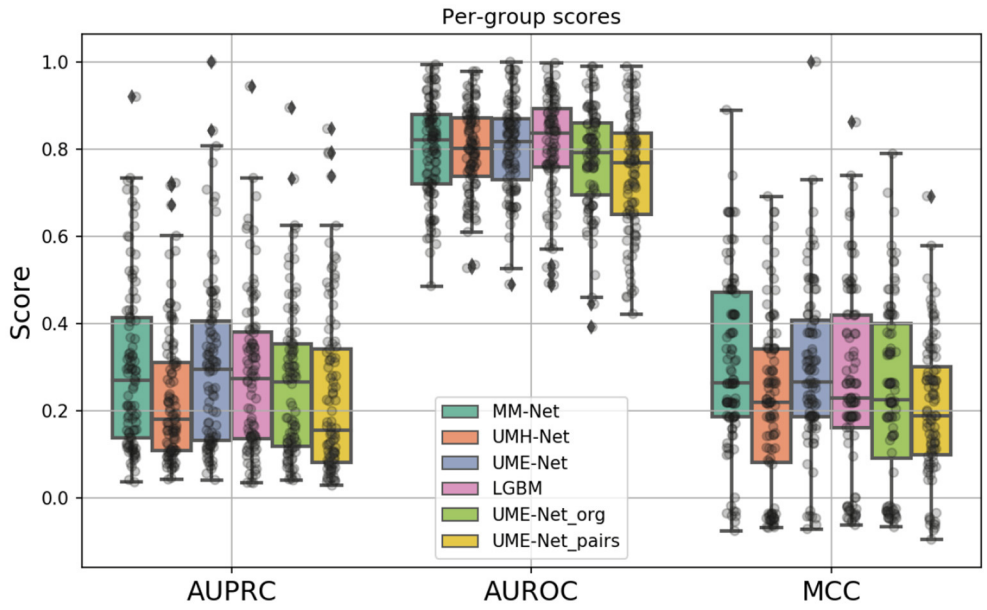


FIGURE 6
Boxplots showing the distribution of scores for the investigated drug response prediction models. The differences between the models are summarized in Table 2.

drug response across multiple cancer types and treatments. We observe that for certain dataset splits, MM-Net outperforms the baselines but for other splits it underperforms, as shown in Figure 7. An in-depth investigation is further required to understand in which cases MM-Net trained with WSI significantly improves prediction generalization.

4. Discussion

In this study, we investigated data augmentation methods and a multimodal architecture for predicting drug response in PDXs. We utilized the PDMR drug response dataset of single-drug and drug-pair treatments that were generated in controlled group experiments with PDX models of multiple cancer types. To assess the utility of the proposed methods, we conducted a large-scale analysis by training MM-Net and three baseline models over 100 data splits that contain GE profiles, histology image tiles, and molecular drug descriptors. We demonstrated that data augmentation methods lead to a significant improvement in drug response predictions across all performance metrics (MCC, AUPRC, and AUROC). Alternatively, the MM-Net model exhibits statistically significant improvement in prediction performance only when measured by the MCC.

The data splitting strategy and the choice of performance metrics play an important role in the downstream analysis when evaluating the utility of prediction models for practical applications. The dataset size and its diversity in terms of PDX models and treatments allowed us to generate multiple data splits while mitigating data leakage between training, validation, and test sets. Since each split comprises unique GE and histology images, we face a challenging prediction problem as opposed to a situation

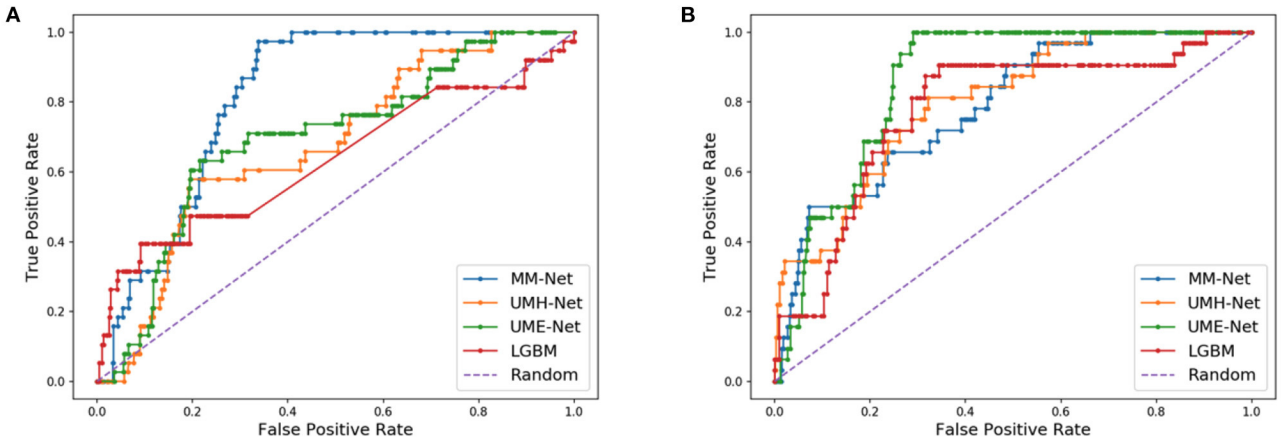


FIGURE 7
Receiver operating characteristic (ROC) curves for two different data splits. **(A)** MM-Net outperforms the baseline models. **(B)** MM-Net underperforms the baseline models.

in which the samples are randomly split. Alternative splitting strategies may involve a careful choice of a single test set with the goal to reduce the distributional shift between training and test sets (49). Instead of carefully assembling the most representative test set, we chose to conduct a large-scale analysis to assess the empirical range of prediction performance with NNs and LGBM. The results show a large spread of scores across the splits, indicating that for certain data splits the models exhibit high prediction performance, while for other splits, the learning of models fails. When we specifically focus on the performance of MM-Net as compared with the baselines, we discover that in 46 out of the 100 splits, the MM-Net outperformed the UME-Net baseline. This observation implies that for certain data splits, the histology images boost the generalization performance of the prediction model, and therefore, its potential utility in preclinical and clinical settings. A further investigation is required to better understand the cases and data characteristics in which histology images improve response prediction.

Technological progress in digital pathology and high-throughput omic profiling have led researchers to generate big data repositories of histology images and omics data, as well as algorithms to jointly analyze these diverse data types. Several papers have explored multimodal architectures that combine histology images with omics data for predicting survival outcomes of cancer patients. Mobadersany et al. demonstrated that a CNN-based supervised learning model combined with cox regression accurately predicts survival outcomes of glioma patients from histology and mutation data (21). Cheerla et al. proposed an unsupervised learning method to learn a low-dimensional representation for each feature type and, consequently, concatenated the learned representations to predict survival outcome of cancer patients (22). They have also demonstrated on 20 cancer types that a custom dropout layer that randomly drops an entire feature vector improves predictions. Building upon existing works, Chen et al. introduced a supervised architecture for multimodal fusion of histology and omics data to predict patient survival and applied their method to glioma and clear cell renal cell carcinoma patients (23). The model uses graph convolutional network (GCN) and CNN to encode histology image data and feed-forward network for mutation data. Each encoded feature vector is passed through an attention mechanism and subsequently fused *via* a Kronecker product. The cox regression is finally used to predict patient survival. While these papers do not consider drug treatments in their analysis, they exploit modern approaches for enhancing predictions of multimodal NNs with histology and omics data and can be further explored for drug response prediction.

A wide spectrum of methods is available in vision applications for inducing changes in images that allow for data augmentation (50). In this study, we exploit the lack of invariance to permutation as the means to augment the sample size. Recently, additional methods have been proposed for augmenting transcriptomic data which can potentially be combined and provide further improvement in predicting drug response (51–53). Considering the scale of existing PDX datasets, data fusion and augmentation provide promising research directions for enhancing predictive capabilities with PDXs. However, special care should be taken because high-dimensional feature sets can often lead to severe overfitting and poor generalization. Presumably to mitigate overfitting, Nguyen et al. (11) and Kim et al. (12) used feature selection methods to reduce the dimensionality of PDX data while considering a single omics feature type at a time. Therefore, multimodal learning exhibits a tradeoff between enriching the feature space *via* multimodal fusion and overfitting. To alleviate this tradeoff, alternative methods can be explored to reduce the feature space while incorporating multiple feature types (54). With the methods proposed in this study and ongoing research into novel augmentation and fusion techniques, PDX pharmaco-omic datasets may become more suitable for modern deep learning techniques and further increase interest for building prediction models to advance precision oncology.

5. Conclusions

Deep learning methods have shown promising results in predicting drug response in cancer cell lines. While PDXs are presumed to better mimic human cancer, drug response datasets with this cancer model are substantially smaller as compared to cell line datasets. We investigate multimodal learning and data augmentation methods to address the challenge of limited drug response sample size. Our results suggest that data augmentation and integration of histology images and gene expressions can improve prediction performance of drug response in PDXs.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://pdmr.cancer.gov/database/default.htm.

Ethics statement

The animal study was reviewed and approved by FNLCR is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International and follows the Public Health Service Policy for the Care and Use of Laboratory Animals. All studies were conducted according to an approved animal care and use committee protocol in accordance with procedures outlined in the Guide for Care and Use of Laboratory Animals 8th Edition.

Author contributions

APa, TB, YZ, and RS: conceptualization. APa, YZ, JDol, SK, MS, and YE: data curation. APa and YZ: formal analysis. RS: funding acquisition. APa and YE: investigation and methodology. TB, APe, YE, and RS: project administration and resources. APa, YZ, and JDol: software. YE and JDor: validation. APa: writing—original draft preparation. APa, TB, YZ, JDol, APe, YE, and RS: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This has been funded in whole or in part with Federal funding by the NCI-DOE Collaboration established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health, Cancer Moonshot Task Order No. 75N91019F00134 and under Frederick National Laboratory for Cancer Research Contract 75N91019D00024. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract 89233218CNA000001, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. Argonne National Laboratory's work on the Preclinical Modeling High-Performance Computing Artificial Intelligence project was supported in whole or in part by Leidos Biomedical Research, Inc., under PRJ1009868.

Conflict of interest

YE was employed by Leidos Biomedical Research, Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1058919/full#supplementary-material

References

1. Ballester PJ, Stevens R, Haibe-Kains B, Huang RS, Aittokallio T. Artificial intelligence for drug response prediction in disease models. *Brief Bioinform*. (2021) 2021:bbab450. doi: 10.1093/bib/bbab450

2. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol*. (2020) 4:19. doi: 10.1038/s41698-020-0122-1

3. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, Hon C, Mammoliti A, Nair SK, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief Bioinform*. (2021) 22:bbab294. doi: 10.1093/bib/bbab294

4. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*. (2013) 108:479–85. doi: 10.1038/bjc.2012.581

5. Seoane J, Mattos-Arruda LD. The challenge of intratumour heterogeneity in precision medicine. *J Internal Med*. (2014) 276:41–51. doi: 10.1111/joim.12240

6. Evrard YA, Srivastava A, Randjelovic J, Doroshow JH, Dean DA, Morris JS, et al. Systematic establishment of robustness and standards in patient-derived xenograft experiments and analysis. *Cancer Res*. (2020) 80:2286–97. doi: 10.1158/0008-5472.CAN-19-3101

7. Goto T. Patient-derived tumor xenograft models: toward the establishment of precision cancer medicine. *J Pers Med*. (2020) 10:64. doi: 10.3390/jpm10030064

8. Yoshida GJ. Applications of patient-derived tumor xenograft models and tumor organoids. *J Hematol Oncol*. (2020) 13:4. doi: 10.1186/s13045-019-0829-z

9. Chiu YC, Chen HIH, Gorthi A, Mostavi M, Zheng S, Huang Y, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. *Brief Bioinform*. (2020) 21:2066–83. doi: 10.1093/bib/bbz144

10. Zhu Y, Brettin T, Evrard YA, Partin A, Xia F, Shukla M, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep*. (2020) 10:1–11. doi: 10.1038/s41598-020-74921-0

11. Nguyen LC, Naulaerts S, Bruna A, Ghislat G, Ballester PJ. Predicting cancer drug response *in vivo* by learning an optimal feature selection of tumour molecular profiles. *Biomedicines*. (2021) 9:319. doi: 10.3390/biomedicines9101319

12. Kim Y, Kim D, Cao B, Carvajal R, Kim M. PDXGEM: patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients. *BMC Bioinform*. (2020) 21:288. doi: 10.1186/s12859-020-03633-z

13. Migliardi G, Sassi F, Torti D, Galimi F, Zanella ER, Buscarino M, et al. Inhibition of MEK and PI3K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clin Cancer Res*. (2012) 18:2515–25. doi: 10.1158/1078-0432.CCR-11-2683

14. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med*. (2015) 21:1318–25. doi: 10.1038/nm.3954

15. Smith SC, Baras AS, Lee JK, Theodorescu D. The COXEN principle: translating signatures of *in vitro* chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer. *Cancer Res*. (2010) 70:1753–8. doi: 10.1158/0008-5472.CAN-09-3562

16. Zhu Y, Brettin T, Evrard YA, Xia F, Partin A, Shukla M, et al. Enhanced co-expression extrapolation (COXEN) gene selection method for building anti-cancer drug response prediction models. *Genes*. (2020) 11:1070. doi: 10.3390/genes11091070

17. Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, et al. Deep learning scaling is predictable, empirically. *arXiv:171200409 [cs, stat*]. (2017). doi: 10.48550/arXiv.1712.00409

18. Rosenfeld JS, Rosenfeld A, Belinkov Y, Shavit N. A constructive prediction of the generalization error across scales. In: *ICLR*. (2020). Available online at: https://iclr.cc/Conferences/2020

19. Partin A, Brettin T, Evrard YA, Zhu Y, Yoo H, Xia F, et al. Learning curves for drug response prediction in cancer cell lines. *BMC Bioinform*. (2021) 22:1. doi: 10.1186/s12859-021-04163-y

20. Xia F, Allen J, Balaprakash P, Brettin T, Garcia-Cardona C, Clyde A, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Brief Bioinform*. (2022) 23:bbab356. doi: 10.1093/bib/bbab356

21. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci USA*. (2018) 115:E2970-E2979. doi: 10.1073/pnas.1717139115

22. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*. (2019) 35:446–54. doi: 10.1093/bioinformatics/btz342

23. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging*. (2019) 44:757–70. doi: 10.1109/TMI.2020.3021387

24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems.*. Curran Associates, Inc. (2017) 30, 3146–3154. Availble online at: http://citebay.com/how-to-cite/light-gradient-boosting-machine/

25. Mer AS, Ba-Alawi W, Smirnov P, Wang YX, Brew B, Ortmann J, et al. Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res*. (2019) 79:4539–50. doi: 10.1158/0008-5472.CAN-19-0349

26. Schwartz LH, Litière S, Vries Ed, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1–Update and clarification: From the RECIST committee. *Eur J Cancer*. (2016) 62:132–7. doi: 10.1016/j.ejca.2016.03.081

27. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. (2014) 32:1202–12. doi: 10.1038/nbt.2877

28. Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Jamie O, et al. Deep learning methods for drug response prediction in cancer: predominant and emerging trends. *Front Med*. (2023) 10:1086097. doi: 10.3389/fmed.2023.1086097

29. Ahmed KT, Park S, Jiang Q, Yeu Y, Hwang T, Zhang W. Network-based drug sensitivity prediction. *BMC Med Genomics*. (2020) 13:193. doi: 10.1186/s12920-020-00829-3

30. Jia P, Hu R, Pei G, Dai Y, Wang YY, Zhao Z. Deep generative neural network for accurate drug response imputation. *Nat Commun*. (2021) 12:1740. doi: 10.1038/s41467-021-21997-5

31. Hostallero DE, Li Y, Emad A. Looking at the BiG picture: incorporating bipartite graphs in drug response prediction. *Bioinformatics*. (2022) 38:3609–20. doi: 10.1093/bioinformatics/btac383

32. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. (2017) 171:1437–52. doi: 10.1016/j.cell.2017.10.049

33. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: improving drug response prediction *via* modeling of drug perturbation effects. *Bioinformatics*. (2019) 35:3743–51. doi: 10.1093/bioinformatics/btz158

34. Deng L, Cai Y, Zhang W, Yang W, Gao B, Liu H. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. *J Chem Inf Model*. (2020) 60:4497–505. doi: 10.1021/acs.jcim.0c00331

35. Zuo Z, Wang P, Chen X, Tian L, Ge H, Qian D. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC Bioinform*. (2021) 22:434. doi: 10.1186/s12859-021-04352-9

36. Kim S, Bae S, Piao Y, Jo K. Graph convolutional network for drug response prediction using gene expression data. *Mathematics*. (2021) 9:772. doi: 10.3390/math9070772

37. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. (2017) 45:D777–83. doi: 10.1093/nar/gkw1121

38. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*. (2020) 36:i911–8. doi: 10.1093/bioinformatics/btaa822

39. Zhu Y, Ouyang Z, Chen W, Feng R, Chen DZ, Cao J, et al. TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics*. (2021) 38:461–8. doi: 10.1093/bioinformatics/btab650

40. Ma T, Liu Q, Li H, Zhou M, Jiang R, Zhang X. DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinform*. (2022) 23:129. doi: 10.1186/s12859-022-04664-4

41. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. (2017) 7:16878. doi: 10.1038/s41598-017-17204-5

42. Dolezal JM, Trzcinska A, Liao CY, Kochanny S, Blair E, Agrawal N, et al. Deep learning prediction of BRAF-RAS gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. *Modern Pathol*. (2021) 34:862–74. doi: 10.1038/s41379-020-00724-3

43. Dolezal JM, Kochanny S, Howard F. Jamesdolezal/slideflow: Slideflow 1.0 - Official Public Release (1.0.4). *Zenodo*. (2021). doi: 10.5281/zenodo.5718806

44. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graphics Appl*. (2001) 21:34–41. doi: 10.1109/38.946629

45. Kaufman S, Rosset S, Perlich C. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discovery Data*. (2012) 6:579. doi: 10.1145/2382577.2382579

46. Chollet F. Xception: deep learning with depthwise separable convolutions. *arXiv:161002357 [cs*]. (2017) doi: 10.1109/CVPR.2017.195

47. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE (2009). p. 248–55.

48. Jiang L, Jiang C, Yu X, Fu R, Jin S, Liu X. DeepTTA: a transformer-based model for predicting cancer drug response. *Brief Bioinform*. (2022) 23:bbac100. doi: 10.1093/bib/bbac100

49. Stacke K, Eilertsen G, Unger J, Lundstrom C. Measuring domain shift for deep learning in histopathology. *IEEE J Biomed Health Inform*. (2021) 25:325–36. doi: 10.1109/JBHI.2020.3032060

50. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. (2019) 6:60. doi: 10.1186/s40537-019-0197-0

51. Moreno-Barea FJ, Jerez JM, Franco L. GAN-Based data augmentation for0prediction improvement using gene expression data in0cancer. In: Groen D, de Mulatier C, Paszynski M, Krzhizhanovskaya VV, Dongarra JJ, Sloot PMA, editors. *Computational Science-ICCS 2022 Lecture Notes in Computer Science*. Cham: Springer International Publishing (2022). p. 28–42.

52. Kircher M, Chludzinski E, Krepel J, Saremi B, Beineke A, Jung K. Augmentation of transcriptomic data for improved classification of patients with respiratory diseases of viral origin. *Int J Mol Sci*. (2022) 23:2481. doi: 10.3390/ijms2305 2481

53. Huang HH, Rao H, Miao R, Liang Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. *BMC Bioinform*. (2022) 23:353. doi: 10.1186/s12859-022-04887-5

54. Zhu Y, Brettin T, Xia F, Partin A, Shukla M, Yoo H, et al. Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep*. (2021) 11:11325. doi: 10.1038/s41598-021-90923-y

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 06 April 2023
DOI 10.3389/fonc.2023.1116210

Check for updates

OPEN ACCESS

EDITED BY
Dinler Amaral Antunes,
University of Houston, United States

REVIEWED BY
Praveen Nekkar Rao,
University of Waterloo, Canada
David Robert Shorthouse,
University College London,
United Kingdom
Dahmane Oukrif,
University College London,
United Kingdom

*CORRESPONDENCE
Hiroshi Haeno
haeno@rs.tus.ac.jp

SPECIALTY SECTION
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

RECEIVED 05 December 2022
ACCEPTED 23 March 2023
PUBLISHED 06 April 2023

CITATION
Abubakar SD, Takaki M and Haeno H (2023) Computational modeling of locoregional recurrence with spatial structure identifies tissue-specific carcinogenic profiles.
Front. Oncol. 13:1116210.
doi: 10.3389/fonc.2023.1116210

COPYRIGHT
© 2023 Abubakar, Takaki and Haeno. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Computational modeling of locoregional recurrence with spatial structure identifies tissue-specific carcinogenic profiles

Sharafudeen Dahiru Abubakar[1], Mitsuaki Takaki[2] and Hiroshi Haeno[1]*

[1]Research Institute for Biomedical Science, Tokyo University of Science, Noda, Japan, [2]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

Introduction: Local and regional recurrence after surgical intervention is a significant problem in cancer management. The multistage theory of carcinogenesis precisely places the presence of histologically normal but mutated premalignant lesions surrounding the tumor - field cancerization, as a significant cause of cancer recurrence. The relationship between tissue dynamics, cancer initiation and cancer recurrence in multistage carcinogenesis is not well known.

Methods: This study constructs a computational model for cancer initiation and recurrence by combining the Moran and branching processes in which cells requires 3 or more mutations to become malignant. In addition, a spatial structure-setting is included in the model to account for positional relativity in cell turnover towards malignant transformation. The model consists of a population of normal cells with no mutation; several populations of premalignant cells with varying number of mutations and a population of malignant cells. The model computes a stage of cancer detection and surgery to eliminate malignant cells but spares premalignant cells and then estimates the time for malignant cells to re-emerge.

Results: We report the cellular conditions that give rise to different patterns of cancer initiation and the conditions favoring a shorter cancer recurrence by analyzing premalignant cell types at the time of surgery. In addition, the model is fitted to disease-free clinical data of 8,957 patients in 27 different cancer types; From this fitting, we estimate the turnover rate per month, relative fitness of premalignant cells, growth rate and death rate of cancer cells in each cancer type.

Discussion: Our study provides insights into how to identify patients who are likely to have a shorter recurrence and where to target the therapeutic intervention.

KEYWORDS

computational modeling, cancer initiation, cancer recurrence, field cancerization, stochastic processes

Introduction

Cancers are dynamic cells whose features favors cellular proliferation, differentiation and movement while restricting cell death and tissue stability (1). Surgery is a potent curative tool for managing cancers, however, local recurrence has remained a clinically significant problem in most cancer types (2–4). Local recurrence rates could be as high as about 85% (5) for ovarian cancer, or 30% in non-small cell lung cancer (NSCLC) (6, 7) and as low as between 8% (8) to 16.5% (9) in breast cancer. Advanced surgical techniques, chemotherapy, radiotherapy (4) as well as endocrine therapy (10) are being used to minimize locoregional recurrence but "minimal" improvements and treatment-related mortality has highlighted the need for a better understanding and strategy for local recurrence (11, 12).

Since its introduction in 1953 (13), field cancerization has been recognized as a major cause of local recurrence (14). Field cancerization is the presence of "histologically normal" cells surrounding cancer cells that have acquired some but not all the genetic and phenotypic traits required for malignancy in a tissue (15, 16). These cancerized cells may have a survival or growth advantage and does serve as a hotbed for recurrent tumors as only a small number of additional steps are needed for cancer initiation. Recent advances in molecular, genomic and bulk sequencing techniques have supported the role of field cancerization (17). In breast cancer, microsatellite markers, epigenetic aberrations, transcriptomic deregulations and hTERT overexpression have been detected in histologically normal mammary tissues (18, 19). In head and neck cancer, loss of heterozygosity of chromosome 9p and telomere dysregulation were commonly observed in benign squamous hyperplasia (20, 21). In colon cancer patients with Crohn's ileocolitis, the same mutations of KRAS, CDKN2A, and TP53 were observed within neoplasia and non-tumor epithelium (22, 23). In Non-small cell lung cancer, miRNA dysfunction has been shown at the level of the tumor and cancerized field (24). Residual tumor (25), anesthesia choice (26) and CTCs (27) have been shown to have minimal impact on cancer recurrence; therefore, a proper understanding of the field cancerization formation process will contribute to the estimation of the risk of locoregional recurrence and the development of optimal treatment in each tissue.

