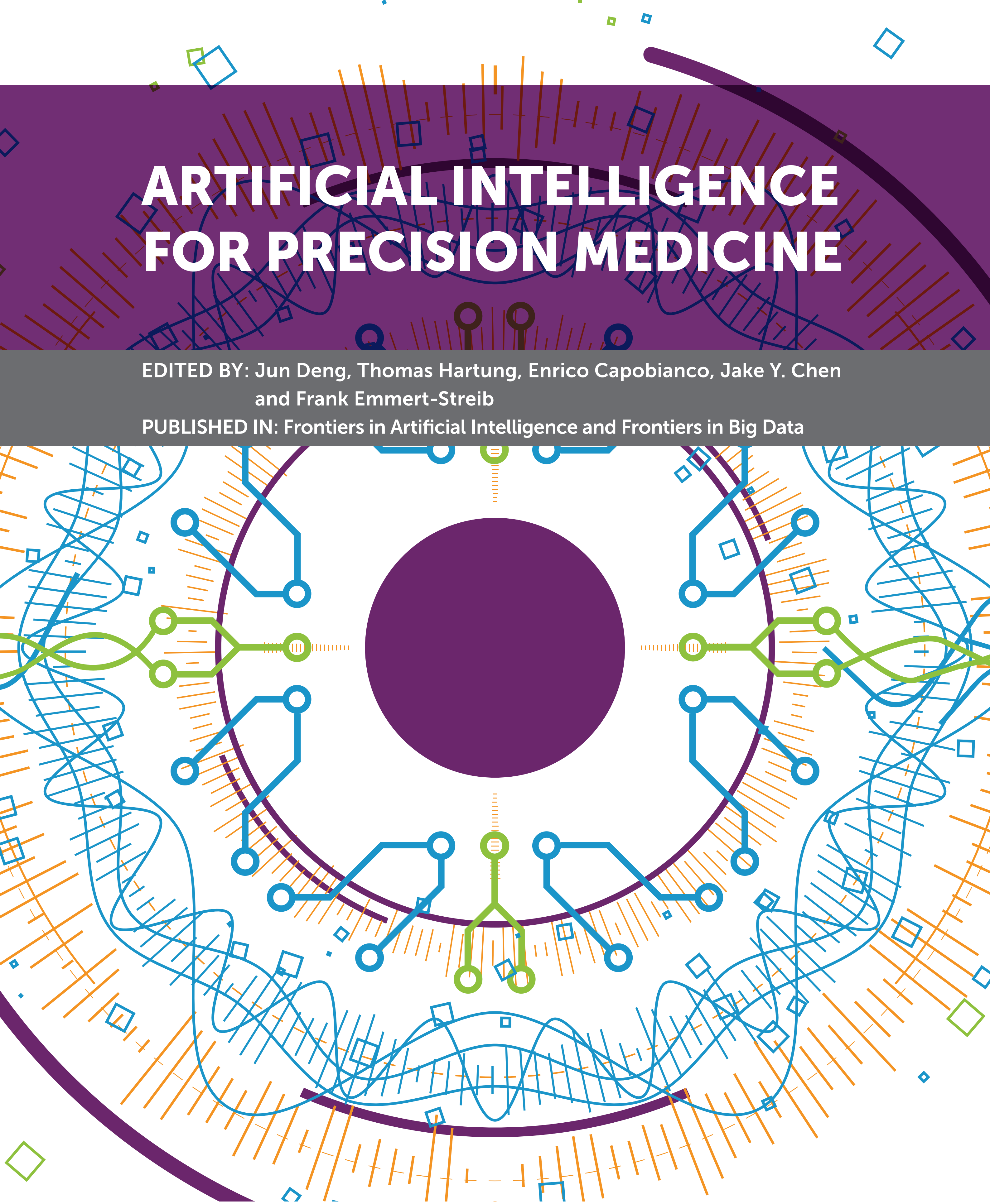


ARTIFICIAL INTELLIGENCE FOR PRECISION MEDICINE

EDITED BY: Jun Deng, Thomas Hartung, Enrico Capobianco, Jake Y. Chen and Frank Emmert-Streib

PUBLISHED IN: Frontiers in Artificial Intelligence and Frontiers in Big Data

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-488-6
DOI 10.3389/978-2-88974-488-6

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

ARTIFICIAL INTELLIGENCE FOR PRECISION MEDICINE

Topic Editors:
Jun Deng, Yale University, United States
Thomas Hartung, Johns Hopkins University, United States
Enrico Capobianco, University of Miami, United States
Jake Y. Chen, University of Alabama at Birmingham, United States
Frank Emmert-Streib, Tampere University, Finland

Citation: Deng, J., Hartung, T., Capobianco, E., Chen, J. Y., Emmert-Streib, F., eds. (2022). Artificial Intelligence for Precision Medicine. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-488-6

Table of Contents

frontiers in Artificial Intelligence

EDITORIAL
published: 21 January 2022
doi: 10.3389/frai.2021.834645

Editorial: Artificial Intelligence for Precision Medicine

Jun Deng[1]*, *Thomas Hartung*[2], *Enrico Capobianco*[3,4], *Jake Y. Chen*[5] *and Frank Emmert-Streib*[6]

[1] Department of Therapeutic Radiology, Yale University, New Haven, CT, United States, [2] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MA, United States, [3] Institute of Data Science and Computing, University of Miami, Coral Gables, FL, United States, [4] National Research Council of Italy (CNR), Institute of Organic Synthesis and Photoreactivity, Bologna, Italy, [5] University of Alabama at Birmingham, Birmingham, AL, United States, [6] Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

Keywords: precision medicine, individualized healthcare, artificial intelligence, big data, machine learning

Editorial on the Research Topic

Editorial: Artificial Intelligence for Precision Medicine

OPEN ACCESS

Edited and reviewed by:
Holger Fröhlich,
Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany

****Correspondence:***
Jun Deng
jun.deng@yale.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *13 December 2021*
Accepted: *29 December 2021*
Published: *21 January 2022*

Citation:
Deng J, Hartung T, Capobianco E, Chen JY and Emmert-Streib F (2022) Editorial: Artificial Intelligence for Precision Medicine. Front. Artif. Intell. 4:834645. doi: 10.3389/frai.2021.834645

SCOPE AND AIM OF THIS RESEARCH TOPIC

Fueled by advances in computing power, algorithms, and big data, the last decade has witnessed widespread applications of artificial intelligence (AI) in every major field, including medicine and healthcare. Generally speaking, AI is expected to help realize the promise of precision medicine in three major areas: (1) disease prevention, (2) personalized diagnosis, and (3) personalized treatment. In this Research Topic, "Artificial Intelligence for Precision Medicine," we aim to set up an open stage in the community where breakthrough application examples of AI for precision medicine are presented. We envisage that AI technologies, if applied openly, fairly, robustly, and in close collaboration with human intelligence, will open new doors for effective and personalized healthcare worldwide.

TOPICS COVERED IN THIS RESEARCH TOPIC

- AI-aided diagnosis and early detection of diseases: Hart et al.
- AI-enhanced treatment and delivery: Chen et al.; Jensen et al.; Mistro et al.; Wang et al.
- Clinical decision support with AI techniques: Barua et al.
- Enhancing patient care via AI applications: Luo
- Radiomics and quantitative imaging: Zhang et al.
- Bioinformatics for more effective healthcare: Kapelner et al.; Namdar et al.
- Innovative AI applications for patient safety: Chan et al.

PAPERS INCLUDED IN THIS RESEARCH TOPIC

In their work, Hart et al. developed seven machine learning algorithms based solely on personal health data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), and compared them with 15 practicing physicians in stratifying endometrial cancer risk for 100 women. The results indicate that their random forest model achieves a testing AUC of 0.96, 2.5 times better at identifying above-average risk women with a 2-fold reduction in the false-positive rate. A novel concept named "Statistical Biopsy" was proposed for the first time.

Chen et al. reported their development of a deep-learning convolutional neural network (DCNN) for enhanced organ-at-risk (OAR) segmentation on cone beam computed tomography

(CBCT), trained with forty post-operative head and neck cancer patients. The developed DCNN improved CBCT in terms of Hounsfield unit (HU) accuracy, image contrast, and OAR delineation accuracy.

Using a cohort of 100 prostate cancer patients, Jensen et al. demonstrated that their novel machine learning model can be used to quickly estimate the Pareto set of feasible dose objectives in cancer radiotherapy, which may directly accelerate the treatment planning process and indirectly improve final plan quality by allowing more time for plan refinement. Their model outperforms the existing machine learning techniques by utilizing optimization priorities and output initialization.

As a first attempt, Mistro et al. have demonstrated that knowledge models can be effectively used as teaching aid to bring inexperienced planners to a level close to experienced planners in fewer than 2 days. The proposed tutoring system can serve as an essential component in an AI ecosystem that will enable clinical practitioners to use knowledge-based planning effectively and confidently for personalized radiation treatment.

Based on 85 training cases and 15 test cases, Wang et al. have demonstrated a novel deep learning framework for pancreas stereotactic body radiation therapy (SBRT) planning, which can predict a fluence map for each beam, hence bypassing the lengthy inverse optimization process.

In their work, Barua et al. demonstrated that a Multivariate Functional Principal Component Analysis (MFPCA) approach can be used to characterize the temporal trajectories of mandibular subvolumes receiving radiation. Their work suggests that temporal trajectories of radiomics features derived from sequential pre- and post-RT CT scans correlate with radiotherapy-induced mandibular injury, which may be used to aid in earlier management of osteoradionecrosis, a major side-effect in radiation therapy of oropharyngeal cancer patients.

In a mini-review, Luo summarized three major approaches currently employed in predicting cervical cancer outcomes: statistical models, medical images, and machine learning, and discussed some of the challenges in making clinical outcome prediction more accurate, reliable, and practical.

Zhang et al. proposed a transfer learning-based prognostication model for overall survival in pancreatic ductal adenocarcinoma patients. The model achieved the area under the receiver operating characteristic curve (AUC) of 0.81, significantly higher than that of the traditional radiomics model of 0.54. Their result suggests that transfer learning-based models may significantly improve prognostic performance in typical small sample size medical imaging studies.

To evaluate the overall effectiveness of personalized medicine, Kapelner et al. introduced and discussed a novel R package called "Personalized Treatment Evaluator (PTE)" developed by them. They combined randomized comparative/controlled trial (RCT) data with a statistical model of the response to estimate outcomes under different treatment allocation protocols. Their PTE package can be used to evaluate personalization models in medicine as well as fields outside of medicine.

In their paper, Namdar et al. presented first a comprehensive review of AUC metric, and then proposed a modified version of AUC that takes confidence of the model into account and incorporated AUC into Binary Cross Entropy (BCE) loss function. They demonstrated the validity of the new concept on MNIST, prostate MRI, and brain MRI datasets.

In a review paper, Chan et al. discussed and summarized the various applications of machine learning approaches in machine-specific and patient-specific quality assurance (QA), a key component in safeguarding patient safety during the radiation treatment of cancer patients.

CONCLUSIONS

Precision medicine is an evolving healthcare approach focused on tailoring medical decisions, treatments, practices, and products to individual patients based on their genetic, environmental, lifestyle, and other factors. In this Research Topic, eleven teams reported promising results from their experience in applying AI for precision medicine. Moving forward, we anticipate that more work needs to be done to eliminate biases in the AI models and make these models interpretable, therefore ultimately achieving the promise of precision medicine, i.e., delivering the right treatment to the right patient at the right time.

AUTHOR CONTRIBUTIONS

JD drafted the editorial. JD, TH, EC, JC, and FE-S revised and approved the final version. All authors contributed to the article and approved the submitted version.

FUNDING

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB022589, by the National Science Foundation under Award Number DMS 1918925, by the National Cancer Institute under Award Number 21X130F, and by the Department of Energy under Award Number DE-SC0021655 to JD. EC was supported by the grant NSF 19-500 under Award Number DMS 1918925 and 1922843 in years 2019-2022.

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of those institutions.

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.

ORIGINAL RESEARCH
published: 28 August 2020
doi: 10.3389/frai.2020.00066

Knowledge Models as Teaching Aid for Training Intensity Modulated Radiation Therapy Planning: A Lung Cancer Case Study

Matt Mistro[1,2], *Yang Sheng*[1]*, *Yaorong Ge*[3], *Chris R. Kelsey*[1], *Jatinder R. Palta*[4], *Jing Cai*[5], *Qiuwen Wu*[1], *Fang-Fang Yin*[1] and *Q. Jackie Wu*[1]

[1] Department of Radiation Oncology, Duke University Medical Center, Durham, NC, United States, [2] Medical Physics Graduate Program, Duke University, Durham, NC, United States, [3] Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, United States, [4] Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, United States, [5] Department of Health Technology and Informatics, Hong Kong Polytechnic University, Hong Kong, China

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Maria F. Chan,
Cornell University, United States
Yibao Zhang,
Peking University Cancer Hospital,
China

***Correspondence:**
Yang Sheng
yang.sheng@duke.edu

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Artificial Intelligence

Received: 25 April 2020
Accepted: 21 July 2020
Published: 28 August 2020

Citation:
Mistro M, Sheng Y, Ge Y, Kelsey CR, Palta JR, Cai J, Wu Q, Yin F-F and Wu QJ (2020) Knowledge Models as Teaching Aid for Training Intensity Modulated Radiation Therapy Planning: A Lung Cancer Case Study. Front. Artif. Intell. 3:66. doi: 10.3389/frai.2020.00066

Purpose: Artificial intelligence (AI) employs knowledge models that often behave as a black-box to the majority of users and are not designed to improve the skill level of users. In this study, we aim to demonstrate the feasibility that AI can serve as an effective teaching aid to train individuals to develop optimal intensity modulated radiation therapy (IMRT) plans.

Methods and Materials: The training program is composed of a host of training cases and a tutoring system that consists of a front-end visualization module powered by knowledge models and a scoring system. The current tutoring system includes a beam angle prediction model and a dose-volume histogram (DVH) prediction model. The scoring system consists of physician chosen criteria for clinical plan evaluation as well as specially designed criteria for learning guidance. The training program includes six lung/mediastinum IMRT patients: one benchmark case and five training cases. A plan for the benchmark case is completed by each trainee entirely independently pre- and post-training. Five training cases cover a wide spectrum of complexity from easy (2), intermediate (1) to hard (2). Five trainees completed the training program with the help of one trainer. Plans designed by the trainees were evaluated by both the scoring system and a radiation oncologist to quantify planning quality.

Results: For the benchmark case, trainees scored an average of 21.6% of the total max points pre-training and improved to an average of 51.8% post-training. In comparison, the benchmark case's clinical plans score an average of 54.1% of the total max points. Two of the five trainees' post-training plans on the benchmark case were rated as comparable to the clinically delivered plans by the physician and all five were noticeably improved by the physician's standards. The total training time for each trainee ranged between 9 and 12 h.

Conclusion: This first attempt at a knowledge model based training program brought unexperienced planners to a level close to experienced planners in fewer than 2 days. The proposed tutoring system can serve as an important component in an AI ecosystem that will enable clinical practitioners to effectively and confidently use KBP.

Keywords: knowledge model, lung cancer, machine learning, tutoring system, intensity modulated radiation therapy

INTRODUCTION

Knowledge models collect and extract important patterns and knowledge from high quality clinical plans and utilize them to predict clinically optimal solutions for new cases. For treatment planning, this comes in the form of selected beam angles, optimized collimator settings, predicted achievable dose-volume histogram (DVH) endpoints for inverse optimization, and combined multiple parameter predictions for a fully automated treatment planning process (Zhu et al., 2011; Breedveld et al., 2012; Zhang et al., 2012, 2018, 2019a,b; Good et al., 2013; Voet et al., 2013; Zarepisheh et al., 2014; Sheng et al., 2015, 2019; Yuan et al., 2015, 2018; Hazell et al., 2016). Knowledge models have been successfully used in the clinical workflow for fully automated planning for some simpler cancer sites like prostate (Voet et al., 2014), but for more complicated sites, there may yet be some hurdles to overcome. Due to the limitation of training samples and other factors, they are often simplified to improve generalizability by regulating the capability of handling a wide array of niche scenarios in which a human planner would be better fit to tackle. Despite this, there is a lot to be gained from investigating the implicit knowledge of these models. The simple, logical principles that most of these models are built upon can not only start a foundation for less experienced users to progress toward clinical reliability but also bridge the gap between human and model knowledge in what to look for in evaluation and identification of planning intricacies. The goal is to make a human-centered artificial intelligence (AI) system to exploit the strengths from both ends and efficiently train competent planners.

While extensive training and arduous hours of practice can certainly cultivate competent and professional planners, more effective training programs are urgently needed to help more planners become proficient in the clinic as technologies continue to become more advanced and more complex. Of course, there are aspects of planning that can only be obtained by years of nuanced planning, but plan quality is not always shown to be better in those who have more experience (Nelms et al., 2012). Some planners with planning experience may encounter a bottle-neck in improving their versatility in planning various scenarios, due to the lack of understanding of the underlying subtlety which can be readily provided and instructed by the knowledge-based models. In addition, training a planner to a highly proficient level in a traditional mentor-tutor fashion is expensive in time and resources, and sometime the limited training resources are dispatched to more entry level learners and/or regional centers. A person can quickly learn how to plan well if the teaching is well-thought out and provides the base for the person to build their own intuition. A training program that introduces the benefit of knowledge-based models can accomplish this and aid in tearing down the notion of these models being entirely a black box which has been restrictive to clinical usage of models. Such a program can be a catalyst to bring more models into routine clinical work by showing how they work and what the best practice is. This study examines the workflow and feasibility of a training program that takes advantage of two knowledge-based models (Yuan et al., 2012, 2018) with carefully developed scoring criteria to facilitate efficient and quality learning of lung IMRT treatment planning to help establish intuition to trainees with no previous clinical planning experience.

The proposed training program lays the foundation for an entirely self-sufficient training module that will be designed as a constraint-based intelligent tutoring system (ITS) (Mitrovic et al., 2007, 2013; Dermeval et al., 2018). The constraint-based approach supports the type of learning problem that does not have an explicit solution or path for a user to follow as is the case of IMRT planning. The constraints are defined in the form of the scoring system, and the end goal is for the user to learn the planning actions that optimize the scoring system to obtain the highest score possible. In this constraint based framework, the user has to forge their own path from the information that is directed to them, and two people can take entirely different strategies and arrive at good solutions. This is a proof-of-concept study to show that there is valuable information to be gained from the knowledge models and they can be effectively and efficiently used in training new planners and give them the ability to utilize these models to generate quality plans. Here, we define new planners as those who have completed adequate medical physics course work but have minimal clinical treatment planning practice. As such, they would have completed classroom instructions of radiation therapy physics and advanced treatment planning. They would have basic operational knowledge of the TPS system, but have no experience in planning real clinical cases.

METHODS AND MATERIALS

Training Program Design

Program Overview

The overall training program design is shown in **Figure 1**. At the core of the training program is the tutoring system which consists of a front-end visualization module powered by KBP models and a plan scoring system. The visualization module

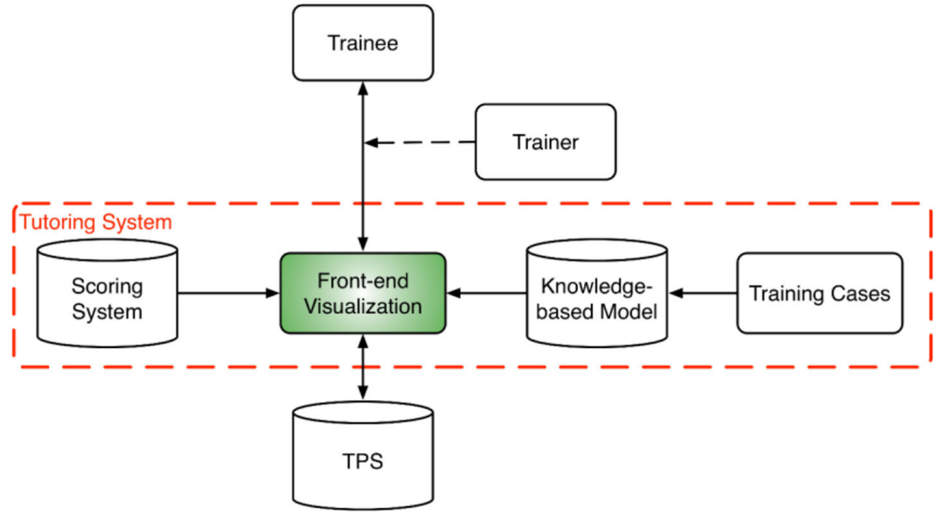


FIGURE 1 | System design diagram for the training program which includes a tutoring system at its core and a host of training cases. The tutoring system brings together the trainee, trainer, and the TPS. A trainer is optional for assisting the interaction between the trainee and the tutoring system. The tutoring system is powered by a scoring system and a set of knowledge models.

(**Figure 2**) provides the vital interactive workspace for the trainee and trainer, while the KBP models and scoring system provides back-end knowledge support. The KBP models currently include a beam bouquet prediction model and a dose-volume histogram (DVH) prediction model, while the scoring system consists of physician chosen criteria for clinical plan evaluation as well as specially designed criteria for learning guidance. These additional specially designed criteria were designed to help trainee understand the full scope of treatment planning and eventually achieve the ability to create a high quality plan, especially focusing on the criteria that are often qualitatively evaluated by the physician such as the overall isodose line conformity. Further, the tutoring system works in concert with the clinical treatment planning system (TPS) as trainees learn to generate clinically plans in a realistic clinical planning environment. In this study, we use the Eclipse® TPS (Varian Medical Systems, Palo Alto, CA) which provides fluence map optimization and dose calculation.

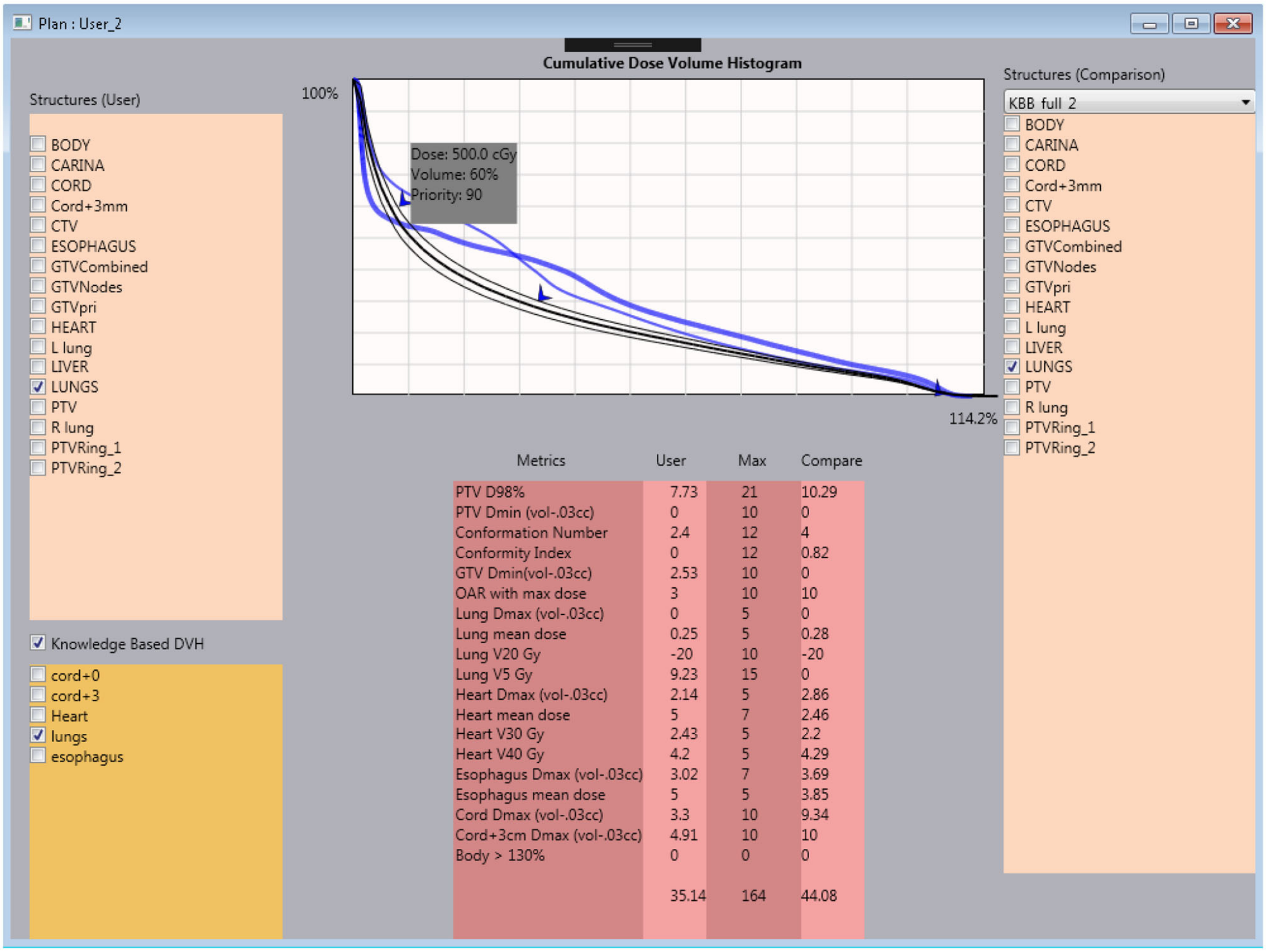
The current training program utilizes six lung/mediastinum IMRT patient cases: one benchmark case (shown in **Figure 3**) and five training cases. Each case is composed of clinical images, structures, and a delivered plan which were de-identified before incorporated into the training program. The benchmark case is used to track skill development. The five training cases cover the complexity from easy (2), intermediate (1) to hard (2) in lung IMRT planning. The difficulty level is determined by an experienced planner who evaluated the prescription, tumor size, complexity of shape, and proximity to organs-at-risk (OARs). The benchmark case, considered "intermediate-to-hard," has a target volume of 762.8 cc and a prescription of 62 Gy; two "easy" training cases have an average target volume of 113.8 cc and prescription of 40 Gy (reduced dose due to prior treatment); the "intermediate" training case has a target volume of 453.0 cc and a prescription of 60 Gy; two "hard" training cases have an average target volume of 845.7 cc and prescriptions of 60 Gy.

Before training begins, each trainee undergoes a benchmarking process to determine baseline score. In this process, the trainee is introduced to the treatment planning system with functionality they might not be familiar with as they have no prior experience. They are provided with the scoring metrics and asked to plan the benchmark case without any intervention from the trainer or the tutoring system (referred to as the baseline plan). The trainee is instructed that they have the choice of 6 or 10 MV beams and could have no more than 11 beams to align with current clinical practice.

Training Workflow

Figure 4 illustrates the typical training workflow (solid lines) for learning to plan one training case. The cases are selected sequentially from the easy ones to the difficult ones. For each case, a trainee goes through two phases of training: the beam selection phase and the fluence map optimization phase. In both phases, each training episode involves three main steps: (1) the trainee makes a decision (or takes an action); (2) the training program generates a plan corresponding to the decision and displays relevant dose metrics; (3) the training program then generates a comparison plan according to predictions from knowledge models and displays the same set of relevant dose metrics for comparison. The majority of interaction centers around the process with which the trainee learns to explain the differences between their plan and the comparison plan, as well as the resulting dosimetric implications of those differences.

During the beam selection phase, the trainee can choose the number of beams (seven to 11) and the direction of beams. The comparison plans are those generated with beams



Structures (User)

□ BODY
□ CARINA
□ CORD
□ Cord+3mm
□ CTV
□ ESOPHAGUS
□ GTVCombined
□ GTVNodes
□ GTVpri
□ HEART
□ L lung
□ LIVER
☑ LUNGS
□ PTV
□ R lung
□ PTVRing_1
□ PTVRing_2

☑ Knowledge Based DVH

□ cord+0
□ cord+3
□ Heart
☑ lungs
□ esophagus

Structures (Comparison)

KBB full 2

□ BODY
□ CARINA
□ CORD
□ Cord+3mm
□ CTV
□ ESOPHAGUS
□ GTVCombined
□ GTVNodes
□ GTVpri
□ HEART
□ L lung
□ LIVER
☑ LUNGS
□ PTV
□ R lung
□ PTVRing_1
□ PTVRing_2

Metrics	User	Max	Compare
PTV D98%	7.73	21	10.29
PTV Dmin (vol-.03cc)	0	10	0
Conformation Number	2.4	12	4
Conformity Index	0	12	0.82
GTV Dmin(vol-.03cc)	2.53	10	0
OAR with max dose	3	10	10
Lung Dmax (vol-.03cc)	0	5	0
Lung mean dose	0.25	5	0.28
Lung V20 Gy	-20	10	-20
Lung V5 Gy	9.23	15	0
Heart Dmax (vol-.03cc)	2.14	5	2.86
Heart mean dose	5	7	2.46
Heart V30 Gy	2.43	5	2.2
Heart V40 Gy	4.2	5	4.29
Esophagus Dmax (vol-.03cc)	3.02	7	3.69
Esophagus mean dose	5	5	3.85
Cord Dmax (vol-.03cc)	3.3	10	9.34
Cord+3cm Dmax (vol-.03cc)	4.91	10	10
Body > 130%	0	0	0
	35.14	164	44.08

FIGURE 2 | Interactive user interface of the tutoring system. Within the system, the trainee is capable of checking the current plan's metrics against the clinical plan and knowledge model DVH prediction.

determined by the knowledge-based beam selection model. Both trainee plans and comparison plans are created by an automatic KBP algorithm. The trainee determines whether they prefer to move along the direction of model prediction or continue with their own direction. At the end of this phase, the final beam comparison provides an assessment of the expected dosimetric differences contributed by trainee's beam design.

When the optimization training phase begins, the trainee creates the initial optimization objectives and finishes the planning process. In parallel, a comparison plan is generated with the KBP beam setting using trainee's dose-volume constraints. Dosimetric comparisons between plans allow the trainee to appreciate whether the results aligned with their expectations during the aforementioned assessment, which builds a forward intuition on beam choice implications.

Following this, the KBP DVH model is imported and the trainee is able to compare their plan's DVHs and dose objectives to where the DVH model predicts they should be able to achieve. The trainee then makes changes based on what they see is obtainable. After the changes are made and the plan is scored, a final comparison is done with the clinically delivered plan. The trainee works backwards by looking at the scoring and DVH of the clinical plan and ponders on how the clinically delivered plan might have been achieved. This is to further ingrain a backwards intuition for the metrics related to certain collective beam arrangements.

As shown in **Figure 4**, the training workflow also includes a few steps (dashed boxes) that are designed for people with little to no knowledge of treatment planning. These steps are optional when trainees are at more advanced stages during the training process. The first beam assessment is an initial guidance with the trainer about the best beam direction to select if they were to make a plan with only a single beam. This step encourages the trainee to think about how each individual beam will contribute to the final dose distribution. The second beam assessment helps the trainee make an optimal plan when only two beams are used. This helps planners understand how multiple beams interact with one another (i.e., the second best beam isn't necessarily the

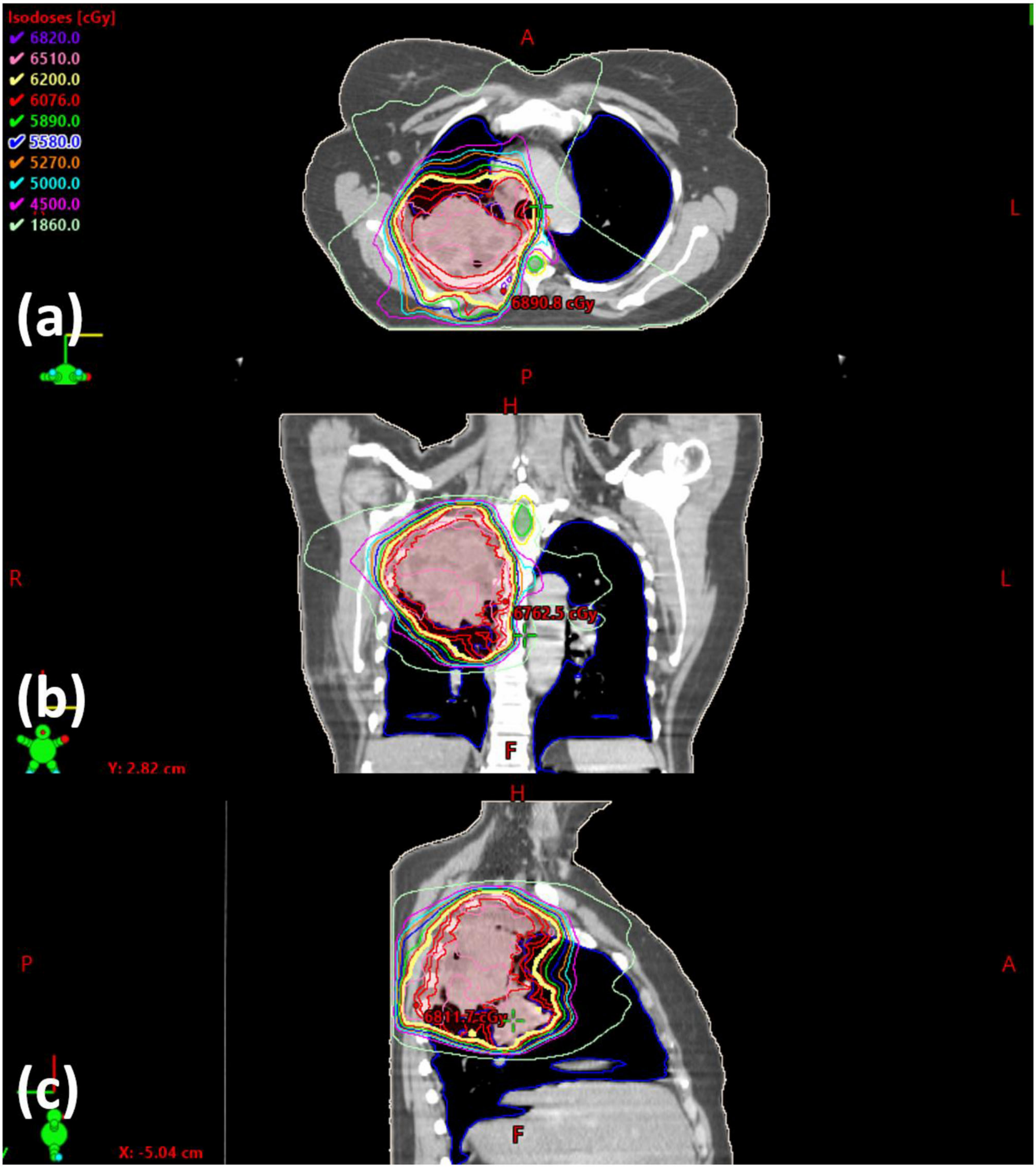


FIGURE 3 | Screenshot of the benchmark case in **(a)** axial, **(b)** coronal, and **(c)** sagittal view. The clinically delivered plan's isodose is displayed.

best beam to work with the first). Lastly, the "basic constraint assessment" step is a simple check to ensure that the trainee has at least one objective for all the relevant structures and two for the target.

The current training program takes the trainee through the workflow described in **Figure 4** five times, one for each training case, in increasing order of difficulty. After completing all five cases, the trainee returns to the benchmark case and creates a

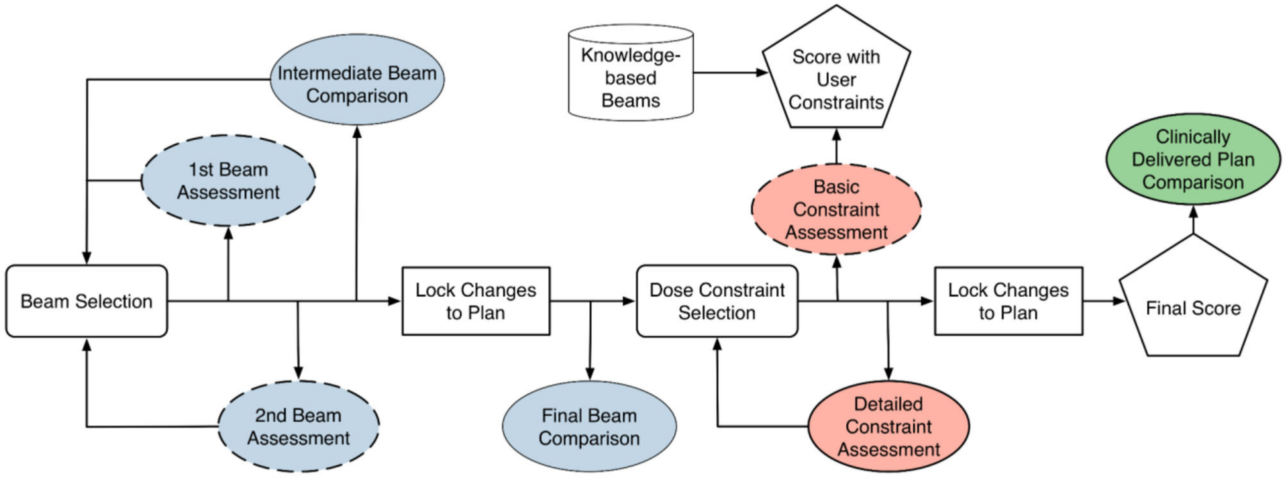

Knowledge-based Beams
Score with User Constraints
Intermediate Beam Comparison
1st Beam Assessment
Clinically Delivered Plan Comparison
Basic Constraint Assessment
Beam Selection
Lock Changes to Plan
Dose Constraint Selection
Lock Changes to Plan
Final Score
2nd Beam Assessment
Final Beam Comparison
Detailed Constraint Assessment

FIGURE 4 | Training diagram that is largely based on comparison between trainee's results and knowledge-based planning (KBP) models. Blue-colored process is geometry-based assessment. Red-colored process is objective-based assessment. Green-colored process is geometry and objective based assessment. Dashed box is considered optional step. Cylindrical block is based on knowledge-based model.

new plan entirely on their own without any intervention from the trainer, knowledge models, or the tutoring system. This post-training plan in comparison to the baseline plan on the same case provides an objective way to assess if there is any significant improvement in their planning ability.

Tutoring System Design

As introduced in the previous section, the current tutoring system includes three major components: a visualization module for user interaction, knowledge models for planning guidance, and a scoring system for plan assessment. The visualization module is integrated with the Eclipse® TPS and is currently implemented as a script using the Eclipse® API. In the following, we provide a brief description of the knowledge models and the scoring system.

Beam Angle Selection Model

The beam model (Yuan et al., 2015, 2018) predicts the best beam configuration for each new case, including the number of beams and the angle of the beams. It operates on a novel beam efficiency index that tries to maximize the dose delivered to a PTV and minimize the dose delivered to OARs based on a number of weighting factors. It also introduces a forced separation among good quality beams to cover sufficient co-planar space. The weighting factors and other parameters of the beam model are learned from a set of high quality prior clinical cases (Yuan et al., 2018). For the purposes of simplicity of introduction to new planners, all beams in the current training program are restricted to co-planar beams.

DVH Prediction Model

The DVH prediction model estimates the best achievable DVH of the OARs based on a number of anatomical features: distance-to-target histogram (DTH) principal components, OAR volume, PTV volume, OAR-PTV overlapping volume, and out-of-field OAR volume (Yuan et al., 2012). The model is trained with a set of prior lung cases with a variety of tumor sizes and locations. For this study, the model predicts DVHs that are useful for the trainees during the learning and planning. Organs-at-risk included in each DVH are cord, cord+3 mm, lungs, heart, and esophagus.

Plan Scoring System

A plan scoring system was designed to help trainees understand the quality of different plans from the choices of beams and DVH parameters. Therefore, the scoring system incorporates both physician's clinical evaluation criteria and planning knowledge. The metrics with their respective max point values are shown in **Table 1**. As noted, since each case has its own unique anatomy and complexity, the most achievable points of a plan is always less than the total max points, while more difficult cases have lower best achievable points. The best achievable points of each plan are not normalized so the trainees are encouraged to rely on the actual planning knowledge to "do their best," rather than to get "100 percent score" or gaming the system. There were a total of 164 points, with which the raw score was normalized to represent the percentage score. A maximum scoring would have 100% percentage score. Normalization was performed after the training was done as a summary of the data. It is worth reiterating that the trainee was unaware of the maximally achievable score for each case so they couldn't game the system. Note that even clinically delivered plans may not be perfect in all categories, and therefore, may not achieve the highest possible scores.

An effective scoring system can be created in many ways. The current system starts with the logic of rewarding dosimetric endpoints that are clinically relevant as explained in RTOG reports (Chun et al., 2017), other clinical considerations (Kong et al., 2011; Baker et al., 2016) and planning competitions powered by ProKnow (ProKnow Systems, Sanford, FL;

TABLE 1 | Metrics chosen to be a part of the scoring system and their respective maximum point value.

Target	Max	Lung	Max	Heart	Max	Esophagus	Max	Spinal cord	Max
PTV D98%	21	Max dose	5	Max dose	5	Max dose	7	Cord max dose	10
PTV min dose	10	Mean dose	5	Mean dose	7	Mean dose	5	Cord+3 mm max dose	10
GTV min dose	10	V20 Gy	10	V30 Gy	5				
CN 95%	12	V5 Gy	15	V40 Gy	5				
CI 50%	12								
Location of max dose	10								

PTV, Planning Target Volume; GTV, Gross Tumor Volume; CN, Conformity Number; CI, Conformity Index.

www.proknowsystems.com). The conformity index (CI) (Knoos et al., 1998) aims to limit the isodose volume. The conformation number (CN) or Paddick conformity index (Paddick, 2000) follows similar logic to CI but focuses on the portion of that isodose volume within the target.

Training Program Assessment

To assess the effectiveness of the training program, five trainees who satisfy the criteria of new planners went through the entire training program. For all five trainees, the baseline and post-training plans of the benchmark case were scored and analyzed for evidence of learning. Furthermore, the post-training plans of all the training cases as well as all the clinically delivered plans were also scored and analyzed for trainee performance and potential knowledge gaps.

Moreover, to assess how the overall scores given by the scoring system closely reflect true plan quality in a real clinical scenario, a physician who specializes in the treatment of lung cancer evaluated each of the plans to provide an expert opinion on their clinical quality. For each trainee, the post-training plan of the benchmark case was first compared with the baseline plan of the same case and then against the clinically delivered plan by the physician. Each comparison was categorized on a simplified 5-point scale of (1) significantly worse, (2) moderately worse, (3) comparable, (4) moderately better, and (5) significantly better. The physician also evaluated the trainee's post-training plans on whether they could be approved for clinical delivery.

RESULTS

Scoring Results

Five trainees went through the training program and their scores are shown in **Figure 5**. All trainees went through multiple classroom courses on radiation physics, anatomy, radiation biology, and treatment planning/dosimetry. They also completed a basic practicum course to learn the essential operations of a treatment planning system. After training, the overall score of all trainees was unanimously improved from the baseline and was much closer to that of the clinically delivered plan (**Figure 5A**). Trainee 1 and 3 received a planning score point that was slightly above that of the clinically delivered plan, with an average of 54.4%, while the other three were marginally lower with an average of 50.1%. In comparison, the score of the clinically delivered plan was 54.1%. Detailed scores are listed in **Table 2**. For the five cases used within the training program (**Figure 5B**), every trainee obtained a score in the final plan that was greater than that of the clinically delivered plan with the exception of case 3 for trainee 5 and there was an overall average of 12.6 raw planning score point improvement over the respective clinically delivered plans. Detailed breakdown of each trainee's performance on each training case is listed in **Table 3**.

Physician Evaluation Results

Table 4 shows the physician evaluation of the trainee's post-training plans as compared to the benchmark plans and the clinically delivered plans for the benchmark case. Two plans designed by trainees #1 and #3 that scored slightly better than the clinical plan per the scoring system were deemed as comparable to the clinical plan by the physician. The other plans were rated as marginally worse. All trainee's post-training plans were rated moderately better than the initial benchmark plans. Only one of the trainee's plan was deemed appropriate for clinical use based on the physician's discretion.

DISCUSSION

This is the first attempt at developing an effective training program for IMRT planning that capitalizes on the implicit planning tactics that is built into knowledge models for lung IMRT. As trainees go through the training program, the prediction from knowledge models provides guidance at multiple steps and the carefully thought-out scoring objectives direct them toward appropriate choices or skills to create a clinically viable plan. The initial assessment indicates that the knowledge model based training program can substantially improve the planning knowledge of novice trainees in a short period of time (9–12 h in this study). Furthermore, for some trainees their knowledge may approach a clinical proficient level within this short period.

This training program demonstrates the feasibility that knowledge models can be effective teaching aids to help human planners understand the key steps toward generating a clinically viable plan. This is an important first attempt to use knowledge models in a human training process. We hypothesize that by giving trainees opportunities to compare and reflect on the predictions from knowledge models and their own understanding of the planning process, these human planners

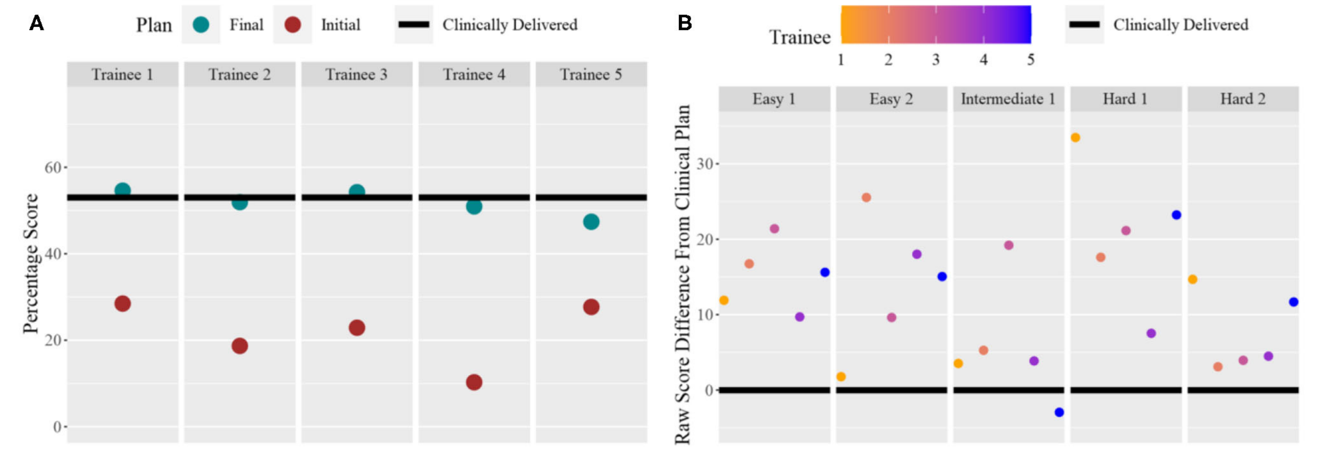


FIGURE 5 | (A) For each trainee (column), the total score for the benchmark case: pre-training plan (purple dot) and post-training plan (green dot) compared to the clinically delivered plan (black line). **(B)** For each training case (column), and for each trainee (color dots), the score difference between the trainee plan and the clinically delivered plan (black line indicating 0).

TABLE 2 | Scores of the benchmark plan for each trainee before and ("Initial") after ("Final") training.

Trainee ID	Plan	Raw score	Percentage score
Trainee 1	Initial	46.69	28.47
Trainee 2	Initial	30.64	18.68
Trainee 3	Initial	37.53	22.88
Trainee 4	Initial	16.83	10.26
Trainee 5	Initial	45.42	27.70
Trainee 1	Final	89.54	54.60
Trainee 2	Final	85.18	51.94
Trainee 3	Final	88.96	54.24
Trainee 4	Final	83.56	50.95
Trainee 5	Final	77.74	47.40

will have a better and more concrete understanding of the knowledge models and thus have confidence in making their planning decisions rather than simply accepting the predicted results. While further research is needed to design more effective mechanisms for incorporating knowledge models in human learning, this study has shown that proper design of a plan scoring system provides one effective approach to helping trainees understand the effects of beams and constraints. Further development and testing of the scoring system are warranted since five cases are not likely to cover the possible case variations and review by only one physician may not be sufficient to cover variations in clinical considerations.

While the beam and DVH prediction models used in this study make for a good foundation, additional and more sophisticated knowledge models are needed to address the skills and knowledge that are currently provided by trainers throughout the training to produce clinically viable plans. Examples of important considerations during planning include collimator optimization and strategies to fine-tune small regions that are less optimal.

In the current implementation, the plan scoring system serves multiple purposes. First, the total score should measure the overall quality of a plan. Second, the less than satisfactory scores should emphasize the most important metrics that require attention. Third, in an indirect way, we want the total score to measure a trainee's mastery of planning knowledge and the difference in scores on the same case to measure the trainee's level of improvement (i.e., learning). Scoring for the first purpose has been studied in quality assurance literature (Mayo et al., 2017). Unfortunately, this scoring will always have an *ad hoc* nature as physicians' preferences will vary, and one scoring system that is in perfect agreement with one physician may not hold true for another. Moreover, some metrics are prioritized conditionally depending on other metrics. One such scoring difficulty is in terms of the metrics that physicians utilize to make decisions based on seemingly minor differences. For example, in some cases, the esophagus may not be prioritized as highly as the lung or the heart, but if the other metrics are at an acceptable level then even small differences in the esophageal metrics may be considered more important than moderate differences in lung dose. This is because most people with locally-advanced lung cancer will experience some degree of esophagitis (Chapet et al., 2005) while a much smaller percentage will experience pneumonitis. This type of conditional prioritization poses significant challenges for scoring system design and require further investigation. Scoring systems for the latter two purposes have not been previously studied. One challenge that we faced is the exploitation of the scoring system by trainees. That is, poorly designed scoring systems tend to allow trainees to attain high scores without actually understanding planning knowledge and actually creating high quality plans. We have improved our scoring system iteratively by adjusting the priority (i.e., max point) assignments based on pilot testing

TABLE 3 | Scores of five plans during training for each of five trainees. Score difference is defined as the difference between trainee's plan's score vs. the clinical plan's score.

Trainee ID	Raw score	Score difference	Plan
1	129.02	11.9	Easy 1
2	133.84	16.74	Easy 1
3	138.49	21.39	Easy 1
4	126.8	9.7	Easy 1
5	132.71	15.61	Easy 1
1	120.51	1.77	Easy 2
2	144.27	25.53	Easy 2
3	128.36	9.62	Easy 2
4	136.75	18.01	Easy 2
5	133.79	15.05	Easy 2
1	54.75	3.53	Intermediate 1
2	56.49	5.27	Intermediate 1
3	70.42	19.2	Intermediate 1
4	55.08	3.86	Intermediate 1
5	48.27	-2.95	Intermediate 1
1	60.66	33.48	Hard 1
2	44.78	17.6	Hard 1
3	48.32	21.14	Hard 1
4	34.7	7.52	Hard 1
5	50.4	23.22	Hard 1
1	83.91	14.68	Hard 2
2	72.32	3.09	Hard 2
3	73.17	3.94	Hard 2
4	73.72	4.49	Hard 2
5	80.91	11.68	Hard 2

TABLE 4 | Physician evaluation of trainee post-training plan on 5-point scale (significantly worse to significantly better) and clinical feasibility rating.

Trainee #	Comparison to clinical	Comparison to benchmark	Clinically feasible
1	Comparable	Moderately better	No
2	Moderately worse	Moderately better	No
3	Comparable	Moderately better	Yes
4	Moderately worse	Moderately better	No
5	Moderately worse	Moderately better	No

results. It is also important not to adjust the scoring priority for every plan or trainee because there will always be new ways to exploit any scoring system. One possible solution to this is to have a progressive scoring system that adjusts priority when reaching certain thresholds. Another approach is to use entirely separate and different mechanisms for the latter two purposes. For example, instead of using a score to measure a trainee's knowledge, we may use a Bayesian model to assess the probability of the trainee's understanding of a case as is done in modern Intelligent Tutoring Systems (Santhi et al., 2013).

The current training program has many limitations. We can observe one example by comparing the left and right of **Figure 5**. As seen in the right figure, after training using the knowledge models, all five trainees were able to generate plans that score higher than clinically delivered plans for all five training cases. However, as shown in the left figure, when the trainees returned to the benchmark case, only two trainees were able to achieve near or just at the level of the clinically delivered plans. It can be inferred that some of the trainees might not have fully absorbed the knowledge that was presented to them through the training program. It is also possible that the benchmark case requires special knowledge that is not well-presented to the trainees. In addition, we noticed that during physician plan evaluation, only one of two plans that outscored the benchmark case's clinical plan was deemed clinically acceptable. It is possible that additional plan quality related metric could be introduced in the scoring system to better quantify a plan's clinical applicability. Further research in all aspects of the program, including the knowledge models, the scoring system, the coverage of essential knowledge, and the selection of training cases, is necessary to improve the effectiveness of the training program. Finally, the current implementation is based on a specific commercial TPS platform and its existing application programming interface. While general principles of training workflow design are applicable to other commercial platforms, methods for adapting the proposed design to other planning technologies and platforms deserve further investigation.

Even though the current training program has shown encouraging results that demonstrate its feasibility, there are clearly much to be done to develop a truly effective training program for knowledge-based IMRT planning. The immediate next stage includes the need to enhance the scoring system, extend knowledge models, and expand to a larger study with more training cases and with a variety of sites beyond just the lung. Another important task is to conduct a larger study with more trainees and more physicians to fully evaluate the benefits of the training program centered around knowledge-based models. As discussed in the introduction, our ultimate goal is to develop the training program into a fully asynchronous intelligent tutoring system as we gain a better understanding of the essential components and algorithms that are required by such a training system. Having a human trainer in the current program will provide important feedback for future designs. With permission of trainees, all conversations can be recorded in order to find where and how best to provide certain learning materials and pertinent hints. An intelligent training system operating asynchronously may be invaluable for reducing costs of planner training, providing an educational resource to graduate programs, tearing down the black box mindset of knowledge models in clinical practices, and improving the quality of care in cancer centers across the world (Zubizarreta et al., 2015).

CONCLUSION

We have demonstrated that knowledge models can be effectively used as teaching aid in a training program to bring unexperienced planners to a level close to experienced planners in a short period of time. The assessments indicate that the knowledge models helped trainees improve their knowledge and skills for producing higher quality plans. We believe this knowledge model based training program can serve as an important component of an AI ecosystem that will enable clinical practitioners to effectively and confidently use KBP in radiation treatment. Further efforts are needed to enhance, validate, and ultimately automate the training program.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MM and YS performed system design, trainee tutoring, and treatment plan comparison. CK reviewed plan as physician expert and provided feedback for plan quality. JC, JP, QW, and F-FY provide consultation, review experiment design, paper content, and statistical analysis. YG and QJW supervised the entire study and revised this paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by NIH R01CA201212 research grant and Varian master research grant.

REFERENCES

Baker, S., Dahele, M., Lagerwaard, F. J., and Senan, S. (2016). A critical review of recent developments in radiotherapy for non-small cell lung cancer. *Rad. Oncol*. 11:115. doi: 10.1186/s13014-016-0693-8

Breedveld, S., Storchi, P. R. M., Voet, P. W. J., and Heijmen, B. J. M. (2012). iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Med. Phys*. 39, 951–963. doi: 10.1118/1.3676689

Chapet, O., Kong, F.-M., Lee, J. S., Hayman, J. A., and Ten Haken, R. K. (2005). Normal tissue complication probability modeling for acute esophagitis in patients treated with conformal radiation therapy for non-small cell lung cancer. *Radiother. Oncol*. 77, 176–181. doi: 10.1016/j.radonc.2005.10.001

Chun, S. G., Hu, C., Choy, H., Komaki, R. U., Timmerman, R. D., Schild, S. E., et al. (2017). Impact of intensity-modulated radiation therapy technique for locally advanced non–small-cell lung cancer: a secondary analysis of the NRG oncology RTOG 0617 randomized clinical trial. *Clin. Trial* 35, 56–62. doi: 10.1200/JCO.2016.69.1378

Dermeval, D., Paiva, R., Bittencourt, I. I., Vassileva, J., and Borges, D. (2018). Authoring tools for designing intelligent tutoring systems: a systematic review of the literature. *Int. J. Artificial Intelligence Edu*. 28, 336–384. doi: 10.1007/s40593-017-0157-9

Good, D., Lo, J., Lee, W. R., Wu, Q. J., Yin, F. F., and Das, S. K. (2013). A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning. *Int. J. Rad. Oncol. Biol. Phys*. 87, 176–181. doi: 10.1016/j.ijrobp.2013.03.015

Hazell, I., Bzdusek, K., Kumar, P., Hansen, C. R., Bertelsen, A., Eriksen, J. G., et al. (2016). Automatic planning of head and neck treatment plans. *J. Appl. Clin. Med. Phys*. 17, 272–282. doi: 10.1120/jacmp.v17i1.5901

Knoos, T., Kristensen, I., and Nilsson, P. (1998). Volumetric and dosimetric evaluation of radiation treatment plans: radiation conformity index. *Int. J. Rad. Oncol. Biol. Phys*. 42, 1169–1176. doi: 10.1016/S0360-3016(98)00239-9

Kong, F.-M., Ritter, T., Quint, D. J., Senan, S., Gaspar, L. E., Komaki, R. U., et al. (2011). Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *Int. J. Rad. Oncol. Biol. Phys*. 81, 1442–1457. doi: 10.1016/j.ijrobp.2010.07.1977

Mayo, C. S., Yao, J., Eisbruch, A., Balter, J. M., Litzenberg, D. W., Matuszak, M. M., et al. (2017). Incorporating big data into treatment plan evaluation: development of statistical DVH metrics and visualization dashboards. *Adv. Radiat. Oncol*. 2, 503–514. doi: 10.1016/j.adro.2017.04.005

Mitrovic, A., Martin, B., and Suraweera, P. (2007). Intelligent tutors for all: the constraint-based approach. *IEEE Intelligent Syst*. 22, 38–45. doi: 10.1109/MIS.2007.74

Mitrovic, A., Ohlsson, S., and Barrow, D. K. (2013). The effect of positive feedback in a constraint-based intelligent tutoring system. *Comp. Educ*. 60, 264–272. doi: 10.1016/j.compedu.2012.07.002

Nelms, B. E., Robinson, G., Markham, J., Velasco, K., Boyd, S., Narayan, S., et al. (2012). Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems. *Pract. Rad. Oncol*. 2, 296–305. doi: 10.1016/j.prro.2011.11.012

Paddick, I. (2000). A simple scoring ratio to index the conformity of radiosurgical treatment plans. *J. Neurosurg*. 93:219. doi: 10.3171/jns.2000.93.supplement_3.0219

Santhi, R., Priya, B., Nandhini, J. M. J. A. (2013). Review of intelligent tutoring systems using bayesian approach. *[arXiv preprint]*. arXiv:abs/1302.7081.