Several Theoretical studies have shed light on field cancerization impacts on cancer initiation (15). Jeon et al. examined the multistage clonal expansion model by employing the Poisson process to consider the effects of premalignant cells on cancer initiation (28). The model was applied clinically to predict the long-term impact of ablative treatments on reducing esophageal adenocarcinoma incidence in Barrett's esophagus (29). Foo et al. developed a spatial evolutionary framework to determine the size distribution of histologically undetectable premalignant fields during diagnosis (30). This model was applied to the head and neck cancer and revealed that the patient's age was a critical predictor of the size and multiplicity of precancerous lesions (31). These findings are in agreement with bulk sequencing data that shows the accumulation of cancer-related mutations as we age (32). A 2-step tumor initiation model provides insights into the relationship between different tissue kinetic parameters and the incidence of recurrent cancers (33) by using public datasets from the cancer genome atlas (TCGA), a valuable resource for genomic and clinical data analysis (34, 35) but fails to account the varying number of mutational hits required for carcinogenesis (36–38). The cancer genome atlas (TCGA) is a rich computational resource for the genomic and mutational data for different cancer types (34, 35) and will be helpful in validating our understanding of field cancerization.

This study developed a novel computational model of multi-stage cancer initiation and recurrence with spatial structure. We employed a combined stochastic model of Moran and a branching process to represent tissue and tumor dynamics, respectively, in order to observe cancer initiation and relapse after surgical resection of the first tumor *in silico*. Particularly, we focused on the relationship between the tissue compositions at the time of surgery and the time until the emergence of recurrent tumors. This model builds upon our previous work (33) by expanding the number of mutation steps for carcinogenesis *via* adding cell types as well as incorporating the spatial structure setting. Moreover, based on the public clinical datasets for locoregional recurrence rates, we succeeded in identifying tissue-specific carcinogenic parameters for various cancer types. Our approach provided insights on how to predict the time of recurrence from the tissue dynamics at the time of surgery and how to intervene patients to prevent the recurrence.

Materials and methods

Computational model

This model employs the multi-stage carcinogenesis concept. As tissues might require anywhere between 2 to 8 driver mutations (denoted by S) for malignant transformation (37), we first identify different cell types that can lead to a malignant transformation based on number of driver mutations. Let us visualize the dynamics of 5 types of cells in a tissue (Figure 1). "Type 0", "Type 1", "Type *K*", "Type *S*-1" and "Type *S*" represent normal healthy cells with no mutation; premalignant cells with one cancer-related mutation; premalignant cells with *K* cancer-related mutation, premalignant cells with *S*-1 cancer-related mutations and cancer cells with *S* cancer-related mutations, respectively. Emergence of cancer cells must be preceded by that of premalignant cells with mutations from Type 1 cell to Type *S*-1 cell. Type *K* cell may or may not be present depending on number of mutations required for carcinogenesis. We assume that a normal healthy tissue consists of Type 0, Type 1, Type *K* and Type *S*-1 cells undergoing cellular turnover with a small probability of a mutation. Moran process is employed to consider the tissue turnover dynamics, where the total number of Type 0,

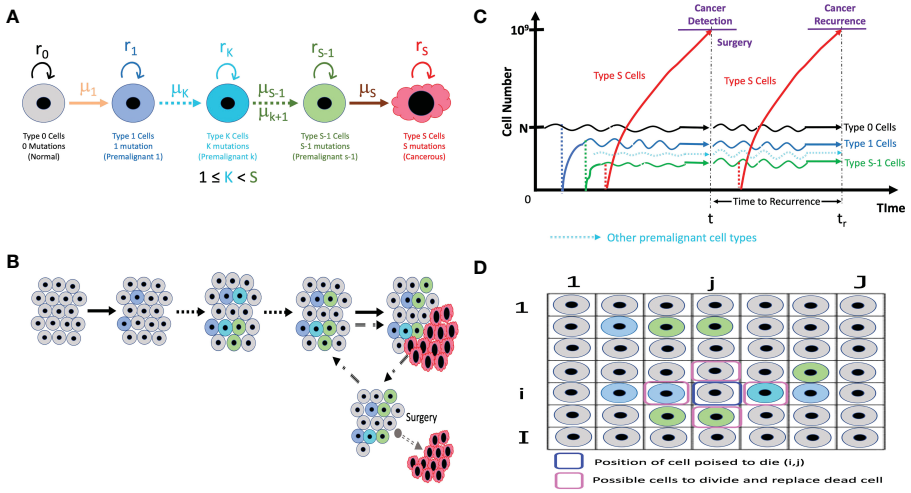


FIGURE 1
The illustrative representation of our models **(A)** The different cell types in our models with its own mutation rate (μ) and fitness (r). Type K cells may not be applicable if only 3 mutations or less are needed for carcinogenesis. **(B)** In a normal tissue composed of Type 0, Type 1, Type K, and Type S-1 cells, cell turnover is conducted according to the Moran process, and the number of cells is kept constant. If a Type S cell emerges, it proliferates without limit and can be detected and primed for surgery when the cancer cell number reaches 10^9. **(C)** At surgical intervention, all the Type S cells are resected while the number of Type 0, Type 1, Type K (if present) and Type S-1 cells remaining in a tissue are preserved. The time until the next Type S population reaches 10^9 is measured as time to recurrence. **(D)** The spatial structure integration in the model accounts for the positional relation between a cell poised to die and the possible cells that can divide to replace the dead cells.

Type 1, Type K and Type S-1 cells is kept constant as N (39). The turnover rate of a whole tissue is defined by d. Type S cells are considered as uncontrolled, highly proliferating cancer cells. The branching process is employed to consider the process of Type S proliferation (40).

Initially, N Type 0 cells occupy the tissue. There is a rare chance of a mutation every time a cell divides, and a daughter cell may change into a Type 1 cell with a mutation rate, μ_1. Mutation rate, μ, refers to the sum total of the genomic or epigenetic factors affecting change from one cell type to another (41). When a cell dies, a cell to be divided in a tissue is selected depending on the cell fitness, r. The fitness of Type 0, Type 1, Type K, and Type S-1 are denoted by r_0, r_1, r_K and $r_{S\text{-}1}$, respectively. Cell fitness, r refers to the transcriptional and metabolic potential of a cell type to "out-compete" other cell types (42). A cell could divide to give rise to the same cell type or mutate to another cell type. When a Type 1 cell divides with a mutation, a daughter cell may change into a Type K cell with mutation rate μ_K (if more mutations are needed) or a Type S-1 cell with a mutation rate, $\mu_{S\text{-}1}$ (if additional mutation steps are not needed). Intermediate cell type, Type K, becomes Type S-1 after sequential accumulation of mutations. Finally, a Type S-1 cell is capable of mutating to become a malignant cell – Type S cell based on the mutation rate from Type S-1 to Type S cell, μ_S. Once a Type S cell appears, the cells proliferate indefinitely based on the growth rate of Type S cells, r_S, disrupting tissue dynamics and homeostasis. Type S cancer cells are "super-competitors" with outstanding metabolic prowess and assumed to increase exponentially with a net growth rate of r_S - $d_S > 0$, where d_S is a death rate.

We propose that the most important premalignant cells are the Type 1 cell that have acquired the first driver mutation and the Type S-1 cell that needs just one more driver mutation to become a cancer cell. These cells look phenotypically normal and are not regarded as important clinically but their genetic features are indispensable in cancer formation. As a result, the cell fitness of these 2 cell types must be taken into account in all computational analysis. For scenarios where additional mutations and more cell types are needed, we approximate intermediate mutational steps between Type 1 and Type S-1 by adjusting the values of $\mu_{S\text{-}1}$ so that a low mutation rate, $\mu_{S\text{-}1}$, represents additional mutational steps. So, our computational analysis will be executed to account for the most important mutational events that affect cell fitness and all the mutation rates that can affect the number of steps required for carcinogenesis. In other words, we skip the state of Type K cell and a mutation rate stands in for the number steps.

The net growth of Type 0, Type 1, and Type S-1 cells is zero (equal frequency of cell division and death), while that of Type S cells is positive. Type 0, Type 1, and Type S-1 cells consist of a healthy tissue based on the Moran process, so r_0, r_1 and $r_{S\text{-}1}$ are parameters to determine status of dividing cell and which daughter cell are obtainable at the time of a cell division. Alternatively, r_S is the growth rate, which determines the average number of increases in Type S cells during a unit time. When the number of Type S cells reaches 10^9 at the first time, all the Type S cells are discarded to represent surgical resection, whereas the number of Type 0, Type 1 and Type S-1 cells in a tissue is preserved so that the time until the emergence of the recurrent tumor is influenced by the frequency of residual Type 0, Type 1 and especially, Type S-1 cells. Since the conversion from the number of cells to the tumor volume is frequently done using the following relationship as 10^9 cells in a 1 cm^3 tumor, the time of surgery in this model is conducted when the size of the tumor becomes 1 cm^3. We describe it as "time of cancer detection". After the first treatment, the simulation continues until the next Type S cell appears from the tissue and the number reaches 10^9 again, representing the recurrence of the tumor after surgery.

10.3389/fonc.2023.1116210

Simulation framework

To integrate the Moran process and branching process, we adopted stochastic simulations based on Gillespie's algorithm (43) as follows: We firstly considered three events: (i) cell turnover in a healthy tissue as per Moran process, (ii) birth of a Type *S* cell as per Branching process, and (iii) death of a Type *S* cell. The rates of each event at time t is given by (i) dN (ii) $r_S X_S(t)$, and (iii) $d_S X_S(t)$, respectively. Here r_S, d_S, and $X_S(t)$ are a proliferation rate, a death rate, and the number of Type *S* cells at time t, respectively. Then an average time until one of the three events happens, ΔT, is given by

$$\Delta T = \frac{1}{dN + r_S X_S(t) + d_S X_S(t)}. \quad (1)$$

Let us first consider the case where a cell turnover happens. The probability that a cell turnover happens in ΔT is given by $dN \cdot \Delta T$. In our model, a cell turnover in a healthy tissue is governed by a cell death. When one of N cells is randomly selected as a cell to die, and another cell is chosen to divide within the same time step to complete cell turnover. In a healthy tissue, there are three types of cells, corresponding to the number of acquired mutations, Type 0, Type 1, and Type *S*-1.The number of each cell type is denoted by X_0, X_1 and $X_{S\text{-}1}$, respectively. In brief, there are several possibilities of tissue composition transitions in the tissue dynamics and we consider the six events that affect the cell type composition of a tissue: (i) a type 0 cell increases by one while a type 1 cell decreases by one (ii) a type 0 cell increases by one while a type *S*-1 cell decreases by one (iii) a type 1 cell increases by one while a type 0 cell decrease by one (iv) a type 1 cell increases by one while a type *S*-1 cell decreases by one (v) a type *S*-1 cell increases by one while a type 0 cell decreases by one; or (vi) a type *S*-1 cell increases by one while a type 1 cell decreases by one. In such a condition, a Type 0 cell can increase by one if either a Type 1 or Type *S*-1 cell dies and a Type 0 cell divides without a mutation. Then the probability for these events leading to an increase in Type 0 cells are given by (i) $\frac{X_1}{N} \cdot \frac{r_0 X_0(1-\mu_1)}{F}$, and (ii) $\frac{X_{S-1}}{N} \cdot \frac{r_0 X_0(1-\mu_1)}{F}$. Here, $F = r_0 X_0 + r_1 X_1 + r_{S-1} X_{S-1}$ is a scaling factor for selecting a dividing cell. The probability of a Type 1 or Type *S*-1 cell death is given by $\frac{X_1}{N}$ and $\frac{X_{S-1}}{N}$, respectively. Taken together, the transition probability that the number of Type 0 cell increases by one and that of Type 1 decreases by one is given by

$$\Pr[X_0 \rightarrow X_0 + 1 \text{ and } X_1 \rightarrow X_1 - 1] = \frac{X_1}{N} \cdot \frac{r_0 X_0(1-\mu_1)}{F}, \quad (2)$$

and the probability that the number of Type 0 cell increases by one and that of Type *S*-1 decreases by one is given by

$$\Pr[X_0 \rightarrow X_0 + 1 \text{ and } X_{S-1} \rightarrow X_{S-1} - 1] = \frac{X_{S-1}}{N} \cdot \frac{r_0 X_0(1-\mu_1)}{F}. \quad (3)$$

A Type 1 cell can increase by one if either a Type 0 or Type *S*-1 cell dies, and either a Type 1 cell divides without mutation or a Type 0 cell divides with mutation to become a Type 1 cell. Then the probabilities for these events leading to an increase in Type 1 cells are given by (iii) $\frac{X_0}{N} \cdot \frac{r_1 X_1(1-\mu_{S-1}) + r_0 X_0 \mu_1}{F}$, and (iv) $\frac{X_{S-1}}{N} \cdot \frac{r_1 X_1(1-\mu_{S-1}) + r_0 X_0 \mu_1}{F}$. Taken together, the transition probability that the number of Type 1 cell increases by one and that of Type 0 decreases by one is given by

$$\Pr[X_1 \rightarrow X_1 + 1 \text{ and } X_0 \rightarrow X_0 - 1] = \frac{X_0}{N} \cdot \frac{r_1 X_1(1-\mu_{S-1}) + r_0 X_0 \mu_1}{F}, \quad (4)$$

and the probability that the number of Type 1 cell increases by one and that of Type *S*-1 decreases by one is given by

$$\Pr[X_1 \rightarrow X_1 + 1 \text{ and } X_{S-1} \rightarrow X_{S-1} - 1] = \frac{X_{S-1}}{N} \cdot \frac{r_1 X_1(1-\mu_{S-1}) + r_0 X_0 \mu_1}{F}. \quad (5)$$

Similarly, a Type *S*-1 cell can increase by one if either a Type 0 or Type 1 cell dies, and either a Type *S*-1 cell divides without mutation or a Type 1 cell divides with mutation. The probabilities for the events leading to an increase in Type *S*-1 cells are given by (v) $\frac{X_0}{N} \cdot \frac{r_{S-1} X_{S-1}(1-\mu_S) + r_1 X_1 \mu_{S-1}}{F}$, and (vi) $\frac{X_1}{N} \cdot \frac{r_{S-1} X_{S-1}(1-\mu_S) + r_1 X_1 \mu_{S-1}}{F}$. Taken together, the transition probability that the number of Type *S*-1 cell increases by one and that of Type 0 decreases by one is given by

$$\Pr[X_{S-1} \rightarrow X_{S-1} + 1 \text{ and } X_0 \rightarrow X_0 - 1] = \frac{X_0}{N} \cdot \frac{r_{S-1} X_{S-1}(1-\mu_S) + r_1 X_1 \mu_{S-1}}{F}, \quad (6)$$

and the probability that the number of Type *S*-1 cell increases by one and that of Type 1 decreases by one is given by

$$\Pr[X_{S-1} \rightarrow X_{S-1} + 1 \text{ and } X_1 \rightarrow X_1 - 1] = \frac{X_1}{N} \cdot \frac{r_{S-1} X_{S-1}(1-\mu_S) + r_1 X_1 \mu_{S-1}}{F}. \quad (7)$$

In addition, a Type *S* cell can increase by one if a Type *S*-1 cell divides with mutation. The probability is given by $\frac{r_{S-1} X_{S-1} \mu_S}{F}$. Since a Type *S* cell is not a component of a tissue, once a Type *S* appears by mutation, another round of selection for a dividing cell is performed according to the transition probabilities described above. This is because malignant Type *S* cell disrupts 2D lattice structure and the Moran process is no longer applicable to it.

Next, let us consider the case where Type *S* cell divides or dies. The probabilities of Type *S* cell division or death is given by $r_S X_S(t) \cdot \Delta T$and $d_S X_S(t) \cdot \Delta T$, respectively.

In summary, the time of one step in our simulation is calculated using Eq. (1) and in one time step, one of the following three processes occurs: (i) a cell turnover in a tissue, (ii) the birth of a Type *S* cell, or (iii) the death of a Type *S* cell. Initially, all the cells are Type 0. Once the number of Type *S* cells reaches 10^9, computational surgical resection sets the number of Type *S* cells to be 0, keeping the cell type composition in a tissue remained and computational carcinogenic process restarts again. After that, the time until the number of Type *S* cells reaches 10^9 is measured as recurrence time.

Spatial structure

Two-dimensional lattice structure ($I \times J$) is introduced to a tissue dynamics in our computational model. The transition probabilities

are basically the same with or without spatial structure. The difference is the choice of a dividing cell. If a cell at position $\langle i,j \rangle$ dies, 4 adjacent cells – $\langle i,j-1 \rangle$, $\langle i,j+1 \rangle$, $\langle i-1,j \rangle$ and $\langle i+1,j \rangle$ can divide to replace it. The transition probabilities are calculated according to the cell type at those positions. We assume wall boundary condition to represent an asymmetric tissue structure.

Deterministic approximation of type *S* cell growth

As for the calculation of the Type *S* growth, we assume that when the number of cells is small, the stochastic effect should be considered. When the number of Type *S* cells exceed twice as large as the size of the normal tissue, 2*N*, growth can be regarded as a deterministic process. Then, the time duration from when the number of Type *S* cells is 2*N* to 10^9, Δt_s, is given by

$$\Delta t_s = \frac{1}{(r_S - d_S)} \ln\left(\frac{10^9}{2N}\right) \quad (8)$$

Clinical data

The data used in our analysis were from TCGA Pan-Cancer Clinical Data Resource (34, 35) and are available in the cBio Cancer Genomics Portal (44, 45). We adopt the clinical data of locoregional recurrence from 8,957 patients with 27 different non-sarcoma, non-hematological cancer types. From these datasets, the inclusion criterium for our study was "disease-free" survival – patients with no detectable malignant disease after surgery or total remission. We excluded data of "progression-free" survival in order to eliminate patients who survived with detectable disease possibly as a result of treatment-resistant clones; and also excluded data containing metastatic progression. We also included data from other independent publications for extra validation. Sarcomas and hematological cancers were excluded due to their non-conformity to a 2-dimensional lattice structure.

Survival time analysis

Survival time analysis of clinical data is calculated using the Kaplan–Meier method from disease-free intervals mentioned in Clinical Data section. In this study, disease-free interval is defined as the survival time without cancer recurrence of each patient, which corresponds to the time to recurrence of each simulation trial.

Statistical analysis

The whole process of our model is conducted on C++. Simulation codes have been deposited in a GitHub repository (https://github.com/sharaf501/Heano-Lab-Codes). The survival time analysis and other statistical analysis is conducted on Prism (version 9.4.1). Mantel-Cox (log-rank) test is used to compare difference between survival curves. A p value less than 0.05 is considered to be statistically significant.

Results

Cancer initiation patterns

Firstly, we conducted stochastic triplicate simulations for the cancer initiation up to the time of cancer detection. We were curious to know what effect the presence or absence of the spatial structure would have on the model. We traced the time course of 4 cell populations – Type 0, Type 1, Type *S*-1 and Type *S* cells using a combination of various parameter sets. Lower mutation rate from Type 1 to Type *S*-1, $\mu_{S\text{-}1}$, was additionally examined to account for additional premalignant cell types between Type 1 and Type *S*-1. In the model without spatial structure, we observed 3 patterns of cancer initiation based on frequency of non-malignant cell population at cancer detection (Figure 2A). Interestingly, all the patterns show a progressive decline of Type 0 cells until the entire tissue is dominated by Type 1, Type *S*-1 or both Type 1 and Type *S*-1 cells. By combining various parameter sets in our simulation, we extrapolated the varying distribution of the cancer initiation patterns (Figure 2B and Supplementary Figure 1). Lower fitness of Type 1 cells, r_1 generally favored Type *S*-1 cells dominance when fitness of Type *S*-1, $r_{S\text{-}1}$, is high. Higher r_1 values favored Type 1 dominance while equal fitness of Type 1 and Type *S*-1 cells yielded Type 1 dominance or Type 1/*S*-1 co-dominance. Mutation rates generally affected time to cancer detection and appearance of dominance. We also extended the mutation rates from Type 1 to Type *S*-1 cell type to denote other additional mutation steps and found a consistent increase in cancer detection times but patterns generally remained the same. In some cases where low fitness of both Type 1 and Type *S*-1 were coupled with lower mutation rates, Type *S* malignant cells failed to appear at extended times and simulations were terminated. Most curiously, mutation rate from Type 0 to Type 1, μ_1, did not affect the pattern of cancer initiation or time to cancer detection (Supplementary Figure 1).

Parameter dependence of recurrence time

Next, we examined the time to recurrence after surgical resection and the proportion of Type *S*-1 cells at the time of surgery in varying parameter sets. We reasoned that since Type *S*-1 cells needs only one more step for malignant transformation; therefore, its proportion was thought to be critical for cancer recurrence. To do this, we ran 1,000 simulations for each parameter set and calculated the mean recurrence time (Figures 3A–K). We also ran similar simulations at higher cell number in a tissue, *N*, between 100 to 1,000 times to assess the effect of tissue size on the parameter dependency (Figures 3B–L). We found that higher fitness of Type 1 cells, r_1 increased the mean

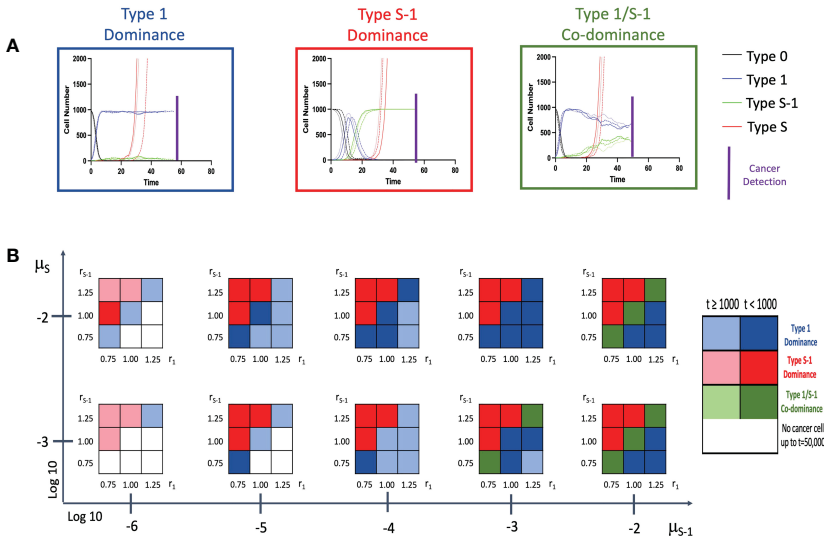


FIGURE 2
Patterns of tissue composition at cancer initiation **(A)** Simulation studies without spatial structure show three patterns of cancer initiation. For each pattern, black, blue, green and red curves indicate Type 0, Type 1, Type *S*-1 and Type *S* cells, respectively. Each parameter set was simulated in triplicate (Joined, dashed, and long-dashed lines). **(B)** Panel showing several patterns of tissue composition and time to detection using combination of various parameter sets. Cell type "Dominance" indicates >90% of a particular cell type at cancer detection. "Co-dominance" refers to 2 cell populations with >40% or 3 cell populations with >30% at cancer detection. Parameter values used are: N = 1,000; $d = d_S$ = 1.0; r_0 = 1.0; r_1 = 0.75, 1.00 and 1.25; $r_{S\text{-}1}$ = 0.75, 1.00 and 1.25; r_S = 1.5; μ_1 = 0.001 and 0.01; $\mu_{S\text{-}1}$ = 0.000001, 0.00001, 0.0001, 0.001 and 0.01; μ_S = 0.001 and 0.01. Note: Mutation rate, μ_1, does not impact the results and is not shown.

recurrence time (Figures 3A, B), while mutation rate from Type 0 to Type 1, μ_1, had no effect on mean recurrence time (Figures 3G, H). Other parameters however, showed a negative correlation to the mean recurrence time – higher parameter values resulted in shorter mean recurrence time. Higher tissue cell number yielded an overall shortening of mean recurrence time but parameter dependency remained the same. We also observed a reduction in the proportion of Type *S*-1 cells at the time of surgery when r_1, r_S and μ_S increases (Figures 3A–L), while $r_{S\text{-}1}$ and μ_S increases in the proportion of Type *S*-1 cells (Figures 3C–J). Mutation rate from Type 0 to Type 1, μ_1, had little effect on the proportion of Type *S*-1 cells (Figures 3G, H).