Sheng, Y., Li, T., Zhang, Y., Lee, W. R., Yin, F. F., Ge, Y., et al. (2015). Atlas-guided prostate intensity modulated radiation therapy (IMRT) planning. *Phys. Med. Biol*. 60, 7277–7291. doi: 10.1088/0031-9155/60/18/7277

Sheng, Y., Zhang, J., Wang, C., Yin, F.-F., Wu, Q. J., and Ge, Y. (2019). Incorporating case-based reasoning for radiation therapy knowledge modeling: a pelvic case study. *Technol. Cancer Res. Treat*. 18:1533033819874788. doi: 10.1177/1533033819874788

Voet, P. W., Dirkx, M. L., Breedveld, S., Fransen, D., Levendag, P. C., and Heijmen, B. J. (2013). Toward fully automated multicriterial plan generation: a prospective clinical study. *Int. J. Rad. Oncol. Biol. Phys*. 85, 866–872. doi: 10.1016/j.ijrobp.2012.04.015

Voet, P. W. J., Dirkx, M. L. P., Breedveld, S., Al-Mamgani, A., Incrocci, L., and Heijmen, B. J. M. (2014). Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients. *Int. J. Rad. Oncol. Biol. Phys*. 88, 1175–1179. doi: 10.1016/j.ijrobp.2013.12.046

Yuan, L., Ge, Y., Lee, W., Yin, F. F., Kirkpatrick, J., and Wu, Q. (2012). Quantitative analysis of the factors which affect the inter-patient organ-at risk dose sparing variation in IMRT plans. *Med. Phys*. 39, 6868–6878. doi: 10.1118/1.4757927

Yuan, L., Wu, Q. J., Yin, F., Li, Y., Sheng, Y., Kelsey, C. R., et al. (2015). Standardized beam bouquets for lung IMRT planning. *Phys. Med. Biol*. 60, 1831–1843. doi: 10.1088/0031-9155/60/5/1831

Yuan, L., Zhu, W., Ge, Y., Jiang, Y., Sheng, Y., Yin, F., et al. (2018). Lung IMRT planning with automatic determination of beam angle configurations. *Phys. Med. Biol*. 63:135024. doi: 10.1088/1361-6560/aac8b4

Zarepisheh, M., Long, T., Li, N., Tian, Z., Romeijn, H. E., Jia, X., et al. (2014). A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Med. Phys*. 41:061711. doi: 10.1118/1.4875700

Zhang, J., Ge, Y., Sheng, Y., Yin, F.-F., and Wu, Q. J. (2019a). Modeling of multiple planning target volumes for head and neck treatments in knowledge-based treatment planning. *Med. Phys*. 46, 3812–3822. doi: 10.1002/mp.13679

Zhang, J., Wu, Q. J., Ge, Y., Wang, C., Sheng, Y., Palta, J., et al. (2019b). Knowledge-based statistical inference method for plan quality quantification. *Technol. Cancer Res. Treat*. 18:1533033819857758. doi: 10.1177/1533033819857758

Zhang, J., Wu, Q. J., Xie, T., Sheng, Y., Yin, F.-F., and Ge, Y. (2018). An ensemble approach to knowledge-based intensity-modulated radiation therapy planning. *Front. Oncol*. 8:57. doi: 10.3389/fonc.2018.00057

Zhang, X., Li, X., Quan, E., Pan, X., and Li, Y. (2012). A methodology for automatic intensity-modulated radiation treatment planning for lung cancer. *Phys. Med. Biol*. 56:9. doi: 10.1088/0031-9155/56/13/009

Zhu, X., Ge, Y., Li, T., Thongphiew, D., Yin, F. F., and Wu, Q. (2011). A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med. Phys*. 38, 719–726. doi: 10.1118/1.3539749

Zubizarreta, E. H., Fidarova, E., Healy, B., and Rosenblatt, E. (2015). Need for radiotherapy in low and middle income countries – the silent crisis continues. *Clin. Oncol*. 27, 107–114. doi: 10.1016/j.clon.2014.10.006

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.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 08 September 2020
doi: 10.3389/frai.2020.00068

Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy

Wentao Wang[1,2], Yang Sheng[1], Chunhao Wang[1], Jiahan Zhang[1], Xinyi Li[1,2], Manisha Palta[1], Brian Czito[1], Christopher G. Willett[1], Qiuwen Wu[1,2], Yaorong Ge[3], Fang-Fang Yin[1,2] and Q. Jackie Wu[1,2]**

[1] Department of Radiation Oncology, Duke University Medical Center, Durham, NC, United States, [2] Medical Physics Graduate Program, Duke University, Durham, NC, United States, [3] Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, United States

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Maria F. Chan,
Cornell University, United States
Xuejun Gu,
University of Texas Southwestern
Medical Center, United States

***Correspondence:**
Wentao Wang
wentao.wang@duke.edu
Q. Jackie Wu
jackie.wu@duke.edu

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Artificial Intelligence

Received: 24 April 2020
Accepted: 27 July 2020
Published: 08 September 2020

Citation:
Wang W, Sheng Y, Wang C, Zhang J, Li X, Palta M, Czito B, Willett CG, Wu Q, Ge Y, Yin F-F and Wu QJ (2020) Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy. Front. Artif. Intell. 3:68. doi: 10.3389/frai.2020.00068

Purpose: Treatment planning for pancreas stereotactic body radiation therapy (SBRT) is a difficult and time-consuming task. In this study, we aim to develop a novel deep learning framework to generate clinical-quality plans by direct prediction of fluence maps from patient anatomy using convolutional neural networks (CNNs).

Materials and Methods: Our proposed framework utilizes two CNNs to predict intensity-modulated radiation therapy fluence maps and generate deliverable plans: (1) Field-dose CNN predicts field-dose distributions in the region of interest using planning images and structure contours; (2) a fluence map CNN predicts the final fluence map per beam using the predicted field dose projected onto the beam's eye view. The predicted fluence maps were subsequently imported into the treatment planning system for leaf sequencing and final dose calculation (model-predicted plans). One hundred patients previously treated with pancreas SBRT were included in this retrospective study, and they were split into 85 training cases and 15 test cases. For each network, 10% of training data were randomly selected for model validation. Nine-beam benchmark plans with standardized target prescription and organ-at-risk constraints were planned by experienced clinical physicists and used as the gold standard to train the model. Model-predicted plans were compared with benchmark plans in terms of dosimetric endpoints, fluence map deliverability, and total monitor units.

Results: The average time for fluence-map prediction per patient was 7.1 s. Comparing model-predicted plans with benchmark plans, target mean dose, maximum dose (0.1 cc), and $D_{95\%}$ absolute differences in percentages of prescription were 0.1, 3.9, and 2.1%, respectively; organ-at-risk mean dose and maximum dose (0.1 cc) absolute differences were 0.2 and 4.4%, respectively. The predicted plans had fluence map gamma indices (97.69 ± 0.96% vs. 98.14 ± 0.74%) and total monitor units (2,122 ± 281 vs. 2,265 ± 373) that were comparable to the benchmark plans.

Conclusions: We develop a novel deep learning framework for pancreas SBRT planning, which predicts a fluence map for each beam and can, therefore, bypass the lengthy inverse optimization process. The proposed framework could potentially change the paradigm of treatment planning by harnessing the power of deep learning to generate clinically deliverable plans in seconds.

Keywords: deep learning, artificial intelligence, fluence map, treatment planning, convolutional neural network, pancreas, SBRT

INTRODUCTION

Pancreatic cancer is an aggressive and lethal malignancy that accounted for an estimated 4.5% of all cancer-related deaths worldwide in 2018 (Bray et al., 2018). Stereotactic body radiation therapy (SBRT) utilizes sophisticated image-guidance and motion-management techniques to allow the delivery of a highly conformal dose of radiation to the target while sparing the surrounding normal tissues. Due to the nature of the higher fractional dose, achieving steeper dose gradients is prioritized to better spare the gastrointestinal (GI) organs at risk (OARs), such as the stomach and duodenum/small bowel. In addition, the highly variable planning target volume (PTV) and OAR geometry make the planning task extremely challenging. Although limiting the OAR maximum dose frequently outweighs target coverage, a trial-and-error process attempts to cover as much of the target with a prescription dose as possible. The consistency of plan quality is hard to maintain due to time pressure and the planner's experience, which may result in suboptimal plans. A system capable of maintaining consistently high plan quality is warranted in modern radiation oncology departments.

Over the last decade, efforts have been made to implement treatment-planning automation. Machine learning (ML) algorithms have been utilized to extract clinical knowledge from existing plans and apply it in various formats to create plans for new patients, which is known as knowledge-based planning (KBP). One KBP approach relies on patient-specific, dose-volume histogram (DVH) prediction to guide inverse optimization. Such modeling is based on the patient's anatomical structures and prior planning knowledge. Traditional ML techniques have seen significant success in DVH prediction for many treatment sites (Zhu et al., 2011; Yuan et al., 2012; Good et al., 2013; Skarpman Munter and Sjolund, 2015). Another approach is voxel-wise dose prediction–based treatment-planning guidance. Over the past several years, a shape-based method (Liu et al., 2015), atlas-selection methods (Sheng et al., 2015; McIntosh and Purdie, 2016, 2017), and artificial neural network methods using handcrafted features (Shiraishi and Moore, 2016; Campbell et al., 2017) were proposed. Recently, convolutional neural networks (CNNs) have shown success in predicting 3-D dose distributions (Kearney et al., 2018; Barragán-Montero et al., 2019; Chen et al., 2019; Fan et al., 2019; Nguyen et al., 2019a,b). This type of model is typically referred to as a deep learning (DL) model. A majority of these models employ network structures similar to U-Net, which was initially developed for biomedical image segmentation (Ronneberger et al., 2015). However, in this approach, a second step of plan generation via inverse optimization is necessary to create a treatment plan aiming to achieve the predictions (McIntosh et al., 2017; Fan et al., 2019), either as DVH-based optimization or as voxel-based dose mimicking.

We contend that high-quality radiotherapy plans with standardized dose constraints and beam settings can be directly created by predicting their fluence maps without optimization or dose mimicking. We refer to this process as direct plan generation (as opposed to the automated planning process used in the literature that generally requires two steps as mentioned above). Few publications have focused on direct fluence map prediction (Lee et al., 2019; Sheng et al., 2019). In the case of whole breast irradiation, fluence prediction was achieved with a random forest model proposed by Sheng et al. (2019). Lee et al. (2019) show that, given the organ contours and the complete set of field-dose distributions, fluence maps for seven-beam prostate IMRT could be reconstructed by a modified U-Net with high accuracy. However, the study did not investigate how to obtain the known field dose. Rather, the authors assumed the field doses were a prerequisite for their technique to work. Indeed, solving the field dose of each beam remains a challenge. We hypothesize that anatomical planning features, together with the physician's planning objectives, could lead to accurate prediction of the field doses of each beam and their corresponding fluence maps. The expert planner incorporates the physician's planning objectives during manual planning. Therefore, these planning objectives are embedded in these plans, and DL models should be able to capture such information in the training data. In this feasibility study, we present a novel deep learning framework for direct fluence map prediction (a.k.a. direct plan generation) and demonstrate its performance using clinical pancreas SBRT cases.

MATERIALS AND METHODS

Patient Selection and Radiation Therapy Plan

One hundred pancreatic cancer patients previously treated with SBRT at Duke University Medical Center between 2014 and 2019 were included in this retrospective study. This study

Abbreviations: BEV, beam's eye view; CNN, convolutional neural network; CT, computed tomography; DAO, direct aperture optimization; DL, deep learning; DVH, dose-volume histogram; FD, field dose; FM, fluence map; GI, gastrointestinal; iGTV, internal gross tumor volume; KBP, knowledge-based planning; MAE, mean absolute error; MU, monitor unit; OAR, organ-at-risk; PTV, planning target volume; ROI, region of interest; SBRT, stereotactic body radiation therapy; TPS, treatment planning system.

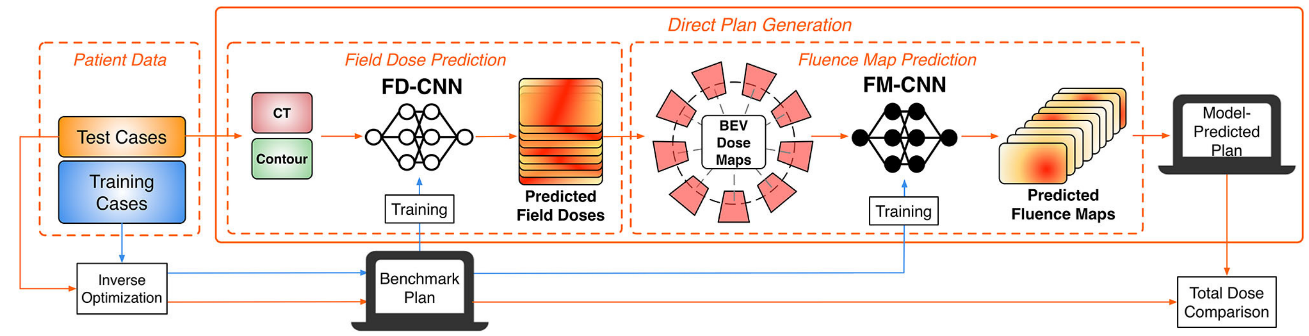


FIGURE 1 | Overall workflow of the DL modeling and validation. The prediction pipeline generates fluence maps from CT data and structure contours.

was approved by the institutional review board. In clinical plans, the dose prescription to the PTV was 25 Gy, often with a simultaneous integrated boost to the internal gross tumor volume (iGTV) with 33 or 40 Gy. The GI OAR (stomach, C-loop/duodenum, and bowels) dose constraints varied in maximum dose and maximum volume according to the different physician preferences. We aim to develop a model that is capable of generating clinical-quality pancreas SBRT IMRT plans. In this feasibility study, each case was replanned by experienced clinical physicists who specialized in GI SBRT using unified planning objectives and a standardized IMRT protocol with a single prescription level. The prescription for both the PTV and iGTV were 33 Gy in five fractions. All plans were designed with nine equally spaced coplanar 10-MV photon beams. Stomach, C-loop/duodenum, and bowels were combined and referred to as the OAR. The maximum dose for the OAR was limited to 25 Gy (0.1 cc). This protocol creates the scenario of an inverted relationship of target and OAR dose prescription, a clinical scenario that often has to be handled manually by an experienced planner for each case. In the following, we refer to the resulting standardized plans as the benchmark plans, which were used to train the model. The same beam orientations, including gantry angles, and beam shape definition via its open field dose, referred to as beam templates, were also included as input for the DL models. The 100 patient cases were divided randomly into an 85:15 training:testing ratio. All treatment plans were generated in the Eclipse® Treatment Planning System (TPS) (Varian Medical Systems, Palo Alto, CA) version 13.7 with the volume dose calculated by the Analytical Anisotropic Algorithm version 13.7.14. A Varian Millennium 120 multi-leaf collimator (MLC) was used to deliver the modulated fluence maps. The leaf-sequencing algorithm used was Smart LMC version 13.7.14.

Study Workflow

The overall study workflow is summarized in **Figure 1**. The proposed framework adopts a pipeline structure, in which two CNNs make consecutive predictions to generate the complete plan with fluence maps. The input into the pipeline includes planning computed tomography (CT) images as well as contours of the PTV and OARs. First, the field-dose CNN (FD-CNN) predicts 9 individual IMRT field dose distributions, i.e., FD-CNN field dose from CT and structure contours. Next, each 3-D field dose is projected along the beam's eye view (BEV), generating the 2-D BEV dose map. Finally, the fluence map CNN (FM-CNN) predicts the fluence map for each beam from the corresponding BEV dose map. The two CNNs were implemented in Keras with the Tensorflow backend and trained separately. The entire model was trained on a workstation with an Intel Xeon E5 v4 processor, 64 GB of RAM, and an NVIDIA Quadro M4000 graphics card. In order to evaluate the proposed framework's performance, we compared the automatically generated plans using the DL technique described in this research study, referred to as "model-predicted plans," against the benchmark plans generated by human experts using the standard inverse planning process.

Data Preprocessing

All plans, including CT images, contours, field doses, and fluence maps, were exported from the Eclipse TPS as DICOM files. As the original plans have different spatial resolutions, resampling was performed on dose and contour images with 1 mm axial resolution and 2 mm slice thickness. Linear interpolation was used to increase the resolution of dose distributions to facilitate more accurate dose prediction. Relative values were used in field doses with the prescription dose of 33 Gy normalized to 100%. Axial slices were cropped to a 192 × 192 pixel image centered at the isocenter. Fluence maps and other BEV projections had a resolution of 2.5 × 2.5 mm^2 at the isocenter plane. All the training data were randomly shuffled before holding out a validation set.

Field Dose Prediction

The objective of FD-CNN is to predict field doses from CT images and structure contours. The network architecture of FD-CNN is illustrated in **Figure 2**, and it operates on a slice-by-slice basis. To predict field dose in one query slice, the main input includes seven PTV slices (the query slice and six adjacent slices) and the OAR query slice, which are all 192 × 192 binary masks. The adjacent PTV slices were included to account for PTV shape change in the superior–inferior direction. In the downsampling block, the contour masks were downsampled three times using strided 2-D convolution to produce 128 channels of 24 × 24 feature images. An upsampling block produced 72-channel

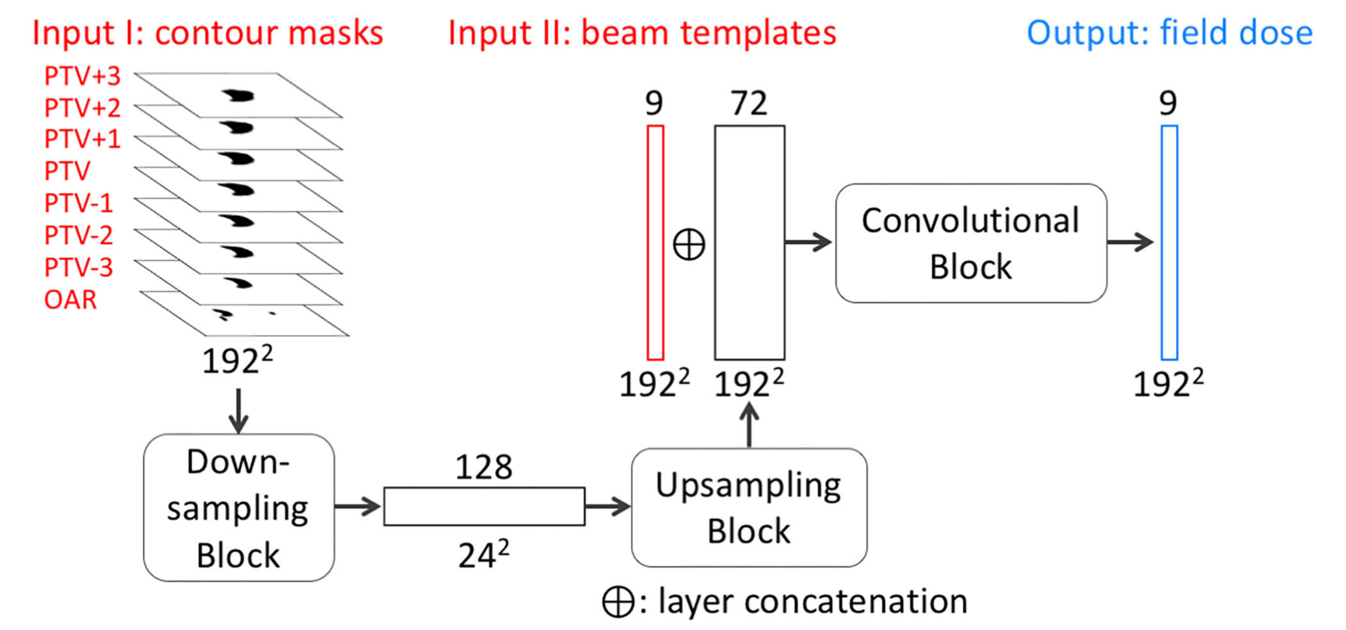


FIGURE 2 | Simplified network architecture of FD-CNN. FD-CNN takes contour masks and beam templates as input and predicts nine field doses in an axial slice. The details of downsampling, upsampling, and convolutional blocks (rounded rectangles) are omitted to highlight the transformation process from the inputs to the output. *PTV±n* refers to the n th PTV slice superior or inferior to the query slice. *OAR* includes only the OAR contour in the query slice. Each rectangle block represents a layer with the number of channels on the top and image dimensions labeled on the bottom of each layer.

feature images, using strided 2-D transposed convolution three times to restore the 192 × 192 resolution. CT images were incorporated in the form of beam templates (Input II in **Figure 2**) calculated by the TPS and concatenated to the 72-channel feature images. A final convolution block was applied to produce nine field doses for the nine equally spaced beams. The prediction region was limited to a region of interest (ROI), which was the PTV expanded by 1 cm. The Swish activation function (Ramachandran et al., 2017) was used in the network to introduce non-linearity. Swish is the product of an identity function and a sigmoid function, which can be expressed as

$$\boldsymbol{Swish}\,(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\boldsymbol{e^{-x}+1}}. \tag{1}$$

In predicting all field doses, the total dose was acquired automatically by summation. The loss function of FD-CNN ($\boldsymbol{L_{FD}}$) was the sum of two parts: field dose (FD) error and total dose (TD) error in the ROI, which is formulated as

$$\boldsymbol{L_{FD}} = \frac{1}{\boldsymbol{N\,(ROI)}} \left[\sum_{\boldsymbol{beam}}\sum_{\boldsymbol{ROI}}\left(\boldsymbol{FD_{bench}} - \boldsymbol{FD_{pred}}\right)^2 + \mu\cdot\sum_{\boldsymbol{ROI}}\left(\boldsymbol{TD_{bench}} - \boldsymbol{TD_{pred}}\right)^2\right] \tag{2}$$

$\boldsymbol{N\,(ROI)}$ is the number of ROI pixels. $\boldsymbol{FD_{bench}}$ and $\boldsymbol{TD_{bench}}$ are the benchmark plan field and total doses. $\boldsymbol{FD_{pred}}$ and $\boldsymbol{TD_{pred}}$ are the predicted field and total doses. The field and total dose error terms were summed with the regularization term of μ as tuned by validation. All slices with ROI were used to predict field dose by FD-CNN. For each patient, all the predicted 2-D dose slices were stacked together to form the predicted 3-D dose distributions of a given beam. In total, there were 3,238 slices from all 85 training cases. The benchmark plan's field dose is used as the ground truth for model training. Ten percent of the training slices were held out for validation. FD-CNN was trained using an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 and early stopping with patience of eight epochs (training terminates when validation loss does not improve for eight epochs).

Fluence Map Prediction

The second DL model is the FM-CNN, which predicts one fluence map from each 3-D field dose. The network architecture of FM-CNN is illustrated in **Figure 3**. It adopts a customized U-Net shape, which includes three resolution hierarchies (96, 48, and 24 pixels). The inputs of FM-CNN are the BEV dose map and the BEV PTV map, and the output is the fluence map. For one beam, the BEV dose map is the projection of the predicted field dose along the BEV, and the BEV PTV map is the binary projection of the PTV contour along the BEV. The upsampling and downsampling were achieved with strided 2-D convolution and strided 2-D transposed convolution, respectively. The BEV dose maps and fluence maps of the benchmark plans serve as ground truth for model training. The loss function of FM-CNN ($\boldsymbol{L_{FM}}$) is a modified mean absolute error (MAE), which is

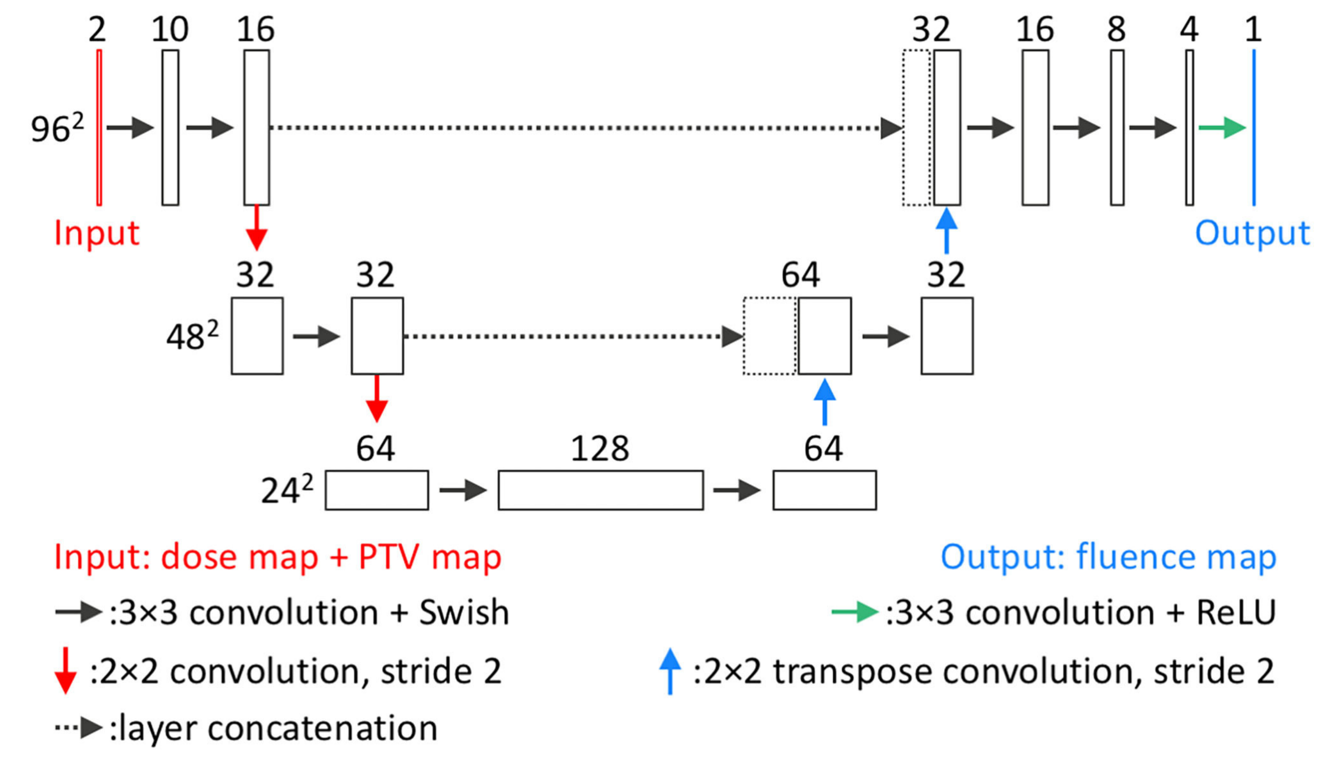

2
10
16
32
16
8
4
1
96^2
Input
Output
32
32
64
32
48^2
64
128
64
24^2
Input: dose map + PTV map
Output: fluence map
→ :3×3 convolution + Swish
→ :3×3 convolution + ReLU
↓ :2×2 convolution, stride 2
↑ :2×2 transpose convolution, stride 2
···▶ :layer concatenation

FIGURE 3 | Network architecture of FM-CNN. For each beam, FM-CNN predicts the fluence map from the dose map and PTV map (concatenated). Three hierarchies of image dimension (96, 48, 24 pixels) are used. Each rectangular block represents a layer with the number of channels on the top and image dimensions labeled on the left of each hierarchy.

formulated as

$$L_{FM} = (1+\lambda)\,\frac{\sum\left|y_{bench}-y_{pred}\right|}{N\left(y_{bench}>0\right)} \qquad (3)$$

where y_{bench} and y_{pred} are the benchmark and predicted values of the fluence map, and $N(y_{bench} > 0)$ is the count of benchmark fluence map pixels with non-zero values. The factor λ is the regularization term to prevent FM-CNN from over- or underestimating the fluence maps overall. It is expressed as

$$\lambda = \frac{\left|N\left(y_{bench}-y_{pred}>0.001\right) - N\left(y_{bench}-y_{pred}<-0.001\right)\right|}{N\left(y_{bench}>0\right)} \qquad (4)$$

Because fluence intensity is directly linked to field dose, the fluence prediction error should have a mean value close to zero in order to avoid overdosing and underdosing. Therefore, this regularization factor is added to control the mean value of prediction error for all pixels and keep the numbers of positive and negative errors at the same level.

The total training data size was 765 for 85 patients, of which 10% were held out for validation. The model was trained using an Adam optimizer with a learning rate of 0.001 and early stopping with patience of 15 epochs.

In the final validation step, these predicted fluence maps were subsequently imported into the TPS for leaf sequencing and dose calculation. The resulting plans are referred as model-predicted plans and are compared to the benchmark plans for overall performance.

Model Assessment

For model evaluation, the benchmark plan is considered as the ground truth. Each of the two models is evaluated separately and then collectively for dosimetric quality and deliverability. The FD-CNN field dose is compared with the corresponding field dose of the benchmark plan to evaluate FD-CNN performance. To evaluate FM-CNN performance, a special plan, the FM-CNN plan, is generated by FM-CNN using the field dose from the benchmark plan, thus eliminating error contamination from the first CNN model. The model-predicted plan is the final plan created with the fluence map predicted by the complete model (i.e., both CNNs) and, thus, evaluates the overall performance of the framework.

The 15 cases not included in model training were used as an independent test set, which consists of 638 slices and 135 fluence maps. For each test case, an FD-CNN field dose, an FM-CNN plan, and a model-predicted plan were created. The voxel-wise

TABLE 1 | Model training and calculation details.

	Trainable parameters	Training data size	Epochs	Training time	Calculation time per image	Calculation time per patient
FD-CNN	3,351,185	3,238	48	3 h	0.026 s	1.100 s
BEV projection	n/a	n/a	n/a	n/a	0.663 s	5.966 s
FM-CNN	203,621	765	134	4 min	0.003 s	0.030 s

The training details of two CNNs include the number of trainable parameters, training epochs, and training time. BEV projection is a deterministic process that requires no training. The calculation times listed are average prediction time of CNNs and average calculation times of BEV projection.

TABLE 2 | Dose differences between all predicted plan groups and benchmark plans.