Effect of spatial structure on cancer initiation patterns

We then incorporated the spatial structure framework into the model to investigate the effect of tissue positional influence in cancer initiation patterns and time of cancer detection. After triplicate simulations using various parameters, we identified seven distinct patterns of cancer initiation (Figure 4A) based on the composition of non-malignant cell population at cancer detection. Figure 4B and Supplementary Figure 2 showed the distribution of the patterns in a wide parameter range. Low r_1 values showed Type 0 dominance at low $r_{S\text{-}1}$ levels with failure to detect cancer cells at very low $\mu_{S\text{-}1}$ levels. With higher r_1 values, Type 1 cell types begin to dominate. When combined with high $\mu_{S\text{-}1}$, we saw Type 1/*S*-1 co-dominance (green region in Figure 2B). When r_1 and $r_{S\text{-}1}$ are equal to fitness of Type 0 (r_0), we saw Type 0/1 co-dominance or Type 0/1/*S*-1 co-dominance depending on $\mu_{S\text{-}1}$ values. We noticed a peculiar pattern of Type 0/*S*-1 co-dominance (black region in Figure 2B) when $r_{S\text{-}1}$, $\mu_{S\text{-}1}$, and μ_S were high with relatively lower r_1 value. Type *S*-1 dominance (red region in Figure 2B) was regarded as the most undesirable scenario due to the abundance of Type *S*-1 cells, indicating shorter recurrence time. We saw this pattern when $r_{S\text{-}1}$, μ_1 and $\mu_{S\text{-}1}$ were high, r_1 was equal to 1.0 and μ_S was relatively small. Some parameter sets with low fitness failed to yield Type *S* cells at extended time points during the simulations. Here, the incorporation of the spatial structure to our simulation framework had remarkable alterations to the cancer initiation patterns and cancer detection time.

Effect of spatial structure on recurrence time

Subsequently, we examined the mean recurrence time after surgical resection and the proportion of Type *S*-1 lesions at the time of surgery in a vast parameter range with the influence of the spatial structure setting (Figure 5). Similarly, we ran 100 to 500 simulations to obtain mean recurrence time (Figures 5A–K) and to check the effect of larger cell numbers (Figures 5B–L). Generally, we saw that the integration of the spatial structure to our simulation framework had noteworthy changes to the parameter dependency to recurrence time. When the size of the normal tissue was small, the effect of fitness advantage on the proportion of Type *S*-1 cells in a tissue became larger. Our simulation results showed that an increase in the cell fitness shortened the mean time to recurrence (Figures 5C–F). However, an increase in r_1 was found to reduce the recurrence

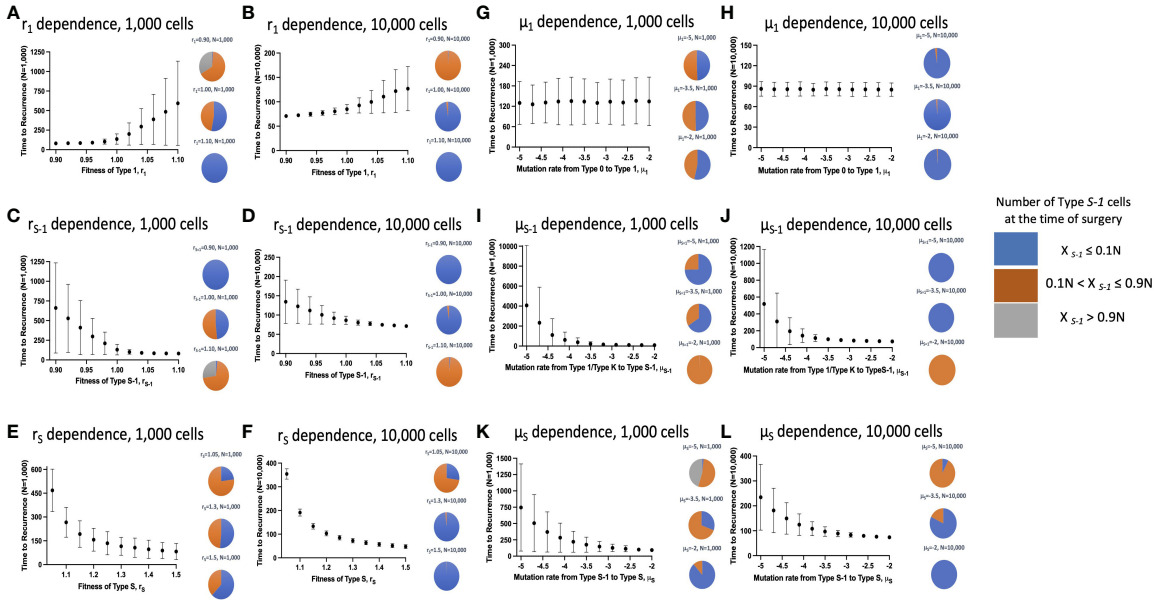


FIGURE 3
Parameter dependence on recurrence time. Simulation studies without spatial structure are shown. Mean values obtained from 100 to 1,000 simulations are shown by dots, and standard deviations are indicated by bars. Pie charts in the panels indicate the proportion of Type *S*-1 cells in a normal tissue at the time of first treatment. Blue, orange and grey represent small ($X_{S\text{-}1} \leq 0.1N$), intermediate ($0.1N < X_{S\text{-}1} \leq 0.9N$), and large ($X_{S\text{-}1} > 0.9N$) proportion of Type *S*-1 cells, respectively. Standard parameter values used in **(A–L)** are $d = d_S = 1.0$, $r_0 = 1.0$, $r_1 = 1.0$, $r_{S\text{-}1} = 1.2$, $r_S = 1.0$, $\mu_1 = 0.001$, $\mu_{S\text{-}1} = 0.001$, $\mu_S = 0.001$; and $N = 1{,}000$ in **(A, C, E, G, I, K)**; and $N = 10{,}000$ in **(B, D, F, H, J, L)**.

time but begin to increase slightly at much higher levels regardless of the tissue size. We also found a consistent reduction in the mean recurrence time as mutation rates μ_1, $\mu_{S\text{-}1}$ and μ_S increased (Figures 5G–L). We also observed a reduction in the proportion of Type *S*-1 cells at the time of surgery with a spatial structure. Especially, when either r_S or μ_S was small, and any of $r_{S\text{-}1}$, μ_1, or $\mu_{S\text{-}1}$ was large, the proportion of Type *S*-1 increased (Figure 5).

Fitting of recurrence time to clinical data

By using our computational model with spatial structure, multiple runs of stochastic simulations were performed with multiple parameter sets and *in silico* Kaplan–Meier curves were made. The data points about the time when 0% to 100% of patients experienced recurrence with an interval of 4% (time when 0%, 4%,

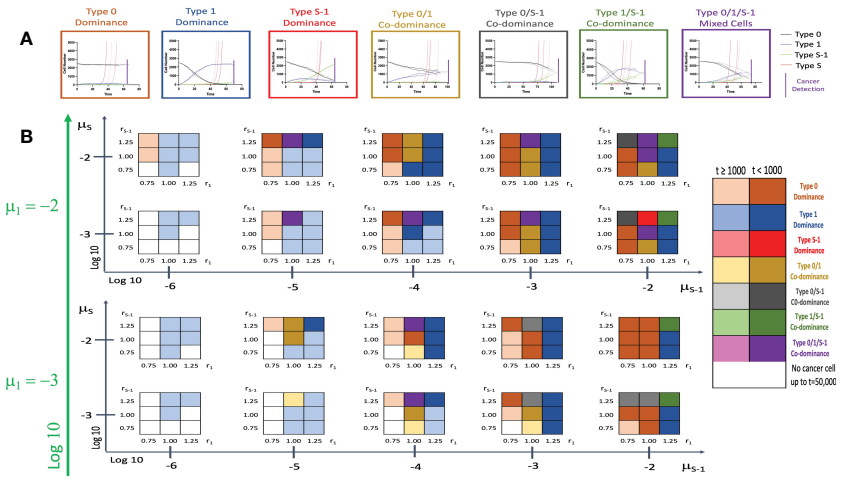


FIGURE 4
Patterns of tissue composition at cancer initiation with spatial structure. **(A)** Simulation studies with spatial structure show 7 patterns of cancer initiation. For each pattern, black, blue, green and red curves indicate Type 0, Type 1, Type *S*-1 and Type *S* cells, respectively. Each parameter set was simulated in triplicate (Joined, dashed, and long-dashed lines). **(B)** Panel showing patterns of tissue composition and time to detection using combination of various parameter sets. The definitions of "Dominance" and "Co-dominance" are the same as those explained in Figure 2. Parameter values used are: $N = 2{,}500$; $d = d_s = 1.0$; $r_0 = 1.0$; $r_1 = 0.75$, 1.00 and 1.25; $r_{S\text{-}1} = 0.75$, 1.00 and 1.25; $r_S = 1.5$; $\mu_1 = 0.001$ and 0.01; $\mu_{S\text{-}1} = 0.000001$, 0.00001, 0.0001, 0.001 and 0.01; $\mu_S = 0.001$ and 0.01.

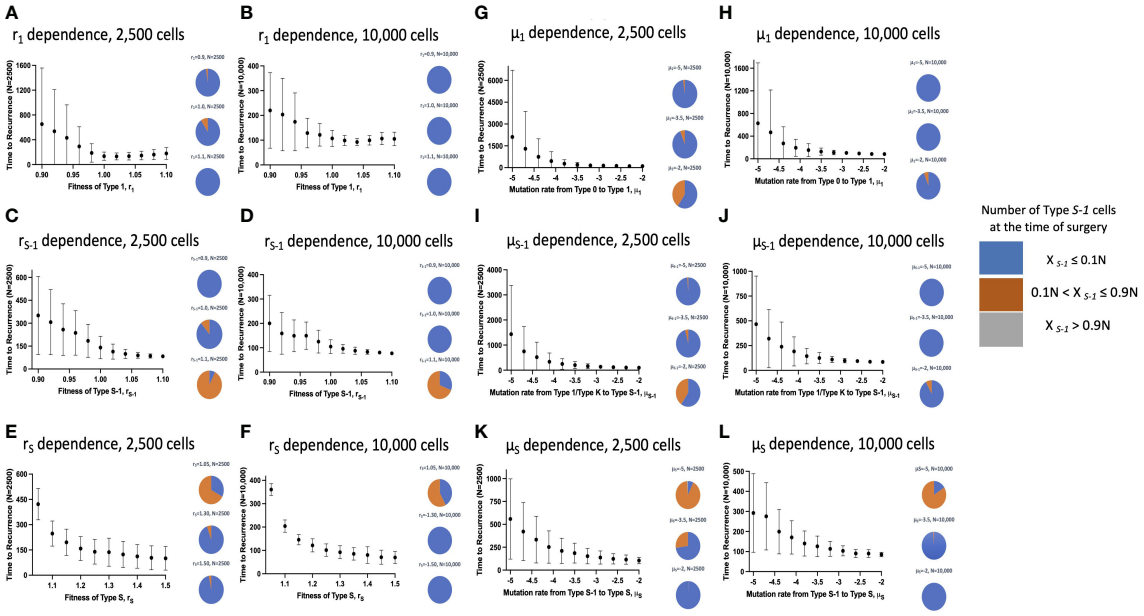


FIGURE 5
Parameter dependence on recurrence time with spatial structure. Simulation results with spatial structure are shown. Mean values obtained from 100 to 1,000 simulations are shown by dots, and standard deviations are indicated by bars. Pie charts in the panels indicate the proportion of Type *S*-1 cells in a normal tissue at the time of first treatment. Blue, orange and grey represent small ($X_{S\text{-}1} \leq 0.1N$), intermediate ($0.1N < X_{S\text{-}1} \leq 0.9N$), and large ($X_{S\text{-}1} > 0.9N$) proportion of Type *S*-1 cells, respectively. Standard parameter values used in **(A–L)** are $d = d_S = 1.0$, $r_0 = 1.0$, $r_1 = 1.0$, $r_{S\text{-}1} = 1.2$, $r_S = 1.0$, $\mu_1 = 0.001$, $\mu_{S\text{-}1} = 0.001$, $\mu_S = 0.001$; and $N = 2{,}500$ in **(A, C, E, G, I, K)**; and $N = 10{,}000$ in **(B, D, F, H, J, L)**.

8%, …, 100% of patients experienced recurrence) were employed to compare between the *in silico* and published clinical data (44, 45) of 27 cancer types. In this analysis, we adopted random sampling for parameters to obtain *in silico* recurrence data and determined the best parameter set for each cancer type that minimized the mean of squared logarithmic residuals (log-MSR) between outputs *in silico* and in public. The accepted parameter set (Table 1) was used to extrapolate recurrence time which were then fitted to clinical data and disease-free survival curves were depicted (Figure 6).

Mantel-Cox test was used to compare between the curves of simulated and clinical data revealing minimal statistical nonconformity. According to the estimated parameters (Table 1), we firstly deduced a tissue-specific turnover per month from d_S. Kidney chromophobe had the fewest cellular turnover cycles per month while bladder urothelial carcinoma and colorectal adenocarcinoma had the highest turnover cycles. Moreover, colorectal adenocarcinoma, kidney chromophobe, renal clear cell carcinoma, thyroid carcinoma, adenoid cystic carcinoma and acral melanoma showed higher fitness of all their premalignant cells than normal cells. Of note, the proliferation rate of the Type *S* malignant cells, r_S, was estimated to be high in cholangiocarcinoma, liver hepatocellular carcinoma, mesothelioma and upper tract urothelial cancer while being relatively low in breast invasive carcinoma, kidney chromophobe and skin cutaneous melanoma. Kidney chromophobe had the lowest mutation rate from the final premalignant cell stage to malignant cells while cervical squamous cell carcinoma and prostate adenocarcinoma had the highest mutation rate. Figure 7 showed the negative correlation between mutational steps required for carcinogenesis (37) and overall mutation rates (μ_I) obtained from our studies by multiplying the mutation rates for all steps.

Discussion

In this study, we constructed computational models with and without spatial structure that described cell population dynamics in both normal and cancer tissues. Using our models, we clearly observed different patterns of cancer initiation and the residual premalignant cells present at the time of cancer detection or surgical intervention. Integrating the spatial structure setting to the model revealed additional patterns of cancer initiation as against just three in the model without spatial structure. Especially, the preservation of intact normal cells was observed in the model with spatial structure (Figures 2–4). According to the comprehensive analysis of parameter dependence, we found that field cancerization at the detection time depended on a combination of fitness and mutation rates.

We also revealed the relationship between the proportion of premalignant cells and recurrence time (Figures 3–5). The model without spatial structure overemphasized the power of Type 1 fitness and its ability to limit Type *S*-1 and Type *S* appearance which led to longer mean recurrence times as r_1 increases (Figures 3A, B). The same effect was seen in the mutation rate from Type 0 to Type 1 which rendered μ_1 impotent in affecting mean recurrence time (Figures 3G, H). All other fitness and mutation parameters led to shorter recurrence time as their

10.3389/fonc.2023.1116210

TABLE 1 Tumor-specific carcinogenic profiles and p values of survival curves (μ values are in $\log_{10}$ while SQ are log-SSR values).

Code	Cancer Type (Data Source)	SQ	r_1	r_{S-1}	r_S	d_S	μ_1	μ_{S-1}	μ_S	μ_I	p value
ACC	Adrenocortical Carcinoma (TCGA, PanCancer Atlas)	0.616	1.003	0.958	5.578	3.436	-3.212	-3.979	-3.006	-10.197	0.9120
BLCA	Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)	0.922	1.097	0.981	5.614	3.939	-2.876	-3.616	-3.223	-9.715	0.8857
BRCA	Breast Invasive Carcinoma (TCGA, PanCancer Atlas)	0.983	0.945	0.948	2.967	1.887	-3.770	-2.228	-2.661	-8.658	0.1707
CESC	Cervical Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	0.897	0.929	0.951	6.684	3.506	-3.842	-3.233	-2.098	-9.172	0.1213
CHOL	Cholangiocarcinoma (TCGA, PanCancer Atlas)	0.847	0.988	1.011	7.170	3.765	-3.504	-2.759	-2.349	-8.612	0.7716
COAD	Colorectal Adenocarcinoma (TCGA, PanCancer Atlas)	0.943	1.086	1.027	5.912	3.859	-3.176	-3.052	-4.347	-10.575	0.2783
ESCA	Esophageal Adenocarcinoma (TCGA, PanCancer Atlas)	0.936	1.048	0.919	6.767	3.655	-4.114	-3.870	-2.411	-10.394	0.8862
HNSC	Head & Neck Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	0.802	0.980	0.988	5.148	3.586	-2.735	-4.886	-2.656	-10.277	0.5908
KICH	Kidney Chromophobe (TCGA, PanCancer Atlas)	0.401	1.070	1.024	2.308	1.558	-3.971	-2.542	-4.658	-11.170	0.9360
KIRC	Kidney Renal Clear Cell Carcinoma (TCGA, PanCancer Atlas)	0.786	1.081	1.064	3.655	2.713	-4.569	-3.445	-4.252	-12.265	0.1519
KIRP	Kidney Renal Papillary Cell Carcinoma (TCGA, PanCancer Atlas)	0.559	0.965	0.958	5.470	3.004	-3.182	-2.748	-3.374	-9.304	0.9880
LIHC	Liver Hepatocellular Carcinoma (TCGA, PanCancer Atlas)	0.846	0.977	1.009	7.675	3.899	-4.620	-3.690	-4.252	-12.562	0.6744
LUAD	Lung Adenocarcinoma (TCGA, PanCancer Atlas)	0.600	0.922	0.977	6.383	3.327	-3.068	-3.471	-2.821	-9.360	0.4484
LUSC	Lung Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	0.524	0.997	0.952	5.220	3.568	-4.174	-2.276	-3.258	-9.708	0.9767
MESO	Mesothelioma (TCGA, PanCancer Atlas)	0.572	0.918	1.030	7.134	3.581	-3.287	-4.222	-3.561	-11.070	0.8411
OV	Ovarian Serous Cystadenocarcinoma (TCGA, PanCancer Atlas)	0.772	1.052	0.939	4.145	2.655	-3.483	-2.415	-3.502	-9.400	0.9710
PAAD	Pancreatic Adenocarcinoma (TCGA, PanCancer Atlas)	0.736	0.988	0.996	3.145	1.654	-3.218	-2.875	-2.416	-8.509	0.8931
PRAD	Prostate Adenocarcinoma (TCGA, PanCancer Atlas)	0.886	0.915	0.918	4.826	3.584	-2.286	-3.987	-2.189	-8.462	0.7336
STAD	Stomach Adenocarcinoma (TCGA, PanCancer Atlas)	0.400	0.936	0.949	5.738	3.177	-3.377	-2.268	-3.389	-9.034	0.9442
SKCM	Skin Cutaneous Melanoma (TCGA, Firehose Legacy)	0.805	1.016	0.980	2.848	1.818	-4.481	-3.607	-2.512	-10.601	0.5770
THCA	Thyroid Carcinoma (TCGA, PanCancer Atlas)	0.764	1.020	0.980	6.541	3.525	-2.824	-4.268	-4.131	-11.222	0.7340
UCEC	Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)	0.892	0.948	0.984	3.965	2.083	-4.276	-3.824	-2.908	-11.007	0.1153
UVM	Uveal Melanoma (TCGA, Firehose Legacy)	0.564	1.022	0.947	6.612	3.535	-3.397	-2.166	-4.000	-9.563	0.2663
ACYC	Adenoid Cystic Carcinoma (MSK, Nat Genet 2013)	0.521	1.082	1.065	4.523	2.492	-2.754	-3.342	-4.046	-10.142	0.6316
MEL	Acral Melanoma (TGEN, Genome Res 2017)	0.524	1.031	1.008	5.199	3.104	-2.572	-3.866	-2.487	-8.925	0.3385
OSCC	Oral Squamous Cell Carcinoma (MD Anderson, Canc. Disc 2013)	1.033	0.901	0.990	8.090	3.700	-2.721	-2.602	-3.640	-8.964	0.9997
UTUC	Upper Tract Urothelial Cancer (MSK, Eur Urol 2015)	0.628	0.938	1.009	7.470	3.800	-2.130	-3.672	-2.815	-8.617	0.5480

effects became larger. Generally, with the spatial structure setting, we found that recurrence time became shorter when mutation rates or fitness of cancer cells were large, while the time became longer when the fitness of premalignant cells or growth rate of cancer cells were low (Figure 5). An exception would be the mean recurrence time with r_1 (Figures 5A, B) which was seen to shorten as r_1 became larger but to get longer as r_1 became much larger. This could be due to a tissue competition between Type 1 and Type *S*-1 cells which subsequently delayed the emergence of Type *S* cells and hence, a more favorable recurrence time.

Moreover, we successfully estimated the characteristic parameter sets of the computational model that best reproduced the clinical data of disease-free survival in each cancer type. All the non-sarcoma cancer types were successfully fitted with no statistical deviation (Table 1). Even though some datasets like ESCA contains 2 different cancer types – esophageal adenocarcinoma and esophageal squamous cell carcinoma, we obtained p values that indicates no statistical difference. At the same time, we obtained valuable information about cellular turnover per month (d_S), relative fitness of premalignant cells (r_1, r_{S-1}), a growth rate of

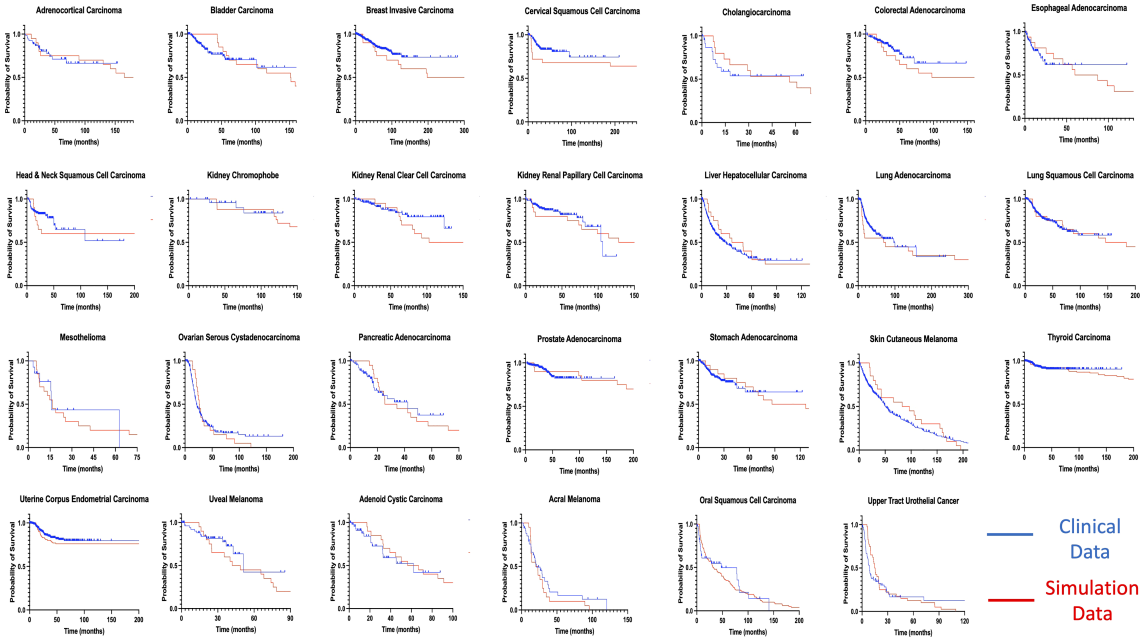


FIGURE 6
Fitting of model-derived *in silico* data to published clinical data for 27 cancer types. Thousands of stochastic runs were used to obtain parameter sets that best fit survival curves of 27 non-sarcoma, non-hematologic cancer types. Blue curves indicate clinical data while red curves indicate simulation data survival curves. *p* values between curves are found in Table 1. ACC, Adrenocortical Carcinoma; BLCA, Bladder Urothelial Carcinoma; BRCA, Breast Invasive Carcinoma; CESC - Cervical Squamous Cell Carcinoma; CHOL – Cholangiocarcinoma; COAD, Colorectal Adenocarcinoma; ESCA, Esophageal Adenocarcinoma; HNSC, Head & Neck Squamous Cell Carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney Renal Clear Cell Carcinoma; KIRP, Kidney Renal Papillary Cell Carcinoma; LIHC, Liver Hepatocellular Carcinoma; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; MESO – Mesothelioma; OV, Ovarian Serous Cystadenocarcinoma; PAAD, Pancreatic Adenocarcinoma; PRAD, Prostate Adenocarcinoma; STAD, Stomach Adenocarcinoma; THCA, Thyroid Carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma; UVM, Uveal Melanoma; ACYC, Adenoid Cystic Carcinoma; MEL, Acral Melanoma; UTUC, Upper Tract Urothelial Cancer; OSCC, Oral Squamous Cell Carcinoma; SKCM, Skin Cutaneous Melanoma.

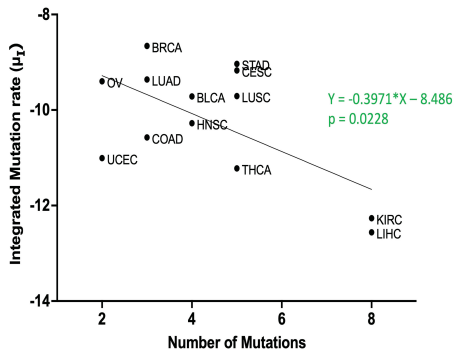
cancer cells (r_S) and mutation rates from one cell type to another (μ_1, $\mu_{S\text{-}1}$, μ_S) for each carcinogenesis. We have specified the growth rate for each cancer using the r_S values from our clinical fitting. Interestingly, we observed relatively high growth rates of malignant cells (Type *S* cells) in some common cancer types like lung and colorectal cancers, whereas a relatively lower growth rate was estimated in breast invasive carcinoma which was also a common cancer type but was relatively asymptomatic in agreement with several studies (46, 47). From the high $\mu_{S\text{-}1}$ values, we elucidated that uveal melanoma, breast invasive carcinoma, stomach adenocarcinoma and lung squamous cell carcinoma had the shortest time to reach late premalignant cell stage from the earliest premalignant cell stage possibly indicating fewer mutational steps. On the other hand, thyroid carcinoma and head and neck carcinoma had small $\mu_{S\text{-}1}$ values, indicating the multiple steps in the carcinogenesis. Our data was in alignment with data that estimated the number of hits required for carcinogenesis (37), where liver, kidney and thyroid cancers had the lowest overall mutational rates indicating more mutational requirements while uterine, ovarian and lung cancers had higher overall mutational rates indicating fewer mutational requirements (Figure 7).

Additionally, our findings successfully revealed average cellular turnover rates per month by inferring our model with published clinical data whose measurements were in months. Kidney chromophobe and pancreatic cancer showed low turnover rates per month (about 1.5 times), possibly indicating low incidence rate. On the other hand, bladder urothelial carcinoma, liver hepatocellular carcinoma, colorectal adenocarcinoma and upper tract urothelial carcinoma had the highest turnover rates (almost 4 times per month) which perhaps explained why they were the most common cancers in men and women (48). This also corresponded with data that suggested that number of cell division was a significant risk factor for cancer (49).

We propose from our findings that certain cell populations, specifically Type S-1 could be targeted to address the threat of locoregional recurrence. With currently available tools and advancements in personalized medicine, it is possible to prevent recurrence by targeting a particular cell type or lesion. An example in case in the outstanding success achieved using PD-1 Blockade in mismatch repair–deficient, locally advanced rectal cancer which recorded a 100% success (50). CRISPR-based mutation can also aid in cell competition studies to identify cell fitness levels among the known and unknown driver mutations to further provide actionable data for more studies.

In this study, we estimated cell fitness as a single numerical value with 1.0 indicating normal cells and other cells with ranges from normal cells. In reality, this is an “oversimplification” as cell fitness is a complex and dynamic concept which can be related to both genetic (51) and non-genetic (52) alterations. Unfortunately, studies on cell fitness with regards to known or even unknown

A



B

Predicted number of mutations	Cancer Type	μ_i
1	PRAD	-8.462
	PAAD	-8.509
	CHOL	-8.612
	UTUC	-8.617
	MEL	-8.925
	OSCC	-8.964
2	KIRP	-9.304
3	UVM	-9.563
4	ACYC	-10.142
	ACC	-10.197
5	ESCA	-10.394
	SKCM	-10.601
7	MESO	-11.070
	KICH	-11.170

FIGURE 7
Relationship between integrated mutation rate and number of mutation hits required for cancer initiation **(A)** Published data for number of mutational hits required for carcinogenesis (37) in some cancer types was plotted against corresponding integrated mutation rate ($\mu_1 \cdot \mu_{5-1} \cdot \mu_5$). The linear regression was performed, and the regression line and the *p* value are shown. **(B)** Predicted number of mutations for unpublished cancer types as per equation from **(A)** ACC, Adrenocortical Carcinoma; BLCA, Bladder Urothelial Carcinoma; BRCA, Breast Invasive Carcinoma; CESC, Cervical Squamous Cell Carcinoma; CHOL – Cholangiocarcinoma; COAD, Colorectal Adenocarcinoma; ESCA, Esophageal Adenocarcinoma; HNSC, Head & Neck Squamous Cell Carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney Renal Clear Cell Carcinoma; KIRP, Kidney Renal Papillary Cell Carcinoma; LIHC, Liver Hepatocellular Carcinoma; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; MESO – Mesothelioma; OV, Ovarian Serous Cystadenocarcinoma; PAAD, Pancreatic Adenocarcinoma; PRAD, Prostate Adenocarcinoma; STAD, Stomach Adenocarcinoma; THCA, Thyroid Carcinoma; UCEC, Uterine Corpus Endometrial Carcinoma; UVM, Uveal Melanoma; ACYC, Adenoid Cystic Carcinoma; MEL, Acral Melanoma; UTUC, Upper Tract Urothelial Cancer; OSCC, Oral Squamous Cell Carcinoma; SKCM, Skin Cutaneous Melanoma.

cancer-related mutations are lacking. Also, order of mutations in premalignant cells and a comprehensive study of cell-based or animal model mutational requirements for certain cancers are unavailable. These limit the tools with which we can perform additional validation of our model. Mutation rates were chosen to include processes involved with DNA repair, epigenetics, infection and role of external agents. Each of these could independently affect the model but we chose to combine them. In the current analysis, hematologic or liquid cancers were not included partly because of their dynamic nature and lack of 2D lattice arrangement but mainly the difficulty in assessing exact cell numbers. Even though certain tumor markers for certain malignancies may be used to quantify cell number, the threshold for detection and overall utility is not fully assured. The model without spatial structure might be applicable in this scenario as well as for sarcomas. Moreover, we did not stratify or independently differentiate demographic information such as age, sex or race for each cancer type. Possible extension of the analysis may be to perform age or other parameter dependent analysis. Furthermore, we did not specify the order of mutations for malignant transformation in our model, which albeit gave us a good fit with clinical data. The order of mutations is quite important as revealed from data accrued from colorectal cancer progression (53). Considering multiple mutational orders could be beneficial especially those leading to histologically 'abnormal' benign lesions. Barrett's Esophagus (BE) is a notable example where whole genome sequencing found similar mutational events between esophageal adenocarcinoma and non-dysplastic BE (54) thereby suggesting different mutational order (55). Reports that prior diagnosis of BE affords a better prognosis (56) with only about 5% of BE patients developing esophageal adenocarcinoma (57) further strengthens the different order of mutation concept.

One challenge for cancer management is late diagnosis. Our model computes a cancer detection stage of $1cm^3$ – 10^9 cells. To evaluate the effect of late diagnosis, we changed the cancer detection time to 10^{10} and assess parameter dependence on recurrence time. We observed a reduction in time to recurrence indicating that late diagnosis might contribute to shorter recurrence time (Supplementary Figure 3). Another challenge to the usage of this model is the variability of proportion of locoregional recurrence out of total recurrence rate among various cancer types. It is common knowledge that recurrence can occur at a distant area from the original issue – metastasis; our model however, does not take this into account. As a result, the utility of this model is high for certain cancer types but unfortunately, subdued for other cancer types. Consequently, malignancies where locoregional recurrence accounts for a high proportion of total recurrence such as thyroid cancer with 94% (58), oral squamous cell carcinoma with 90% (59), cholangiocarcinoma with 85% (60), prostate cancer with 81% (61), liver cancer with 78% (62), mesothelioma with 74% (63), head & neck squamous cell carcinoma with 69% (64), and ovarian cancer with 68% (65) could reap great benefit from this model. On the other hand, cancers where distant metastasis accounts for a major proportion of total recurrence such as kidney cancers with 73% (66), skin cutaneous melanoma with 71% (67) and bladder urothelial cancer with 66% (68) might feel the need to complement our model with additional tools to increase its precision. Interestingly, we can get some insight from recurrence pattern of breast cancer. In patients undergoing conservative breast

surgery only, locoregional recurrence accounts for 62% of all cancer recurrence (69). However, in a study with data for different surgical intervention types, locoregional recurrence rates were 42.9% and 19% of total recurrence in breast conservative surgery and total mastectomy respectively (70). This could perhaps be due to the elimination of the cancerized field by total mastectomy which conservative surgery is unable to achieve.

In conclusion, this model reveals parameter combinations that fit clinical data and contributes to the ever-growing knowledge about cancer initiation and recurrence. The model shows elucidate cancers which have premalignant cells with high fitness are likely to have a short recurrence time. This approach can be a valuable tool in the management of cancer especially in the field of personalized molecular medicine to target patients who are at highest risk of recurrence.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

SDA, MT, and HH conceived the idea of the study. SDA, MT, and HH developed simulation codes. SDA conducted computational simulations. SDA and HH contributed to the interpretation of the results. HH supervised the conduct of this study. All authors reviewed the manuscript draft and revised it. All authors approved the final version of the manuscript to be published.

Funding

This work was supported by JSPS KAKENHI Grant Number 16H06279 (PAGS), AMED under Grant Number JP22ck0106700, and National Cancer Research Fund 2020-A-7 (HH).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2023.1116210/full#supplementary-material

References

1. Tlsty TD, Coussens LM. Tumor stroma and regulation of cancer development. *Annu Rev Pathol* (2006) 1:119–50. doi: 10.1146/annurev.pathol.1.110304.100224

2. Uramoto H, Tanaka F. Recurrence after surgery in patients with nsclc. *Transl Lung Cancer Res* (2014) 3(4):242–9. doi: 10.3978/j.issn.2218-6751.2013.12.05

3. Hiller JG, Perry NJ, Poulogiannis G, Riedel B, Sloan EK. Perioperative events influence cancer recurrence risk after surgery. *Nat Rev Clin Oncol* (2018) 15(4):205–18. doi: 10.1038/nrclinonc.2017.194

4. Mahvi DA, Liu R, Grinstaff MW, Colson YL, Raut CP. Local cancer recurrence: The realities, challenges, and opportunities for new therapies. *CA Cancer J Clin* (2018) 68(6):488–505. doi: 10.3322/caac.21498

5. Corrado G, Salutari V, Palluzzi E, Distefano MG, Scambia G, Ferrandina G. Optimizing treatment in recurrent epithelial ovarian cancer. *Expert Rev Anticancer Ther* (2017) 17(12):1147–58. doi: 10.1080/14737140.2017.1398088

6. Karacz CM, Yan J, Zhu H, Gerber DE. Timing, sites, and correlates of lung cancer recurrence. *Clin Lung Cancer* (2020) 21(2):127–35.e3. doi: 10.1016/j.cllc.2019.12.001

7. Sonoda D, Matsuura Y, Kondo Y, Ichinose J, Nakao M, Ninomiya H, et al. Comparison of local therapy in patients with lung oligo-recurrence of non-Small-Cell lung cancer. *J Surg Oncol* (2021) 123(8):1828–35. doi: 10.1002/jso.26453

8. Holleczek B, Stegmaier C, Radosa JC, Solomayer EF, Brenner H. Risk of loco-regional recurrence and distant metastases of patients with invasive breast cancer up to ten years after diagnosis - results from a registry-based study from Germany. *BMC Cancer* (2019) 19(1):520. doi: 10.1186/s12885-019-5710-5

9. Holloway CMB, Shabestari O, Eberg M, Forster K, Murray P, Green B, et al. Identifying breast cancer recurrence in administrative data: Algorithm development and validation. *Curr Oncol* (2022) 29(8):5338–67. doi: 10.3390/curroncol29080424

10. Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P, et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N Engl J Med* (2017) 377(19):1836–46. doi: 10.1056/NEJMoa1701830

11. Kim HG, Kim HS, Yang SY, Han YD, Cho MS, Hur H, et al. Early recurrence after neoadjuvant chemoradiation therapy for locally advanced rectal cancer: Characteristics and risk factors. *Asian J Surg* (2021) 44(1):298–302. doi: 10.1016/j.asjsur.2020.07.014

12. Nakauchi M, Vos E, Tang LH, Gonen M, Janjigian YY, Ku GY, et al. Outcomes of neoadjuvant chemotherapy for clinical stages 2 and 3 gastric cancer patients: Analysis of timing and site of recurrence. *Ann Surg Oncol* (2021) 28(9):4829–38. doi: 10.1245/s10434-021-09624-5

13. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* (1953) 6 (5):963–8. doi: 10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q

14. Sinjab A, Han G, Wang L, Kadara H. Field carcinogenesis in cancer evolution: What the cell is going on? *Cancer Res* (2020) 80(22):4888–91. doi: 10.1158/0008-5472.CAN-20-1956

15. Curtius K, Wright NA, Graham TA. An evolutionary perspective on field cancerization. *Nat Rev Cancer* (2018) 18(1):19–32. doi: 10.1038/nrc.2017.102

16. Willenbrink TJ, Ruiz ES, Cornejo CM, Schmults CD, Arron ST, Jambusaria-Pahlajani A. Field cancerization: Definition, epidemiology, risk factors, and outcomes. *J Am Acad Dermatol* (2020) 83(3):709–17. doi: 10.1016/j.jaad.2020.03.126

17. Rahal Z, Sinjab A, Wistuba II, Kadara H. Game of clones: Battles in the field of carcinogenesis. *Pharmacol Ther* (2022) 237:108251. doi: 10.1016/j.pharmthera.2022.108251

18. Heaphy CM, Griffith JK, Bisoffi M. Mammary field cancerization: Molecular evidence and clinical importance. *Breast Cancer Res Treat* (2009) 118(2):229–39. doi: 10.1007/s10549-009-0504-0

19. Gadaleta E, Thorn GJ, Ross-Adams H, Jones LJ, Chelala C. Field cancerization in breast cancer. *J Pathol* (2022) 257(4):561–74. doi: 10.1002/path.5902

20. Califano J, Ahrendt SA, Meininger G, Westra WH, Koch WM, Sidransky D. Detection of telomerase activity in oral rinses from head and neck squamous cell carcinoma patients. *Cancer Res* (1996) 56(24):5720–2.

21. Boscolo-Rizzo P, Da Mosto MC, Rampazzo E, Giunco S, Del Mistro A, Menegaldo A, et al. Telomeres and telomerase in head and neck squamous cell carcinoma: From pathogenesis to clinical implications. *Cancer Metastasis Rev* (2016) 35(3):457–74. doi: 10.1007/s10555-016-9633-1

22. Galandiuk S, Rodriguez-Justo M, Jeffery R, Nicholson AM, Cheng Y, Oukrif D, et al. Field cancerization in the intestinal epithelium of patients with crohn's ileocolitis. *Gastroenterology* (2012) 142(4):855–64.e8. doi: 10.1053/j.gastro.2011.12.004

23. Cappellesso R, Lo Mele M, Munari G, Rosa-Rizzotto E, Guido E, De Lazzari F, et al. Molecular characterization of "Sessile serrated" adenoma to carcinoma transition in six early colorectal cancers. *Pathol Res Pract* (2019) 215(5):957–62. doi: 10.1016/j.prp.2019.02.001

24. Pirlog R, Cismaru A, Nutu A, Berindan-Neagoe I. Field cancerization in nsclc: A new perspective on micrornas in macrophage polarization. *Int J Mol Sci* (2021) 22 (2):746-761. doi: 10.3390/ijms22020746

25. Syk E, Torkzad MR, Blomqvist L, Ljungqvist O, Glimelius B. Radiological findings do not support lateral residual tumour as a major cause of local recurrence of rectal cancer. *Br J Surg* (2006) 93(1):113–9. doi: 10.1002/bjs.5233

26. Sessler DI, Pei L, Huang Y, Fleischmann E, Marhofer P, Kurz A, et al. Recurrence of breast cancer after regional or general anaesthesia: A randomised controlled trial. *Lancet* (2019) 394(10211):1807–15. doi: 10.1016/S0140-6736(19) 32313-X

27. Zhou KQ, Sun YF, Cheng JW, Du M, Ji Y, Wang PX, et al. Effect of surgical margin on recurrence based on preoperative circulating tumor cell status in hepatocellular carcinoma. *EBioMedicine* (2020) 62:103107. doi: 10.1016/ j.ebiom.2020.103107

28. Jeon J, Meza R, Moolgavkar SH, Luebeck EG. Evaluation of screening strategies for pre-malignant lesions using a biomathematical approach. *Math Biosci* (2008) 213 (1):56–70. doi: 10.1016/j.mbs.2008.02.006

29. Curtius K, Hazelton WD, Jeon J, Luebeck EG. A multiscale model evaluates screening for neoplasia in barrett's esophagus. *PloS Comput Biol* (2015) 11(5): e1004272. doi: 10.1371/journal.pcbi.1004272

30. Foo J, Leder K, Ryser MD. Multifocality and recurrence risk: A quantitative model of field cancerization. *J Theor Biol* (2014) 355:170–84. doi: 10.1016/ j.jtbi.2014.02.042

31. Ryser MD, Lee WT, Ready NE, Leder KZ, Foo J. Quantifying the dynamics of field cancerization in tobacco-related head and neck cancer: A multiscale modeling approach. *Cancer Res* (2016) 76(24):7078–88. doi: 10.1158/0008-5472.CAN-16-1054

32. Evans EJJr., DeGregori J. Cells with cancer-associated mutations overtake our tissues as we age. *Aging Cancer* (2021) 2(3):82–97. doi: 10.1002/aac2.12037

33. Takaki M, Haeno H. Mathematical modeling of locoregional recurrence caused by premalignant lesions formed before initial treatment. *Front Oncol* (2021) 11:743328. doi: 10.3389/fonc.2021.743328

34. Wang Z, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (Tcga). *Methods Mol Biol* (2016) 1418:111–41. doi: 10.1007/978-1-4939-3578-9_6

35. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* (2018) 173(2):400–16.e11. doi: 10.1016/j.cell.2018.02.052

36. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. *Cell* (2017) 171(5):1029–41 e21. doi: 10.1016/j.cell.2017.09.042

37. Anandakrishnan R, Varghese RT, Kinney NA, Garner HR. Estimating the number of genetic mutations (Hits) required for carcinogenesis based on the distribution of somatic mutations. *PloS Comput Biol* (2019) 15(3):e1006881. doi: 10.1371/journal.pcbi.1006881

38. Iranzo J, Martincorena I, Koonin EV. Cancer-mutation network and the number and specificity of driver mutations. *Proc Natl Acad Sci U.S.A.* (2018) 115 (26):E6010–E9. doi: 10.1073/pnas.1803155115

39. Moran P. *The statistical processes of evolutionary theory*. Oxford: Clarendon (1962).

40. Athreya KB, Ney PE, Ney P. *Branching processes*. Courier Corporation, New York: Dover Publications (2004).

41. Piraino SW, Furney SJ. Beyond the exome: The role of non-coding somatic mutations in cancer. *Ann Oncol* (2016) 27(2):240–8. doi: 10.1093/annonc/mdv561

42. Lawlor K, Perez-Montero S, Lima A, Rodriguez TA. Transcriptional versus metabolic control of cell fitness during cell competition. *Semin Cancer Biol* (2020) 63:36–43. doi: 10.1016/j.semcancer.2019.05.010

43. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys* (1976) 22(4):403–34.

44. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* (2012) 2(5):401–4. doi: 10.1158/2159-8290.CD-12-0095

45. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci Signal* (2013) 6(269):pl1. doi: 10.1126/scisignal.2004088

46. Nakabayashi N, Hirose M, Suzuki R, Suzumiya J, Igawa M. How asymptomatic are early cancer patients of five organs based on registry data in Japan. *Int J Clin Oncol* (2018) 23(5):999–1006. doi: 10.1007/s10147-018-1287-2

47. Zeng H, Ran X, An L, Zheng R, Zhang S, Ji JS, et al. Disparities in stage at diagnosis for five common cancers in China: A multicentre, hospital-based, observational study. *Lancet Public Health* (2021) 6(12):e877–e87. doi: 10.1016/S2468-2667(21)00157-2

48. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin* (2022) 72(1):7–33. doi: 10.3322/caac.21708

49. Tomasetti C, Vogelstein B. Cancer etiology. variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* (2015) 347 (6217):78–81. doi: 10.1126/science.1260825

50. Cercek A, Lumish M, Sinopoli J, Weiss J, Shia J, Lamendola-Essel M, et al. Pd-1 blockade in mismatch repair-deficient, locally advanced rectal cancer. *N Engl J Med* (2022) 386(25):2363–76. doi: 10.1056/NEJMoa2201445

51. Jain IH, Calvo SE, Markhard AL, Skinner OS, To TL, Ast T, et al. Genetic screen for cell fitness in high or low oxygen highlights mitochondrial and lipid metabolism. *Cell* (2020) 181(3):716–27.e11. doi: 10.1016/j.cell.2020.03.029

52. Fennell KA, Vassiliadis D, Lam EYN, Martelotto LG, Balic JJ, Hollizeck S, et al. Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature* (2022) 601(7891):125–31. doi: 10.1038/s41586-021-04206-7

53. Yang L, Wang S, Lee JJ, Lee S, Lee E, Shinbrot E, et al. An enhanced genetic model of colorectal cancer progression history. *Genome Biol* (2019) 20(1):168. doi: 10.1186/s13059-019-1782-4

54. Weaver JMJ, Ross-Innes CS, Shannon N, Lynch AG, Forshew T, Barbera M, et al. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat Genet* (2014) 46(8):837–43. doi: 10.1038/ng.3013

55. Smyth EC, Lagergren J, Fitzgerald RC, Lordick F, Shah MA, Lagergren P, et al. Oesophageal cancer. *Nat Rev Dis Primers* (2017) 3:17048. doi: 10.1038/nrdp.2017.48

56. Thrift AP. Global burden and epidemiology of Barrett oesophagus and oesophageal cancer. *Nat Rev Gastroenterol Hepatol* (2021) 18(6):432–43. doi: 10.1038/s41575-021-00419-3

57. Sharma P. Barrett Esophagus: A review. *JAMA* (2022) 328(7):663–71. doi: 10.1001/jama.2022.13298

58. Kim H, Kim TH, Choe JH, Kim JH, Kim JS, Oh YL, et al. Patterns of initial recurrence in completely resected papillary thyroid carcinoma. *Thyroid* (2017) 27 (7):908–14. doi: 10.1089/thy.2016.0648

59. Wang B, Zhang S, Yue K, Wang XD. The recurrence and survival of oral squamous cell carcinoma: A report of 275 cases. *Chin J Cancer* (2013) 32(11):614–8. doi: 10.5732/cjc.012.10219

60. Hu LS, Zhang XF, Weiss M, Popescu I, Marques HP, Aldrighetti L, et al. Recurrence patterns and timing courses following curative-intent resection for intrahepatic cholangiocarcinoma. *Ann Surg Oncol* (2019) 26(8):2549–57. doi: 10.1245/s10434-019-07353-4

61. Wilt TJ, Jones KM, Barry MJ, Andriole GL, Culkin D, Wheeler T, et al. Follow-up of prostatectomy versus observation for early prostate cancer. *N Engl J Med* (2017) 377(2):132–42. doi: 10.1056/NEJMoa1615869

62. Zhang H, Liu F, Wen N, Li B, Wei Y. Patterns, timing, and predictors of recurrence after laparoscopic liver resection for hepatocellular carcinoma: Results from a high-volume hpb center. *Surg Endosc* (2022) 36(2):1215–23. doi: 10.1007/s00464-021-08390-5

63. Kostron A, Friess M, Crameri O, Inci I, Schneiter D, Hillinger S, et al. Relapse pattern and second-line treatment following multimodality treatment for malignant pleural mesothelioma. *Eur J Cardiothorac Surg* (2016) 49(5):1516–23. doi: 10.1093/ ejcts/ezv398

64. Murakami N, Matsumoto F, Yoshimoto S, Ito Y, Mori T, Ueno T, et al. Patterns of recurrence after selective postoperative radiation therapy for patients with head and neck squamous cell carcinoma. *BMC Cancer* (2016) 16:192. doi: 10.1186/s12885-016-2229-x

65. Harter P, Sehouli J, Vergote I, Ferron G, Reuss A, Meier W, et al. Randomized trial of cytoreductive surgery for relapsed ovarian cancer. *N Engl J Med* (2021) 385 (23):2123–31. doi: 10.1056/NEJMoa2103294

66. Speed JM, Trinh QD, Choueiri TK, Sun M. Recurrence in localized renal cell carcinoma: A systematic review of contemporary data. *Curr Urol Rep* (2017) 18(2):15. doi: 10.1007/s11934-017-0661-3

67. Cho SI, Lee J, Jo G, Kim SW, Minn KW, Hong KY, et al. Local recurrence and metastasis in patients with malignant melanomas after surgery: A single-center analysis of 202 patients in south Korea. *PloS One* (2019) 14(3):e0213475. doi: 10.1371/ journal.pone.0213475

68. Ozbir S, Girgin C, Kara C, Dincel C. Local and systemic recurrence patterns of urothelial cancer after radical cystectomy. *Kaohsiung J Med Sci* (2014) 30(10):504–9. doi: 10.1016/j.kjms.2014.03.011

69. Elsayed M, Alhussini M, Basha A, Awad AT. Analysis of loco-regional and distant recurrences in breast cancer after conservative surgery. *World J Surg Oncol* (2016) 14:144. doi: 10.1186/s12957-016-0881-x

70. Huang J, Tong Y, Chen X, Shen K. Prognostic factors and surgery for breast cancer patients with locoregional recurrence: An analysis of 5,202 consecutive patients. *Front Oncol* (2021) 11:763119. doi: 10.3389/fonc.2021.763119

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 04 July 2023
DOI 10.3389/frai.2023.1060879

Check for updates

OPEN ACCESS

EDITED BY
Kathleen Wilkie,
Toronto Metropolitan University, Canada

REVIEWED BY
Frederick R. Adler,
The University of Utah, United States
Kevin Leder,
University of Minnesota Twin Cities, United States

*CORRESPONDENCE
Thomas Hillen
✉ thillen@ualberta.ca

RECEIVED 03 October 2022
ACCEPTED 05 June 2023
PUBLISHED 04 July 2023

CITATION
Deutscher K, Hillen T and Newby J (2023) A computational model for the cancer field effect. *Front. Artif. Intell.* 6:1060879. doi: 10.3389/frai.2023.1060879

COPYRIGHT
© 2023 Deutscher, Hillen and Newby. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A computational model for the cancer field effect

Karl Deutscher, Thomas Hillen* and Jay Newby

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

Introduction: The Cancer Field Effect describes an area of pre-cancerous cells that results from continued exposure to carcinogens. Cells in the cancer field can easily develop into cancer. Removal of the main tumor mass might leave the cancer field behind, increasing risk of recurrence.