Plan type	Dose type	Region	Voxel dose difference [%]	D_{mean} difference [%]	D_{max} difference [%]
FD-CNN dose	Total dose (CNN)	ROI	1.79 ± 2.21	0.41 ± 0.28	0.48 ± 0.31
		PTV	0.91 ± 0.79	0.57 ± 0.25	0.48 ± 0.31
		ROI–PTV	2.65 ± 2.75	0.51 ± 0.34	0.49 ± 0.54
	Field dose (CNN)	ROI	1.25 ± 1.11	0.51 ± 0.42	1.82 ± 1.44
FM-CNN plan	Total dose (TPS)	PTV	1.22 ± 0.96	0.88 ± 0.65	1.46 ± 1.19
		OAR	0.86 ± 0.75	0.30 ± 0.17	0.86 ± 0.52
Model-predicted plan	Total dose (TPS)	PTV	2.41 ± 1.87	1.24 ± 0.74	4.10 ± 2.35
		OAR	2.70 ± 2.45	0.94 ± 0.65	4.77 ± 2.84

Model-predicted plans exhibit larger dose differences than FD-CNN doses and FM-CNN plans.

percentage dose difference ΔD is calculated as

$$\Delta D(V) = \frac{1}{N(V)} \sum_{i \in V} \left| \frac{D_{bench}^{(i)} - D_{pred}^{(i)}}{D_{prescription}} \right| \times 100 \qquad (5)$$

V is the calculation volume, and $N(V)$ is the number of voxels in this volume. Several dosimetric endpoints were also used for assessment. These include PTV max dose (0.1 cc), mean dose, $D_{95\%}$ for the PTV, and mean and max doses (0.1 cc) for the OARs. To provide a direct assessment of fluence map prediction, MAEs were calculated between FM-CNN and benchmark fluence maps.

In Eclipse TPS, optimal fluence maps, generated by inverse optimization or deep learning models, are converted to actual fluence maps by leaf-sequencing algorithms to enable delivery on the machine. Unrealistic optimal fluence map features, such as extremely heterogeneous regions or high transmission value at a single pixel, could potentially result in a large discrepancy between optimized and delivered doses. Therefore, the deliverability of fluence maps was measured by the gamma index between optimal (before leaf sequencing) and actual fluence maps (after leaf sequencing) for both benchmark and predicted plans. We employed the gamma analysis in a similar fashion and intent as IMRT quality assurance. Here, a high gamma passing rate indicates that the optimal fluence map is physically realistic and could be achieved by the leaf-sequencing algorithm. Gamma analysis was performed using an in-house program with a 3%/3 mm criterion. Total monitor units (MUs) from benchmark and model-predicted plans were compared.

After the DL framework was completely trained and tested, we reduced the training cases for both CNNs and calculated the loss values on the test set. In addition, a series of ablation studies were conducted, in which certain CNN components were removed to test the model performance. For the FD-CNN model, we removed the input of one, two, or three pairs of adjacent PTV slices or beam templates. For the FM-CNN model, we removed the input of the PTV map. The reduced models were evaluated on the same test set and compared with the original models.

RESULTS

Model Training

The model training details are summarized in **Table 1**. FD-CNN has 3.35 million trainable parameters and took 3 h to train. FM-CNN has a much less complex architecture with 0.20 million trainable parameters and took 4 min to train. The projection of field dose and PTV along the BEV is relatively time-consuming compared to CNN predictions. On average, prediction of nine fluence maps for each patient took 7.1 s, including 1.10 s for FD-CNN prediction, 5.97 s for BEV projection, and 0.03 s for FM-CNN prediction. In the entire workflow, the computation time of the model is typically less than that of TPS dose calculation. A DL model-predicted plan was generated within 1 to 2 min, including calculating the model-predicted plan dose in TPS, as compared to the traditional manual planning, which takes between 1 and 3 h.

Model Assessment

The dosimetric evaluation results are summarized in **Table 2**. Here, the ground truth is the dose from the benchmark plans. Model-predicted plans have the largest dose differences among the three evaluation plans, and they represent the overall performance of the workflow. In the deliverable plans, i.e., FM-CNN and model-predicted plans, PTV and OAR (stomach, C-loop/duodenum, and bowels combined) maximum dose errors are larger than mean dose errors. **Figure 4** compares the total dose distribution between the model-predicted and benchmark plans of an example case. **Figure 5** compares the DVH for the same case. As shown in the figure, the predicted fluence map

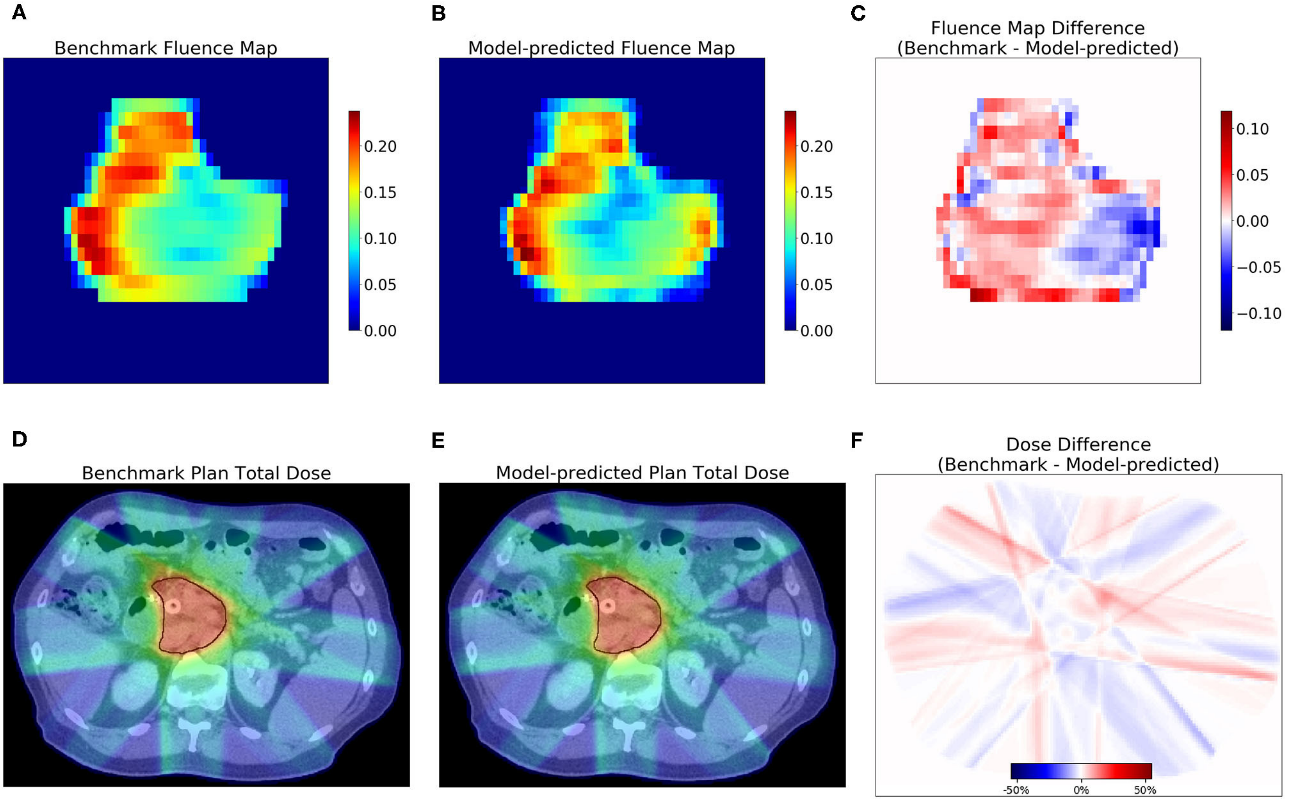

A
Benchmark Fluence Map
0.20
0.15
0.10
0.05
0.00
B
Model-predicted Fluence Map
0.20
0.15
0.10
0.05
0.00
C
Fluence Map Difference
(Benchmark - Model-predicted)
0.10
0.05
0.00
−0.05
−0.10
D
Benchmark Plan Total Dose
E
Model-predicted Plan Total Dose
F
Dose Difference
(Benchmark - Model-predicted)
-50%
0%
50%

FIGURE 4 | Examples of fluence map and dose comparisons with the benchmark in one test case. The model-predicted fluence map recreated the fluence contrast in the benchmark. The model-predicted plan achieved a similar total dose as the benchmark. The first row shows the benchmark **(A)** and model-predicted **(B)** fluence maps of one beam, and the difference **(C)**. The second row shows one axial slice of the dose distribution of benchmark plan **(D)** and model-predicted plan **(E)** and the dose difference **(F)**. PTV contour is marked with black lines in **(D,E)**.

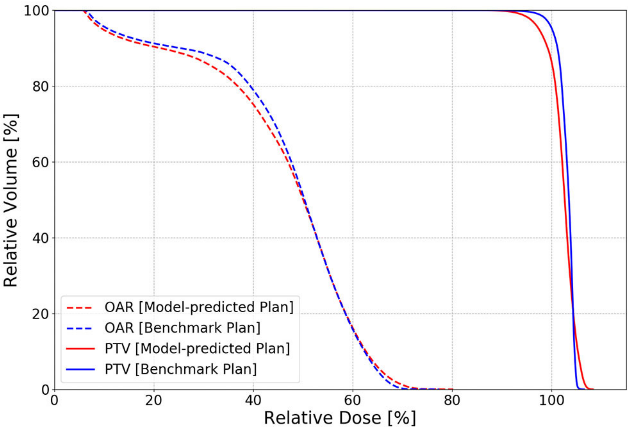


FIGURE 5 | An example of PTV (solid) and OAR (dashed) DVH comparison in one test case between the benchmark (blue) and model-predicted (red) plans. The benchmark plan has slightly better PTV homogeneity than the model-predicted plan with the FM-CNN plan in between.

achieves similar fluence modulation as the fluence map of the benchmark plan. Further, the TPS-calculated dose distribution of the predicted plan exhibits small differences from the corresponding benchmark plan, indicating highly similar plan quality. The distributions of PTV and OAR dose metrics of benchmark and model-predicted plans are plotted in **Figure 6**.

In terms of fluence map deliverability, the average ± standard deviation gamma passing rate was 98.14% ± 0.74% for benchmark plans and 97.69% ± 0.96% for model-predicted plans, respectively, which demonstrates highly similar deliverability. The average ± standard deviation of total MU per patient is 2,122 ± 281 in model-predicted plans and 2,265 ± 373 in benchmark plans.

The model performance of FD-CNN and FM-CNN were plotted against the number of training cases used, as shown in **Figure 7**. It can be seen that the testing loss of FD-CNN plateaued after 55 cases although FM-CNN required only 35 cases to achieve reasonably good performance. The ablation study showed that, for FD-CNN, removing the beam template input would increase the testing loss by 20%; using only four, two, and zero adjacent PTV slices would increase the testing loss by 7, 25, and 66%, respectively. For FM-CNN, removing the PTV map input would only slightly increase the testing loss by 1%.

DISCUSSION

We develop a novel deep learning framework to generate clinical-quality pancreas SBRT plans in seconds. It offers the advantage

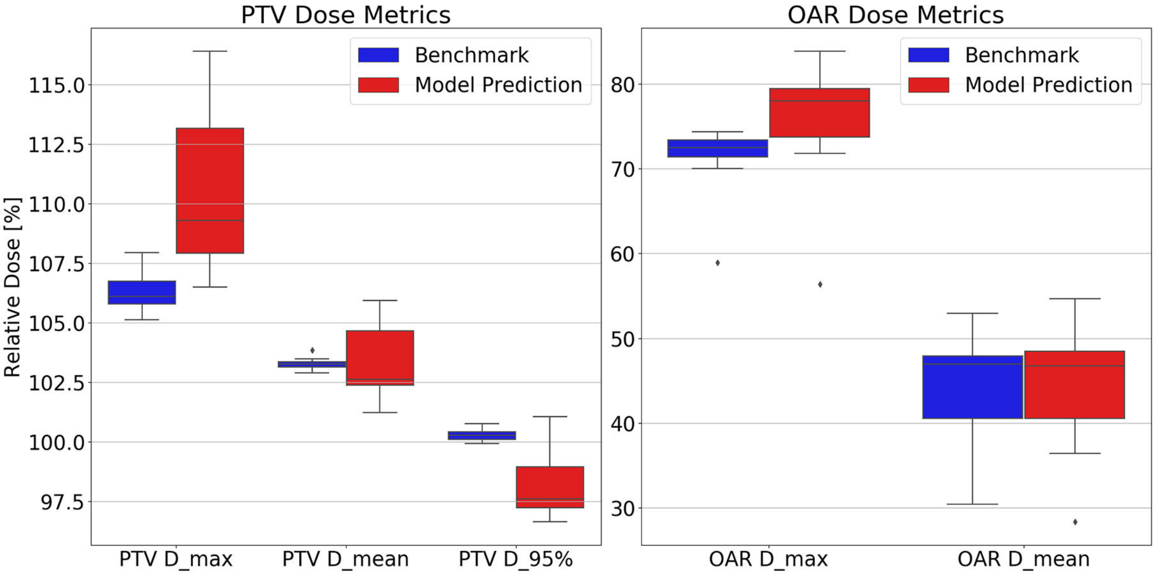


FIGURE 6 | Test set distributions of PTV **(Left)** and OAR **(Right)** dose metrics comparing benchmark and model-predicted plans. Model-predicted plans have higher PTV and OAR maximum dose and lower $D_{95\%}$ than the other plan groups. Dose values are reported as percentage of the prescription dose. D_{max}, maximum dose; D_{mean}, mean dose; $D_{95\%}$, minimum dose received by 95% of the volume.

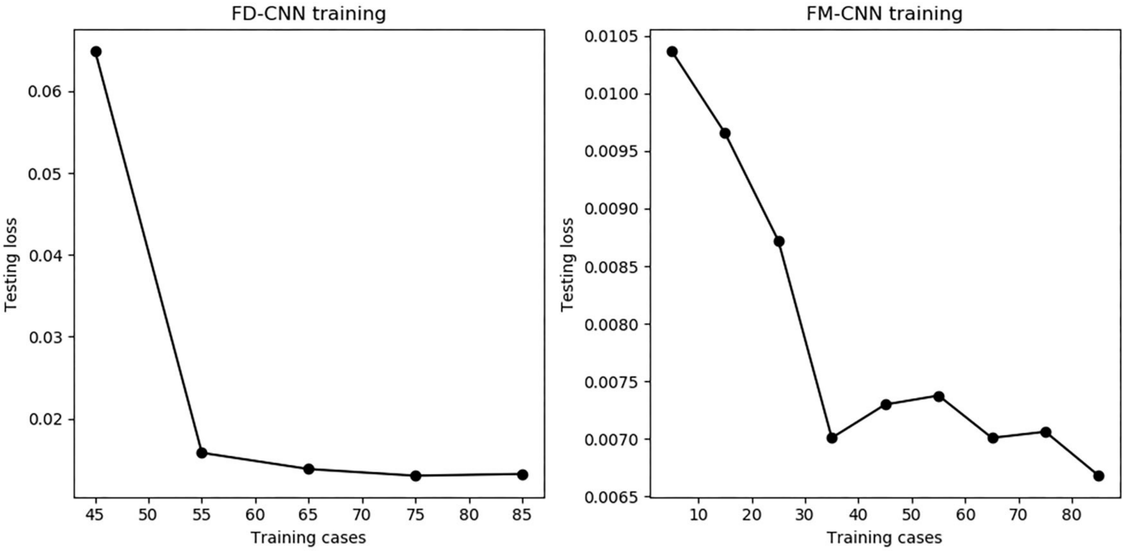


FIGURE 7 | The number of training cases vs. testing loss for both CNNs. The testing loss stabilized when using 55 or more training cases for FD-CNN **(Left)** and 35 or more cases for FM-CNN **(Right)**.

of bypassing lengthy optimization, during which the planner needs to adjust optimization objectives and aims to achieve similar performance as the human expert exercising inverse optimization. This study demonstrates the novel approach of AI-driven treatment planning via predicting fluence maps, thus providing a more complete approach to generating deliverable high-quality plans, which has not been sufficiently addressed in previous studies (Liu et al., 2015; Skarpman Munter and Sjolund, 2015; Kearney et al., 2018; Barragán-Montero et al., 2019; Chen et al., 2019; Nguyen et al., 2019a,b). Translating predictions from previous KBP models, either DVH-based or 3-D dose guidance, to the final deliverable plan has been challenging and remains a key implementation bottleneck in clinics. Efforts have been made to complement KBP models to arrive at the final plan (McIntosh et al., 2017; Long et al., 2018; Fan et al., 2019). The aim of the proposed DL solution is to garner knowledge from existing plans and generate deliverable plans for new patients, which falls under the broad KBP vision. Our approach directly predicts fluence maps rather than predicting achievable DVH/doses in other KBP approaches. More specifically, we use CNNs to establish the correlation between patient anatomy patterns and each individual beam's dose/fluence map, which has not been investigated in previous KBP studies. This approach is built upon beamlet-based fluence optimization, with which a subsequent leaf-sequencing process converts the fluence maps to MLC motion parameters. By replacing the FM-CNN, the proposed approach could also be employed along with direct aperture optimization (Shepard et al., 2002) in step-and-shoot IMRT, which would offer the advantage of fewer segments and MUs. This is an area of potential study that warrants future effort. It would also be of interest to compare the proposed approach with other KBP-based plan-generation methods in future studies.

We redesigned the entire radiation therapy treatment-planning workflow by incorporating DL models for dose prediction and fluence map generation, thereby completing AI-driven plan generation in seconds. Our results demonstrate that such AI-driven plans have similar quality when compared to manually generated inversely optimized plans although, more importantly, a ready-to-deliver plan is generated with no further human intervention needed. In dose prediction, the total dose in PTV predicted by FD-CNN achieved a similar level of accuracy as existing deep learning–based dose-prediction models. The input of adjacent PTV slices provided superior–inferior contour change information efficiently, which significantly reduced the testing loss while maintaining lower memory consumption than 3-D networks. The second step of our framework, i.e., fluence map prediction from an existing field dose, directly converts an individual field dose into its corresponding fluence map, which eliminates interplay among beams and require no optimization or intermediate dose calculation. With the existing ground truth field dose, we achieved similar fluence map MAE (mean value: 2.06×10^{-3}) as Lee et al. (2019) (median value: 9.95×10^{-4}). With similar fluence map prediction accuracy, our proposed framework is capable of directly predicting a fluence map from contour and CT alone, which Lee et al. (2019) has yet to achieve.

We used standardized nine-beam IMRT plans as a benchmark in this study, and this increased consistency in plan quality and reduced the need for a large amount of training data. We argue that the training data meticulously generated by human experts is optimal in terms of the endpoints of target coverage and luminal structure maximum dose. One limitation of the model is that the training and testing cases must have the same beam arrangement, dose prescription level, and physician preferences. Substantially more training data are anticipated to be required to train a model that incorporates different beam arrangements and dose constraints. This study focuses on pancreas SBRT although we are modifying and testing the model for other disease sites. With the current model, we do not think it is generically applicable to other disease sites. We anticipate that data from each specific site are required to train a robust model. Further study is underway to address these challenges.

CONCLUSION

We develop a deep learning framework utilizing two CNNs to directly generate a clinical-quality IMRT plan from CT images and contours for pancreas SBRT. This framework changes the traditional approach of inverse treatment planning by replacing the inverse optimization engine with the intelligent neural networks. The proposed method has great potential to improve clinical efficiency and plan quality consistency for challenging treatment sites.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Duke University Health System Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

WW and YS performed model development and plan comparison. CW and JZ performed patient selection and benchmark plan generation. MP, BC, and CW are GI physicians from whom we learned expert domain knowledge. QW, F-FY, and YG reviewed experiment design, paper content, and statistical analysis. QJW supervised the entire study and revised this paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partially supported by an NIH grant (#R01CA201212).

REFERENCES

Barragán-Montero, A. M., Nguyen, D., Lu, W., Lin, M., Norouzi-Kandalan, R., Geets, X., et al. (2019). Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. *Med. Phys.* 46, 3679–3691. doi: 10.1002/mp.13597

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Campbell, W. G., Miften, M., Olsen, L., Stumpf, P., Schefter, T., Goodman, K. A., et al. (2017). Neural network dose models for knowledge-based planning in pancreatic SBRT. *Med. Phys.* 44, 6148–6158. doi: 10.1002/mp.12621

Chen, X., Men, K., Li, Y., Yi, J., and Dai, J. (2019). A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Med. Phys.* 46, 56–64. doi: 10.1002/mp.13262

Fan, J., Wang, J., Chen, Z., Hu, C., Zhang, Z., and Hu, W. (2019). Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Med. Phys.* 46, 370–381. doi: 10.1002/mp.13271

Good, D., Lo, J., Lee, W. R., Wu, Q. J., Yin, F. F., and Das, S. K. (2013). A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning. *Int. J. Radiat. Oncol. Biol. Phys.* 87, 176–181. doi: 10.1016/j.ijrobp.2013.03.015

Kearney, V., Chan, J. W., Haaf, S., Descovich, M., and Solberg, T. D. (2018). DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Phys. Med. Biol.* 63:235022. doi: 10.1088/1361-6560/aaef74

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:*1412.6980.

Lee, H., Kim, H., Kwak, J., Kim, Y. S., Lee, S. W., Cho, S., et al. (2019). Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep-neural-network. *Sci. Rep.* 9:15671. doi: 10.1038/s41598-019-52262-x

Liu, J., Wu, Q. J., Kirkpatrick, J. P., Yin, F. F., Yuan, L., and Ge, Y. (2015). From active shape model to active optical flow model: a shape-based approach to predicting voxel-level dose distributions in spine SBRT. *Phys. Med. Biol.* 60, N83–N92. doi: 10.1088/0031-9155/60/5/N83

Long, T., Chen, M., Jiang, S., Lu, W. (2018). Threshold-driven optimization for reference-based auto-planning. *Phys. Med. Biol.* 63:04NT1. doi: 10.1088/1361-6560/aaa731

McIntosh, C., and Purdie, T. G. (2016). Contextual atlas regression forests: multiple-atlas-based automated dose prediction in radiation therapy. *IEEE Trans. Med. Imaging* 35, 1000–1012. doi: 10.1109/TMI.2015.2505188

McIntosh, C., and Purdie, T. G. (2017). Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys. Med. Biol.* 62, 415–431. doi: 10.1088/1361-6560/62/2/415

McIntosh, C., Welch, M., McNiven, A., Jaffray, D. A., and Purdie, T. G. (2017). Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys. Med. Biol.* 62, 5926–5944. doi: 10.1088/1361-6560/aa71f8

Nguyen, D., Jia, X., Sher, D., Lin, M. H., Iqbal, Z., Liu, H., et al. (2019a). 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Phys. Med. Biol*. 64:065020. doi: 10.1088/1361-6560/ab039b

Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., et al. (2019b). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci. Rep.* 9:1076. doi: 10.1038/s41598-018-37741-x

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv [Preprint]. arXiv:*1710.05941.

Ronneberger, O., Fischer, P., and Brox, T. (eds.). (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer). doi: 10.1007/978-3-319-24574-4_28

Sheng, Y., Li, T., Yoo, S., Yin, F. F., Blitzblau, R., Horton, J. K., et al. (2019). Automatic planning of whole breast radiation therapy using machine learning models. *Front. Oncol.* 9:750. doi: 10.3389/fonc.2019.00750

Sheng, Y., Li, T., Zhang, Y., Lee, W. R., Yin, F. F., Ge, Y., et al. (2015). Atlas-guided prostate intensity modulated radiation therapy (IMRT) planning. *Phys. Med. Biol.* 60, 7277–7291. doi: 10.1088/0031-9155/60/18/7277

Shepard, D. M., Earl, M. A., Li, X. A., Naqvi, S., Yu, C. (2002). Direct aperture optimization: a turnkey solution for step-and-shoot IMRT. *Med. Phys*. 29, 1007–18. doi: 10.1118/1.1477415

Shiraishi, S., and Moore, K. L. (2016). Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med. Phys.* 43:378. doi: 10.1118/1.4938583

Skarpman Munter, J., and Sjolund, J. (2015). Dose-volume histogram prediction using density estimation. *Phys. Med. Biol.* 60, 6923–6936. doi: 10.1088/0031-9155/60/17/6923

Yuan, L., Ge, Y., Lee, W. R., Yin, F. F., Kirkpatrick, J. P., and Wu, Q. J. (2012). Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med. Phys.* 39, 6868–6878. doi: 10.1118/1.4757927

Zhu, X., Ge, Y., Li, T., Thongphiew, D., Yin, F. F., and Wu, Q. J. (2011). A. planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med. Phys*. 38, 719–726. doi: 10.1118/1.3539749

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.

frontiers
in Artificial Intelligence

REVIEW
published: 29 September 2020
doi: 10.3389/frai.2020.577620

Integration of AI and Machine Learning in Radiotherapy QA

Maria F. Chan[1]*, *Alon Witztum*[2] *and Gilmer Valdes*[2]

[1] *Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, United States,* [2] *Department of Radiation Oncology, University of California, San Francisco, San Francisco, CA, United States*

The use of machine learning and other sophisticated models to aid in prediction and decision making has become widely popular across a breadth of disciplines. Within the greater diagnostic radiology, radiation oncology, and medical physics communities promising work is being performed in tissue classification and cancer staging, outcome prediction, automated segmentation, treatment planning, and quality assurance as well as other areas. In this article, machine learning approaches are explored, highlighting specific applications in machine and patient-specific quality assurance (QA). Machine learning can analyze multiple elements of a delivery system on its performance over time including the multileaf collimator (MLC), imaging system, mechanical and dosimetric parameters. Virtual Intensity-Modulated Radiation Therapy (IMRT) QA can predict passing rates using different measurement techniques, different treatment planning systems, and different treatment delivery machines across multiple institutions. Prediction of QA passing rates and other metrics can have profound implications on the current IMRT process. Here we cover general concepts of machine learning in dosimetry and various methods used in virtual IMRT QA, as well as their clinical applications.

Keywords: artificial intelligence, machine learning, radiotherapy, quality assurance, IMRT, VMAT

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Li Zhou,
Sichuan University, China
Ying Zhang,
Medical College of Wisconsin,
United States

****Correspondence:***
Maria F. Chan
chanm@mskcc.org

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *29 June 2020*
Accepted: *24 August 2020*
Published: *29 September 2020*

Citation:
Chan MF, Witztum A and Valdes G (2020) Integration of AI and Machine Learning in Radiotherapy QA. Front. Artif. Intell. 3:577620. doi: 10.3389/frai.2020.577620

INTRODUCTION

Machine learning (ML) has the potential to revolutionize the field of radiation oncology in many processes and workflows to improve the quality and efficiency of patient care (Feng et al., 2018). The delivery of radiotherapy is complex and each step in the integrated process requires quality assurance (QA) to prevent errors and to ensure patients receive the prescribed treatment correctly. The recent research in machine learning efforts in the QA has produced a variety of proofs-of-concept, many with promising results (Kalet et al., 2020). In this article, we review the machine learning applications in radiotherapy QA.

The first question we seek to answer is why we want to integrate ML in radiotherapy QA. The term, machine learning, refers to the automated detection of meaningful patterns in data. In the past few years, it has become a major area of research and a common tool in many processes in radiotherapy (Feng et al., 2018). In this review paper, we will focus on machine learning applications to QA. As medical physicists, we perform an increasing number of QA tasks in our daily work, and prioritizing those that will help deliver the safest treatment is of paramount importance as stated in the American Association of Physicists in Medicine (AAPM) Task Group (TG) 100 (Huq et al., 2016). As such, learning from our QA data to choose those tasks that need early intervention is essential for our profession as more complex treatments are adopted. Currently, most of the data acquired during QA is utilized only as a one-time evaluation measurement but there is a lot of QA data available from which we can "learn" using machine learning methods and utilize past experience as knowledge.

This review will begin by introducing some general machine learning concepts for those who are not as familiar with this field. We will then combine these descriptions with explanations of their direct applications to QA. We also provide a non-exhaustive analysis of the literature on the applications of ML to QA data. This article hopes to demonstrate the power of machine learning and the advantages it offers to our QA programs.

Artificial Intelligence and Machine Learning

Machine Learning maybe somewhat misleadingly referred as Artificial intelligence (AI), is already part of our everyday lives. The easiest way to explain the relationship of AI and ML is to visualize them as concentric circles with AI - the idea that came first, the largest; then ML - which blossomed later. In AI a general purpose algorithm that can reason about different problems is sought while in ML this idea is abandoned to search for a specific model that maps an input to an output using statistical learning techniques. Many classes of algorithms exist within ML that fit different functions, such as linear models like Lasso and Ridge regression (Hastie et al., 2009), Decision Trees (Luna et al., 2019); ensembles like Random Forrest and Gradient Boosting (Hastie et al., 2009), and Neural Networks (Rumelhart et al., 1986). All these algorithms are needed because one cannot guarantee a priori that an algorithm will be better than another in a random problem, a theorem knows as a no free lunch theorem (Wolpert, 1996). In practice, certain classes of algorithms work better than others in classes of problems. For instance, for the analysis of images, Convolutional Neural Network (CNN), a deep learning network, excels while for the analysis of tabular data Gradient Boosting has the lead. In the majority of problems. CNN uses convolution filters to extract general concepts that are later combined with other concepts that resemble how the visual cortex in animals works and puts emphasis in the local importance of each pixel (Le Cun and Bengio, 2002). Additionally, max pooling layers that take average of pixel and data augmentation techniques make it somewhat independent of translation and rotations of the images, all important part of their success. For the analysis of tabular data, however, this customization is not needed and an algorithm that is better at handling missing values, performs automatic feature selection, does not depend on monotonic transformations of the input variables and it is easy to train and regularize is more important. This is the case for Gradient Boosting (Friedman, 2001).

Types of Learning

Machine learning algorithms use computational methods to "learn" information directly from data. There are two main types of learning: unsupervised learning and supervised learning. In unsupervised learning, the training data does not include label responses or desired outputs and the objective is to model the probability distribution of the given inputs. On the other hand, in supervised learning the training data does include labels or desired outputs and allows for the learning of a mapping between the input variables and the output (e.g., classification, regression, etc.).

Unsupervised Learning

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity (MathWorks.com, 2020). Li et al. utilized unsupervised learning tools of K-means and hierarchical clustering algorithms to analyze patients' breathing curves extracted from 4D radiotherapy data (Li et al., 2017). The authors classified patients' breathing patterns into sub-groups, such as perfect, regular, and irregular breathers. The breathing signals and frequency spectrum were extracted from 341 real-time position management (RPM) datasets. Correlation plots of 6 features (frequency, amplitude, standard deviation of amplitude, spread of frequency spectrum) were chosen for the clustering task. Two clustering algorithms were used by the authors: hierarchical clustering and k-means. Hierarchical clustering generates more consistent results than k-means but requires a more (and usually prohibiting) training time than k-means (Li et al., 2017). This could lead to inefficiency in large datasets. K-means is extremely sensitive to cluster center initialization; therefore, some degrees of prior knowledge about the data is required for its effective usage. We will also demonstrate that the same RPM data could be used for both unsupervised and supervised learning to achieve different goals, although this topic might not be directly related to radiotherapy QA.

Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs (Russell and Norvig, 2010). A supervised learning algorithm takes a known set of input data and responses (output) to learn the regression/classification model. A learning algorithm is then used to train a model and generate a prediction for the response to new data or the test dataset. When statistical learning algorithms are used (e.g., Random Forest, Gradient Boosting, Decision Trees) features that are expected to describe the output need to be defined and calculated (Shobha and Rangaswamy, 2018). Therefore, for each observation features are extracted and associated with the label sought to be predicted. We can then use these features and output to learn a mapping from one to the other using ML algorithms. Thus, when a new IMRT plan is generated, the same features can be extracted to be used in the trained predictive model to show the expected label such as pass/fail (classification) or passing rate (regression). This is the approach first proposed in Virtual IMRT QA (Valdes et al., 2016) and further validated (Valdes et al., 2017).

Supervised learning was also used with the same RPM data described in section Unsupervised Learning above (Lin et al., 2019). With over 1,700 RPM data from 3 institutions, a Long short-term memory (LSTM) model was built by Lin et al. to predict different types of patients' respiratory motions in real-time (Lin et al., 2019). LSTM is a recurrent neural network (RNN) recently designed to alleviate the issues with vanishing gradients seen in earlier RNN. LSTM is specifically useful for the analysis of sequence data like text or this RPM data (Lin et al., 2019). In this study, the authors used a sliding window technique to partition the RPM data into the input and supervised output. This study demonstrates the potential of using deep learning models in respiratory signal prediction and incorporating the motion into treatments. This example, though slightly removed from radiotherapy QA, is chosen to emphasize the fact that applying different learning algorithms on the same dataset could serve different purposes.

Semi-supervised Learning

Semi-supervised learning falls between unsupervised learning and supervised learning. In semi-supervised learning, part of the training data does not contain a label. However, by leveraging the correlation structure of the input variables, a model that explains the label portion is obtained. Naqa et al. performed a multi-institutional study with data from eight Linacs and seven institutions (El Naqa et al., 2019). The authors investigated the use of machine learning methods for the automation of machine QA. A total of 119 EPID (electronic portal imaging device) images of a special QA phantom were fed into the support vector data description (SVDD) clustering algorithm (unsupervised learning). QA test data was first mapped to a higher dimensional space to identify the minimal enclosing sphere. This sphere was then mapped back to the input space to detect outliers. The separate clusters generated were further used to evaluate the tolerance boundaries and limits as indicated in the AAPM TG-142. The prediction tests included gantry sag, radiation field shift, and multileaf collimator (MLC) offset data. This study demonstrated that machine learning methods with SVDD clustering are promising for developing automated QA tools.

Validation of ML Models

In machine learning, model validation is referred to as the process where a trained model is evaluated with a testing data set. The test data set should be a separate portion of the same data set from which the training set is derived. The main purpose of using the testing data set is to validate the generalizability of a trained model (Alpaydin, 2010). Validation of a predictive model is an essential part of the model building process, and is used to assess the quality of a model. When conducting a machine learning study, commonly used validation methods include: (1) using different machine learning algorithms on the same data to compare the results, (2) using cross-validation to obtain an error estimation on out of sample data, (3) using a hold-out sample for testing, (4) comparing with other well-established models that are not necessary machine-learning models, (5) validating using a sample not from the training period but acquired at a later time, (6) validating using a sample that is selected from a different population than that used to build the model (e.g., different clinic). Model validation is usually carried out after model training to find the optimal model with the best performance. The two most popular types of validation methods used in predictive models of radiotherapy QA are splitting training/test/holdout datasets and k-fold cross-validation. There are multiple ways to split the data. One method is to split the data pool into roughly 70% used for training the model and 30% for testing the model, and another method splits the data into three with, for example, 60% for training, 30% for testing, and the remaining 10% for holdout. Validating on the holdout set is done to check if the model suffers from overfitting due to optimization of the model hyperparameters. Instead of the data splitting as described above, k-fold cross-validation splits the data into k folds, then trains the data on k-1 folds and tests on the remaining fold to evaluate the model (Alpaydin, 2010; Russell and Norvig, 2010). The procedure is repeated k times with a different group of observations treated as a validation set each time. The most frequently used in radiotherapy QA applications is either 5- or 10-fold cross-validation. The model accuracy can be evaluated using a variety of metrics including, but not limited to, the mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), receiver operating characteristic (ROC), correlation coefficient, regression plot, residual error histogram, sensitivity and specificity.

MACHINE LEARNING APPLICATIONS IN MACHINE QA

In this section, we will focus on the general applications of ML to Linac QA before discussing IMRT QA. There have been many studies of machine learning applications in Linac QA including prediction of machine dosimetry as well as discrepancies of MLC positioning and their impact on the actual dose delivery.

ML Model Built From Dosimetric QA or Beam Data

Another application, Li and Chan developed a model to predict the performance of Linac over time (Li and Chan, 2017). The study applied Artificial Neural Networks (ANNs) time-series prediction modeling to the longitudinal data of 5-years of daily Linac QA. A set of one hidden layer, six hidden neurons, and two input delays were chosen after a trial-and-error process to form the network architecture. The predictive model was compared with a well-established model, autoregressive integrated moving average (ARIMA). The ANN time-series mode was found to be more accurate than the ARIMA techniques to predict the Linac beam symmetry accurately (Li and Chan, 2017). Zhao et al. (in press) utilized 43 sets of commissioning and annual QA beam data from water tank measurements to build a machine learning model that could predict the percent depth doses (PDD) and profiles of other field sizes such as 4×4 cm^2, 30×30 cm^2 accurately within 1% accuracy with 10×10 cm^2 data input. This application would potentially streamline the data acquisition for the entire commissioning process in TPS as well as optimize periodic QA of Linacs to a minimum set of measurements.

ML Model Built From Delivery Log Files

Carlson et al. were the first to use machine learning techniques to train models to predict these discrepancies (Carlson et al., 2016). Predictive leaf motion parameters such as leaf position and speed were calculated for the models. Differences in positions between synchronized DICOM-RT files and Dynalog files from 74 VMAT plans were used as a target response for training the models. Three machine learning algorithms were used—linear regression, random forest, and a cubist model. They found that the cubist model outperformed all other models in terms of accuracy to predict MLC position errors. The objective of these predictions was to incorporate them into the TPS and provide clinicians with a more realistic view of the dose distribution as it will truly be delivered to the patient. Osman et al. (2020) collected 400 delivery log files and trained a model with feed-forward ANN architecture mapping the input parameters with the output to predict the MLC leaf positional deviations with a train/test split of 70 and 30%. The ANN model achieved a maximum MSE of 0.0001 mm^2 in predicting the leaf positions for each leaf in the test data. The results of the study could be extended to utilizing this information in the dose calculation/optimization algorithm. Chuang et al. developed a machine learning model using prior trajectory log files generated from 116 IMRT and 125 VMAT plans to predict the MLC discrepancies during delivery and provide feedback of dosimetry (Chuang et al., in press). A workflow was developed to extract discrepancies and mechanical parameters from trajectory logs and use the proposed machine learning algorithm to predict discrepancy. The authors used multiple machine learning models including linear regression, decision tree, and ensemble methods.

ML Model Built From Proton Fields

Sun et al. used 1,754 proton fields with various range and modulation width combinations to train an output factor (OF) model in three different algorithms (Random Forest, XGBoost, and Cubist) with a train/test split of 81 and 19% (Sun et al., 2018). They found that the Cubist—based solution outperformed all other models with a mean absolute discrepancy of 0.62% and maximum discrepancy of 3.17% between the measured and predicted OF. They concluded that machine learning methods can be used for a sanity check of output measurements and has the potential to eliminate time-consuming patient-specific measurements. Similarly, Grewal et al. utilized 4,231 QA measurements with a train/test split of 90 and 10% to build models to predict OF and MU for uniform scanning proton beams with two learning algorithms—Gaussian process regression and shallow neural network (Grewal et al., 2020). They found that the prediction accuracy of machine and deep learning algorithms is higher than the empirical model currently used in the clinic. They have used these models in the clinic as a secondary check of MU or OF.

Table 1 lists the studies on radiotherapy machine QA using machine learning techniques. All of these studies showed that machine learning techniques can give physicists insights into past QA data and to predict potential machine failures. This would alert physicists to take proactive actions and make informed decisions.

MACHINE LEARNING APPLICATIONS IN IMRT/VMAT QA

This section will now focus on describing the applications of Machine Learning to IMRT QA. Features can be extracted from each IMRT plan and compute multiple complexity metrics associated with passing rates. These features can be used to build a model that can predict the passing rate for any new IMRT plan.

ML Applied to IMRT QA

Early ML Models

Valdes et al. developed the first virtual IMRT QA using a Poisson regression machine learning model to predict passing rates (Valdes et al., 2016). The initial dataset contained 498 clinical IMRT plans from the University of Pennsylvania, with QA results from a MapCHECK (Sun Nuclear Corporation, Melbourne, FL) QA device. An additional dataset was obtained containing 203 clinical IMRT beams also planned from Eclipse (Varian Medical Systems, Palo Alto, CA) but QA results were obtained using portal dosimetry. The plans from the University of Pennsylvania were used to identify 78 important features. Additionally, 10 further features were added to take into account the specific characteristics of portal dosimetry (Valdes et al., 2017). All parameters of each IMRT beam were automatically extracted from Eclipse with SQL queries and scripts were written to read the MLC positions and collimation rotation from the files. Matlab (The MathWorks Inc., Natick, MA) functions were developed to calculate the features for each beam. For MapCHECK, the important features extracted included the fraction of area delivered outside a circle with a 20 cm radius (to capture symmetry disagreements), duty cycle, the fraction of opposed MLCs with an aperture smaller than 5 mm (to quantify the effects of rounded leaves in the MLC), etc. For portal dosimetry, the important features included the CIAO (Complete Irradiated Area Outline) area, the fraction of MLC leaves with gaps smaller than 20 or 5 mm, the fraction of area receiving $<50\%$ of the total calibrated MUs, etc. (Valdes et al., 2017).

A machine learning algorithm was trained to learn the relationship between the plan characteristics and the passing rates. There are 80 complexity metrics being used in the calculation in the initial modeling with Penn data using the MapCHECK QA data. A learning curve for the initial model was established to show that around 200 composite plans are needed to adequately train the model. A strong correlation between the MapCHECK measurement and virtual IMRT predicted passing rates for data that the algorithm had not seen was obtained. All predictions of passing rates were within ±3% error.

For the portal dosimetry model, a learning curve was also performed to estimate the number of IMRT fields needed, and it was shown that close to 100 individual IMRT fields are sufficient to build a reliable predictive model. In total there were 90 continuous variables used for the virtual IMRT QA model which predicted EPID panel passing rates. The authors presented the residual errors of the passing rates prediction for the two institutions (the University of Pennsylvania and Memorial Sloan Kettering Cancer Center). Although the passing rates are site-dependent, different models were not built for each site because, conditional on the plan characteristics, this dependency disappears.

In order to implement virtual IMRT QA in a clinic the following workflow should be followed: (1) collect or access IMRT QA data, (2) extract all the parameters of the IMRT fields from plan files, (3) extract the features for the calculation of all

TABLE 1 | Summary of studies on machine QA using machine learning techniques in a chronological order.

References	QA Source	Data Source	ML Model	Task
Carlson et al. (2016)	DICOM_RT, Dynalog files	74 VMAT plans	Regression, Random Forest, Cubist	MLC Position Errors Detection
Li and Chan (2017)	Daily QA Device	5-year Daily QA Data	ANN Time-Series, ARIMA Models	Symmetry Prediction
Sun et al. (2018)	Ion Chamber	1,754 Proton Fields	Random Forrest, XGBoost, Cubist	Output for Compact Proton Machine
El Naqa et al. (2019)	EPID	119 Images from 8 Linacs	Support Vector Data Description, Clustering	Gantry Sag, Radiation Field Shift, MLC Offset
Grewal et al. (2020)	Ion Chamber	4,231 Proton Fields	Gaussian Processes, Shallow NN	Output and Patient QA Proton Machine
Osman et al. (2020)	log files	400 machine delivery log files	ANN	MLC Discrepancies during Delivery & Feedback
Chuang et al. (in press)	Trajectory log files	116 IMRT plans, 125 VMAT plans	Boosted Tree Outperformed LR	MLC Discrepancies during Delivery & Feedback
Zhao et al. (in press)	Water Tank Measurement	43 Truebeam PDD, Profiles	Multivariate Regression (Ridge)	Modeling of Beam Data Linac Commissioning

complexity metrics affecting the passing rates, (4) use a machine learning algorithm to build a virtual IMRT QA model. During this process, we identify the most impactful features that affect the passing rate.

Deep Learning Models

The process described in the previous section Early ML Models requires carefully designing features that describe the correlation between plan characteristics and passing rates. Using an algorithm capable of designing their own features, Dr. Valdes and his group compared a Deep Neural Network against their own Poisson regression model using the same patient QA data previously described (Interian et al., 2018). The input to the CNN, a special type of neural network designed to analyze images, was the fluence map for each plan without the need of expert designed features. The models were trained to predict IMRT QA gamma passing rates using TensorFlow and Keras. The authors concluded that CNNs with transfer learning can predict IMRT QA passing rates by automatically designing features from the fluence maps without human expert supervision. The predictions from the CNNs were comparable to the virtual IMRT QA system described above which was carefully designed by physicist experts.

Tomori et al. built a prediction model for gamma evaluation of IMRT QA based on deep learning (Tomori et al., 2018) using sixty IMRT QA plans. Fifteen-layer CNN were developed to learn the planar dose distributions from a QA phantom. The gamma passing rate was measured using EBT3 film. The input training data also included the volume of PTV, rectum, and overlapping region, and the monitor unit for each field. The network produced predicted gamma passing rates at four criteria: 2%/2 mm, 3%/2 mm, 2%/3 mm, and 3%/3 mm. Five-fold cross-validation was applied to validate the performance. A linear relationship was found between the measured and predicted values for all criteria. These results also suggested that deep learning methods may provide a useful prediction model for gamma evaluation of patient-specific QA.

Lam et al. applied 3 tree-based machine learning algorithms (AdaBoost, Random Forest, and XGBoost) to train the models and predict gamma passing rates using a total of 1,497 IMRT beams delivered with portal dosiemtry (Lam et al., 2019). They reported that both AdaBoost and Random Forest had 98 ± 3% of predictions within 3% of the measured 2%/2 mm gamma passing rates with a maximum error $< 4\%$ and a MAE $< 1\%$. XGBoost showed a slightly worse prediction accuracy with 95% of the predictions $< 3\%$ of the measured gamma passing rates and a maximum error of 4.5%. The three models identified the same nine features in the top 10 most important ones that are related to plan complexity and maximum aperture displacement from the central axis or the maximum jaw size in a beam. Their results demonstrated that portal dosimetry IMRT QA gamma passing rates can be accurately predicted using tree-based ensemble learning models.

Nyflot et al. investigated a deep learning approach to classify potential treatment delivery errors and predict QA results using image and texture features from 186 EPID images (Nyflot et al., 2019). Three sets of planar doses were exported from each QA plan corresponding to (a) the error-free case, (b) a random MLC error case, and (c) a systematic MLC error case. Each plan was delivered to an EPID panel and gamma analysis was performed using the EPID dosimetry software. Two radiomic approaches (image and texture features) were used. The resulting metrics from both approaches were used as input into four machine learning classifiers in order to determine whether images contained the introduced errors. After training, a single extractor is used as a feature extractor for classification. The performance of the deep learning network was superior to the texture features approach, and both radiomic approaches were better than using gamma passing rates in order to predict the clinically relevant errors.

ML Applied to VMAT QA

ML applications have been extended to volumetric modulated arc therapy (VMAT) QA. Granville et al. built a ML model with 1620 VMAT plans (Elekta) to predict the results of VMAT QA measurements using not only treatment plan characteristics but also Linac performance metrics (Granville et al., 2019). They trained a linear Support Vector Classifier (SVC) to classify the results of VMAT QA. The outputs in this model were simple classes representing the median dose difference (±1%) between measured and expected dose distributions rather than passing rates. In the model development phase, a recursive feature elimination (RFE) cross-validation technique was used to eliminate unimportant features. Of the ten features found to be most predictive of VMAT QA measurement results, half were derived from treatment plan characteristics and a half from Linac QA metrics. Such a model has the potential to provide more timely failure detection for patient-specific QA. Ono et al. utilized 600 VMAT plans and their corresponding ArcCHECK measurements to build prediction models using three machine learning algorithms—regression tree analysis, multiple regression analysis, and neural network. They found that the neural networks model achieved slightly better results among the 3 models in terms of prediction error (Ono et al., 2019).

Li et al. investigated the impact of delivery characteristics on the dose accuracy of VMAT (Li et al., 2019a). Ten metrics reflecting VMAT delivery characteristics were extracted from 344 QA plans. The study found that leaf speed is the most important factor affecting the accuracy of gynecologic, rectal, and head and neck plans, while the field complexity, small aperture score, and MU are the most important factors influencing the accuracy of prostate plans. Li et al. also studied the accuracy of prediction using machine learning for VMAT QA (Li et al., 2019b). The authors presented the workflows for two prediction models; the classic Poisson regression model, and the newly constructed Random Forest classification model. To test the prediction accuracy, 255 VMAT plans (Varian) with 10-fold cross-validation were used to explore the model performance under different gamma criteria and action limits. In clinical validation, independent 48 VMAT plans without cross-validation were used to further validate the reliability of models. The authors also showed the absolute prediction error with both technical and clinical validations. The prediction accuracy was greatly affected by the absolute value of the measured gamma passing rates and gamma criteria. The regression model was able to accurately predict those passing rates for the majority VMAT plans, but the classification model had a much better sensitivity to accurately detect failed QA plans. Later the same group further improved their prediction model using autoencoder based classification-regression (ACLR) to generate gamma passing rates predictions for three different gamma criteria from 54 complexity metrics as input (Wang et al., in Press). With an additional 150 VMAT plans available for clinical validation to evaluate the generalized performance of the model, the group reported that such a hybrid model significantly improved prediction accuracy over their early model, Poisson Lasso regression.

Wall and Fontenot used 500 VMAT and MapCHECK2 data to build predictive models using four different machine learning

TABLE 2 | Summary of studies on patient-specific QA using machine learning techniques.

Group	TPS/Delivery	QA Source	Data Source	ML Model	Research Highlight
Valdes et al. (2016)	Eclipse/Varian	MapCHECK2	498 IMRT Plans	Poisson Regression	Founding Paper
Valdes et al. (2017)	Eclipse/Varian	Portal Dosimetry	203 IMRT Beams	Poisson Regression	Multi-sites Validation
Interian et al. (2018)	Eclipse/Varian	MapCHECK2	498 IMRT Plans	Convolutional Neural Network	Fluence Maps as Input
Tomori et al. (2018)	iPlan/Varian	EBT3 film	60 IMRT Plans	Convolutional Neural Network	Planar Dose, Volumes, MU
Lam et al. (2019)	Eclipse/Varian	Portal Dosimetry	1,497 IMRT Beams	AdaBoost, Random Forest, XGBoost	Tree-based High Accuracy
Nyflot et al. (2019)	Pinnacle/Elekta	EPID	186 IMRT Beams	Convolutional Neural Network	Image, Texture Features
Granville et al. (2019)	Monaco/Elekta	Delta4	1,620 VMAT Beams	Support Vector Classifier	1st VMAT & w/ QC Metrics
Ono et al. (2019)	RayStation, Eclipse/Vero, Varian	ArcCHECK	600 VMAT Plans	Regression Tree, Multiple Regression, Neural Network	ML Models Comparison
Li et al. (2019b)	Eclipse/Varian	MatriXX	255 VMAT Beams	Poisson Lasso & Random Forest	Specificity & Sensitivity
Wang et al. (in Press)	Eclipse/Varian	MatriXX	576 VMAT Beams	Hybrid Model ACLR	High Prediction Accuracy
Wall and Fontenot (2020)	Pinnacle/Elekta	MapCHECK2	500 VMAT Plans	Linear Regression, SVM, Tree-based, ANN	ML Models Comparison
Hirashima et al. (2020)	RayStation, Eclipse/ Vero, Varian	ArcCHECK	1,255 VMAT Plans	Hybrid Model XGBoost	Plan Complexity & Dosiomics

algorithms and then compared their performance (Wall and Fontenot, 2020). They found that the SVM model, trained using the 100 most important features selected using the linear regression method, gave the lowest cross-validated testing MAE of 3.75% as compared to linear models, tree-based models, and neural networks. More recently, Hirashima et al. (2020) used Gradient Boosting, the most accurate algorithm up to date for the analysis of tabular data, to create a model to predict ArcCHECK measurements using plan complexity and dosiomic features extracted from 1,255 VMAT plans, also showing the validity of virtual VMAT QA.

Table 2 lists the studies on virtual IMRT/VMAT QA. In short, there have been multiple studies that all find similar conclusions independent of the brand of Linac, TPS, and QA tool used: QA results can be predicted accurately using machine learning.

SUMMARY AND FUTURE DIRECTIONS

Since the early ML models applied to machine and patient-specific QA were reported in early 2016, a significant improvements have been seen in more recent models as machine learning techniques in radiotherapy QA matured. The models grew from simple Poisson regressions to deep learning classification models, and then to complex hybrid models which improved prediction accuracy. Therefore, it is expected that future ML models built on the foundation of existing knowledge can continue to be refined. With deep learning models, there is a greater potential to make QA processes more efficient and effective in clinical settings. In the meantime, it is very important to fully understand the limitations of virtual QA. Kalet et al. has highlighted some of the unique challenges of ML applications in radiotherapy QA including data quality, model adaptability, and model limitations (Kalet et al., 2020). Data quality is by far the most basic and essential requirement for building an accurate prediction model. Not only can incomplete data, such as small sample size, lead to wrong conclusions, but "true" QA data from detectors, especially for extremely small/large field size or large low dose regions, can also lead to imperfect prediction models due to detector system limitations (Valdes et al., 2017). Multi-institutional validation is often helpful to validate and generalize the ML models. In addition to the challenges of data integrity, Kearney et al. raised awareness of some persistent misuse of deep learning in the field (Kearney et al., 2018).

To date, many applications of ML to radiotherapy QA have focused on predicting machine performance and IMRT/VMAT QA results. Fully understanding and dissecting all factors that govern delivery accuracy is extremely important for clinical physicists to be able to implement a risk-based program as suggested in the AAPM TG-100 report. Further developments could lead to QA predictions being included in the treatment planning optimizer so that all QA could pass. We could also know ahead of time that we need to run a clinically-relevant QA on those plans with the lowest expected passing rates. It is clear that prediction of QA results could have profound implications on the current radiotherapy process. Before implementing in-house or commercial ML models to perform sanity check, second check, and automated or virtual QA in any clinical setting, we should carefully assess and address the limitations of both data and ML models.

AUTHOR'S NOTE

The materials were presented in a SAM Therapy Educational Course at the 61st AAPM Annual Meeting in San Antonio, TX, in July 2019.

AUTHOR CONTRIBUTIONS

MC, AW, and GV have contributed to writing this review article. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

REFERENCES

Alpaydin, E. (2010). *Introduction to Machine Learning*. Cambridge: MIT Press.

Carlson, J. N., Park, J. M., Park, S. Y., Park, J. I., Choi, Y., and Ye, S. J. (2016). A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys. Med. Biol.* 61:2514. doi: 10.1088/0031-9155/61/6/2514

Chuang, K. C., Adamson, J., and Giles, W. M. (in press). A tool for patient specific prediction of delivery discrepancies in machine parameters using trajectory log files. *Med. Phys*.

El Naqa, I., Irrer, J., Ritter, T. A., DeMarco, J., Al-Hallaq, H., Booth, J., et al. (2019). Machine learning for automated quality assurance in radiotherapy: a proof of principle using EPID data description. *Med. Phys.* 46, 1914–1921. doi: 10.1002/mp.13433

Feng, M., Valdes, G., Dixit, N., and Solberg, T. D. (2018). Machine learning in radiation oncology: opportunities, requirements, and needs. *Front. Oncol.* 8:110. doi: 10.3389/fonc.2018.00110

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232. doi: 10.1214/aos/1013203451

Granville, D. A., Sutherland, J. G., Belec, J. G., and La Russa, D. J. (2019). Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. *Phys. Med. Biol.* 64:095017. doi: 10.1088/1361-6560/ab142e

Grewal, H. S., Chacko, M. S., Ahmad, S., and Jin, H. (2020). Prediction of the output factor using machine and deep learning approach uniform scanning proton therapy. *J. Appl. Clin. Med. Phys*. 21, 128–134. doi: 10.1002/acm2.12899

Hastie, T., Tibshirani, R., and Friedman, J. (eds). (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, *2nd Edn*. New York, NY: Springer.

Hirashima, H., Ono, T., Nakamura, M., Miyabe, Y., Mukumoto, N., Iramina, H., et al. (2020). Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features. *Radiat. Oncol*. doi: 10.1016/j.radonc.2020.07.031. [Epub ahead of print].

Huq, M. S., Fraass, B. A., Dunscombe, P. B., Gibbons, J. P. Jr, Ibbott, G. S., Mundt, A. J., et al. (2016). The report of task group 100 of the AAPM: application of risk analysis methods to radiation therapy quality management. *Med. Phys.* 43, 4209–4262. doi: 10.1118/1.4947547

Interian, Y., Rideout, V., Kearney, V. P., Gennatas, E., Morin, O., Cheung, J., et al. (2018). Deep nets vs expert designed features in medical physics: An IMRT QA case study. *Med. Phys.* 45, 2672–2680. doi: 10.1002/mp.12890

Kalet, A. M., Luk, S. M. H., and Phillips, M. H. (2020). Radiation therapy quality assurance tasks and tools: the many roles of machine learning. *Med. Phys.* 47, e168–e177. doi: 10.1002/mp.13445

Kearney, V., Valdes, G., and Solberg, T. D. (2018). Deep learning misuse in radiation oncology. *Int. J. Radiat. Oncol. Biol. Phys.* 102:S62. doi: 10.1016/j.ijrobp.2018.06.174

Lam, D., Zhang, X., Li, H., Deshan, Y., Schott, B., Zhao, T., et al. (2019). Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. *Med. Phys.* 46:46666–44675. doi: 10.1002/mp.13752

Le Cun, Y., and Bengio, Y. (2002). "World-level training of a handwritten word recognizer based on convolutional neural networks." in *IEEE. Proceedings of the 12th IAPR International Conference on Pattern Recognition. Vol. 3-Conference C: Signal Processing* (Niagara Falls, ON).

Li, J., Wang, L., Zhang, X., Liu, L., Li, J., Chan, M. F., et al. (2019b). Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy. *Int. J. Rad. Oncol. Biol. Phys.* 105, 893–902. doi: 10.1016/j.ijrobp.2019.07.049

Li, J., Zhang, X., Li, J., Jiang, R., Sui, J., Chan, M. F., et al. (2019a). Impact of delivery characteristics on dose accuracy of volumetric modulated arc therapy for different treatment sites. *J. Radiat. Res.* 60, 603–611. doi: 10.1093/jrr/rrz033

Li, Q., and Chan, M. F. (2017). Predictive time series modeling using artificial neural networks for Linac beam symmetry – an empirical study. *Ann. N. Y. Acad. Sci*. 1387, 84–94. doi: 10.1111/nyas.13215

Li, Q., Chan, M. F., and Shi, C. (2017). "Clustering breathing curves in 4D radiotherapy by using multiple machine learning tools: K-means and Hierarchical clustering algorithms." in *Proceedings of the 11th Annual Machine Learning Symposium* (New York, NY), 28–29.

Lin, H., Shi, C., Wang, B., Chan, M. F., Tang, X., and Ji, W. (2019). Towards real-time respiratory motion prediction based on long short-term memory neural networks. *Phys. Med. Biol.* 64:085010. doi: 10.1088/1361-6560/ab13fa

Luna, J. M., Gennatas, E. D., Ungar, L. H., Eaton, E., Diffenderfer, E. S., Jensen, S. T., et al. (2019). Building more accurate decision trees with the additive tree. *PNAS* 116, 19887–19893. doi: 10.1073/pnas.1816748116

MathWorks.com (2020). *Unsupervised Learning*. Available online at: https://www.mathworks.com/discovery/unsupervised-learning.html (accessed July 9, 2020).

Nyflot, M. J., Thammasorn, P., Wootton, L. S., Ford, E. C., and Chaovalitwongse, W. A. (2019). Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. *Med. Phys.* 46, 456–464. doi: 10.1002/mp.13338

Ono, T., Hirashima, H., Iramina, H., Mukumoto, N., Miyabe, Y., Nakamura, M., et al. (2019). Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning. *Med. Phys.* 46:382303832. doi: 10.1002/mp.13669

Osman, A. F., Maalej, N. M., and Jayesh, K. (2020). Prediction of the individual multileaf collimator positional deviations during dynamic IMRT delivery priori with artificial neural network. *Med. Phys.* 47, 1421–1430. doi: 10.1002/mp.14014

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Russell, S. J., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, *3rd Edn.* Harlow: Prentice Hall.

Shobha, G., and Rangaswamy, S. (2018). "Computational analysis and understand of natural languages: principles, methods and applications." in *Handbook of Statistics*, eds V. Gudivada and C. R. Rao (Amsterdam: North Holland), 2–515.