Methods: The model we propose for the cancer field effect is a hybrid cellular automaton (CA), which includes a multi-layer perceptron (MLP) to compute the effects of the carcinogens on the gene expression of the genes related to cancer development. We use carcinogen interactions that are typically associated with smoking and alcohol consumption and their effect on cancer fields of the tongue.

Results: Using simulations we support the understanding that tobacco smoking is a potent carcinogen, which can be reinforced by alcohol consumption. The effect of alcohol alone is significantly less than the effect of tobacco. We further observe that pairing tumor excision with field removal delays recurrence compared to tumor excision alone. We track cell lineages and find that, in most cases, a polyclonal field develops, where the number of distinct cell lineages decreases over time as some lineages become dominant over others. Finally, we find tumor masses rarely form via monoclonal origin.

KEYWORDS

cancer field effect, field cancerization, carcinogenesis, head and neck squamous cell carcinoma, computational modeling, hybrid cellular automaton, smoking

1. Introduction

The idea of field cancerization was first mentioned by Slaughter et al. (1953) in 1953 when histologically observing 783 squamous-cell tumors in oral cancers. Within the entire patient population it was found that benign epithelium surrounding the malignant tumor was abnormal. As well, some of the patients had multiple separate tumors occur in the same area of the oral cavity. From these observations, Slaughter et al. proposed a process termed *field cancerization*, in which a carcinogenic agent preconditions an area of epithelium toward cancer. If an area of epithelium is exposed to a carcinogenic agent for a sufficient amount of time and with enough intensity then it produces irreversible changes in cells and cell groups, such that the process toward cancer becomes inevitable (Slaughter et al., 1953). In Slaughter et al. (1953), it is also noted that a field of preconditioned epithelium may develop cancer at multiple points and possibly lead to multiple tumors. As a result, cancer does not arise from one cell that suddenly becomes malignant but instead from areas of precancerous change. Local recurrence after surgery or radiation occurs due to left-over benign epithelium that is preconditioned toward cancer, i.e., from the remaining pre-cancer field. Many papers were written following (Slaughter et al., 1953) that showed field cancerization can be found in colon carcinoma (Galandiuk et al., 2012; Alonso et al., 2015), gastric carcinoma (Kang et al., 1997; Zaky et al., 2008; Takeshima et al., 2015), esophageal carcinoma (Cense et al., 1997; Oka et al., 2009; Lee et al., 2011), non-small-cell lung squamous carcinoma (Franklin et al., 1997; Steiling et al., 2008; Kadara and Wistuba, 2012), non-small-cell lung adenocarcinoma (Gomperts et al., 2013; Kadara et al., 2014),

head and neck squamous cell carcinoma (HNSCC) (oral, oropharynx, hypopharynx, larynx) (Slaughter et al., 1953; Califano et al., 1996; Braakhuis et al., 2003; Angadi et al., 2012), breast carcinoma (Trujillo et al., 2011; Rivenbark and Coleman, 2012; Foschini et al., 2013), cervix (Chu et al., 1999), prostate carcinoma (Nonn et al., 2009; Trujillo et al., 2012), bladder carcinoma (Hafner et al., 2002), and skin carcinoma (Kanjilal et al., 1995; Hu et al., 2012; Szeimies et al., 2012).

Biomarkers that were discovered to correlate with the presence of a cancer field are loss of heterozygosity (LOH) (Tabor et al., 2001), micro-satellite alterations (Tabor et al., 2001), chromosomal instability (Hittelman, 2001), and mutations in the TP53 gene (Brennan et al., 1995; van Houten et al., 2002). Braakhuis et al. (2003) related field cancerization to genetics and identified: "*growth of one or more genetically altered cell(s) that produces a field of cells predisposed to subsequent tumor growth*." Based on genetic evidence, there currently exists two main hypotheses that explain the underlying cellular basis of field cancerization: polyclonal origin and monoclonal origin. Polyclonal origin proposes that mutations occur in multiple sites of the epithelium due to continuous carcinogen exposure, which leads to multi-focal carcinomas or lesions of independent origin (van Oijen and Slootweg, 2000). Monoclonal origin proposes that the mutant cells from the initial lesion migrate and develop multiple lesions that share a common clonal origin (Braakhuis et al., 2003). Here, we show that while all cases are possible in our model, a polyclonal field is by far the most common outcome of our simulations.

Another breakthrough in biology since (Slaughter et al., 1953) was the discovery of cancer stem cells CSCs and their importance in cancer initiation, progression, and treatment. Simple et al. (2015) explain field cancerization using Braakhuis model of genetic alterations (Braakhuis et al., 2003) plus the addition of CSCs. They consider both monoclonal and polyclonal origin within their model. In Simple's model (Simple et al., 2015) for oral cancer, a continuous exposure of the oral mucosa to carcinogens results in molecular alterations that lead to the induction of CSC-like behavior in a step-wise manner. CSCs originate either by transformation of normal stem cells (NSCs), or by dedifferentiation of the tumor cells (TC) and migration through normal mucosa to develop the field. Repeated mutations at 17p (the location of the TP53 gene) and 3p/9p (p16/FHIT gene) lead to transformation of the NSCs into transit amplifying cells (TACs). These transformed cells divide and expand to create a field of neoplastic cells. Finally, a genetic hit in the cells within the field at 13q (the location of the Rb gene) allows a carcinoma to develop. Note that alteration to the Rb gene is known to release CSCs from their quiescent stage such that proliferation, self-renewal and formation of tumors can occur. The work of Simple motivated us to include gene expression levels explicitly in our model, and to use a neural network to describe those changes.

Recently Curtius et al. (2017) studied field cancerization from an evolutionary perspective. They define a cancerized field to be a single cell or group of cells that are further along the evolutionary path toward cancer. Driver mutations have been found in both the carcinoma and the cancerized field thus indicating that a driver mutation may also be a field cancerization characteristic. As a result, field cancerization can occur because of multiple independent clonal expansions, i.e., polyclonal origin. Thus, both Simple et al. (2015) and Curtius et al. (2017) consider that a cancerized field can be formed via monoclonal or polyclonal origin.

1.1. Our model

In our model, a cancer field will be considered as a region of tissue that has genetic and phenotypic change that preconditions it toward the possible formation of one or more tumors within it. The genetic and phenotypic change can be caused by carcinogenic onslaught, genetic defects at birth, mutations later in life, or a combination thereof. We focus on the effect of a certain body region or organ—such as the mouth or the tongue—and we do not consider a system-wide pre-mutation as a cancer field.

The steps of the process of field cancerization that will be considered here are as follows:

- A region of tissue is repeatedly affected by one or more carcinogens over time, for example through smoking tobacco.
- The carcinogen(s) cause genetic mutations in the cells of the tissue which in turn influence the phenotype of the cell;
- As the cells start to proliferate and differentiate, the field expands;
- Eventually a CSC will be created, which will finally create the first TC and consequently a tumor.

We develop the model in the framework of a hybrid cellular automaton as introduced by Gerlee and Anderson (2007). The details will be explained in Section 2. With this model we try to answer:

1. What degree of carcinogenic onslaught is necessary for field cancerization to occur? Which carcinogens are the most aggressive, smoking related carcinogens or alcohol?
2. How long before a cancer field is formed? How long before the first tumor cell within the field is formed?
3. Is the field formed via monoclonal origin, polyclonal origin, or a mixture of both? Which type of origin is the most common?
4. How long does it take for a tumor to be large enough such that it is detectable by physicians? Once a tumor is detected, what size is the surrounding field?
5. How long does it take a recurrence to occur after removal of the tumor vs. the removal of tumor and field?
6. What are the dynamics of different cell lineages in an established cancer field?

Though field cancerization is found in many types of tissue throughout the body, the most commonly studied case is head and neck squamous cell carcinoma (HNSCC), which we consider here. Alcohol and smoking are the most commonly associated carcinogens to HNSCC. These two carcinogens typically enter the body through smoking and/or chewing tobacco and drinking alcohol, respectively. We use parameter values that are typically associated to smoking and alcohol drinking.

We find in our model, that a continued, long-time onslaught of carcinogen on tissue of the mouth inevitably leads to a formation

of a cancer field. Our model confirms the general understanding (Hashibe et al., 2009; LoConte et al., 2018) that smoking induced carcinogens are much more potent than alcohol. The timing of the first tumor cell is usually quite long, in the order of 10s of years, hence for many people, a cancer will not arise. In almost all situations the field is polyclonal. Monoclonal fields are only seen in very small domains. A removal of the tumor, in our model, leads to a quick recurrence if the field is left behind. If the field is removed as well, then recurrence takes very long time. When we follow the cell lineages, we see that new lineages form and several of them will die out over time. However, some lineages establish themselves and become cancerous. In that case we observe a polyclonal cancer field, and also polyclonal cancers.

1.2. Carcinogenesis

Most carcinogenesis models consider that cancer is initialized from the result of a multi-step process (Frank, 2007). A normal cell does not become a cancer cell until multiple genetic alterations accumulate within it. The number of genetic alterations in a cancer cell is an indicator of the level of malignancy of the cell.

Gatenby and Gillies (2008) found six micro-environmental barriers for a malignant phenotype: apoptosis with loss of basement membrane contact, inadequate growth promotion, senescence (deterioration of a cells' power of division and growth with age), hypoxia (deficiency in the amount of oxygen reaching the tissues), acidosis (excessively acidic condition of the body fluids or tissues), and ischaemia (restriction of blood supply to tissues, causing hypoxia). The development of cancer occurs when a normal cell overcomes at least one of these barriers. Thus, the micro-environment is an important factor to consider in cancer initialization.

A normal cell lineage can acquire mutations (Curtius et al., 2017), that are positively selected in the micro-environment of a healthy organ. A driver mutation is one that confers growth or survival advantages for tumor cells within the appropriate micro-environment (Greaves et al., 2003; Calabrese et al., 2004; Stratton et al., 2009). A passenger (neutral) mutation is one that passively accumulates in cell lineages (Greaves et al., 2003; Calabrese et al., 2004; Stratton et al., 2009). It may be that some driver mutations are not currently affecting cancer growth but instead had previously driven the growth of an lineage (Curtius et al., 2017). Progression to cancer usually requires the accumulation of multiple driver mutations (Weaver et al., 2014). A mutant lineage/clone, can grow to produce large patches, or fields, of cells that are predisposed to eventually progress to neoplasm.

It has been reported by Knopf et al. (2015) that at least 232 genes are directly involved in HNSCC of young patients. Here, we focus on 10 genes of importance: TP53, TP73, RB, TP21, T16, EGFR, CCDN1, MYC, PIK3CA, and RAS (Knopf et al., 2015). Some of these genes are oncogenes (EGFR, CCDN1, MYC, PIK3CA, RAS), i.e., supporting cancer development if over expressed, and some are tumor suppressor genes (TP53, TP73, RB, TP21, TP16), i.e., support tumor growth if expression is inhibited.

1.3. Cancer stem cells

Before discussing cancer stem cells (CSCs) it should be noted that there is no single standardized definition of CSCs. Instead, many slightly different and sometimes contradictory definitions have emerged, each suited to a particular study. In general, CSCs are not normal stem cells (NSC), they are cells that have some of the characteristics of NSC. We consider CSCs as multipotent cells in a tumor that like NSCs have self-renewal ability, but in addition, have the abilities of tumor initiation, migration and metastasis (Biddle et al., 2011; Bu and Cao, 2012).

The origin of CSCs is explained by three possible processes. The first process states that a NSC undergoes several genetic as well as epigenetic alterations to give rise to a CSC (Feller et al., 2013). The second process states that CSCs originate from NSCs that acquire a precancerous phenotype during their development stage (Bjerkvig et al., 2005; Feller et al., 2013; González-Moles et al., 2013). The third process states that the CSC originate from mature tumor cells (Moon et al., 2011; Kumar et al., 2012; Herreros-Villanueva et al., 2013; Di Fiore et al., 2014) or epithelial cells (Bjerkvig et al., 2005; Feller et al., 2013; González-Moles et al., 2013) that undergo dedifferentiation into a CSC through modifications in signaling pathways and regulatory mechanisms. Note that the first and second processes only differ in whether an NSC acquires a genetic alteration when it is fully developed or still in development.

1.4. Previous mathematical models

There exists an extensive amount of literature that studies cancer initiation (Gentry and Jackson, 2013; Durrett et al., 2016; Paterson et al., 2020), progression (Beerenwinkel et al., 2014; Enderling and Chaplain, 2014), metastasis (Franssen et al., 2019), treatment (chemotherapy, immunotherapy, radiation) (de Pillis et al., 2009; Enderling and Chaplain, 2014; Radunskaya et al., 2018), and effects of various micro-environmental and external factors on cancer development (Gerlee and Anderson, 2007). However, the only mathematical model for field cancerization, that could be found at the time of writing, is the model by Foo et al. and its follow-up studies (Foo et al., 2014, 2020, 2022; Ryser et al., 2016). Here, we will further extend Foo's model using the hybrid cellular automaton approach of Gerlee and Anderson (2007).

Foo et al. (2014) describe field cancerization as a spatial Moran process on a square lattice. The cells are classified into three phenotypes, $k = 0$ healthy, $k = 1$ pre-cancerous, $k = 2$ cancer cells. These phenotypes have different fitness, with healthy having the lowest fitness, pre-cancer cells an intermediate fitness, and cancer cells the highest fitness. Random mutations allow cells to transit from $k = 0$ to $k = 1$ and $k = 2$. Foo et al. (2014) consider the dependence of their model on the parameters, and they identify three regimes. A first regime where pre-cancer cells quickly become cancerous. In that case no field is generated, as the progression to cancer is fast. A second regime where progress from pre-cancerous to cancerous is slow. This leads to a significant sized cancer field and multiple lesions in the tissue. And a third regime, where Field development and cancer development are on the same time scale. More recently, Foo et al. (2020, 2022) extended the model to include

a three dimensional tissue structure consisting of epithelial layers, and more than $k = 2$ phenotypes.

We use these models by Foo et al.'s (2014, 2020, 2022) as a starting point for our model and we extend it by (i) allowing multiple mutations at different genes, which can or cannot lead to cancer development, (ii) allow for non-constant micro-environments, (iii) consider carcinogens as mutational driver, (iv) follow cell lineages, and consider additional phenotypic action such as apoptosis, quiescence, transit amplifying cells, and de-differentiation.

Ryser et al. (2016) applied the model described in Foo et al. (2014) to head and neck squamous cell carcinoma (HNSCC). The three histopathological stages of epithelial dysplasia (precancerous stages) are mild, moderate, and severe (carcinoma *in situ* [CIS]). Ryser et al. (2016) consider the following four type of cells: normal cells (type 0), mildly dysplastic cells (type 0*), moderately dysplastic cells (type 1), and severely dysplastic cells (type 2). They use the stochastic Moran model on a regular two-dimensional lattice as described in Foo et al. (2014). To estimate and compute the parameters for their model they use age-specific incidence rates from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (18 registries, 2000–2012) in a Bayesian framework. Ryser et al. (2016) computed the survival function, the probability density function of the local field radius, and the probability of harboring at least two clonally unrelated fields in the head and neck region with respect to the mean age at smoking initiation to diagnosis with invasive cancer. They found that there is a strong dependence of the local field size on age at diagnosis, with a doubling of the expected field diameter between ages at diagnosis of 50 and 90 years. Further the probability of harboring multiple clonally unrelated fields at the time of diagnosis were found to increase substantially with patient age. As a result of these discoveries they suggest that patient age at diagnosis is a critical predictor of the size and multiplicity of precancerous lesions.

Our extensions of the model of Foo et al. (2014) will be in the framework of a hybrid cellular automaton as developed by Gerlee and Anderson (2007). Gerlee and Anderson (2007) created a hybrid cellular automaton to model the effect of various micro-environmental factors on solid tumor growth. Their model is a hybrid cellular automaton because the rule of the automaton depends upon the output of a neural network and partial differential equations. The cellular automaton is comprised of two cell types: an empty cell (normal cell) and a tumor cell. The neural network is used to approximate the relationship between the micro-environmental variables and the phenotype of a cell. The partial differential equations are used to model the spread of the various micro-environmental variables in the domain of consideration. While Gerlee and Anderson (2007) considered tumor growth, they did not specifically study the cancer field effect.

For the neural network they use a multi-layer perceptron (MLP), with input being the output of the partial differential equation for the cell at a location (x, y) and output being a vector of likelihoods of a phenotype and movement occurring at a time-step. The hidden layer of the MLP represents the genes and hence the neural network attempts to replicate the genotype-phenotype relationship. They consider the phenotypes proliferation (P), quiescence (Q), and apoptosis (A). Each time-step represents a cell cycle so that a single phenotypic action will occur once per time step for each cell. The maximum of the likelihoods between P, Q, and A determines which phenotypic action occurs. If the likelihood of movement is sufficiently large then the cell is allowed to move. The quiescent state is used to describe any normal activity of the cell which is not one of the three actions explicitly modeled above.

We extend Gerlee's model by (i) considering six cell types (normal tissue cells NTC, mutated normal tissue cells MNTC, normal stem cells NSC, mutated normal stem cells MNSC, cancer stem cells CSC, tumor cells TC, and empty cells) (ii) including transit amplifying cells (iv) including four phenotypic actions (proliferation, quiescence, apoptosis, differentiation) (v) including 10 gene expression levels for TP53, TP73, RB, TP21, T16, EGFR, CCDN1, MYC, PIK3CA, and RAS. (vi) following cell lineages, and (vii) considering full or partial surgical removal of cancer and cancer field cells.

2. The hybrid cellular automaton model

Here, we develop a hybrid cellular automaton (CA) model for the cancer field effect. We model on three distinct levels, a model for the carcinogen distribution, a neural network for the gene expression and a model for the cell dynamics over time.

We use carcinogen concentration function $c_i(x, y)$ to describe the carcinogen input on a two dimensional domain, where the index i is used to distinguish different carcinogens, for example $i = 1$ for smoking induced carcinogens, and $i = 2$ for alcohol induced carcinogens. In our simulations we consider a wide variety of carcinogen distributions $c_i(x, y)$, including a uniform, smooth spatial patterns, and random distributions. We found that the Gaussian distribution described below gave the most instructive results, hence we only report those. We choose a Gaussians centered in the domain middle as

$$c_i(x,y) = \exp\left(-\frac{1}{2}\frac{(x-\mu)^2+(y-\mu)^2}{\sigma^2}\right),\ \mu = \frac{N}{2}-1,\ \sigma = \frac{N}{15}; \tag{1}$$

where N is the domain size (number of grid cells), assuming a square domain is used.

2.1. Gene expression neural network

In this section, we describe a deep neural network with one hidden layer to account for varying levels of gene expressions through cell age and carcinogenic onslaught. A schematic of the layers of the neural network is given in Figure 1. We consider $G \in \mathbb{N}$ genes that are biomarkers to the considered cancer type. Later we will model the gene expressions for HNSCC of the 10 genes TP53, TP73, RB, TP21, TP16, EGFR, CCDN1, MYC, PIK3CA, and RAS. Here, we formulate the model in general terms first.

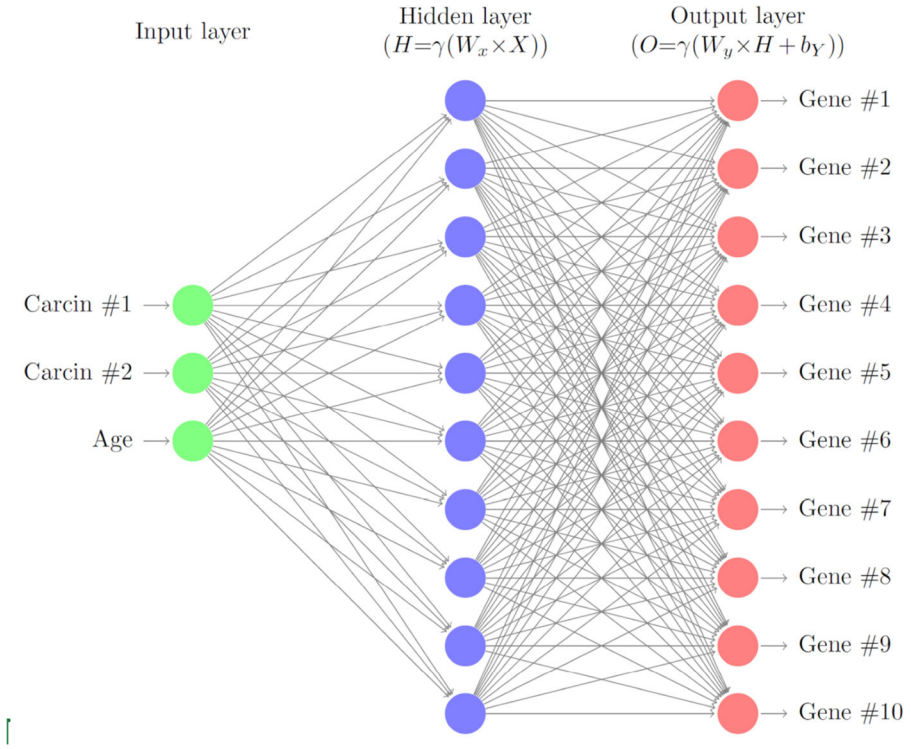


FIGURE 1
Schematic of the gene neural network. As input we use the carcinogen concentrations and the cell age. The hidden layer relates the carcinogens to the gene expression levels, which are the output of the neural network.

2.1.1. General case

A good general reference for neural networks can be found in Hastie et al. (2009). The gene expression of each gene is represented by the function

$$e_j(t) \in \mathbb{R}, \qquad j = 1, 2, ..., G. \tag{2}$$

The gene expression is a non-dimensional value that is zero when the expression is normal, negative when it is under-expressed, and positive when it is over-expressed. The gene expression of each gene changes over time based upon a simple multi-layer perceptron (MLP). The input of the MLP is the vector

$$\boldsymbol{X}(t) := [\{c_i(t-1)\}_{i=1,...,C}, \alpha(t-1)]^T \in \mathbb{R}_+^{C+1}, \tag{3}$$

where $c_i(t)$ are the carcinogen concentrations, C denotes the number of carcinogen considered and $\alpha(t)$ is the age of the cell. The cell age α is measured for each cell after mitosis. Changes in gene expression are based upon the carcinogens in the environment of the cell and the age, which essentially means we are looking at the effects of the carcinogens and replication errors as a cell ages. The output of the MLP is given by

$$\boldsymbol{Y}(t) := [\{\overline{\delta}_j(t)\}_{j=1,2,...,G}]^T \in \mathbb{R}^G, \tag{4}$$

where $\overline{\delta}_j(t)$ is the computed maximum possible change in gene expression for gene j. The amount the gene j will be mutated in a time-step is a random sample from the uniform distribution multiplied by $\overline{\delta}_j(t)$.

$\boldsymbol{Y}(t)$ is computed using matrix multiplication, addition and application of a non-linear transform. The hidden layer is computed by

$$\boldsymbol{H}(t) := \gamma(W_X \boldsymbol{X}(t)) \in \mathbb{R}^G, \tag{5}$$

where

$$\gamma(\xi) := \frac{\xi}{\sqrt{1+\nu\xi^2}}, \in \left(\frac{-1}{\sqrt{\nu}}, \frac{1}{\sqrt{\nu}}\right) \tag{6}$$

is the non-linear activation function that is applied element wise to a vector, and $W_X \in \mathbb{R}^{G\times C+1}$ is a weight matrix. Note that the activation function is chosen to ensure $|\overline{\delta}_j(t)| < \frac{1}{\sqrt{\nu}}$, hence allowing us to control the maximum amount the expression of gene j can change in a time-step via ν. After the hidden layer is computed the output is computed by

$$\boldsymbol{Y}(t) = \gamma(W_Y \boldsymbol{H}(t) + \boldsymbol{b}_Y(t)), \tag{7}$$

where $W_Y \in \mathbb{R}^{G\times G}$ is a weight matrix and $\boldsymbol{b}_Y(t) \in \mathbb{R}^G$ is a bias vector.

Biologically speaking $W_X^{(i,j)}, i \in [1, G], j \in [1, C]$ represents how carcinogen i influences gene j, $W_X^{(i,C+1)}$ represents whether cell age influences gene i, $W_Y^{(i,j)}$ represents whether gene i influences gene j, and $b_Y^{(i)}(t)$ denotes whether gene i has a higher or lower chance of gene expression changes relative to other genes. Note that if a value in the weight matrices is negative it means there is a negative relationship, if it is positive it means there is a positive relationship, and finally if it is zero it means there is no relationship.

We assume that as the cell ages, replication errors increase in frequency and cause random changes in gene expression. In the case of determining how age affects each gene, the values of $W_X^{(i,C+1)}, i \in [1, G]$ are randomly made positive or negative at every time-step by sampling a Bernoulli random variable $b_{i,t}$, with success probability parameter $p = 0.5$, and scaling $W_X^{(i,C+1)}, i \in [1, G]$ by $\tilde{\alpha}_{i,t} \in \{\pm 1\}$ where

$$b_{i,t} \sim \text{Bernoulli}(p = 0.5) \tag{8}$$

$$\tilde{\alpha}_{i,t} = 2b_{i,t} - 1. \tag{9}$$

Let $z_j \sim \text{Uniform}(0, 1), j = 1, \ldots, G$, be independent uniform random variables. The gene expression, $e_j(t)$, of a gene is updated by

$$e_j(t) = e_j(t-1) + z_j \overline{\delta}_j(t). \tag{10}$$

A gene j is considered to be mutated if its gene expression is above the threshold value $\overline{M} \in \mathbb{R}_+$, i.e., $|e_j(t)| \geq \overline{M}$. The bias for a gene j, $b_Y^{(j)}(t)$, is updated through the relation

$$b_Y^{(j)}(t) = \begin{cases} \beta, & e_j(t-1) \geq \overline{M} \\ -\beta, & e_j(t-1) \leq -\overline{M} \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

where $\beta \in \mathbb{R}_+$ is preset parameter.

2.1.2. Neural network parameters for the HNSCC case

For the case of HNSCC we make the following choices. We consider the $G = 10$ genes TP53, TP73, RB, TP21, TP16, EGFR, CCDN1, MYC, PIK3CA, and RAS. The type of gene is summarized in Table 1.

The weight matrix associated with the input of the neural network (5) is given by:

$$W_X(t) = \begin{bmatrix} 1 & -1 & \tilde{\alpha}_{1,t}10^{-7} \\ 0 & 0 & \tilde{\alpha}_{2,t}10^{-7} \\ 0 & -1 & \tilde{\alpha}_{3,t}10^{-7} \\ 1 & -1 & \tilde{\alpha}_{4,t}10^{-7} \\ 1 & 0 & \tilde{\alpha}_{5,t}10^{-7} \\ 1 & 1 & \tilde{\alpha}_{6,t}10^{-7} \\ 1 & 1 & \tilde{\alpha}_{7,t}10^{-7} \\ 0 & 1 & \tilde{\alpha}_{8,t}10^{-7} \\ 0 & 1 & \tilde{\alpha}_{9,t}10^{-7} \\ 1 & 1 & \tilde{\alpha}_{10,t}10^{-7} \end{bmatrix}, \tag{12}$$

where $\tilde{\alpha}_{i,t}$ is defined by (9). As sufficient data was unavailable we assumed that each carcinogen has a weight of 1, −1, or 0 for each gene depending on how the carcinogen effects that gene. For example since ethanol tends to upregulate TP53 then $W_X^{11} = 1$. The impact of alcohol and smoking induced carcinogens on gene mutations was taken from the large online data bases (PubChem, 2023a,b), and are listed in the last two columns in Table 1.

We assume that each gene has the same mutation rate which causes the last column in W_X, that is associated with mutations caused by transcription errors due to cell age, to have one value. The mutation rate was chosen based upon the human genomic mutation rate being approximately 2.5×10^{-8} per base per generation (Nachman and Crowell, 2000). The weight matrix associated with the output of the neural network (7) is given by:

$$W_Y = \begin{bmatrix} 1.00 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0.1 & 0 & 0 & 0.01 & -0.01 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0.01 & 0 & 0.01 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0.01 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3 & 0 & 0.01 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0 \\ 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0.3 \end{bmatrix}. \tag{13}$$

The main diagonal of the above matrix gives the main weights for each gene with W_Y^{11} being the highest as it is TP53. Each diagonal value was given a default of 0.1 and it is increased by 0.1 for each gene it calls or is related to, so TP53 gets a value of 1 because it is assumed all the genes relate to TP53. Each column describes the relations between the other genes and the gene associated with the main diagonal value of that column, where if the gene is upregulated by the diagonal gene it gets a value of 0.01 and when it downregulates the gene it gets a value of −0.01. The magnitude of the values in the matrix were chosen by trial and error since there is not sufficient data to complete the matrix with accurate values.

The activation function (6) parameter is given by $\nu = 10^6$. The value of ν results in the neural network outputting values in the range $(\frac{-1}{\sqrt{\nu}}, \frac{1}{\sqrt{\nu}}) = (-10^{-3}, 10^{-3})$ and was chosen so to keep the maximum amount each gene can change to a reasonable figure.

Finally the mutation bias vector update function (11) parameter is given by $\beta = 10^{-3}$. The value of β was chosen to correspond with the maximum output value of the neural network, so that when a gene is mutated, the neural network will always output the maximum value.

2.2. Cellular automaton

We consider a two dimensional regular grid with N grid cells. Each cell can be occupied by a cell, or be empty. The six cell classes that we consider are normal tissue cells (NTC, brown), mutated normal tissue cells (MNTC, green), normal stem cells (NSC, blue), mutated normal stem cells (MNSC, yellow), cancer stem cells (CSC, purple), and tumor cells (TC, red), plus an empty cell (white) (see Table 2).

The cell class in the CA is represented by $s(t) \in \{0, 1, ..., 6\}$ with 0 = NTC, 1 = MNTC, 2 = NSC, 3 = MNSC, 4 = CSC, 5 = TC, 6 = empty.

Evolution of the model occurs in the following basic steps:

1. Given a carcinogen exposure, changes in gene expressions of a cell are computed by the neural network causing gene mutations to occur.
2. Based on the genetic profile, the cell executes one of four phenotypic actions: proliferation, quiescence, apoptosis, differentiation. If empty space is available in the neighborhood, the cell moves with a certain probability.

TABLE 1 List of 10 genes that are relevant for HNSCC.

Index	Gene	Gene-type	Regulation	Phenotypes	Carcinogen 1	Carcinogen 2
1	TP53	Tumor-suppressor	Down	↑: p ↓: a, q	Up	Down
2	TP73	Tumor-suppressor	Down	↓: a		
3	RB	Tumor-suppressor	Down	↑: p, d ↓: q		Down
4	TP21	Tumor-suppressor	Down	↑: p	Up	Down
5	TP16	Tumor-suppressor	Down	↑: p	Up	
6	EGFR	Oncogene	Up	↑: p	Up	Up
7	CCDN1	Oncogene	Up	↓: a	Up	Up
8	MYC	Oncogene	Up	↑: p, d ↓: a		Up
9	PIK3CA	Oncogene	Up	↓: a		Up
10	RAS	Oncogene	Up	↑: p, d ↓: a	Up	Up

We indicate the type of gene as tumor suppressor or oncogene and show how the phenotypes change as the gene is up-or down regulated. The abbreviations p, a, q, d stand for the phenotypic actions of proliferation, apoptosis, quiescence, and differentiation, respectively. The column "regulation" indicates the expression level changes that make the cell more cancerous. The column "Phenotypes" indicates the relative change in p, a, q, d. In the last two columns we list how the given carcinogen changes the corresponding gene expression. Carcinogen 1 and 2 reflect the typical effects of smoking and alcohol consumption, respectively.

TABLE 2 CA cell classes.

Cell class	$s(t)$	Color
Normal tissue cell (NTC)	0	**brown**
Mutated normal tissue cell (MNTC)	1	**green**
Normal stem cell (NSC)	2	**blue**
Mutated normal stem cell (MNSC)	3	**yellow**
Cancer stem cell (CSC)	4	**purple**
Tumor cell (TC)	5	**red**
Empty cell	6	**white**

Here, we give a shortened description of the rules of the cellular automaton. A full description, involving mathematical formulas for all possible transitions, is given in the Appendix 1 (Supplementary material).

Each cell in the CA tracks the gene expression of the G genes in a vector defined by

$$\boldsymbol{E}(t) = [\{e_j(t)\}_{j=1,...,G}]. \quad (14)$$

The phenotype of a cell is tracked by a vector that contains probabilities for each type of phenotypic action occurring in a given time-step and is defined by

$$\boldsymbol{P}(t) = [p(t), q(t), a(t), d(t)], \quad (15)$$

where $p(t)$ represents proliferation, $q(t)$ represents quiescence, $a(t)$ represents apoptosis, and $d(t)$ represents differentiation. The probabilities are set such that $\boldsymbol{P}(t)$ generates a probability distribution, so that

$$\sum_{i=1}^{4} P_i(t) \equiv p(t) + q(t) + a(t) + d(t) = 1 \text{ and } P_i(t) \geq 0, \forall t. \quad (16)$$

At a time-step in the CA a phenotypic action is chosen to occur by sampling from the probability distribution generated from $\boldsymbol{P}(t)$. Since we do not want a cell to reproduce more than once in a time-step, each time-step represents the length of a typical cell cycle for the type of tissue under consideration.

When a NSC, MNSC, or CSC differentiate the resultant cell initially is a transit amplifying cell (TAC) for a set number of generations, Θ, after which it turns respectively into a NTC, MNTC, or TC. As a result of this each cell has two parameters $\overline{\tau}(t) \in \{0, 1\}$ and $\overline{n}(t) \in \{0, ..., \Theta\}$, where $\overline{\tau}(t)$ is a binary parameter used to determine if a cell is currently a TAC or not and $\overline{n}(t)$ is the number of generations a TAC cell lineage has produced. The parameters $\overline{\tau}(t)$ and $\overline{n}(t)$ are copied from parent to child cell and once $\overline{n}(t) = \Theta$ then $\overline{\tau}(t+1) = 0, \overline{n}(t+1) = 0$.

The final aspect of the cell that is tracked and represented in the overall cell state is the age of the cell, $\alpha(t) \in \mathbb{N}$. Hence, the complete state of a cell in the CA is given by the vector that contains cell type s, age α, gene expression E, phenotye vector P, TAC state $\bar{\tau}$, and TAC generation $\bar{n}$ as

$$\boldsymbol{S}(t) = [s(t), \alpha(t), \boldsymbol{E}(t), \boldsymbol{P}(t), \overline{\tau}(t), \overline{n}(t)] \quad (17)$$

Each cell has a neighborhood that contains itself, the cardinal directions around it, and the cells directly NE, SE, SW, and NW of the cell. In CA theory this is called the Moore neighborhood (Gray, 2003). The boundary conditions of the grid are standard periodic

boundary conditions for convenience. Other boundary conditions could be easily implemented too.

2.2.1. Cell mutations

The chosen G genes are known genes related to the type of cancer being studied. For the case of HNSCC, (see Table 1). We define the vector $\boldsymbol{T} \in \{0,1\}^G$, where $T_j = 0$ represents a tumor suppressor gene and $T_j = 1$ represents an oncogene. A gene j is positively mutated toward cancer (positively mutated) if either it is a tumor suppressor gene and its gene expression is downregulated, $e_j(t) \leq -\overline{M}$, or it is an oncogene and its gene expression is upregulated, $e_j(t) \geq \overline{M}$, where $\overline{M}$ is the given threshold. At each time-step the gene expression of each gene is updated from the results of the gene expression neural network from Section 2.1. The changes in the gene expression allow the gene to become mutated or even go from mutated to non mutated (normally expressed).

We also consider that a mutated gene can influence another gene, where we assume that a positively mutated gene will cause a positive mutation of a related gene. A non-positively mutated gene will cause a negative mutation (mutation that regulates a gene toward normal expression) of a related gene. We express this interdependence through a matrix $R \in \{0,1\}^{G\times G}$, where each entry, R_{ij}, represents whether gene i is related to gene j with $0 =$ unrelated and $1 =$ related. Note that the matrix R is not necessarily symmetric as a gene i might regulate gene j but not vice versa.

We show in the Appendix 1 (Supplementary material) how this update is done mathematically.

2.2.2. Update rules for phenotypic action

When a gene is mutated it can modify the probability of a phenotypic action occurring, $P_i(t)$. Hence the phenotypic actions P_i need to be updated for each cell at each time step. We define the matrix $\overline{U} \in \mathbb{R}^{4\times G}$, where each entry, $\overline{U}_{ij}$, is an increment to the probability of phenotypic action i, $P_i(t)$, under the circumstance that gene j is mutated and its' expression is upregulated. Similarly, we define the matrix $\overline{D} \in \mathbb{R}^{4\times G}$, where each entry, $\overline{D}_{ij}$, is an increment to the probability of phenotype action i, $P_i(t)$, under the circumstance that gene j is mutated and its' expression is downregulated. We define detailed rules for updates to the phenotypic action probabilities in Appendix 1 (Supplementary material). Note that we often choose $\overline{U} = -\overline{D}$ for symmetry. The sum of the phenotype vector equaling one is maintained by balancing the probability of each phenotype action against the probability of quiescence and quiescence equally against all the other phenotypic actions (see Appendix 1 in Supplementary material for more details).

2.2.3. Update rules for cell class

A cell is considered mutated if it has more than or equal to $\Upsilon \in \mathbb{N}$ positively mutated genes. If a transition from a non-mutated cell to a mutated cell occurs, the phenotype vector is updated as described in Appendix 1 (Supplementary material).

Dedifferentiation is the process of a specialized cell reverting back to a non-specialized cell. In our model this is accomplished by a non stem cell becoming a stem cell. Dedifferentiation depends on a complex interaction of positive and negative feedback mechanisms involved in cell proliferation and gene expression. A recent review (Hillen and Shyntar, 2023) (and the references therein) discuss detailed modeling of these aspects. Here, dedifferentiation is used to help maintain the proper ratio of stem cells to non stem cells in the grid by dedifferentiating whenever the number of stem cells in the neighborhood of a non stem cell is less than or equal to some chosen value, $\hat{S}$, or if the number of empty cells in the neighborhood of a non stem cell is less than or equal to some chosen value, $\hat{E}$. To help reduce the number of cells dedifferentiating, the process is completed only when a random sample from the uniform distribution is less than or equal to some threshold, $\hat{D} \in \mathbb{R}_+(0,1)$.

We associate a fitness value to each cell, so that the cells can compete and the population contains only the healthiest, or in the case of mutated cells, the most positively mutated cells. The characteristics that affect the fitness are based upon work by Bowling et al. (2019). High fitness is characterized by a low apoptotic rate, a high proliferation rate, young age. Also cancer cells are considered to have a high fitness.

In addition to the four phenotypic actions, quiescent cells can move with a specified probability into a neighboring cell. A CSC or TC can move into an occupied neighboring cells, killing the occupant in the process. In this case, the probability of moving into an occupied cell is lower than the probability of moving into an empty cell.

CSCs and TCs are the only class of cells that can kill other cells when moving during quiescence. If the parent cell is a CSC or TC and the chosen cell has a higher fitness then the phenotypic action is accomplished only if a sample from some random variable is less than a threshold to kill, $\kappa \in \mathbb{R}_+(0,1)$. A CSC can kill a TC and TC a CSC only if the fitness is lower.

At each time step, each non-empty cell in the CA grid chooses an action to execute and attempts to complete such action. Consider that the cell that is performing the action is located at $\boldsymbol{x}^{(p)} \in \Omega$. In certain cases the action will be performed upon a (randomly chosen) neighboring target cell, located at $\boldsymbol{x}^{(c)} \in \Omega$. As mentioned above, the target cell need not be empty. In the case of proliferation, and differentiation, the action can take place when the occupant of target cell has a lower fitness.

The lineage of each cell is tracked for the purpose of following tumor cell lineages from their origin, checking how many independent tumor masses form throughout the simulation, and whether the origin is monoclonal or polyclonal.

2.2.4. CA model timeline

Each time-step has the following order of actions (1) Run the gene expression neural network, (2) Update the gene expressions based upon the output of the neural network in step 2, (3) Update the gene expressions via the gene instability process, (4) Update the phenotype vector based upon the gene expressions of each gene, (5) Update the states of each cell using the state transition process, (6) Apply the dedifferentiation process, (7) Apply the phenotypic action chosen by the cell for that time-step, and apply possible random walk to a neighboring empty cell, (8) Possibly perform tumor excision.

TABLE 3 Shows which genes are activated by certain genes.

Gene	Gene activation's
TP53	TP21, TP16, RB
RB	TP53, CCDN1
CCDN1	TP21
MYC	TP21 (de-activates), Ras
RAS	CCDN1, MYC

2.2.5. CA parameters for HNSCC

The initial seed is set such that the domain has the following breakdown of each cell type: 64.5% normal tissue cells (NTC; brown), 6.5% normal stem cells (NSC; yellow), and 29% empty cells (white). The maximum number of TAC generations is given by $\Theta = 2$. The chance a cell moves when it is quiescent is 0.25. The chance a tumor cell (TC; red) or cancer stem cell (CSC; purple) randomly kills another cell during movement, proliferation, or differentiation is 0.2. The chance that an SC or MSC becomes a CSC is 2.5×10^{-6}. The chance a non stem cell becomes a stem cell through dedifferentiation is 10^{-4}. If either there are no stem cells or there are at least six empty cells in the neighborhood of a non stem cell, then the process of dedifferentiation will be attempted. When an excision is performed the number of neighbors around a TC removed is two.

We consider 10 genes which are given in Table 1. We set the mutation threshold to $\overline{M} = 0.1$ and the minimum number of positively mutated genes for a cell to be considered mutated to be four (Anandakrishnan et al., 2019). Using the last two columns of Table 1 and assuming each phenotypic action is modified at the same magnitude we obtain the phenotypic action increment matrices (see Section 2.2.2 and Appendix 1 in Supplementary material) given by:

$$\overline{D} = \begin{bmatrix} 10^{-6} & -10^{-6} & -10^{-6} & 0 \\ 0 & 0 & -10^{-6} & 0 \\ 10^{-6} & -10^{-6} & 0 & 10^{-6} \\ 10^{-6} & 0 & 0 & 0 \\ 10^{-6} & 0 & 0 & 0 \\ -10^{-6} & 0 & 0 & 0 \\ 0 & 0 & 10^{-6} & 0 \\ -10^{-6} & 0 & 10^{-6} & -10^{-6} \\ 0 & 0 & 10^{-6} & 0 \\ -10^{-6} & 0 & 10^{-6} & -10^{-6} \end{bmatrix}, \quad \overline{U} = -\overline{D}. \tag{18}$$

Using Table 1, we can create the gene type vector, $\boldsymbol{T}$, that is used in (A6), (A7), and (A28) which is given by:

$$\boldsymbol{T} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T, \tag{19}$$

where 0 means the gene is a tumor-suppressor and 1 means it is an oncogene.

Using Table 3, we can create the gene relationship matrix, R, that is used in (A7) which is given by:

$$R = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \tag{20}$$

where 0 means the genes are not related and 1 means the genes are related. Note that in the above matrix we assumed that TP53 is related to all the genes. The main diagonal is zero so that genes cannot modify themselves during the genetic instability phase of the model. The chance that a gene modifies the gene expression of another or that the body tries to repair the gene over or under expression is 0.45. The maximum amount a gene expression can be changed during the gene instability stage is $\frac{1}{\sqrt{\nu}}$. We let $\overline{a}_1 = \frac{\tilde{c}}{\bar{c}_1}$ and $\overline{a}_2 = \frac{\tilde{c}}{\bar{c}_2}$ be the initial probabilities of apoptosis for a normal tissue cell and normal stem cell. Where $\tilde{c}$ is the length of the cell cycle in hours, $\bar{c}_1$ is the life span of a cell, and $\bar{c}_2$ is the life span of a stem cell. The initial phenotype matrix that is used in Equation (A25) is given by:

$$\tilde{P} = \begin{bmatrix} \overline{p}_1\overline{a}_1 & \overline{a}_1 & 1-\overline{a}_1(\overline{p}_1+1) & 0 \\ \overline{p}_1\overline{a}_1 & \frac{\overline{a}_1}{\overline{\alpha}} & 1-\overline{a}_1(\overline{\alpha}^{-1}+\overline{p}_1) & 0 \\ \overline{p}_2\overline{a}_2 & \overline{a}_2 & 1-\overline{a}_2(\overline{p}_2+1)-\vec{\overline{d}}\tilde{d} & \vec{\overline{d}}\tilde{d} \\ \overline{p}_2\overline{a}_2 & \frac{\overline{a}_2}{\overline{\alpha}} & 1-\overline{a}_2(\overline{\alpha}^{-1}+\overline{p}_2)-\vec{\overline{d}}\tilde{d} & \vec{\overline{d}}\tilde{d} \\ \overline{p}_2\overline{a}_2 & \frac{\overline{a}_2}{5\overline{\alpha}^2} & 1-\overline{a}_2((5\overline{\alpha})^{-2}+\overline{p}_2)-\vec{\overline{d}}\tilde{d} & \vec{\overline{d}}\tilde{d} \\ \overline{p}_1\overline{a}_1 & \frac{\overline{a}_1}{5\overline{\alpha}^2} & 1-\overline{a}_1((5\overline{\alpha})^{-2}+\overline{p}_1) & 0 \end{bmatrix}, \tag{21}$$

where $\overline{p}_1$ is the proliferation factor for normal tissue cell types, $\overline{p}_2$ is the proliferation factor for normal stem cell types, $\overline{\alpha}$ is the apoptotic factor, $\vec{\overline{d}}$ is the differentiation factor, and $\tilde{d}$ is the probability of differentiation occurring neglecting competition between cells. The cell cycle length can range anywhere between 8 and 24 h for the various cells in the body, since we are analyzing the tongue we will use $\tilde{c} = 10$h (Beidler and Smallman, 1965). The lifespan of a taste bud is 250±50 h (Beidler and Smallman, 1965), so $\bar{c}_1 = 250$h. The lifespan of a typical stem cell is around 25,550 h (Sieburg et al., 2011), so $\bar{c}_2 = 25550$h. We set $\overline{\alpha} = 1.625$, $\overline{p}_1 = 0.65$, $\overline{p}_2 = 14.75$, and $\vec{\overline{d}} = 1.485$ so that equilibrium in the tissue is maintained when there are no carcinogens in the domain. Note that $\overline{p}_1$ is less than 1, since we want most of the new cells to come from TACs created by SCs, because, biologically speaking, normal tissue cells rarely proliferate. Since each TAC produces a certain number of generations, given by Θ, then it will produce $2^{\Theta+1} - 2$ new cells so we set

$$\tilde{d} = \frac{1}{2^{\Theta+1}-2}. \tag{22}$$

When a cell is a TAC the probability of proliferation increases by $\frac{1}{3}$, so that it will create its Θ generations in as few time-steps as possible, assuming there is enough available space. The chance that a gene modifies the probability of a phenotypic action is given by

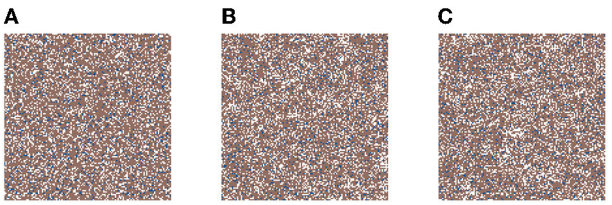


FIGURE 2
This figure includes three time-steps. The grid size is 128 × 128 and no active carcinogens are present. Using the color map for the cell classes as provided in Table 2. These show **(A)** the initial seed of the simulation $t = 0$, in **(B)** the domain (tissue) at the halfway point of the simulation $t = 4,383$ steps = 5 years, and in **(C)** the final time-step at $t = 10$ years.