Sun, B., Lam, D., Yang, D., Grantham, K., Zhang, T., Mutic, S., et al. (2018). A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Med. Phys.* 45, 2243–2251. doi: 10.1002/mp.12842

Tomori, S., Kadoya, N., Takayama, Y., Kajikawa, T., Shima, K., Narazaki, K., et al. (2018). A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. *Med. Phys.* 45, 4055–4065. doi: 10.1002/mp.13112

Valdes, G., Chan, M. F., Lim, S., Scheuermann, R., Deasy, J. O., and Solberg, T. D. (2017). IMRT QA using machine learning: A multi-institutional validation. *J. Appl. Clin. Med. Phys.* 18, 278–284. doi: 10.1002/acm2.12161

Valdes, G., Scheuermann, R., Hung, C. Y., Olszanski, A., Bellerive, M., and Solberg, T. D. (2016). A mathematical framework for virtual IMRT QA using machine learning. *Med. Phys.* 43, 4323–4334. doi: 10.1118/1.4953835

Wall, P. D. H., and Fontenot, J. D. (2020). Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning. *Inform. Med. Unlocked*. 18:100292. doi: 10.1016/j.imu.2020.100292

Wang, L., Li, J., Zhang, S., Zhang, X., Zhang, Q., Chan, M. F., et al. (in Press). Multi-task autoencoder based classification-regression (ACLR) model for patient-specific VMAT QA. *Phys. Med. Biol.*

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Comp.* 8, 1341–1390. doi: 10.1162/neco.1996.8.7.1341

Zhao, W., Schüler, E., Patil, I., Han, B., Yang, Y., and Xing, L. (in press). Beam data modeling of linear accelerators (linacs) through machine learning and its potential applications in fast and robust linac commissioning and quality assurance. *Radiat Oncol*.

Conflict of Interest: MC has no commercial or financial interests to declare. AW and GV report an ownership stake in Foretell Med LLC, which is developing machine learning models in medicine. However, this work was done before and outside of that company.

ORIGINAL RESEARCH
published: 05 October 2020
doi: 10.3389/frai.2020.550890

Prognostic Value of Transfer Learning Based Features in Resectable Pancreatic Ductal Adenocarcinoma

Yucheng Zhang[1], Edrise M. Lobo-Mueller[2], Paul Karanicolas[3], Steven Gallinger[4], Masoom A. Haider[4,5†] and Farzad Khalvati[1,6,7†]*

[1] Department of Medical Imaging, University of Toronto, Toronto, ON, Canada, [2] Department of Diagnostic Imaging and Department of Oncology, Faculty of Medicine and Dentistry, Cross Cancer Institute, University of Alberta, Edmonton, AB, Canada, [3] Department of Surgery, Sunnybrook Health Sciences Centre, Toronto, ON, Canada, [4] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada, [5] Joint Department of Medical Imaging, Sinai Health System, University Health Network, University of Toronto, Toronto, ON, Canada, [6] Research Institute, The Hospital for Sick Children, Toronto, ON, Canada, [7] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

OPEN ACCESS

Edited by:
Frank Emmert-Streib,
Tampere University, Finland

Reviewed by:
Junjie Zhang,
Royal Bank of Canada, Canada
Shailesh Tripathi,
Tampere University of Technology, Finland

****Correspondence:***
Farzad Khalvati
Farzad.Khalvati@utoronto.ca

† These authors share senior authorship

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *10 April 2020*
Accepted: *24 August 2020*
Published: *05 October 2020*

Citation:
Zhang Y, Lobo-Mueller EM, Karanicolas P, Gallinger S, Haider MA and Khalvati F (2020) Prognostic Value of Transfer Learning Based Features in Resectable Pancreatic Ductal Adenocarcinoma. Front. Artif. Intell. 3:550890. doi: 10.3389/frai.2020.550890

Background: Pancreatic Ductal Adenocarcinoma (PDAC) is one of the most aggressive cancers with an extremely poor prognosis. Radiomics has shown prognostic ability in multiple types of cancer including PDAC. However, the prognostic value of traditional radiomics pipelines, which are based on hand-crafted radiomic features alone is limited.

Methods: Convolutional neural networks (CNNs) have been shown to outperform radiomics models in computer vision tasks. However, training a CNN from scratch requires a large sample size which is not feasible in most medical imaging studies. As an alternative solution, CNN-based transfer learning models have shown the potential for achieving reasonable performance using small datasets. In this work, we developed and validated a CNN-based transfer learning model for prognostication of overall survival in PDAC patients using two independent resectable PDAC cohorts.

Results: The proposed transfer learning-based prognostication model for overall survival achieved the area under the receiver operating characteristic curve of 0.81 on the test cohort, which was significantly higher than that of the traditional radiomics model (0.54). To further assess the prognostic value of the models, the predicted probabilities of death generated from the two models were used as risk scores in a univariate Cox Proportional Hazard model and while the risk score from the traditional radiomics model was not associated with overall survival, the proposed transfer learning-based risk score had significant prognostic value with hazard ratio of 1.86 (95% Confidence Interval: 1.15–3.53, *p*-value: 0.04).

Conclusions: This result suggests that transfer learning-based models may significantly improve prognostic performance in typical small sample size medical imaging studies.

Keywords: transfer learning, radiomics, prognosis, pancreatic cancer, survival analysis

INTRODUCTION

Pancreatic Ductal Adenocarcinoma (PDAC) is one of the most aggressive malignancies with poor prognosis (Stark and Eibl, 2015; Stark et al., 2016; Adamska et al., 2017). Evidence suggested that surgery can improve overall survival in resectable PDAC cohorts (Stark et al., 2016; Adamska et al., 2017). However, the 5-year survival rate of patients who went through surgery is still low (Fatima et al., 2010). Thus, it is important to identify high-risk and low-risk surgical candidates so that healthcare providers can make personalized treatment decisions (Khalvati et al., 2019a). In resectable patients, clinicopathologic factors such as tumor size, margin status at surgery, and histological tumor grade have been studied as biomarkers for prognosis (Ahmad et al., 2001; Ferrone et al., 2012; Khalvati et al., 2019a). However, many of these biomarkers can only be assessed after the surgery and thus, the opportunity for patient-tailored neoadjuvant therapy is lost. Recently, quantitative medical imaging biomarkers have shown promising results in prognostication of the overall survival for cancer patients, providing an alternative solution (Kumar et al., 2012; Parmar et al., 2015; Lambin et al., 2017).

As a rapidly developing field in medical imaging, radiomics is defined as the extraction and analysis of a large number of quantitative imaging features from medical images including CT and MRI (Kumar et al., 2012; Lambin et al., 2012; Khalvati et al., 2019b). The conventional radiomic analysis pipeline consists of four steps as shown in **Figure 1**. Following this pipeline, several radiomic features have been shown to be significantly associated with clinical outcomes including overall survival or recurrence in different cancer sites such as lung, head and neck, and pancreas (Aerts et al., 2014; Coroller et al., 2015; Carneiro et al., 2016; Cassinotto et al., 2017; Chakraborty et al., 2017; Eilaghi et al., 2017; Lao et al., 2017; Zhang et al., 2017; Attiyeh et al., 2018; Yun et al., 2018; Sandrasegaran et al., 2019). Using these radiomic features, patients can be categorized into low-risk or high-risk groups guiding clinicians to design personalized treatment plans (Chakraborty et al., 2018; Varghese et al., 2019). Although limited work has been done in the context of PDAC, recent studies have confirmed the potential of new quantitative imaging biomarkers for resectable PDAC prognosis (Eilaghi et al., 2017; Khalvati et al., 2019a).

Despite recent progress, radiomics analytics solutions have a major limitation in terms of performance. The performance of radiomics models relies on the amount of information that radiomics features can capture from medical images (Kumar et al., 2012). Most radiomics features represent morphology, first order, or texture information from the regions of interest (Van Griethuysen et al., 2017). The equations of these radiomic features are often manually designed. This is a sophisticated and time-consuming process, requiring prior knowledge of image processing and tumor biology. Consequently, a poor design of the feature bank may fail to extract important information from medical images, having a significant negative impact on the performance of prognostication. In contrast, the ability of deep learning for automatic feature extraction has been proven and shown to achieve promising performances in different medical imaging tasks (Shen et al., 2017; Yamashita et al., 2018; Yasaka et al., 2018).

A convolutional neural network (CNN) (Schmidhuber, 2014; LeCun et al., 2015) performs a series of convolution and pooling operations to get comprehensive quantitative information from input images (LeCun et al., 2015). Compared to hand-crafted radiomic features that are predesigned and fixed, the coefficients of CNNs are modified in the training process. Hence, the final features generated from a successfully trained CNN are tuned to be associated with the target outcomes (e.g., overall survival, recurrence). It has been shown that CNN architectures are effective in different medical imaging tasks such as segmentation for head and neck anatomy and diagnosis for the retinal disease (Dalmiş et al., 2017; De Fauw et al., 2018; Nikolov et al., 2018; Irvin et al., 2019).

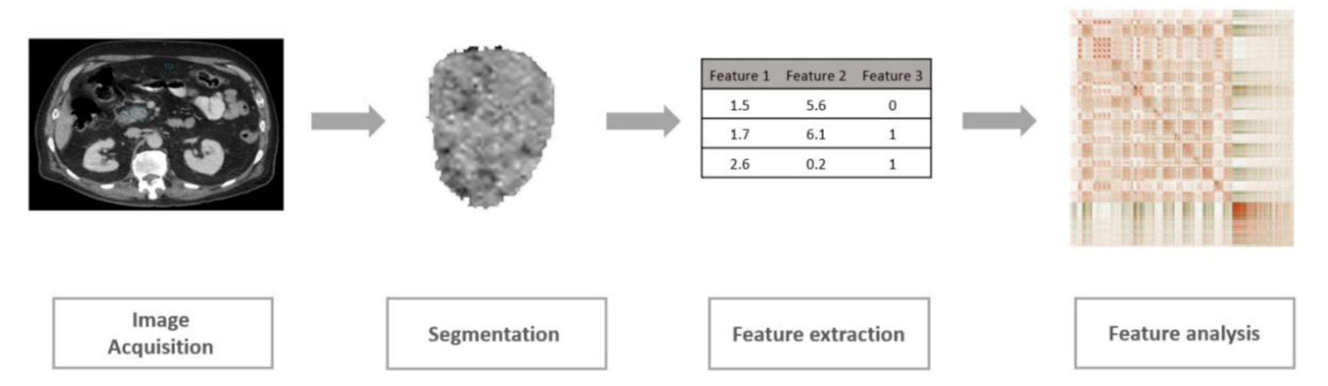
However, to train a CNN from scratch, millions of parameters need to be tuned. This requires a large sample size which is not feasible to collect in most medical imaging studies (Du et al., 2018). As an alternative solution, CNN-based transfer learning is more suitable for medical imaging tasks since it can achieve a comparable performance using a limited amount of data (Pan and Yang, 2010; Chuen-Kai et al., 2015).

CNN-based transfer learning is defined as taking images from a different domain such as natural images (e.g., ImageNet) to build a pretrained model and then apply the pretrained model to target images (e.g., CT images of lung cancer) (Ravishankar et al., 2017). The idea of transfer learning is based on the assumption that the structure of a CNN is similar to the human visual cortex as both are composed of layers of neurons (Pan and Yang, 2010; Tan et al., 2018). Top layers of CNNs can extract general features from images while deeper layers are able to extract information that is more specific to the outcomes (Yosinski et al., 2014).

Transfer learning utilizes this property, training top layers using another large dataset while finetuning deeper layers using data from the target domain. For example, the ImageNet dataset contains more than 14 million images (Russakovsky et al., 2015). Hence, pretraining a model using this dataset would help the model learn how to extract general features using initial layers. Given that many image recognition tasks are similar, top (shallower) layers of the pretrained network can be transferred to another CNN model. In the next step, deeper layers of the CNN model can be trained using the target domain images (Torrey and Shavlik, 2009). Since the deeper layers are more target-specific, finetuning them using the images from the target domain may help the model quickly adapt to the target outcome, and hence, improve the overall performance.

In medical imaging, the target dataset is often so small that it is impractical to properly finetune the deeper layers. Consequently, in practice, a pretrained CNN can be used as a feature extractor (Hertel et al., 2015; Lao et al., 2017). Given that convolution layers can capture high-level and informative details from images, passing the target domain images through these layers allows

Abbreviations: ROC, Receiver operating characteristic; AUC, Area under the ROC curve; CT, Computed tomography; CI, Confidence interval; CNN, Convolutional neural network; GLCM, Gray-Level Co-occurrence matrix; NSCLC, Non-small-cell lung cancer; PDAC, Pancreatic ductal adenocarcinoma; ROI, Region of interest; SMOTE, Synthetic minority over-sampling technique.



Feature 1	Feature 2	Feature 3
1.5	5.6	0
1.7	6.1	1
2.6	0.2	1

FIGURE 1 | Conventional radiomics analytics pipeline.

extractions of features. These features can be further used to train a classifier for the target domain, enabling building a high-performance transfer learning model using a small dataset.

In this study, using two independent small sample size resectable PDAC cohorts, we evaluated the prognosis performance of a transfer learning model and compared its performance to that of a traditional radiomics model. The goal of the prognostication was to dichotomize PDAC patients who were candidates for curative-intent surgery to high-risk and low-risk groups. We found that the transfer learning model provides better prognostication performance compared to the conventional radiomics model, suggesting the potential of transfer learning in a typical small sample size medical imaging study.

METHODS

Dataset

Two cohorts from two independent hospitals consisting of 68 (Cohort 1) and 30 (Cohort 2) patients were enrolled in this retrospective study. All patients underwent curative intent surgical resection for PDAC from 2007–2012 to 2008–2013 in Cohort 1 and Cohort 2, respectively, and they did not receive other neoadjuvant treatment. Preoperative portal venous phase contrast-enhanced CT images were used. Overall survival (including survival as duration and death as the event) was collected as the primary outcome and it was calculated as the duration from the date of preoperative CT scan until death. To exclude the confounding effect of postoperative complications, patients who died within 90 days after the surgery were excluded. Institutional review board approval was obtained for this study from both institutions (Khalvati et al., 2019a).

An in-house developed Region of Interest (ROI) contouring tool (ProCanVAS Zhang et al., 2016) was used by a radiologist with 18 years of experience who completed the contours blind to the outcome (overall survival). Following the protocol, the slices were contoured with the largest visible 2D cross-section of the tumor on the portal venous phase. When the boundary of the tumor was not clear, it was defined by the presence of pancreatic or common bile duct cut-off and the review of pancreatic phase images (Khalvati et al., 2019a). An example of the contour is shown in **Figure 2**.

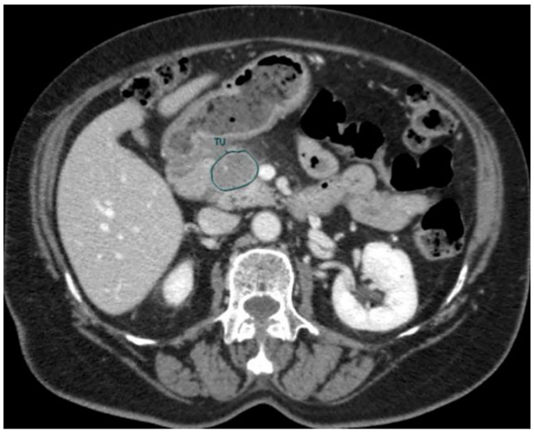


FIGURE 2 | A manual contour of CT scan from a representative patient in cohort 2.

Radiomics Feature Extraction

Radiomics features were extracted using the PyRadiomics library (Van Griethuysen et al., 2017) (version 2.0.0) in Python. Voxels with Hounsfield unit under−10 and above 500 were excluded so that the presence of fat and stents will not affect the values of the features. The bin width (number of gray levels per bin) was set to 25. In total, 1,428 radiomic features were extracted from CT images within the ROI for both cohorts. **Table 1** lists different classes of features used in this study (Khalvati et al., 2019a).

Transfer Learning

We developed a transfer learning model (LungTrans) pretrained by CT images from non-small-cell lung cancer (NSCLC) patients. The Lung CT dataset was published on Kaggle for Lung Nodule Analysis (LUNA16), containing CT images from 888 lung cancer patients and the outcome (malignancy or not) (Armato et al., 2011). All input ROIs were resized to 32×32 greyscale. An

8-layer CNN was trained from scratch using LUNA16 CT images with batch size 16 and learning rate 0.001 (**Figure 3**). This configuration was shown to have high performance in differentiating malignancy vs. normal tissue in the LUNA16 competition (De Wit, 2017). In addition, given small ROI sizes of data in this study (32×32) and the fact that images are grayscale instead of RGB color, off-the-shelf deep CNNs such as ResNet (He et al., 2015) do not provide adequate performance. Each convolutional layer except for Conv_5 has Kernel size as 3×3 with stride of 1 with zero padding. Conv_5 has 2×2 kernel size and stride of 1 without padding. All the Max Pooling layers have 2×2 kernel size. Previous research has shown that top layers in the CNN extract generic features from the image, while bottom layers can extract features specific to the tasks (Yosinski et al., 2014; Paul et al., 2019). Since our pretrained domain (lung CT) and target domain (PDAC CT) are rather similar, we extracted features from the bottom layer. In addition, the number of features (coefficients) in the CNN significantly decreases as the layers become deeper, due to Max pooling. If we picked a layer above the final layer, the number of extracted features would increase significantly. Considering the sample size of our training (68) and test (30) datasets, all the convolution layers were frozen and features were extracted from the end of the CNN (Conv_5). As a result, for each ROI from PDAC CT images, 64 features were extracted. This was the ideal number of intermediate features tested in LUNA16 dataset (De Wit, 2017).

TABLE 1 | List of radiomic feature classes and filters.

First-order features	**Histogram-based features**
Second-order texture features	Features extracted from Gray-Level Co-Occurrence matrix (GLCM)
Morphology features	Features based on the shape of the region of interest
Filters	No filter, exponential, gradient, logarithm, square, square-root, local binary pattern, wavelet

TABLE 2A | Summary of models' performances in AUC.

	Training cohort (n = 68) (5-Fold cross validation)	**Test cohort (n = 30)**
PyRadiomics model	0.57 (95% CI: 0.42–0.73)	0.54 (95% CI: 0.32–0.76)
Transfer learning model	0.72 (95% CI: 0.58–0.86)	0.81 (95% CI: 0.64–0.98)

__Tables 2B__,__C__ show Confusion Matrix for Random Forest models using PyRadiomics and LungTrans features, respectively, in the test cohort.

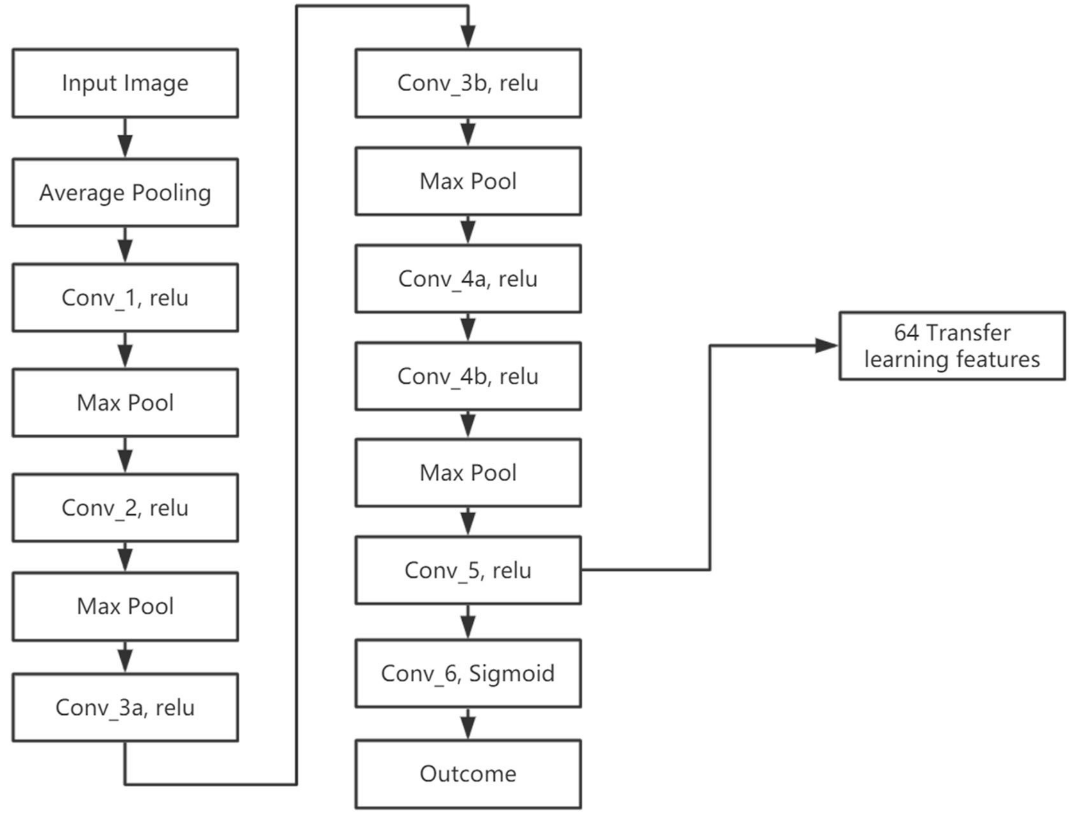


FIGURE 3 | Architecture for pretrained CNN using LUNA16 data.

Prognostic Models

To have a proper and robust validation, training and test datasets were collected from two different institutions. In Cohort 1 (training cohort, $n = 68$), two prognostic models for overall survival were trained using features extracted from conventional radiomics feature bank (PyRadiomics) and transfer learning model (LungTrans). The prognosis models were built using the Random Forest classifier, which is a common classifier in radiomics analytic pipeline, with 500 decision trees (Chen and Ishwaran, 2012; Zhang et al., 2017). Random Forest classifier is highly data-adaptive, which have shown the potential to handle large P small N problem by choosing the best subset of features for classification (Chen and Ishwaran, 2012). The "data-adaptive" characteristic makes the random forest a good candidate for our study where transfer learning and PyRadiomics offered different numbers of features. The number of variables available for splitting at each tree node (mtry) was determined by the best performing mtry option in the training cohort. Due to the imbalanced outcome in the training data, (Cohort 1: 52 Deaths vs. 16 Survivals), a data balancing algorithm, SMOTE (Ryu et al., 2002), was applied in the training process to artificially balance the training data.

The prognostic values of these two models were evaluated in Cohort 2 ($n = 30$, 15 Deaths vs. 15 Survivals) using the area under the receiver operating characteristic (ROC) curve (AUC). DeLong test, as one of the common comparison tests, was used to test the difference between the two ROC curves (DeLong et al., 1988). To further assess the prognosis values, the predicted probabilities of death generated from the two classifiers were used as risk scores in survival analyses. These risk scores were tested in Cohort 2 using univariate Cox Proportional Hazards Model for their Hazard Ratio and Wald test *p*-value (Cox, 1972). These analyses were done in R (version 3.5.1) using "caret," "pROC," and "survival" packages (Kuhn, 2008; Therneau, 2020).

TABLE 2B | Confusion Matrix of PyRadiomics model in the test cohort.

Test cohort	Deceased patients	Survived patients
Predicted death	12	10
Predicted survival	3	5

Accuracy: 0.57, Sensitivity: 0.8, Specificity: 0.33, Precision: 0.55.

TABLE 2C | Confusion matrix of transfer learning model in the test cohort.

Test cohort	Deceased patients	Survived patients
Predicted Death	13	4
Predicted Survival	2	11

Accuracy: 0.80, Sensitivity: 0.87, Specificity: 0.73, Precision: 0.76.

RESULTS

Prognostic Models Performance

Using features from the PyRadiomics feature bank, the Random Forest model yielded AUC of 0.54 [95% Confidence Interval (CI): 0.32–0.76] in the test cohort (Cohort 2) (mtry: 2). In contrast, using LungTrans features, the AUC of the Random Forest model reached 0.81 (95% CI: 0.64–0.98) in the test cohort (mtry: 17). The performances of these two models for both training and test cohorts are listed in **Table 2A**. We performed a 5-fold cross-validation to produce AUCs for the training cohort. The AUCs for the test cohort were generated using the models trained by the training cohort.

To investigate the prognostic value of each PyRadiomics features, variable importance indices were calculated using the Caret Package in R. The top ten features were first order entropy, first order uniformity, first order interquartile range, GLSZM gray level non-uniformity normalized, GLRLM run length non-uniformity normalized, GLCM cluster tendency, NGTDM busyness, GLSZM small area high gray level emphasis, GLSZM low gray level zone emphasis, and GLSZM large area high gray level emphasis. This confirming previous studies in

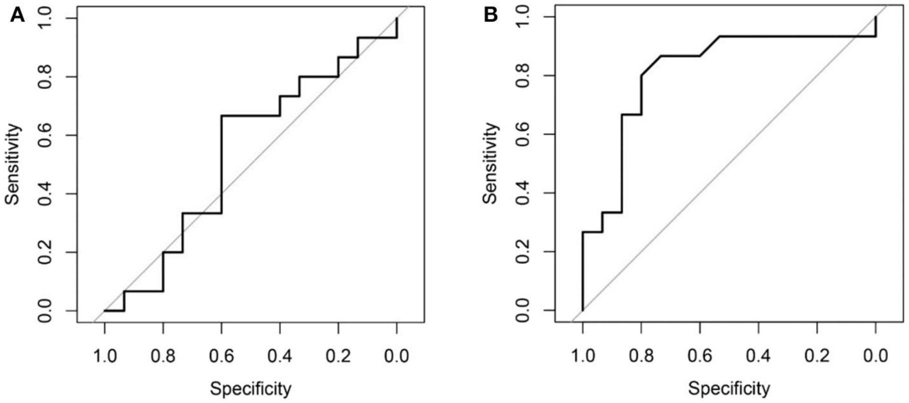


FIGURE 4 | (A) ROC curve for the test cohort for PyRadiomics model (AUC = 0.54). **(B)** ROC curve for the test cohort for Transfer Learning (LungTrans) model (AUC = 0.81).

this field where similar radiomic features have been reported to be prognostic of PDAC (Eilaghi et al., 2017; Chu et al., 2019; Khalvati et al., 2019a; Li et al., 2020). It is worth noting that morphologic features were not ranked as top features in the list. This may be attributed to the challenges associated with contouring the PDAC regions of interest, leading to the low robustness of morphology features.

Comparing the ROC curves using Delong ROC test (DeLong et al., 1988), the LungTrans (Transfer Learning) prognosis model had significantly higher performance than that of PyRadiomics feature bank with a p-value of 0.0056 (AUC of 0.81 vs. 0.54). This result indicated that the transfer learning model based on lung CT images (LungTrans) significantly improved the prognostic performance compared to that of the traditional radiomics methods (PyRadiomics). **Figure 4** shows the ROC curves for the two models for the test cohort.

Risk Score

In univariate Cox Proportional Hazard analysis, the risk score from the PyRadiomics model was not associated with overall survival. In contrast, the risk score from the LungTrans model had significant prognostic value with a Hazard Ratio of 1.86 [95% Confidence Interval (CI): 1.15–3.53], p-value: 0.04 as shown in **Table 3**.

Using the risk scores, patients can be categorized into low-risk or high-risk groups based on the median values. As shown in Kaplan-Meier plots in **Figure 5**, the LungTrans model was able to differentiate patients with high risk from those with low risk. This result further confirms that the transfer learning feature extractor pretrained by NSCLC CT images is capable of providing prognostic information for PDAC patients.

TABLE 3 | Performance of risk score models in Cox Proportional Hazard analysis.

	Hazard ratio and CI	p
PyRadiomics based risk score	1.03 (95% CI: 0.60–1.76)	0.91
Transfer learning based risk score	1.86 (95% CI: 1.15–3.53)	0.04

DISCUSSION

In this study, we developed and compared two prognostic models for overall survival for resectable PDAC patients using the PyRadiomics and transfer learning features banks pretrained by lung CT images (LungTrans). The LungTrans model achieved significantly better prognosis performance compared to that of the traditional radiomics approach (AUC of 0.81 vs. 0.54). This result suggested that the transfer learning approach has the potential of significantly improving prognosis performance in the resectable PDAC cohort using CT images.

Previous transfer learning studies in medical imaging research often utilized ImageNet pretrained models (Chuen-Kai et al., 2015; Lao et al., 2017). In our study, we used a lung CT pretrained CNN (LungTrans) as feature extractor and showed the potential of transfer learning in a typical small sample size setting. Although CNNs are capable of achieving high performance in image recognition tasks, training these networks needs a large sample size. If a CNN with the same architecture as LungTrans was trained from scratch in the training cohort (Cohort 1), it could not provide any prognostic value in the test cohort (Cohort 2) (AUC of ∼0.50). Transfer learning, unlike conventional deep learning methods which need large datasets, can achieve reasonable performance using a limited number of samples, making it suitable for most medical imaging studies. Although the training cohort in our study was small ($n = 68$), in the PDAC test cohort, our transfer learning model had positive predictive value (Precision) of 76%, demonstrating its prognostic value in finding high-risk patients. This may significantly benefit patients by providing personalized neoadjuvant or adjuvant therapy for better prognosis.