0.35. The maximum value a gene can modify the phenotypic action by is 10^{-6}.

3. Results

In this section, we present the results of simulations of the CA model as applied to HNSCC. We will explore what impact the following have on the results: grid size, number of carcinogens, carcinogen concentration, and excising the tumor vs. excising the entire field. We will discuss field growth, changes in probabilities of phenotypic actions over time, mutation spread rates, and the number of lineages. By tracking lineages we will also check monoclonal vs. polyclonal origins.

3.1. Equilibrium

We first ensure that the model can maintain a healthy tissue structure, if not exposed to a carcinogen. This will show that in our base model random mutations alone cannot cause cancer on the time scales considered here due to the low mutational rate of genes in the healthy body, and the fact that the body is well adept at repairing mutations as they occur. We run the simulation on a 128×128 grid for 8,766 time-steps (about 120 months), and as stated above, with no carcinogens.

In Figure 2, we present three time-steps from a simulation where no carcinogens were included, with normal tissue cells (NTC) as the brown cells, normal stem cells (NSC) as the blue cells, and empty cells are white. Figure 2 shows (Figure 2A) the initial seed, Figure 2B the domain (tissue) at the halfway point of the simulation (5 years), and Figure 2C the final time-step (10 years). We observe that the tissue stayed in equilibrium. The changes through time are due to cell movement and the natural birth and death processes. The figures show that, as desired, no mutated cells (green, yellow) arise and thus no cancer stem cells (purple) or tumor cells (red) are formed.

3.2. Results with carcinogens

Now we study simulations where carcinogens are present and cause mutations and, ultimately, cancer. Figure 3 illustrates the development of a cancer field and tumors within it, where smoking and alcohol consumption are simulated using carcinogen spatial distribution (1). The various time-steps show (Figure 3A) the initial seed, Figure 3B the cancer field at its early development, Figure 3C the cancer field further developing prior to cancer, in Figures 3D–F the multiple stages of cancer development. The color map for the cell classes is as provided in Table 2. The cancer field (green) is initially minimal and undeveloped, but over time it evolves and matures, eventually forming tumors (red and purple). These tumors grow and outpace the growth rate of the cancer field.

We have uploaded four video files to youtube that show the dynamics of our hybrid CA model. Each video shows from left to right the carcinogen, the CA dynamics, and the 20 largest cell lineages. The grid size is 256 × 256 in all these simulations (see Table 4).

3.2.1. Field development

Regardless of changes to parameters we observe that the field begins to form where the carcinogen is most concentrated; this can be verified with Figure 3B. Initially, the field is made up of only mutated normal tissue cells (green cells) and mutated normal stem cells (yellow cells). We defined the cancer field as the areas of the domain that contain cells from the mutated cell classes thus, the figure shows the cancer field growth over time. Typically, the first mutated cell is a MNTC due to there being a higher number of NTC compared to NSC. The field grows outwards as it takes over normal tissue. Once the first CSC arises, cancer grows quickly. We observe a number of focal regions of cancer development, and if untreated, it will fully overgrow the domain. The different focal points in Figure 3E correspond to different cell lineages, as we show later.

In Figure 4A, we show the time evolution of the fraction of all cell classes. The cancer field begins to from once the green curve starts to grow. Once the field has developed and grown large enough, the odds of a NSC or MNSC becoming a CSC increases. Soon after the emergence of the first CSC, TCs begin to form and grow in a logistic fashion into the domain. As cancer grows the normal tissue (brown) and field tissue (green) is reduced.

3.2.2. Mutational evolution

Recall we use the term "positive mutation" for mutations that promote cancer (i.e., upregulation of an oncogene or downregulation of a tumor suppressor gene). In Figure 5, we show various graphs that represent the mutational evolution of the genes over time. In Figures 5A, B the average gene expression is illustrated for first the tumor suppressors and secondly the oncogenes. In Figure 5C, we show the time evolution of the fraction of genes that are positively mutated. In Figure 5A all the tumor suppressor genes are downregulated, hence positively mutated. Other than RB, which decreases at a faster rate, the gene expressions of all the other tumor suppressor genes decrease at a similar rate. Figure 5B displays that the oncogenes are upregulated, therefore positively mutated. The gene expressions of the oncogenes increase at a similar rate except CCDN1 and RAS, which increase at a faster rate. We see that the gene expressions between the genes can vary significantly, principally with P21 and CCDN1. These two genes mutate because

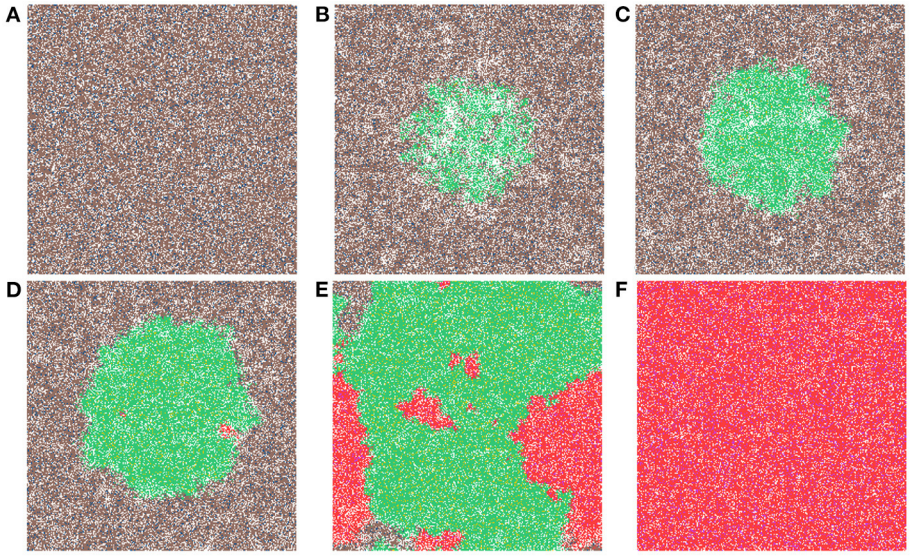


FIGURE 3
This figure includes time-steps illustrating the development of a cancer field and cancer cells using the color map for the cell classes as provided in Table 2. These show **(A)** the initial seed, **(B)** the cancer field at its early development, **(C)** the cancer field further developing but prior to cancer, **(D)** the first stages of cancer development, **(E)** further cancer growth, and **(F)** the final time-step. Parameters are as follows: grid size 256 × 256, carcinogen spatial distribution 2, both carcinogens activated. A video of this simulation can be found on youtube: https://youtu.be/eKxsrSoDiKs.

they are related to the most genes, and therefore have a higher weight in the MLP output weight matrix, W_y, in Equation (13).

Figure 5C shows a lag of time before the first positively mutated genes occur, this is due to the low mutational rate. The initial spike in mutational rate at the onset of the first mutated cell is due to the relative size of the domain vs. the mutated cells. Once multiple genes become positively mutated the progression accelerates, due to changes in the expression of other genes caused by genetic instability, as displayed in the period starting at about 20 to 25 months. We observe that all the genes are positively mutated by about 25 months.

TABLE 4 List of youtube videos for the CA dynamics, where we show from left to right the carcinogen, the CA simulation, and the 20 largest cell lineages.

Youtube video	Showing this effect
https://youtu.be/	
eKxsrSoDiKs	General scenario
Gtf6MoxXCkM	Different carcinogen distribution
zngGzjSlPwU	Excision of cancer cells
EOFI4Ai1A9U	Excision of cancer and field

3.2.3. Phenotypic evolution

Next we consider the evolution of a phenotypic actions of proliferation, quiescence, apoptosis, and differentiation as functions of time. Figure 4C illustrates the phenotypic evolution of these actions as the time evolution of the fraction of cells that underwent each phenotypic action at a given time step. The probability of apoptosis occurring decreases as the cell population moves toward being cancerous. This occurs since the majority of the genes become positively mutated causing apoptosis to decrease. While the probability of apoptosis decreases, the chance of proliferation and differentiation increases, which again is caused by positively mutated genes. Probability of differentiation increases at a slower rate than proliferation because fewer of the genes we consider influence differentiation. Finally, for the most part, the probability of quiescence remains stable—it goes slightly up and down, due to being balanced against the other phenotypic actions, and not many genes are influencing it, but otherwise it is at equilibrium. Figure 4 shows us that apoptosis and proliferation change the most over time, in particular, as apoptosis decreases, we see that proliferation increases, due to less cells dying before they can become more cancerous.

3.3. Grid size comparisons

In Figure 6, we show a sample time-step for each grid size we compare. In Figure 6A the grid size is 64 × 64, in Figure 6B the grid size is 128 × 128, in Figure 6C the grid size is 256 × 256, and in Figure 6D the grid size is 512 × 512. When comparing the grid sizes all the other parameters were the same, both carcinogens were activated and carcinogen spatial distribution (1) was used. We see that as the domain size increases, the tumor masses within it also increase.

The various grid sizes show slight differences in four ways, all predominantly due to the increase in the number of cells. Most of the events in the CA are probabilistic, as a result, almost automatically, as we increase the size of the grid, the chance of a probabilistic event increases as well. Therefore, the overall

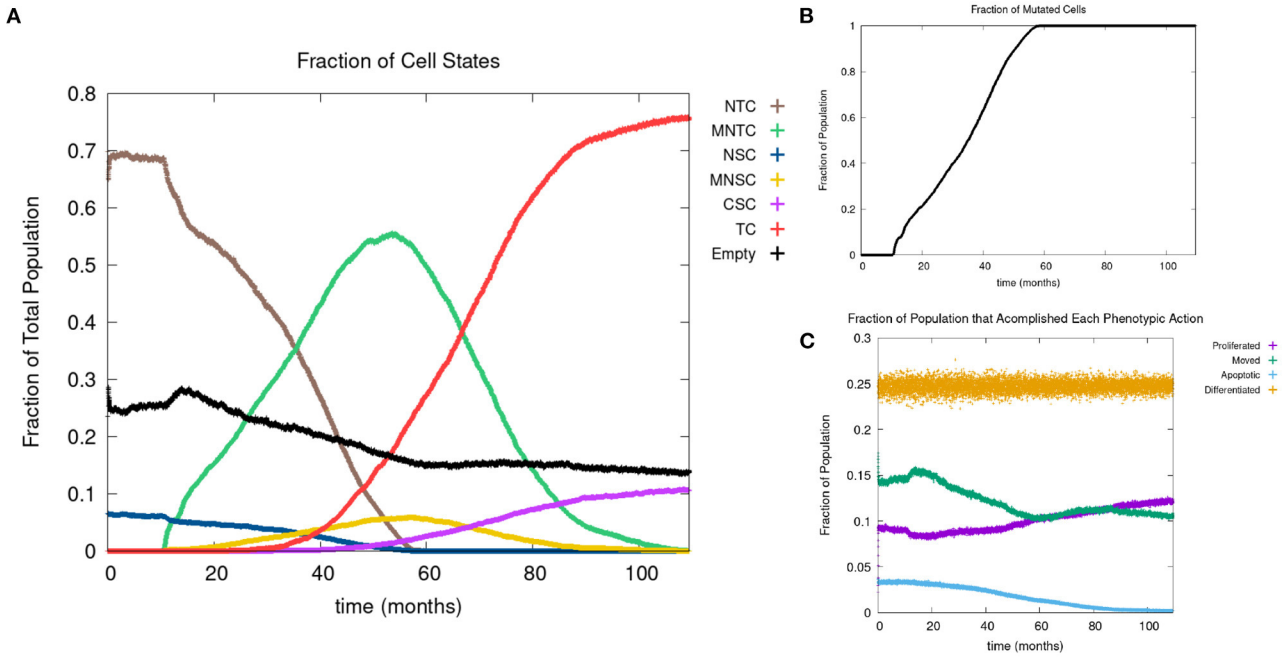


FIGURE 4
(A) Fraction of the various sub-populations as function of time. **(B)** Time course of the fraction of mutated cells, thus illustrating the cancer field growth over time. **(C)** Evolution of phenotypic probabilities of proliferation, apoptosis, quiescence, and differentiation. Parameters are as follows: grid size 256 × 256, carcinogen spatial distribution 2, both carcinogens activated shows the fraction of cancer cells over time. Parameters are as follows: grid size 256 × 256, carcinogen spatial distribution 2, both carcinogens activated.

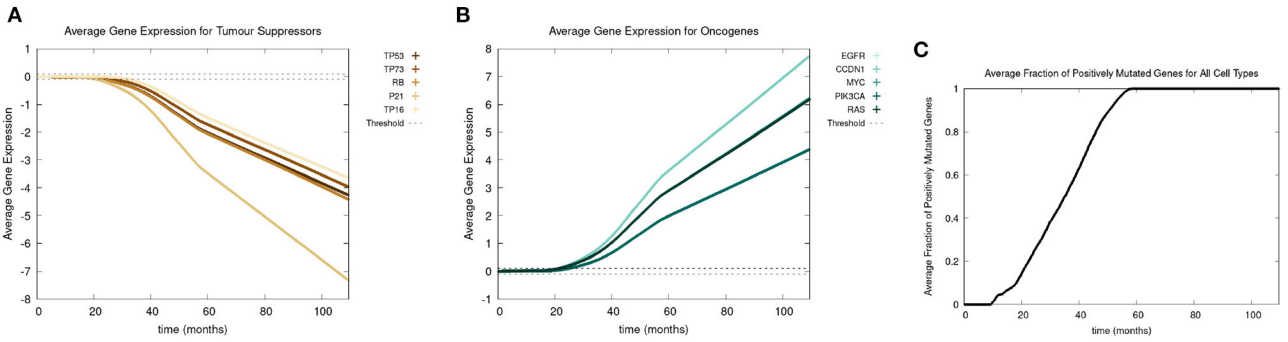


FIGURE 5
In this figure, we show the mutational evolution of the genes. The time course of the average gene expression are shown for in **(A)** the tumor suppressor genes, in **(B)** the oncogenes. In **(C)**, we show the time course of the fraction of genes that are positively mutated. Parameters were chosen as follows: grid size of 256 × 256, carcinogen spatial distribution 2 was used, both carcinogens were activated.

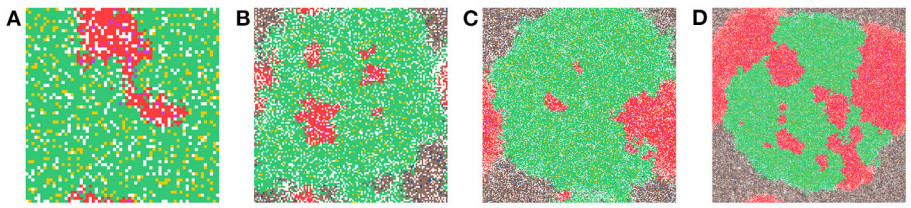


FIGURE 6
This figure shows a time-step from each of the grid sizes that were considered for comparison. In **(A)** the grid size is 64 × 64, in **(B)** the grid size is 128 × 128, in **(C)** the grid size is 256 × 256, and in **(D)** the grid size is 512 × 512. Parameters: both carcinogens were activated and carcinogen spatial distribution 2 was used.

TABLE 5 In this table, we compare the time-step the first mutated cell forms, the time step where the first cancer stem cell (CSC) arises, and the number of cancer cell lineages at the end of the simulations, between the different grid sizes.

Grid size	Time-step first mutated cell forms	Time step first CSC	# tumor cell lineages
64 × 64	817	2,872	1
128 × 128	748	1,402	1
256 × 256	744	1,650	6
512 × 512	735	1,082	24

dynamics are the same for each grid size, but the timing of various main events differ slightly as will be illustrated with the following table.

In Table 5, we show the time-step at which the first mutated cell forms. The times are comparable and the first mutated cell is always an MNTC, due to the fact that a higher ratio of NTC than SC exists in the domain. We also compare the time-step the first CSC forms for each grid size. The variation between the smallest grid size as compared to the remaining three is substantial, due to the fact that the probability of a CSC forming is minuscule, attempting such an action within such a small grid size reduces the chances of these formations drastically in comparison to the larger grid sizes. We also note that the time for a first TC to develop from a CSC is short at about 3–4 time steps in all our simulations.

A notable difference arises if we count the cell lineages inside the cancerous tissue, since a larger domain simply has more space for different cell lineages to thrive. Table 5 shows the number of tumor cell lineages at the end of the simulation for each grid size.

3.4. Various carcinogen schedules

In this section, we consider the carcinogens from alcohol and smoking individually and in combination with one another. We consider several administration schedules for heavy and light smokers and drinkers.

In Figure 7, we present the time evolution of the fraction of cells in the different cell classes for alcohol alone (Figure 7A), smoking alone (Figure 7B), and alcohol and smoking together (Figure 7C). In our simulations, alcohol alone does not cause a cancer field to develop (case A), while smoking alone and smoking with alcohol generates a cancer field (cases B and C). It would, however, be premature to conclude that alcohol has no effect on carcinogenesis. Alcohol causes cancer, but the mechanism might be different and not fully captured by the 10 genes that we consider here. Also, it is reported in Hashibe et al. (2009) and LoConte et al. (2018) that the carcinogenic effect of alcohol is significantly lower than the effect of smoking.

3.5. Tumor excision

In this section, we examine the dynamics of the field development after a tumor excision, including the recurrence time. We consider two types of excisions: (A) killing only the tumor and cancer stem cells but not the cancer field and (B) killing all mutated cell classes, including cancer and cancer field cells. Treatment schedule A corresponds to a targeted treatment such as an immunotherapy or an oncolytic viral therapy, while scenario B corresponds to a more global attack such as through radiation treatment or standard chemotherapy.

In Figure 8, we show the time evolution of the fraction of cells in the different cell classes. In Figure 8A, we consider the case of removing the cancer cells but not the cancer field cells and in Figure 8B, we show the case of removing the cancer and cancer field cells. Videos for the scenarios of keeping the field and removing the field are provided at https://youtu.be/zngGzjSlPwU and https://youtu.be/EOFI4Ai1A9U, respectively. Each video shows from left to right the carcinogen spatial distribution, the CA grid, and a visualization that shows the top 20 cell lineages. The excision occurs in the period of 40–60 months, prior to this period we observe normal cancer field and tumor development. We can follow the various sub populations in Figure 8. As the field develops, the number of normal tissue cells decreases as the number of mutated cells increases, with TC just starting to form and accelerate its growth and a very small uptake in CSCs beginning. At the point of excision there is a spike in the number of empty cells (black), which is more prominent in case B, and the number of tumor cells is set to be zero (red and purple). In the case B, the number of mutated cells (yellow) is reduced to zero as well. After an extended lag the field restarts its growth at about the same rate as originally. Whereas, when the field is kept intact (as in case A), the cancer comes back very quickly.

3.6. Cell lineages

The last feature that we include is an identification of cell lineages. In Figure 9, we show time-steps that are relevant for the major developmental stages of field cancerization. We display the top 20 cell lineages at each time step. The time steps include the initial seed (Figure 9A), the early development of the cancer field (Figures 9B, C), the first emergence of cancer cells (Figure 9D), full cancer development (Figure 9E) where all 20 leading lineages are cancer lineages, and finally a polyclonal tumor (Figure 9F) comprised of the fittest cancer lineages. The videos that are listed in Table 4 show the cell lineage developments over time. Figure 9 shows that the largest cell lineages in the domain are contained in the cancer field, which is expected because it contains the most fit cells in the domain.

In all experiments on the 256 × 256 grid where we followed the lineages we found polyclonal cancer fields. Monoclonal cancers were only observed at small grid sizes such as 64 × 64 and 128 × 128.

4. Conclusion

We developed a sophisticated cellular automata model for the cancer field effect. The model is an extension to existing cellular automata models (Gerlee and Anderson, 2007; Foo et al., 2014) as we include the effect of two carcinogens, one related to smoking and one related to alcohol. The impact on gene expression of oncogenes

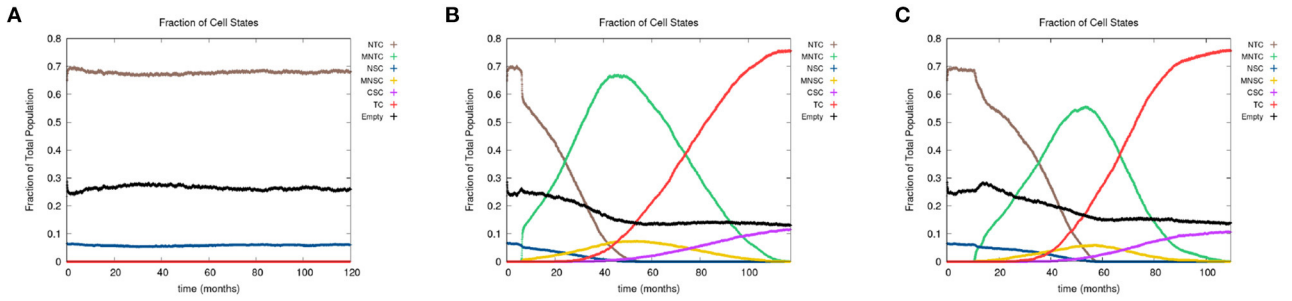


FIGURE 7
We show the time course of the fraction of cells in the different cell classes NTC, MNTC, NSC, MNSC, CSC, TC, and empty for **(A)** alcohol alone, **(B)** smoking alone, and **(C)** alcohol and smoking together. Grid size 256 × 256.

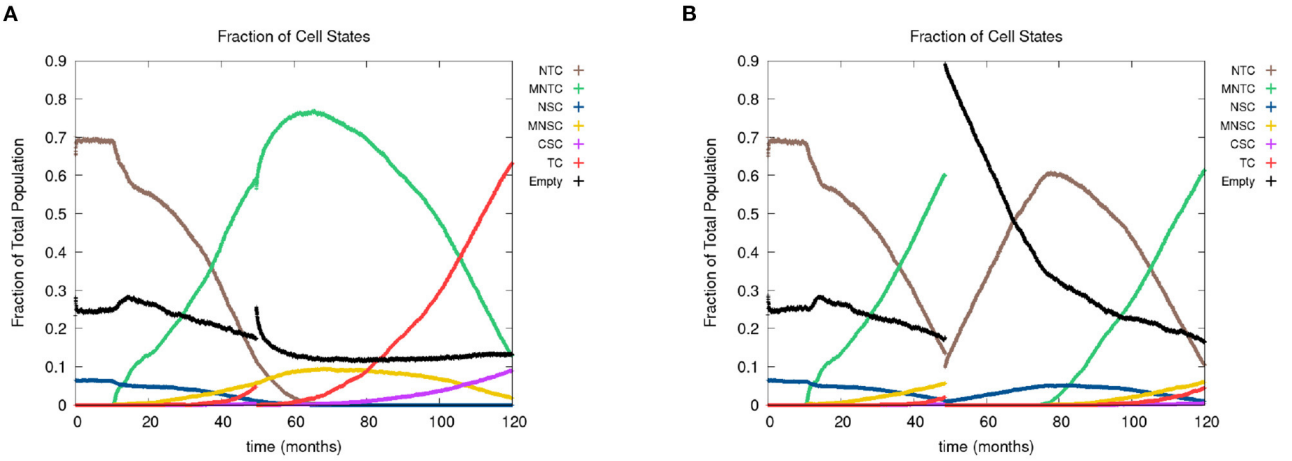


FIGURE 8
In figures **(A, B)**, we show the time course of the fraction of cells in the different cell classes NTC, MNTC, NSC, MNSC, CSC, TC, empty. In the plots, we consider the case where we **(A)** remove only TCs (keeping the field) and **(B)** we remove all mutated cells (removing the field). Parameters are as follows: grid size 256 × 256, both carcinogens activated, carcinogen spatial distribution 2, and time elapse of excision following first TC appearance was 18 months.

and tumor suppressor genes are of significant importance to the given cancer. Based on the existing literature we also find that smoking is a potent carcinogen, while the effect of alcohol is minimal. The gene expression was modeled by a multi-layer neural network, which can be trained once more data is available. We admit that our modeling of gene mutations and their impact on phenotype is simplified. To train the neural network for the gene expression dynamics, we would need single cell genomic data. These data need to explain the mutational changes in gene expression based on carcinogen exposure, plus an understanding of how mutated genes change the phenotypic actions of proliferation, apoptosis, quiescence and differentiation. The former could be obtained from large scale cell profiling. The second question, how do genes impact phenotype, is still an unresolved holy grail of genomics.