Although the proposed transfer learning model outperformed the conventional radiomics model, this was not an indication to discard radiomic features altogether. These hand-crafted features have been shown to be prognostic for survival and recurrence in different cancer sites (Kumar et al., 2012; Balagurunathan et al., 2014; Haider et al., 2017). In the PDAC radiomics field, more than forty features have been found to be significantly associated with tissue classification or overall survival for PDAC patients (e.g., sum entropy, cluster tendency, dissimilarity, uniformity,

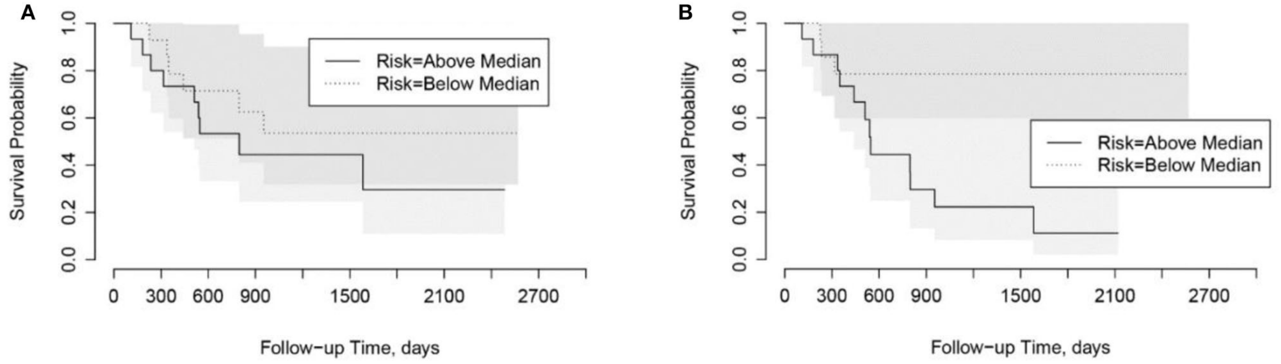


FIGURE 5 | Kaplan-Meier plots for overall survival in Cohort 2. **(A)** PyRadiomics based risk score ($p = 0.91$). **(B)** Transfer Learning (LungTrans) based risk score ($p = 0.04$).

and busyness) (Cassinotto et al., 2017; Chakraborty et al., 2017; Attiyeh et al., 2018; Yun et al., 2018; Chu et al., 2019; Sandrasegaran et al., 2019; Li et al., 2020; Park et al., 2020). Furthermore, a few radiomics features have been found to be associated with tumor heterogeneity and genomics profile (Lambin et al., 2012; Itakura et al., 2015; Rizzo et al., 2016; Li et al., 2018). Hence, radiomics features can provide unique information about the lesions. Thus, studying the associations between radiomics and transfer learning features, together with feature fusion analysis, may further improve the prognostication performance in future research.

Despite achieving promising results, we should also note that the differences between NSCLC and PDAC are substantial, in terms of their biological profiles and prognoses, and thus, they may not have similar appearances in CT images. This is a limitation of the present study. A larger PDAC dataset would allow us to address these differences and test different transfer learning approaches in the context of PDAC prognosis. For example, finetuning a few layers of the CNN pretrained by NSCLS CT images using PDAC CT images would allow the network extract features that may further adapt to the PDAC images and lead to better performance.

In this study, we aimed to improve the accuracy of the survival model using the transfer learning approach. For diseases with poor prognosis, including PDAC, providing binary survival classifications offers limited information for clinicians for decision making since the survival rates are usually low. It would be more beneficial to provide time vs. risk information, e.g., identify the high-risk time intervals for a resectable PDAC patient using CT images. Future studies may choose to combine the transfer learning-based features extraction methods with the recent work on deep learning-based survival models (e.g., DeepSurv Katzman et al., 2018) to provide more practical prognosis information for personalized care.

CONCLUSION

Deep transfer learning has the potential to improve the performance of prognostication for cancers with limited sample sizes such as PDAC. In this work, the proposed transfer learning model outperformed a predefined radiomics model for prognostications in resectable PDAC cohorts.

DATA AVAILABILITY STATEMENT

The datasets of Cohort 1 and Cohort 2 analyzed during the current study are available from the corresponding author on reasonable request pending the approval of the institution(s) and trial/study investigators who contributed to the dataset.

ETHICS STATEMENT

This study was reviewed and approved by the research ethics boards of University Health Network, Sinai Health System, and Sunnybrook Health Sciences Centre. For this retrospective study the informed consent was obtained for Cohort 1 and the need for informed consent was waived for Cohort 2.

AUTHOR CONTRIBUTIONS

YZ, MAH, and FK contributed to the design of the concept. EML, SG, PK, MAH, and FK contributed in collecting and reviewing the data. YZ and FK contributed to the design and implementation of quantitative imaging feature extraction and machine learning modules. All authors contributed to the writing and reviewing of the paper and read and approved the final manuscript.

FUNDING

This study was conducted with support of the Ontario Institute for Cancer Research (PanCuRx Translational Research Initiative) through funding provided by the Government of Ontario, the Wallace McCain Centre for Pancreatic Cancer supported by the Princess Margaret Cancer Foundation, the Terry Fox Research Institute, the Canadian Cancer Society Research Institute, and the Pancreatic Cancer Canada Foundation. This study was also supported by charitable donations from the Canadian Friends of the Hebrew University (Alex U. Soyka). The funding bodies had no role in the design of the study, collection, analysis, and interpretation of data, or in writing the manuscript.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at arXiv (Zhang et al., 2019).

REFERENCES

Adamska, A., Domenichini, A., and Falasca, M. (2017). Pancreatic ductal adenocarcinoma: current and evolving therapies. *Int. J. Mol. Sci.* 18:1338. doi: 10.3390/ijms18071338

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. doi: 10.1038/ncomms5006

Ahmad, N. A., Lewis, J. D., Ginsberg, G. G., Haller, D. G., Morris, J. B., Williams, N. N., et al. (2001). Long term survival after pancreatic resection for pancreatic adenocarcinoma. *Am. J. Gastroenterol.* 96, 2609–2615. doi: 10.1111/j.1572-0241.2001.04123.x

Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., et al. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* 38, 915–931. doi: 10.1118/1.3528204

Attiyeh, M. A., Chakraborty, J., Doussot, A., Langdon-Embry, L., Mainarich, S., Gönen, M., et al. (2018). Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis. *Ann. Surg. Oncol.* 25, 1034–1042. doi: 10.1245/s10434-017-6323-3

Balagurunathan, Y., Kumar, V., Gu, Y., Kim, J., Wang, H., Liu, Y., et al. (2014). Test–retest reproducibility analysis of lung CT image features. *J. Digit. Imaging* 27, 805–823. doi: 10.1007/s10278-014-9716-x

Carneiro, G., Oakden-Rayner, L., Bradley, A. P., Nascimento, J., and Palmer, L. (2016). "Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography," in *Autom. 5-year*

Mortal. Predict. Using Deep Learn. Radiomics Featur. from Chest Comput. Tomogr. doi: 10.1109/ISBI.2017.7950485

Cassinotto, C., Chong, J., Zogopoulos, G., Reinhold, C., Chiche, L., Lafourcade, J. P., et al. (2017). Resectable pancreatic adenocarcinoma: Role of CT quantitative imaging biomarkers for predicting pathology and patient outcomes. *Eur. J. Radiol.* 90, 152–158. doi: 10.1016/j.ejrad.2017.02.033

Chakraborty, J., Langdon-Embry, L., Cunanan, K. M., Escalon, J. G., Allen, P. J., Lowery, M. A., et al. (2017). Preliminary study of tumor heterogeneity in imaging predicts two year survival in pancreatic cancer patients. *PLoS ONE.* 12:e0188022. doi: 10.1371/journal.pone.0188022

Chakraborty, J., Midya, A., Gazit, L., Attiyeh, M., Langdon-Embry, L., Allen, P. J., et al. (2018). CT radiomics to predict high-risk intraductal papillary mucinous neoplasms of the pancreas. *Med. Phys.* 45, 5019–5029. doi: 10.1002/mp.13159

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics* 99, 323–329. doi: 10.1016/j.ygeno.2012.04.003

Chu, L. C., Park, S., Kawamoto, S., Fouladi, D. F., Shayesteh, S., Zinreich, E. S., et al. (2019). Utility of CT radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue. *Am. J. Roentgenol.* 213, 349–357. doi: 10.2214/AJR.18.20901

Chuen-Kai, S., Chung-Hisang, C., Chun-Nan, C., Meng-Hsi, W., and Edward, Y. C. (2015). "Transfer representation learning for medical image analysis," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

Coroller, T. P., Grossmann, P., Hou, Y., Rios Velazquez, E., Leijenaar, R. T., Hermann, G., et al. (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* 114, 345–350. doi: 10.1016/j.radonc.2015.02.015

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc*. 34, 187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x

Dalmiş, M. U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., et al. (2017). Using deep learning to segment breast and fibroglandular tissue in MRI volumes: *Med. Phys.* 44, 533–546. doi: 10.1002/mp.12079

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* 24, 1342–1350. doi: 10.1038/s41591-018-0107-6

De Wit, J. (2017). Kaggle datascience bowl 2017. *Github/kaggle_ndsb* (2017).

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Du, S. S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R., and Singh, A. (2018). "How many samples are needed to estimate a convolutional neural network?" in *Conference on Neural Information Processing Systems*.

Eilaghi, A., Baig, S., Zhang, Y., Zhang, J., Karanicolas, P., Gallinger, S., et al. (2017). CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma – a quantitative analysis. *BMC Med. Imaging* 17:38. doi: 10.1186/s12880-017-0209-5

Fatima, J., Schnelldorfer, T., Barton, J., Wood, C. M., Wiste, H. J., Smyrk, T. C., et al. (2010). Pancreatoduodenectomy for ductal adenocarcinoma: Implications of positive margin on survival. *Arch. Surg.* 145, 167–172. doi: 10.1001/archsurg.2009.282

Ferrone, C. R., Pieretti-Vanmarcke, R., Bloom, J. P., Zheng, H., Szymonifka, J., Wargo, J. A., et al. (2012). Pancreatic ductal adenocarcinoma: long-term survival does not equal cure. *Surgery* 152, S43–S49. doi: 10.1016/j.surg.2012.05.020

Haider, M. A., Vosough, A., Khalvati, F., Kiss, A., Ganeshan, B., Bjarnason, G. A., et al. (2017). CT texture analysis: a potential tool for prediction of survival in patients with metastatic clear cell carcinoma treated with sunitinib. *Cancer Imaging* 17:4. doi: 10.1186/s40644-017-0106-8

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Deep residual learning for image," in *Recognition* 770–77. doi: 10.1109/CVPR.2016.90

Hertel, L., Barth, E., Käster, T., and Martinetz, T. (2015). "Deep convolutional neural networks as generic feature extractors," in *2015 International Joint Conference on Neural Networks (IJCNN)*.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., et al. (2019). "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison," in *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Itakura, H., Achrol, A. S., Mitchell, L. A., Loya, J. J., Liu, T., Westbroek, E. M., et al. (2015). Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci. Transl. Med.* 7:303ra138. doi: 10.1126/scitranslmed.aaa7582

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18:24. doi: 10.1186/s12874-018-0482-1

Khalvati, F., Zhang, Y., Baig, S., Lobo-Mueller, E. M., Karanicolas, P., Gallinger, S., et al. (2019a). Prognostic value of CT radiomic features in resectable pancreatic ductal adenocarcinoma. *Nat. Sci. Reports*. 9:5449. doi: 10.1038/s41598-019-41728-7

Khalvati, F., Zhang, Y., Wong, A., and Haider, M. A. (2019b). Radiomics. *Encycloped Biomed Eng*. 2, 597–603. doi: 10.1016/B978-0-12-801238-3.99964-1

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: the process and the challenges. *Magn. Reson. Imaging* 30, 1234–1248. doi: 10.1016/j.mri.2012.06.010

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14, 749–762. doi: 10.1038/nrclinonc.2017.141

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 48, 441–446. doi: 10.1016/j.ejca.2011.11.036

Lao, J., Chen, Y., Li, Z. C., Li, Q., Zhang, J., Liu, J., et al. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci. Rep.* 7:10353. doi: 10.1038/s41598-017-10649-8

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, K., Yao, Q., Xiao, J., Li, M., Yang, J., Hou, W., et al. (2020). Contrast-enhanced CT radiomics for predicting lymph node metastasis in pancreatic ductal adenocarcinoma: a pilot study. *Cancer Imaging* 20:12. doi: 10.1186/s40644-020-0288-3

Li, Y., Qian, Z., Xu, K., Wang, K., Fan, X., Li, S., et al. (2018). MRI features predict p53 status in lower-grade gliomas via a machine-learning approach. *NeuroImage Clin.* 17, 306–311. doi: 10.1016/j.nicl.2017.10.030

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv*:1809.04430v1.

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Park, S., Chu, L., Hruban, R. H., Vogelstein, B., Kinzler, K. W., Yuille, A. L., et al. (2020). Differentiating autoimmune pancreatitis from pancreatic ductal adenocarcinoma with CT radiomics features. *Diagn. Interv. Imaging*. 1, 770–778. doi: 10.1016/j.diii.2020.03.002

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., and Aerts, H. J. W. L. (2015). Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* 5:13087. doi: 10.1038/srep13087

Paul, R., Schabath, M., Balagurunathan, Y., Liu, Y., Li, Q., Gillies, R., et al. (2019). Explaining deep features using radiologist-defined semantic features and traditional quantitative features. *Tomogr.* 5, 192–200. doi: 10.18383/j.tom.2018.00034

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., et al. (2017). "Understanding the mechanisms of deep transfer learning for medical images," in *Deep Learning and Data Labeling for Medical Applications*. 188–196. doi: 10.1007/978-3-319-46976-8_20

Rizzo, S., Petrella, F., Buscarino, V., De Maria, F., Raimondi, S., Barberis, M., et al. (2016). CT radiogenomic characterization of EGFR, K-RAS, and ALK mutations in non-small cell lung cancer. *Eur. Radiol.* 26, 32–42. doi: 10.1007/s00330-015-3814-0

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Ryu, S., Lee, H., Lee, D. K., Kim, S. W., Kim, C. E., Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Sandrasegaran, K., Lin, Y., Asare-Sawiri, M., Taiyini, T., and Tann, M. (2019). CT texture analysis of pancreatic cancer. *Eur. Radiol.* 29, 1067–1073. doi: 10.1007/s00330-018-5662-1

Schmidhuber, J. (2014). Deep learning in neural networks: an overview. *Neural Networks* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Stark, A., and Eibl, G. (2015). "Pancreatic Ductal Adenocarcinoma," in *Pancreapedia: The Exocrine Pancreas Knowledge Base*. Version 1.0 (Ann Arbor, MI: Michigan Publishing; University of Michigan Library), 1–9. doi: 10.3998/panc.2015.14

Stark, A. P., Ikoma, N., Chiang, Y. J., Estrella, J. S., Das, P., Minsky, B. D., et al. (2016). Long-term survival in patients with pancreatic ductal adenocarcinoma. *Surgery* 159, 1520–1527. doi: 10.1016/j.surg.2015.12.024

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., et al. (2018). "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning*. 270–279. doi: 10.1007/978-3-030-01424-7_27

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 32–3.

Torrey, L., and Shavlik, J. (2009). "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, eds E.S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, A. J. S. López.

Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* 77, e104–e107. doi: 10.1158/0008-5472.CAN-17-0339

Varghese, B., Chen, F., Hwang, D., Palmer, S. L., De Castro Abreu, A. L., Ukimura, O., et al. (2019). Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. *Sci. Rep.* 9:1570. doi: 10.1038/s41598-018-38381-x

Yamashita, R., Nishio, M., Do, R. K. G., Togashi, K., Yamashita, R., Nishio, M., et al. (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9, 611–629. doi: 10.1007/s13244-018-0639-9

Yasaka, K., Akai, H., Abe, O., Kiryu, S., Yasaka, K., Akai, H., et al. (2018). Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 286, 887–896. doi: 10.1148/radiol.2017170706

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27* (NIPS 2014).

Yun, G., Kim, Y. H., Lee, Y. J., Kim, B., Hwang, J. H., Choi, D. J., et al. (2018). Tumor heterogeneity of pancreas head cancer assessed by CT texture analysis: association with survival outcomes after curative resection. *Sci. Rep.* 8:7226. doi: 10.1038/s41598-018-25627-x

Zhang, J., Baig, S., Wong, A., Haider, M. A., and Khalvati, F. (2016). A local ROI-specific Atlas-based segmentation of prostate gland and transitional zone in diffusion MRI. *J. Comput. Vis. Imaging Syst.* 2, 2–4. doi: 10.15353/vsnl.v2i1.113

Zhang, Y., Lobo-Mueller, E. M., Karanicolas, P., Gallinger, S., Haider, M. A., Khalvati, F., et al. (2019). *Prognostic Value of Transfer Learning Based Features in Resectable Pancreatic Ductal Adenocarcinoma*. arXiv.

Zhang, Y., Oikonomou, A., Wong, A., Haider, M. A., and Khalvati, F. (2017). Radiomics-based prognosis analysis for non-small cell lung cancer. *Nat. Sci. Rep.* 7:46349. doi: 10.1038/srep46349

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.

The reviewer JZ declared a past co-authorship with one of the authors FK to the handling editor.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/frai.2020.539879

Population-Based Screening for Endometrial Cancer: Human vs. Machine Intelligence

Gregory R. Hart[1], Vanessa Yan[2], Gloria S. Huang[3], Ying Liang[1], Bradley J. Nartowt[1], Wazir Muhammad[1] and Jun Deng[1]*

[1]Department of Therapeutic Radiology, Yale University, New Haven, CT, United States, [2]Department of Statistics and Data Science, Yale University, New Haven, CT, United States, [3]Department of Obstetrics, Gynecology and Reproductive Sciences, Yale University, New Haven, CT, United States

Incidence and mortality rates of endometrial cancer are increasing, leading to increased interest in endometrial cancer risk prediction and stratification to help in screening and prevention. Previous risk models have had moderate success with the area under the curve (AUC) ranging from 0.68 to 0.77. Here we demonstrate a population-based machine learning model for endometrial cancer screening that achieves a testing AUC of 0.96.

We train seven machine learning algorithms based solely on personal health data, without any genomic, imaging, biomarkers, or invasive procedures. The data come from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). We further compare our machine learning model with 15 gynecologic oncologists and primary care physicians in the stratification of endometrial cancer risk for 100 women.

We find a random forest model that achieves a testing AUC of 0.96 and a neural network model that achieves a testing AUC of 0.91. We test both models in risk stratification against 15 practicing physicians. Our random forest model is 2.5 times better at identifying above-average risk women with a 2-fold reduction in the false positive rate. Our neural network model is 2 times better at identifying above-average risk women with a 3-fold reduction in the false positive rate.

Our machine learning models provide a non-invasive and cost-effective way to identify high-risk sub-populations who may benefit from early screening of endometrial cancer, prior to disease onset. Through statistical biopsy of personal health data, we have identified a new and effective approach for early cancer detection and prevention for individual patients.

Keywords: endometrial cancer, cancer screening, early detection, machine learning, statistical biopsy

OPEN ACCESS

Edited by:
Katrina M. Waters,
Pacific Northwest National Laboratory (DOE), United States

Reviewed by:
Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory (DOE), United States
Maria F. Chan,
Cornell University, United States

****Correspondence:***
Jun Deng
jun.deng@yale.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *02 March 2020*
Accepted: *29 October 2020*
Published: *24 November 2020*

Citation:
Hart GR, Yan V, Huang GS, Liang Y, Nartowt BJ, Muhammad W and Deng J (2020) Population-Based Screening for Endometrial Cancer: Human vs. Machine Intelligence. Front. Artif. Intell. 3:539879. doi: 10.3389/frai.2020.539879

INTRODUCTION

Endometrial cancer is the fourth most common cancer among women (Howlader et al., 2017). Symptoms such as bleeding or spotting often manifest early in the disease, resulting in the early detection of most cancers and a relatively high 5-years survival rate of 82% (American Cancer Society, 2017). The standard method for detecting endometrial cancer is endometrial biopsy, although transvaginal ultrasounds are sometimes used for detection as well (Smith et al., 2001;

Smith et al., 2018). Screening recommendations from the American Cancer Society (ACS) have remained constant since 2001 (Smith et al., 2018). Women with average or elevated risk are not recommended to get screened; instead, they should discuss with their doctor about the risks and symptoms of endometrial cancer at the onset of menopause. For very high-risk women such as those with Lynch syndrome, a high likelihood of being a mutation carrier, or families with suspected autosomal-dominant predisposition to colon cancer, ACS recommends annual screening (Smith et al., 2001).

While the 5-years survival rate for endometrial cancer is high, incidence and death rates of endometrial cancer have increased from 2010 (Howlader et al., 2017) and are expected to continue to increase (Arnold et al., 2015). It is expected that endometrial cancer will soon surpass ovarian cancer as the leading cause of gynecological cancer death. This has led to academic interest in improving endometrial cancer detection and prevention.

Two previous studies have been carried out to predict endometrial cancer risk (Pfeiffer et al., 2013; Hüsing et al., 2016). Both studies use traditional epidemiological models and non-invasive data for decision-making on targeted screening and preventive procedures. Hüsing *et al* trained a model on a dataset of 201,811 women (mostly aged 30–65 years), with 855 positive cases of endometrial cancer (0.4%). This model achieved an AUC of 0.77 (Hüsing et al., 2016). Pfeiffer et al's model (Pfeiffer et al., 2013), which produced an AUC of 0.68, was trained on the same PLCO dataset that we used, in addition to the NIH-AARP dataset. Their full dataset had a total of 304,950 women with 1,559 positive cases of endometrial cancer (0.51%). Noting the moderate performance of endometrial risk stratification models that were previously created (Hüsing et al., 2016), and the promising results of our previous work in using machine learning for cancer risk stratification (Hart et al., 2018; Roffman et al., 2018a; Roffman et al., 2018b; Muhammad et al., 2019), we decided to develop machine learning models to achieve greater performance in predicting endometrial cancer risk. We were able to surpass the performance of both these models with an AUC of 0.96.

Finally, a recent review suggests that a risk prediction model that divides the population up into low-, medium-, and high-risk groups would be useful for developing tailored cancer prevention strategies for each patient (Kitson et al., 2017). Such a model can help clinicians target high-risk populations, for whom clinicians could suggest interventions to modulate endometrial cancer risk, such as dietary and exercise changes, progestin or anti-estrogen therapy, insulin-lowering therapy, and scheduled endometrial biopsies. This is why we further applied our machine learning model to stratify patients into low-, medium, and high-risk groups. We compared our model's performance on the 3-tier risk stratification with physicians' judgment and achieved promising results.

METHODS

The PLCO Dataset

In this study we developed our machine learning models based on the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) dataset (Kramer et al., 1993). PLCO was a randomized, controlled trial investigating the effectiveness of various screenings for prostate, lung, colorectal, and ovarian cancers. It was a prospective study that enrolled participants from November 1993 through July 2001. Participants were between 55 and 75 years old. Shortly after enrollment, participants completed a baseline survey detailing their health history and current health condition. They were then followed until they were diagnosed with cancer or died, or when 13 years had passed. From the PLCO dataset, we sub-selected the 78,215 female participants for whom we have data on whether they developed endometrial cancer within 5 years of enrolling in the PLCO trial. 961 of these females developed endometrial cancer within five years of enrolling. This gave 77,254 non-cancer cases (98.8%) and 961 cancer cases (1.2%) on which we train our model. For full details about this data and its collection see Kramer et al., 1993.

With authorization from the National Cancer Institute (NCI) to access PLCO trial data (PLCO-365), we used the following inputs for our model: age (Howlader et al., 2017), BMI(Renehan et al., 2008; Crosbie et al., 2010), weight (20 years, 50 years, present) (Hosono et al., 2011; Aune et al., 2015), race (Howlader et al., 2017), smoking habits (Zhou et al., 2008), diabetes (Anderson et al., 2001), emphysema, stroke, hypertension (Aune et al., 2017), heart disease, arthritis (Parikh-Patel et al., 2009), another cancer, family history of breast, ovary, and endometrial cancer, ovarian surgery (Dossus et al., 2010), menarche age (Dossus et al., 2010), parity (Dossus et al., 2010), use of birth control (Dossus et al., 2010), and age at menopause (Dossus et al., 2010). Many of these inputs, such as BMI, diabetes and family history, were selected because they correlate strongly with endometrial cancer incidence (Anderson et al., 2001; Dossus et al., 2010; Aune et al., 2015) and were also used as inputs in other works on endometrial cancer risk prediction (Pfeiffer et al., 2013; Hüsing et al., 2016; Kitson et al., 2017). Other factors, such as smoking habits, emphysema and heart disease, were included because they contributed to good performance in our past works. (Hart et al., 2018; Roffman et al., 2018a; Roffman et al., 2018b; Muhammad et al., 2019). There are other known risk factors such as Hereditary Non-polyposis Colorectal Cancer (HNPCC) or Lynch Syndrome which would be good to include in a model but are not in the PLCO dataset. All inputs were scaled to within the range of [0, 1].

To evaluate the different algorithms, we randomly split the dataset 70%/30% into training and testing sets, keeping the proportion of those with and without cancer constant between the two datasets. The final model was trained on the full training set and evaluated on the testing set. This gives our model a Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) level 2a of robustness (Collins et al., 2015).

Machine Learning Algorithms

In creating our risk prediction models, we trained algorithms that produce continuous output from 0 to 1, which indicated the probability that a woman would develop endometrial cancer within five years since the input data was gathered. The algorithms we used were: logistic regression (LR), neural network (NN), support vector machine (SVM), decision tree (DT), random forest (RF), linear discriminant analysis (LDA), and naive Bayes (NB) (Bishop, 2006). The logistic regression was fit using the NN code with 0 hidden layers. The NN was fit using the in-house MATLAB code we developed for previous works (Hart et al., 2018; Roffman et al., 2018a; Roffman et al., 2018b; Muhammad et al., 2019). It was a multilayer perceptron consisting of two hidden layers with 12 neurons each and a logistic activation function. We then used the built-in MATLAB function "fitrsvm" with a Gaussian kernel to fit the SVM, and we used the function "fitctree" to create the decision tree. The random forest was fit with the built-in MATLAB function "TreeBagger" with 50 trees. The LDA was fit using the built-in MATLAB function "fitcdiscr," with "discrimType" set to "diaglinear". Lastly, the NB was fit using the built-in MATLAB function "fitcnb," with "OptimizeHyperparameters" set to "auto". For "fitctree", "fitcdiscr", and "fitcnb," the "score" from the "predict" function was used to get continuous output, akin to that returned by the LR, NN, and SVM. We used these six algorithms because they are well-established machine learning techniques.

In selecting the algorithm for the final model(s), we used 10-fold cross-validation within the training and testing sets to determine the mean AUC of each algorithm. We identified the two models that achieved the highest testing mean AUCs between training and testing performance. These two models were then trained on the full training set and evaluated on the testing set. Afterward, for each of the two best models, we selected a threshold for determining the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), by maximizing the sum of the training sensitivity and specificity, i.e., maximizing the balanced accuracy.

Risk Stratification

Once we selected the two best models, we used them to stratify the population into below, at, or above-average risk, to facilitate a comparison of our models' prediction to physicians' judgment in the clinic. Specifically, in selecting the boundaries based on the training data, we considered the bottom 15.9% of risks as below average, the top 15.9% of risks as above average, and the middle 68.2% as average risks.

Human Intelligence (HI) vs. Artificial Intelligence (AI)

For comparison of the models' prediction against physicians' judgment, we created an online survey (https://yalesurvey.ca1.qualtrics.com/jfe/form/SV_3TVh1XP27eaktud) with a sub-data set of 100 women from our original dataset. The survey presented the information used by our model to physicians and asked them to rate each woman as below, at, or above-average for endometrial cancer risk. Clinicians were given no instructions on how to classify individuals, so that we would get results representative of what would happen in practice. In an effort to get high-quality data, we limited the length of the survey by only showing each physician a random subset of 20 of the 100 women. The answers from the various physicians were aggregated and averaged for each woman. We then used our model to stratify the same group of women. We invited physicians from Yale, Harvard, and University of Michigan Departments of Obstetrics, Gynecology, and Reproductive Science/Biology, as well as primary care physicians from INOR Cancer Hospital (Abbottabad, Pakistan) and Yale Health Center to participate. We received usable responses from 15 physicians.

RESULTS

We evaluated seven different algorithms: logistic regression (LR), neural network (NN), support vector machine (SVM), decision tree (DT), random forest (RF), linear discriminant analysis (LDA), and naïve Bayes (NB). **Table 1** presents the mean average area under the receiver operating characteristic (ROC) curve (AUC) with one standard deviation, from the 10-fold cross-validation on both the training and testing datasets. The training AUCs range from 0.68 to 0.99 and the testing AUCs range from 0.68 to 0.95. There is no significant difference in the training and testing performance for four of the algorithms (LR, NN, LDA and NB), but SVM, DT, and RF have a significant drop in performance going from training to testing. The highest testing performance was for the RF, although NN and RF testing performance are within one standard deviation of each other. For the remainder of this paper we will be focusing on the random forest and neural network models because these two models achieved the highest mean testing AUCs during cross-validation.

Selecting the random forest and neural network as the top models, we then trained them on the full training dataset and evaluated them on the testing dataset. When calculating the models' performance on both the training and testing datasets, we calculated 95% confidence intervals of the AUC, sensitivity, specificity, PPV, and NPV (Hanley and McNeil, 1982).

Figure 1A shows the sensitivity and specificity as a function of the decision threshold for the random forest on both the training and testing datasets. The same is done for the neural network in **Figure 1B**. For the random forest, the sensitivity decreases as the threshold value increases, while the specificity is above 99.9% on both the training and testing datasets. For the neural network model, the sensitivity hovers around 60% and the specificity remains above 99.9% for most threshold values on both the training and testing datasets. Given a threshold value that maximizes the sum of the training sensitivity and specificity, the random forest model's sensitivity is 98.4% for the training set and 75.7% for testing. The specificity is 98.9 and 98.3% for training and testing respectively. The neural network model's sensitivity is 77.2% for the training set and 67.7% for testing. The specificity is 91.2 and 91.1% for training and testing, respectively.

Using the sensitivity, specificity, and prevalence of endometrial cancer, we calculate the PPV and NPVs as well. For random forest, the PPV is 28.2 and 16.3% for the training and testing datasets respectively. The NPV is 99.9 and 99.9% for training and testing respectively. For the neural network, the PPV

TABLE 1 | Mean AUC (standard deviation) over 10 cross-validation folds for the seven algorithms tested.

	LR	NN	SVM	DT	RF	LDA	NB
Training	0.68 (0.11)	**0.89 (0.05)**	0.99 (0.00)	0.98 (0.00)	**0.99 (0.01)**	0.81 (0.00)	0.72 (0.12)
Testing	0.68 (0.10)	**0.88 (0.07)**	0.80 (0.03)	0.85 (0.04)	**0.95 (0.01)**	0.81 (0.03)	0.72 (0.12)

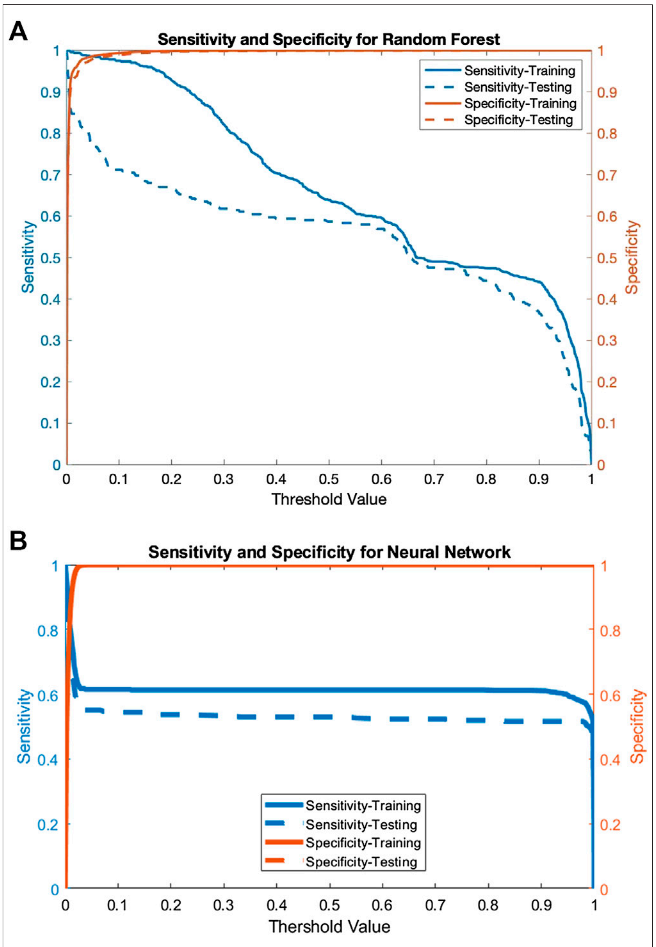


FIGURE 1 | A) The sensitivity and specificity of the random forest for both the training and testing data as a function of the threshold value and **(B)** The sensitivity and specificity of the neural network for both the training and testing data as a function of the threshold value.

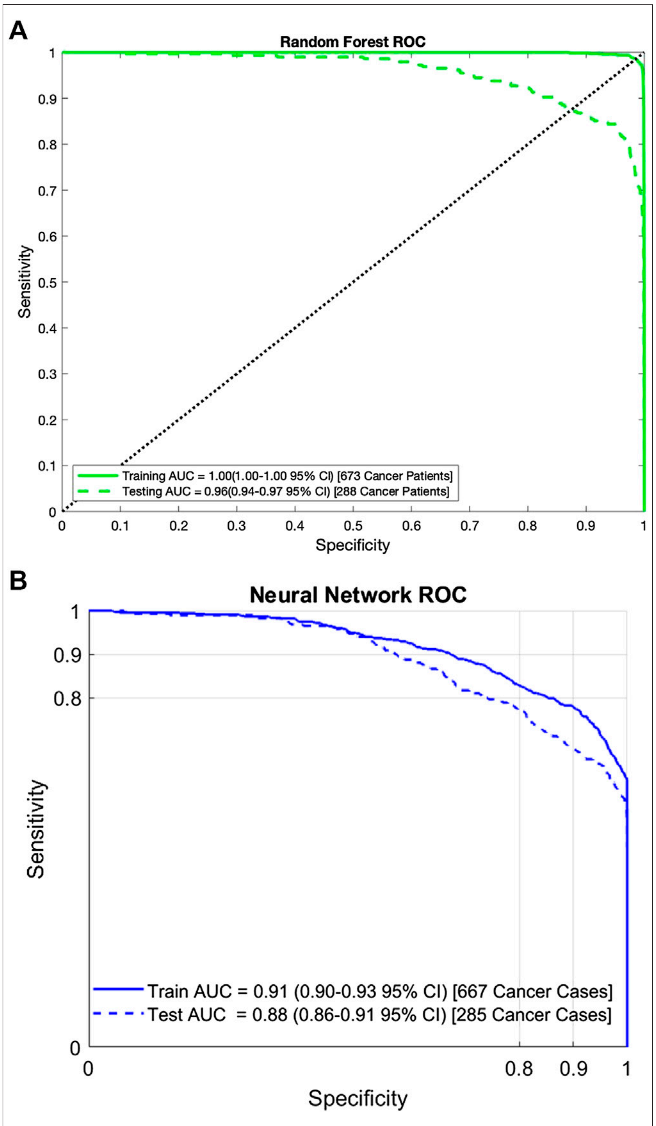


FIGURE 2 | A) Area under the ROC curve for the random forest on both the training and testing data. Similar performance on both datasets indicates that the random forest has no overfit and **(B)** Area under the ROC curve for the neural network on both the training and testing data. Similar performance on both datasets indicates that the neural network has no overfit.

is 3.8 and 3.3%, and the NPV is 99.9 and 99.8%, for the training and testing datasets respectively. The ROC curves for the random forest and neural network are shown in **Figures 2A,B**. For the random forest, the AUC for training and testing are, respectively, 0.99 (95% CI: 0.99–1.00) and 0.96 (95% CI: 0.94–0.97). For the neural network, the AUC for the training data is 0.91 (95% CI: 0.90–0.93) and for testing it is 0.88 (95% CI: 0.86–0.91).

Following the recommendation of Ref 8, we used the random forest and neural network models to create a 3-tiered risk stratification scheme. Based on the risk boundaries selected using the training data, we stratified the testing data into three groups: below, at, and above-average risk. **Figures 3A,B** show Kaplan-Meier plots for these three groups over the full 13 years they were followed. The figures clearly show that women classified as above-average risk have the highest chance of developing endometrial cancer. This is supported further by the hazard ratio (HR) between the above-average group and the two other groups.

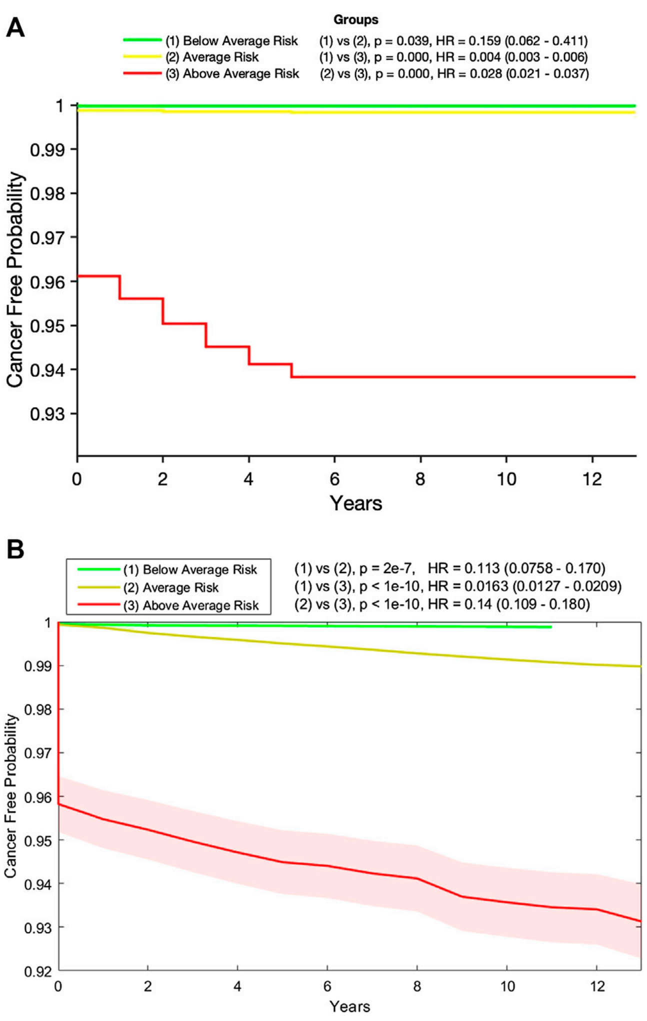


FIGURE 3 | (A) Kaplan-Meier plot of the below- (green), at- (yellow), and above- (red) average risk groups created from the testing data by our random forest model. Also shown are the *p*-value and hazard ratio (HR) between each group. Those in the above-average risk group clearly have the highest chance of developing cancer and **(B)** Kaplan-Meier plot of the below- (green), at- (yellow), and above- (red) average risk groups created from the testing data by our neural network model with 95% confidence intervals (shaded). Also shown are the *p*-value and hazard ratio (HR) between each group. Those in the above-average risk group clearly have the highest chance of developing cancer.

In fact, as shown in **Table 2**, 90.3% of those in the testing set who developed endometrial cancer during the next 5 years were labeled by the random forest model as above-average risk and 15.7% of those who did not develop cancer were labeled as below-average risk. The incidence rates in the below-average, average, and above-average risk groups are 0.03, 0.17, and 6.17%, respectively. Similar performances were observed for the neural network as shown in Table 3

TABLE 2 | Stratifying the testing data into three risk groups by the random forest.

	Below average risk		Average risk		Above average risk	
	Number	%	Number	%	Number	%
Cancer	1	0.3	27	9.4	260	90.3
No cancer	3,628	15.7	15,592	67.3	3,956	17.1

The percentages in each row sum up to 1.

TABLE 3 | Stratifying the testing data into three risk groups by the neural network.

	Below average risk		Average risk		Above average risk	
	Number	%	Number	%	Number	%
Cancer	3	1.0	76	26.7	206	72.3
No cancer	3,705	16.0	15,920	68.7	3,553	15.3

The percentages in each row sum up to 1.

Tables 4, **5** show the comparison of our models with practicing clinicians in assessing risk for 100 women. In the below-average risk population, the physicians identified 2.8 times as many women who did not develop cancer as being below-average risk, compared to our random forest model (39.5 vs. 14.0%), and 1.65 times as many as our neural network model (39.5 vs. 24.0%). However, the physicians misidentified twice as many women who did not develop cancer as being high risk, compared to the random forest model (27.9 vs. 14.0%), and 3.5 times as many compared to the neural network (27.9 vs. 8.0%). Furthermore, our model was much better than physicians at aptly stratifying patients who would develop endometrial cancer. Physicians misidentified 22% of those who did develop cancer as having below average risk, whereas our random forest and neural network models predicted none. Additionally, compared to physicians' predictions, our random forest model identified 2.5 times as many women who did develop cancer (94.0 vs. 38.0%) as above-average risk, and our neural network model identified almost twice as many as the physicians did (70.0 vs. 38.0%). Finally, there is a large inter-observer variability on the physicians' assessments, while our models return the same predictions every time.

DISCUSSION

We created seven different models to predict the probability of an individual woman developing endometrial cancer in five years based on readily available personal health data. Of these seven models we found that the random forest model performed best in terms of testing AUC, and the neural network performed second best. We then used both models to stratify the population into three risk categories. The above-average risk category captured the majority of those who developed cancer in five years. This above-average risk population could benefit from regular screening procedures such as endometrial biopsy and/or transvaginal ultrasounds.

Of our seven models, logistic regression and naive Bayes performed the worst and had the most variation between cross-validation folds. We think that the relatively poor performance of logistic regression and naive Bayes is due to

TABLE 4 | Random forest model vs. physician stratification of 50 women with cancer (ground truth positives) and 50 women without cancer (ground truth negatives) into below-, at-, or above-average risk groups.

	Below average risk		Average risk		Above average risk	
	Model	**Physicians**	**Model**	**Physicians**	**Model**	**Physicians**
Ground truth positives	0.0	22.0% (17%)	6.0	40.0% (16%)	94.0	38.0% (24%)
Ground truth negatives	14.0	39.5% (22%)	72.0	32.6% (16%)	14.0	27.9% (20%)

Inter-observer variability for the physicians is captured by a standard deviation in parenthesis.

TABLE 5 | Neural network model vs. physician stratification of 50 women with cancer (ground truth positives) and 50 women without cancer (ground truth negatives) into below-, at-, or above-average risk groups.

	Below average risk		Average risk		Above average risk	
	Model	**Physicians**	**Model**	**Physicians**	**Model (%)**	**Physicians**
Ground truth positives	0.0	22.0% (17%)	30.0	40.0% (16%)	70.0	38.0% (24%)
Ground truth negatives	24.0	39.5% (22%)	68.0	32.6% (16%)	8.0	27.9% (20%)

Inter-observer variability for the physicians is captured by a standard deviation in parenthesis.

the lack of interaction terms in these models. Without interactions between the input factors, these models have no advantage over traditional epidemiological models. A neural network with at least one hidden layer allows for mixing of the input parameters, which may explain its outperforming the other algorithms we tested.

Four of the models (LR, NN, LDA and NB) generalized well with similar training and testing AUCs, while SVM and DT overfit on the training data. Even though SVM, DT and RF achieved near-perfect AUC on the training data, they still performed better on the testing data than previous works; a phenomenon we also saw with lung cancer (Hart et al., 2018). The neural network achieved an AUC of 0.88 on both the 10-fold cross-validation and the testing set. The random forest achieved a testing AUC of 0.96. Both our random forest and neural network models significantly outperformed two previous risk prediction models, including the model introduced by Pfeiffer et al which achieved an AUC of 0.68 (Pfeiffer et al., 2013). This improvement is particularly interesting because Pfeiffer et al trained their model on not only the PLCO data, but also data from the National Institutes of Health-AARP Diet and Health Study. Although our model outperforms their model, theirs is more robust since it has been validated on an external data set, making it TRIPOD level 3 compared to our level 2a. Another previous work, by Hüsing *et al*, achieved an AUC of 0.77 (Hüsing et al., 2016). Their improvement was made by explicitly adding interaction terms to the epidemiological model. They used cross-validation making it TRIPOD level 1b. We are seeking access to the datasets used in these other works as external testing on our model.

With our random forest and neural network models outperforming previous works, we turn our attention to comparing our models with clinical judgment. The ultimate goal of this and our previous work is to create a risk prediction tool that can support physicians in their clinical decision-making about cancer prevention and screening for individuals prior to disease onset. In stratifying 100 women into below-, at-, and above-average risk groups, the physicians' true negative rate in the below-average group was 1.6 times better than that of our neural network model (39.5 vs. 24.0%). However, physicians' judgment resulted in a worse false negative rate in the below-average group (22 vs. 0%) and lower true positive rate in the above-average group, compared to both our random forest and neural network models. Thus, we have shown that our machine learning models are better than practicing physicians at identifying high-risk above average risk women.

While current guidelines only recommend screening for very high-risk women, our models may be capable of identifying a larger population who would benefit from screening. In fact, when stratifying the population based on stricter criteria (Hart et al., 2018; Roffman et al., 2018a; Roffman et al., 2018b; Muhammad et al., 2019) than what was used in this paper, our neural network model identifies a high-risk group in which 47% of women developed endometrial cancer within 5 years, among whom most developed the cancer under a year (data not shown). In addition to informing women and their physicians in their discussion of the potential pros and cons of screening, our models can help prompt life-style changes and other preventive measures or intervention (see Arnold et al., 2015). Admittedly, the downside to our models for this application is that understanding the contribution of individual input factors to the overall risk is not intuitive. So, while the current model can stratify the population and suggest the above-average risk group to participate in preventive strategies, it does not offer much help in deciding which strategies (e.g., diet and exercise, progestin or anti-estrogen therapy, and insulin-lowering therapy etc.) would be most effective. We will carry out this study in our future works. Nevertheless, our machine learning approach shows great promise in aiding early detection of endometrial cancer, as the approach yields high-accuracy predictions based solely on personal health information prior to disease onset, without need for any invasive or costly procedures like endometrial biopsy or transvaginal ultrasounds. Furthermore, it could be integrated into existing electronic medical record systems, giving risk predictions directly to primary care physicians when they see patients.

Compared with clinical judgment, the strong performance of our models, combined with other strongly discriminatory models for non-melanoma skin cancer (Roffman et al., 2018a), prostate cancer (Roffman et al., 2018b), lung cancer (Hart et al., 2018), and pancreatic cancer (Muhammad et al., 2019), presents a real opportunity to perform a "statistical biopsy" on individuals prior to disease onset. Analogous to traditional biopsy, which analyzes cells from a specimen, and the recently developed liquid biopsy, which evaluates circulating DNA from a blood sample to diagnose cancer, our machine learning approach to cancer prediction is essentially a statistical biopsy that mines personal health data of an individual for early cancer detection and prevention. Different from traditional biopsy and liquid biopsy, statistical biopsy seeks to decipher the invisible correlations and inter-connectivity between multiple medical conditions and health parameters via statistical modeling. By mining personal health data via statistical biopsy, it is possible to generate a holistic profile of an individual's risk for a variety of cancer types, with little cost in time or money and no side effects. Furthermore, if integrated into a modern electronic medical record (EMR) system, statistical biopsy may help inform preventive interventions and/or screening decisions in real time. As we test our models on external datasets and expand the types of cancer covered, we hope to build a comprehensive model available to primary care physicians worldwide, allowing for statistical biopsies during routine clinical care for the general public.

CONCLUSION

In this work we construct machine learning models to predict the five-year risk of developing endometrial cancer for individual women based solely on personal health data, without any genomic or imaging biomarkers, or invasive procedures. We test seven different algorithms and find that the random forest performs optimally and outperforms previous models. We further demonstrate that the random forest is superior to the 15 physicians in stratifying the population into three risk groups, with a 2.5-fold increase in true positive rate, 2-fold reduction in false positive rate, and reduction to zero in false negative rate. With strong discriminatory power, our random forest offers a cost-effective and non-invasive method to population-based screening for endometrial cancer prior to disease onset and is capable of targeting the sub-population with above-average risk. The ability to identify female patients with above-average risk can in turn inform the adoption of early cancer prevention strategies, including both immediate actions like screening and long-term preventative measures such as chemoprevention.

DATA AVAILABILITY STATEMENT

The data analyzed in this study was obtained from the NCI database https://cdas.cancer.gov/plco/ (PLCO-392). Requests to access the PLCO datasets should be directed to NCI Cancer Data Access System (CDAS) at cdas@imsweb.com.

AUTHOR CONTRIBUTIONS

GRH analyzed data, produced results, and wrote technical details and the manuscript. VY wrote parts of the manuscript and the code, generated results for a new RF algorithm, and responded to reviewers comments. GRH and VY made equal contributions to this work. GSH contributed to the study design, provided clinical consultation and interpretation of results, and wrote parts of the manuscript. YL did preliminary data exploration, provided technical consultation, and reviewed the manuscript. BN and WM provided technical consultation and reviewed the manuscript. JD generated research ideas, contributed to the study design, provided technical consultation, and reviewed the manuscript.

FUNDING

This study was supported by a National Institutes of Health (NIH) National Institute of Biomedical Imaging and Bioengineering Award (R01EB022589) to JD. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

ACKNOWLEDGMENTS

We thank Caren M. Stalburg, Richard W. Lieberman, Farivaz Seifi, Sajid Zaman, and the other 11 physicians from Yale, Harvard, University of Michigan, and INOR Cancer Hospital who participated in our online survey.

REFERENCES

American Cancer Society (2017). Cancer facts and figures 2017. Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf.

Anderson, K. E., Anderson, E., Mink, P. J., Hong, C. P., Kushi, L. H., Sellers, T. A., et al. (2001). Diabetes and endometrial cancer in the Iowa women's health study. *Cancer Epidemiol. Biomarkers Prev.* 10, 611–616.

Arnold, M., Pandeya, N., Byrnes, G., Renehan, A. G., Stevens, G. A., Ezzati, M., et al. (2015). Global burden of cancer attributable to high body-mass index in 2012: a population-based study. *Lancet Oncol.* 16, 36–46. doi:10.1016/s1470-2045(14)71123-4.

Aune, D., Sen, A., and Vatten, L. J. (2017). Hypertension and the risk of endometrial cancer: a systematic review and meta-analysis of case-control and cohort studies. *Sci. Rep.* 7, 44808. doi:10.1038/srep44808.

Aune, D., Navarro Rosenblatt, D. A., Chan, D. S. M., Vingeliene, S., Abar, L., Vieira, A. R., et al. (2015). Anthropometric factors and endometrial cancer risk: a systematic review and dose-response meta-analysis of prospective studies. *Ann. Oncol.* 26, 1635–1648. doi:10.1093/annonc/mdv142.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin, Germany: Springer, 738.

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual

Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* 162, 55–63. doi:10.7326/m14-0697.

Crosbie, E. J., Zwahlen, M., Kitchener, H. C., Egger, M., and Renehan, A. G. (2010). Body mass index, hormone replacement therapy, and endometrial cancer risk: a meta-analysis. *Cancer Epidemiol. Biomark. Prev.* 19, 3119–3130. doi:10.1158/1055-9965.epi-10-0832.

Dossus, L., Allen, N., Kaaks, R., Bakken, K., Lund, E., Tjonneland, A., et al. (2010). Reproductive risk factors and endometrial cancer: the European prospective investigation into cancer and nutrition. *Int. J. Cancer.* 127, 442–451. doi:10.1002/ijc.25050

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi:10.1148/radiology.143.1.7063747.

Hart, G. R., Roffman, D. A., Decker, R., and Deng, J. (2018). A multi-parameterized artificial neural network for lung cancer risk prediction. *PloS One.* 13, e0205264. doi:10.1371/journal.pone.0205264.

Hosono, S., Matsuo, K., Hirose, K., Ito, H., Suzuki, T., Kawase, T., et al. (2011). Weight gain during adulthood and body weight at age 20 are associated with the risk of endometrial cancer in Japanese women. *J. Epidemiol.* 21, 466–473. doi:10.2188/jea.je20110020.

Howlader, N., Noone, A., Krapcho, M., Miller, D., Bishop, K., Kosary, C. L., et al. (2017). SEER cancer statistics review, 1975-2014. Available from: https://seer.cancer.gov/csr/1975_2014/ (Accessed May 26, 2007).

Hüsing, A., Dossus, L., Ferrari, P., Tjønneland, A., Hansen, L., Fagherazzi, G., et al. (2016). An epidemiological model for prediction of endometrial cancer risk in Europe. *Eur. J. Epidemiol.* 31, 51–60. doi:10.1007/s10654-015-0030-9.

Kitson, S. J., Evans, D. G., and Crosbie, E. J. (2017). Identifying high-risk women for endometrial cancer prevention strategies: proposal of an endometrial cancer risk prediction model. *Cancer Prev. Res.* 10, 1–13. doi:10.1158/1940-6207.capr-16-0224.

Kramer, B. S., Gohagan, J., Prorok, P. C., and Smart, C. (1993). A National Cancer Institute sponsored screening trial for prostatic, lung, colorectal, and ovarian cancers. *Cancer* 71, 589–593. doi:10.1002/cncr.2820710215.

Muhammad, W., Hart, G. R., Nartowt, B. J., Farrell, J. J., Johung, K., Liang, Y., et al. (2019). Pancreatic cancer prediction through an artificial neural network. *Front. Artif. Intell.* 2, 2. doi:10.3389/frai.2019.00002.

Parikh-Patel, A., White, R. H., Allen, M., and Cress, R. (2009). Risk of cancer among rheumatoid arthritis patients in California. *Cancer Causes Control.* 20, 1001–1010. doi:10.1007/s10552-009-9298-y.

Pfeiffer, R. M., Park, Y., Kreimer, A. R., Lacey, J. V., Pee, D., Greenlee, R. T., et al. (2013). Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. *PLoS Med.* 10, e1001492. doi:10.1371/journal.pmed.1001492.

Renehan, A. G., Tyson, M., Egger, M., Heller, R. F., and Zwahlen, M. (2008). Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* 371, 569–578. doi:10.1016/s0140-6736(08)60269-x.

Roffman, D., Hart, G. R., Girardi, M., Ko, C. J., and Deng, J (2018a). Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Sci. Rep.* 8, 1701. doi:10.1038/s41598-018-19907-9.

Roffman, D. A., Hart, G. R., Leapman, M. S., Yu, J. B., Guo, F. L., Ali, I., et al. (2018b). Development and validation of a multiparameterized artificial neural network for prostate cancer risk prediction and stratification. *JCO Clin. Cancer Inform.* 2, 1–10. doi:10.1200/CCI.17.00119.

Smith, R. A., Andrews, K. S., Brooks, D., Fedewa, S. A., Manassaram-Baptiste, D., Saslow, D., et al. (2018). Cancer screening in the United States, 2018: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA A Cancer J. Clin.* 68, 297–316. doi:10.3322/caac.21446.

Smith, R. A., von Eschenbach, A. C., Wender, R., Levin, B., Byers, T., Rothenberger, D., et al. (2001). American cancer society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers: also: update 2001–testing for early lung cancer detection. *CA A Cancer J. Clin.* 51, 38–75. doi:10.3322/canjclin.51.1.38.

Zhou, B., Yang, L., Sun, Q., Cong, R., Gu, H., Tang, N., et al. (2008). Cigarette smoking and the risk of endometrial cancer: a meta-analysis. *Am. J. Med.* 121, 501–508. doi:10.1016/j.amjmed.2008.01.044.

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.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 11 February 2021
doi: 10.3389/frai.2020.614384

Clinical Enhancement in AI-Based Post-processed Fast-Scan Low-Dose CBCT for Head and Neck Adaptive Radiotherapy

Wen Chen[1,2], Yimin Li[2,3], Nimu Yuan[4], Jinyi Qi[4], Brandon A. Dyer[5], Levent Sensoy[2], Stanley H. Benedict[2], Lu Shang[2], Shyam Rao[2] and Yi Rong[2,6]**

[1]Department of Radiation Oncology, Xiangya Hospital, Central South University, Changsha, China, [2]Department of Radiation Oncology, University of California Davis Medical Center, Sacramento, CA, United States, [3]Department of Radiation Oncology, Xiamen Cancer Center, The First Affiliated Hospital of Xiamen University, Xiamen, China, [4]Department of Biomedical Engineering, University of California, Davis, CA, United States, [5]Department of Radiation Oncology, University of Washington, Seattle, WA, United States, [6]Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix, AZ, United States

Purpose: To assess image quality and uncertainty in organ-at-risk segmentation on cone beam computed tomography (CBCT) enhanced by deep-learning convolutional neural network (DCNN) for head and neck cancer.

Methods: An in-house DCNN was trained using forty post-operative head and neck cancer patients with their planning CT and first-fraction CBCT images. Additional fifteen patients with repeat simulation CT (rCT) and CBCT scan taken on the same day (oCBCT) were used for validation and clinical utility assessment. Enhanced CBCT (eCBCT) images were generated from the oCBCT using the in-house DCNN. Quantitative imaging quality improvement was evaluated using HU accuracy, signal-to-noise-ratio (SNR), and structural similarity index measure (SSIM). Organs-at-risk (OARs) were delineated on o/eCBCT and compared with manual structures on the same day rCT. Contour accuracy was assessed using dice similarity coefficient (DSC), Hausdorff distance (HD), and center of mass (COM) displacement. Qualitative assessment of users' confidence in manual segmenting OARs was performed on both eCBCT and oCBCT by visual scoring.

Results: eCBCT organs-at-risk had significant improvement on mean pixel values, SNR ($p < 0.05$), and SSIM ($p < 0.05$) compared to oCBCT images. Mean DSC of eCBCT-to-rCT (0.83 ± 0.06) was higher than oCBCT-to-rCT (0.70 ± 0.13). Improvement was observed for mean HD of eCBCT-to-rCT (0.42 ± 0.13 cm) vs. oCBCT-to-rCT (0.72 ± 0.25 cm). Mean COM was less for eCBCT-to-rCT (0.28 ± 0.19 cm) comparing to oCBCT-to-rCT (0.44 ± 0.22 cm). Visual scores showed OAR segmentation was more accessible on eCBCT than oCBCT images.

Conclusion: DCNN improved fast-scan low-dose CBCT in terms of the HU accuracy, image contrast, and OAR delineation accuracy, presenting potential of eCBCT for adaptive radiotherapy.

Keywords: deep convolutional neural network, image quality, cone beam CT, head and neck cancer, adaptive radiotherapy

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Yuliang Huang,
Peking University Cancer Hospital, China
Rui Zhang,
Louisiana State University, United States

****Correspondence:***
Shyam Rao
sdrao@ucdavis.edu
Yi Rong
rong.yi@mayo.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *06 October 2020*
Accepted: *28 December 2020*
Published: *11 February 2021*

Citation:
Chen W, Li Y, Yuan N, Qi J, Dyer BA, Sensoy L, Benedict SH, Shang L, Rao S and Rong Y (2021) Clinical Enhancement in AI-Based Post-processed Fast-Scan Low-Dose CBCT for Head and Neck Adaptive Radiotherapy.
Front. Artif. Intell. 3:614384.
doi: 10.3389/frai.2020.614384

INTRODUCTION

Head and neck cancer (HNC) is reported as the eighth leading cause of cancer-related death worldwide (Parkin et al., 2005). HNC can have heterogeneous responses to definitive chemoradiotherapy regarding locoregional control and overall survival (Yan et al., 2012). Anatomic changes due to tumor response or weight loss may lead to under- or over-dosage to target volumes or overdosage to organs at risk (OARs) during radiotherapy. Changes in the plan dosimetry may result in increased risk of toxicity and/or impact tumor control (Chen et al., 2014; Castelli et al., 2015). In recent years, adaptive radiation therapy (ART) has been proposed to account for changes in tumor and normal organs to enhance the therapeutic ratio (Castadot et al., 2010; Schwartz, 2012). However, ART requires re-segmentation of OARs and treatment target volumes on each re-planning CT image. This process, if performed manually, is time-consuming with high intra- and inter-observer segmentation variability (Brouwer et al., 2012; Nelms et al., 2012; Lim and Leech, 2016).

Cone beam CT (CBCT) is the most common and readily available onboard