We demonstrated that when an excision is performed that removes only the tumor cells but leaves the remaining surrounding tissue intact, the cancer recurs faster than when removing the entire field of mutated tissue. When the field is not removed during excision, the cancer that recurs is more aggressive than before the field was removed. We observed, by tracking cell lineages, that the tumor masses mostly form from polyclonal origins.

There are a number of possible extensions to our model that might yield additional insights. A dynamic mutation threshold could be considered that depends on a number of factors such as the number of mutated genes or cell age. This mutation threshold could also be specific for each gene. For example, in our model we assumed TP53 is related to all other genes, and, as a result, once it is mutated, all other genes become mutated as well. However, it might be useful to consider a specific order of gene mutations that lead to cancer.

Telomeres are at the end of the DNA strands and with each cell division they get cut shorter, eventually becoming so short that the cell can no longer proliferate and so will enter senescence. Senescent cells are similar to quiescent cells except they can not perform any actions and eventually they undergo apoptosis. Therefore, the model could be enhanced by introducing telomeres.

It would be interesting to include viral infections to the model such as human papillomavirus (HPV) as input to the gene expression neural network (Lee and Tameru, 2012).

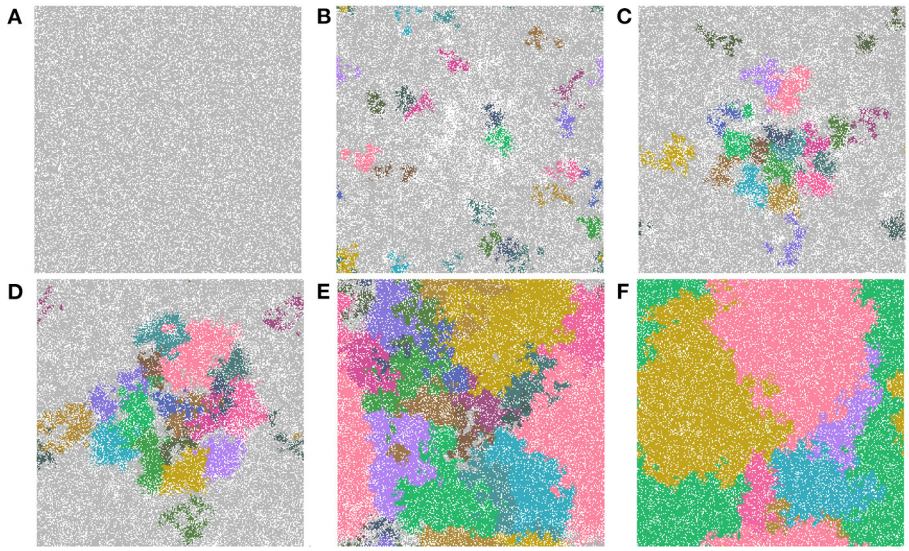


FIGURE 9
Important time-steps that show the top 20 lineages throughout the development stages of field cancerization. In the figures we show **(A)** the initial seed, **(B)** early cancer field formation, **(C)** later cancer field development, **(D–F)** cancer development. Note that light gray means the cell is not in any of the top 20 lineages and each color represents a different cell lineage. Parameters are as follows: grid size 256 × 256 and both carcinogens activated. See also video https://youtu.be/eKxsrSoDiKs.

One of the questions we originally wanted to answer was how long it would take for a tumor to become large enough to be detected by physicians, however, we were not able to answer this question due to the size of the cells requiring at least a domain size 1024 × 1024 to represent the required $1cm$ detection size. A few simulations at 1024 × 1024 were run and we found it would take more than 10 years to fill in the space, thus it would take at least 10 years for the tumor to be detectable.

With regards to efficiency of running the model, as the complexity increases, the speed of the calculations involved in the gene expression neural network could be improved with linear algebra libraries available in CUDA. Using texture memory in the GPU to store cell neighborhoods would make calculations both faster and easier, as it has faster bandwidth and built in boundary conditions. The code could be made more cross compatible by allowing parallel computation on the CPU and switching from CUDA to OpenCL.

Finally, we note that there are many aspects of cancer biology that are not included here. Chemical signaling and feedback mechanisms among cancer cell sub populations are an important aspect of cancer growth, as well as the interactions with the immune response, mechanical aspects of the tissue, and angiogenesis. The seminal "hallmarks" papers of Hanahan and Weinberg (2000), Hanahan and Weinberg (2011), and Hanahan (2022) give a rather complete picture of effects that are important to cancer growth. As our model is already quite complicated, we did not include all of these effects here. But they are interesting extensions for future versions of the model.

We are grateful for the comments of two anonymous referees. One of them made the following interesting observation. People with a hereditary carcinogenic birth defect are predestined to form cancers, even early in their lives. For example, Rb mutations can lead to Retinoblastoma and the BRCA gene mutations can increase breast cancer risks. In such a case, would the entire body be considered a cancer field? This is an interesting question of further thought, which certainly exceeds the abilities of our computational model.

Data availability statement

Publicly available datasets were analyzed in this study. The datasets can be accessed via the reference list.

Author contributions

TH and JN led the study. KD developed the hybrid cellular automaton. TH, JN, and KD wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Natural Sciences and Engineering Research Council (NSERC) Grant Numbers RGPIN-2017-04158 (TH) and RGPIN-2019-06435 (JN).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1060879/full#supplementary-material

References

Alonso, S., Dai, Y., Yamashita, K., Horiuchi, S., Dai, T., Matsunaga, A., et al. (2015). Methylation of MGMT and ADAMTS14 in normal colon MUCOSA: biomarkers of a field defect for cancerization preferentially targeting elder African-Americans. *Oncotarget* 6, 3420–3431. doi: 10.18632/oncotarget.2852

Anandakrishnan, R., Varghese, R. T., Kinney, N. A., and Garner, H. R. (2019). Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations. *PLoS Comput. Biol.* 15, e1006881. doi: 10.1371/journal.pcbi.1006881

Angadi, P. V., Savitha, J. K., Rao, S. S., and Sivaranjini, Y. (2012). Oral field cancerization: current evidence and future perspectives. *Oral Maxillofacial Surg.* 16, 171–180. doi: 10.1007/s10006-012-0317-x

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2014). Cancer evolution: mathematical models and computational inference. *Syst. Biol.* 64, e1–e25. doi: 10.1093/sysbio/syu081

Beidler, L. M., and Smallman, R. L. (1965). Renewal of cells within taste buds. *J. Cell Biol.* 27, 263–272. doi: 10.1083/jcb.27.2.263

Biddle, A., Liang, X., Gammon, L., Fazil, B., Harper, L. J., Emich, H., et al. (2011). Cancer stem cells in squamous cell carcinoma switch between two distinct phenotypes that are preferentially migratory or proliferative. *Cancer Res.* 71, 5317–5326. doi: 10.1158/0008-5472.CAN-11-1059

Bjerkvig, R., Tysnes, B. B., Aboody, K. S., Najbauer, J., and Terzis, A. J. A. (2005). The origin of the cancer stem cell: current controversies and new insights. *Nat. Rev. Cancer* 5, 899–904. doi: 10.1038/nrc1740

Bowling, S., Lawlor, K., and Rodriguez, T. A. (2019). Cell competition: the winners and losers of fitness selection. *Development* 146, 1–12. doi: 10.1242/dev.167486

Braakhuis, B. J. M., Tabor, M. P., Kummer, J. A., Leemans, C. R., and Brakenhoff, R. H. (2003). A genetic explanation of slaughter's concept of field cancerization. *Cancer Res.* 63, 1727–1730.

Brennan, J. A., Mao, L., Hruban, R. H., Boyle, J. O., Eby, Y. J., Koch, W. M., et al. (1995). Molecular assessment of histopathological staging in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* 332, 429–435. doi: 10.1056/NEJM199502163320704

Bu, Y., and Cao, D. (2012). The origin of cancer stem cells. *Front. Biosci.* 4, 819–830. doi: 10.2741/s302

Calabrese, P., Tavare, S., and Shibata, D. (2004). Pretumor progression: clonal evolution of human stem cell populations. *Am. J. Pathol.* 164, 1337–1346. doi: 10.1016/S0002-9440(10)63220-8

Califano, J., van der Riet, P., Westra, W., Nawroz, H., Clayman, G., Piantadosi, S., et al. (1996). Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer Res.* 56, 2488–2492.

Cense, H. A., van Lanschot, J. J. B., Fockens, P., Obertop, H., and Offerhaus, G. J. A. (1997). A patient with seven primary tumors of the upper aerodigestive tract: the process of field cancerization versus distant monoclonal expansion. *Dis. Esophagus* 10, 139–142.

Chu, T., Shen, C., Lee, H., and Liu, H. (1999). Monoclonality and surface lesion-specific microsatellite alterations in premalignant and malignant neoplasia of uterine cervix: a local field effect of genomic instability and clonal evolution. *Genes Chromosomes Cancer* 24, 127–134.

Curtius, K., Wright, N. A., and Graham, T. A. (2017). An evolutionary perspective on field cancerization. *Nat. Rev. Cancer* 18, 19–32. doi: 10.1038/nrc.2017.102

de Pillis, L., Renee Fister, K., Gu, W., Collins, C., Daub, M., Gross, D., et al. (2009). Mathematical model creation for cancer chemo-immunotherapy. *Comput. Math. Methods Med.* 10, 165–184.

Di Fiore, R., Marcatti, M., Drago-Ferrante, R., D'Anneo, A., Giuliano, M., Carlisi, D., et al. (2014). Mutant p53 gain of function can be at the root of dedifferentiation of human osteosarcoma MG63 cells into 3AB-OS cancer stem cells. *Bone* 60, 198–212. doi: 10.1016/j.bone.2013.12.021

Durrett, R., Foo, J., and Leder, K. (2016). Spatial Moran models, II: cancer initiation in spatially structured tissue. *J. Math. Biol.* 72, 1369–1400. doi: 10.1007/s00285-015-0912-1

Enderling, H., and Chaplain, M. (2014). Mathematical modeling of tumor growth and treatment. *Curr. Pharm. Design* 20, 4934–4940. doi: 10.2174/1381612819666131125150434

Feller, L. L., Khammissa, R. R., Kramer, B. B., and Lemmer, J. J. (2013). Oral squamous cell carcinoma in relation to field precancerisation: pathobiology. *Cancer Cell Int.* 13, 1–8. doi: 10.1186/1475-2867-13-31

Foo, J., Gunnarsson, E., Leder, K., and Storey, K. (2022). Spread of premalignant mutant clones and cancer initiation in multilayered tissue. *Ann. Appl. Prob.* 33, 299–343. doi: 10.1214/22-AAP1816

Foo, J., Leder, K., and Schweinsberg, J. (2020). Mutation timing in a spatial model of evolution. stochastic processes and their applications. 130, 6388–6413. doi: 10.1016/j.spa.2020.05.015

Foo, J., Leder, K. Z., and Ryser, M. D. (2014). Multifocality and recurrence risk: a quantitative model of field cancerization. *J. Theor. Biol.* 355, 170–184. doi: 10.1016/j.jtbi.2014.02.042

Foschini, M. P., Morandi, L., Leonardi, E., Flamminio, F., Ishikawa, Y., Masetti, R., et al. (2013). Genetic clonal mapping of *in situ* and invasive ductal carcinoma indicates the field cancerization phenomenon in the breast. *Human Pathol.* 44, 1310–1319. doi: 10.1016/j.humpath.2012.09.022

Frank, S. (2007). *The Dynamics of Cancer*. Princeton, NY: Princeton University Press.

Franklin, W. A., Gazdar, A. F., Haney, J., Wistuba, I. I., La Rosa, F. G., Kennedy, T., et al. (1997). Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J. Clin. Investig.* 100, 2133–2137. doi: 10.1172/JCI119748

Franssen, L. C., Lorenzi, T., Burgess, A. E., and Chaplain, M. A. (2019). A mathematical framework for modelling the metastatic spread of cancer. *Bull. Math. Biol.* 81, 1965–2010. doi: 10.1007/s11538-019-00597-x

Galandiuk, S., Rodriguez-Justo, M., Jeffery, R., Nicholson, A. M., Cheng, Y., Oukrif, D., et al. (2012). Field cancerization in the intestinal epithelium of patients with Crohn's ileocolitis. *Gastroenterology* 142, 855–864. doi: 10.1053/j.gastro.2011.12.004

Gatenby, R. A., and Gillies, R. J. (2008). A microenvironmental model of carcinogenesis. *Nat. Rev. Cancer* 8, 56–61. doi: 10.1038/nrc2255

Gentry, S. N., and Jackson, T. L. (2013). A mathematical model of cancer stem cell driven tumor initiation: implications of niche size and loss of homeostatic regulatory mechanisms. *PLoS ONE* 8, e71128. doi: 10.1371/journal.pone.0071128

Gerlee, P., and Anderson, A. R. (2007). An evolutionary hybrid cellular automaton model of solid tumour growth. *J. Theor. Biol.* 246, 583–603. doi: 10.1016/j.jtbi.2007.01.027

Gomperts, B. N., Walser, T. C., Spira, A., and Dubinett, S. M. (2013). Enriching the molecular definition of the airway "field of cancerization": establishing new paradigms for the patient at risk for lung cancer. *Cancer Prev. Res.* 6, 4–7. doi: 10.1158/1940-6207.CAPR-12-0470

González-Moles, M. A., Scully, C., Ruiz-Ávila, I., and Plaza-Campillo, J. J. (2013). The cancer stem cell hypothesis applied to oral carcinoma. *Oral Oncol.* 49, 738–746. doi: 10.1016/j.oraloncology.2013.04.002

Gray, L. (2003). A mathematician looks at Wolfram's new kind of science. *Notices Am. Math. Soc.* 50, 200–211.

Greaves, M. F., Maia, A. T., Wiemels, J. L., and Ford, A. M. (2003). Leukemia in twins: lessons in natural history. *Blood* 102, 2321–2333. doi: 10.1182/blood-2002-12-3817

Hafner, C., Knuechel, R., Stoehr, R., and Hartmann, A. (2002). Clonality of multifocal urothelial carcinomas: 10 years of molecular genetic studies. *Int. J. Cancer* 101, 1–6. doi: 10.1002/ijc.10544

Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer Discov.* 12, 31–46. doi: 10.1158/2159-8290.CD-21-1059

Hanahan, D., and Weinberg, R. (2000). The hallmarks of cancer. *Cell* 100, 57–70.

Hanahan, D., and Weinberg, R. (2011). Hallmarks of cancer: The next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hashibe, M., Brennan, P., Chuang, S. C., Boccia, S., Castellsague, X., Chen, C., et al. (2009). Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. *Cancer Epidemiol. Biomark. Prev.* 18, 541–550. doi: 10.1158/1055-9965.EPI-08-0347

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. 2. New York, NY: Springer.

Herreros-Villanueva, M., Zhang, J. S., Koenig, A., Abel, E. V., Smyrk, T. C., Bamlet, W. R., et al. (2013). SOX2 promotes dedifferentiation and imparts stem cell-like features to pancreatic cancer cells. *Oncogenesis* 2, 1–12. doi: 10.1038/oncsis.2013.23

Hillen, T., and Shyntar, A. (2023). "Modelling of cancer stem cell driven solid tumors," in *Problems in Mathematical Biophysics - a volume in Memory of Alberto Gandolfi*, eds A. Fasano, A. d'Onofrio, F. Papa, and C. Sinisgalli (Heidelberg: Springer), 1–21.

Hittelman, W. N. (2001). Genetic instability in epithelial tissues at risk for cancer. *Ann. N. Y. Acad. Sci.* 952, 1–12. doi: 10.1111/j.1749-6632.2001.tb02723.x

Hu, B., Castillo, E., Harewood, L., Ostano, P., Reymond, A., Dummer, R., et al. (2012). Multifocal epithelial tumors and field cancerization from loss of mesenchymal CSL signaling. *Cell* 149, 1207–1220. doi: 10.1016/j.cell.2012.03.048

Kadara, H., Fujimoto, J., Yoo, S.-Y., Maki, Y., Gower, A. C., Kabbout, M., et al. (2014). Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer. *J. Natl. Cancer Inst.* 106, 1–9. doi: 10.1093/jnci/dju004

Kadara, H., and Wistuba, I. I. (2012). Field cancerization in non-small cell lung cancer: implications in disease pathogenesis. *Proc. Am. Thor. Soc.* 9, 38–42. doi: 10.1513/pats.201201-004MS

Kang, G. H., Kim, C. J., Kim, W. H., Kang, Y. K., Kim, H. O., and Kim, Y. I. (1997). Genetic evidence for the multicentric origin of synchronous multiple gastric carcinoma. *Lab. Investig.* 76, 407–417.

Kanjilal, S., Strom, S. S., Clayman, G. L., Weber, R. S., El-Naggar, A. K., Kapur, V., et al. (1995). p53 mutations in nonmelanoma skin cancer of the head and neck: molecular evidence for field cancerization. *Cancer Res.* 55, 3604–3609.

Knopf, A., Lempart, J., Bas, M., Slotta-Huspenina, J., Mansour, N., and Fritsche, M. (2015). Oncogenes and tumor suppressor genes in squamous cell carcinoma of the tongue in young patients. *Oncotarget* 6, 3443–51.

Kumar, S. M., Liu, S., Lu, H., Zhang, H., Zhang, P. J., Gimotty, P. A., et al. (2012). Acquired cancer stem cell phenotypes through Oct4-mediated dedifferentiation. *Oncogene* 31, 4898–4911. doi: 10.1038/onc.2011.656

Lee, S. L., and Tameru, A. M. (2012). A mathematical model of human papillomavirus (HPV) in the united states and its impact on cervical cancer. *J. Cancer* 3, 262–268. doi: 10.7150/jca.4161

Lee, Y.-C., Wang, H.-P., Wang, C.-P., Ko, J.-Y., Lee, J.-M., Chiu, H.-M., et al. (2011). Revisit of field cancerization in squamous cell carcinoma of upper aerodigestive tract: better risk assessment with epigenetic markers. *Cancer Prev. Res.* 4, 1982–1992. doi: 10.1158/1940-6207.CAPR-11-0096

LoConte, N. K., Brewster, A. M., Kaur, J. S., Merrill, J. K., and Alberg, A. J. (2018). Alcohol and cancer: a statement of the American Society of Clinical Oncology. *J. Clin. Oncol.* 36, 83–93. doi: 10.1200/JCO.2017.76.1155

Moon, J.-H., Kwon, S., Jun, E. K., Kim, A., Whang, K. Y., Kim, H., et al. (2011). Nanog-induced dedifferentiation of p53-deficient mouse astrocytes into brain cancer stem-like cells. *Biochem. Biophys. Res. Commun.* 412, 175–181. doi: 10.1016/j.bbrc.2011.07.070

Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304. doi: 10.1093/genetics/156.1.297

Nonn, L., Ananthanarayanan, V., and Gann, P. H. (2009). Evidence for field cancerization of the prostate. *Prostate* 69, 1470–1479. doi: 10.1002/pros.20983

Oka, D., Yamashita, S., Tomioka, T., Nakanishi, Y., Kato, H., Kaminishi, M., et al. (2009). The presence of aberrant DNA methylation in noncancerous esophageal mucosae in association with smoking history: a target for risk diagnosis and prevention of esophageal cancers. *Cancer Interdiscipl. Int. J. Am. Cancer Soc.* 115, 3412–3426. doi: 10.1002/cncr.24394

Paterson, C., Clevers, H., and Bozic, I. (2020). Mathematical model of colorectal cancer initiation. *Proc. Natl. Acad. Sci. U.S.A.* 117, 20681–20688. doi: 10.1073/pnas.2003771117

PubChem (2023a). Ethanol (compound). Available online at: https://pubchem.ncbi.nlm.nih.gov/compound/Ethanol#section=Protein-Bound-3D-Structures

PubChem (2023b). Nicotine. Available online at: https://pubchem.ncbi.nlm.nih.gov/compound/89594#section=Chemical-Target-Interactions&fullscreen=true

Radunskaya, A., Kim, R., and Woods, T. II (2018). Mathematical modeling of tumor immune interactions: a closer look at the role of a PD-L1 inhibitor in cancer immunotherapy. *Spora J. Biomath.* 4, 25–41. doi: 10.30707/SPORA4.1Radunskaya

Rivenbark, A. G., and Coleman, W. B. (2012). Field cancerization in mammary carcinogenesis-implications for prevention and treatment of breast cancer. *Exp. Mol. Pathol.* 93, 391–398. doi: 10.1016/j.yexmp.2012.10.018

Ryser, M. D., Lee, W. T., Ready, N. E., Leder, K. Z., and Foo, J. (2016). Quantifying the dynamics of field cancerization in tobacco-related head and neck cancer: a multiscale modeling approach. *Cancer Res.* 76, 7078–7088. doi: 10.1158/0008-5472.CAN-16-1054

Sieburg, H. B., Rezner, B. D., and Muller-Sieburg, C. E. (2011). Predicting clonal self-renewal and extinction of hematopoietic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4370–4375. doi: 10.1073/pnas.1011414108

Simple, M., Suresh, A., Das, D., and Kuriakose, M. A. (2015). Cancer stem cells and field cancerization of oral squamous cell carcinoma. *Oral Oncol.* 51, 643–651. doi: 10.1016/j.oraloncology.2015.04.006

Slaughter, D. P., Southwick, H. W., and Smejkal, W. (1953). "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer* 6, 963–968.

Steiling, K., Ryan, J., Brody, J. S., and Spira, A. (2008). The field of tissue injury in the lung and airway. *Cancer Prev. Res.* 1, 396–403. doi: 10.1158/1940-6207.CAPR-08-0174

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724. doi: 10.1038/nature07943

Szeimies, R. M., Torezan, L., Niwa, A., Valente, N., Unger, P., Kohl, E., et al. (2012). Clinical, histopathological and immunohistochemical assessment of human skin field cancerization before and after photodynamic therapy. *Brit. J. Dermatol.* 167, 150–159. doi: 10.1111/j.1365-2133.2012.10887.x

Tabor, M. P., Brakenhoff, R. H., van Houten, V. M. M., Kummer, J. A., Snel, M. H. J., Snijders, P. J. F., et al. (2001). Persistence of genetically altered fields in head and neck cancer patients: biological and clinical implications. *Clin. Cancer Res.* 7, 1523–1532.

Takeshima, H., Niwa, T., Takahash, T., Wakabayashi, M., Yamashita, S., Ando, T., et al. (2015). Frequent involvement of chromatin remodeler alterations in gastric field cancerization. *Cancer Lett.* 357, 328–338. doi: 10.1016/j.canlet.2014.11.038

Trujillo, K. A., Hines, W. C., Vargas, K. M., Jones, A. C., Joste, N. E., Bisoffi, M., et al. (2011). Breast field cancerization: isolation and comparison of telomerase-expressing cells in tumor and tumor adjacent, histologically normal breast tissue. *Mol. Cancer Res.* 9, 1209–1221. doi: 10.1158/1541-7786.MCR-10-0424

Trujillo, K. A., Jones, A. C., Griffith, J. K., and Bisoffi, M. (2012). Markers of field cancerization: proposed clinical applications in prostate biopsies. *Prostate Cancer* 2012, 1–12.

van Houten, V. M. M., Tabor, M. P., van den Brekel, M. W. M., Kummer, J. A., Denkers, F., Dijkstra, J., et al. (2002). Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. *J. Pathol.* 198, 476–486. doi: 10.1002/path.1242

van Oijen, M. G. C. T., and Slootweg, P. J. (2000). Oral field cancerization: carcinogen-induced independent events or micrometastatic deposits? *Cancer Epidemiol. Prev. Biomark.* 9, 249–256.

Weaver, J. M. J., Ross-Innes, C. S., Shannon, N., Lynch, A. G., Forshew, T., Barbera, M., et al. (2014). Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* 46, 837–843. doi: 10.1038/ng.3013

Zaky, A. H., Watari, J., Tanabe, H., Sato, R., Moriichi, K., Tanaka, A., et al. (2008). Clinicopathologic implications of genetic instability in intestinal-type gastric cancer and intestinal metaplasia as a precancerous lesion: proof of field cancerization in the stomach. *Am. J. Clin. Pathol.* 129, 613–621. doi: 10.1309/DFLELPGPNV5LK6B1

Frontiers in Immunology

Explores novel approaches and diagnoses to treat immune disorders.

The official journal of the International Union of Immunological Societies (IUIS) and the most cited in its field, leading the way for research across basic, translational and clinical immunology.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact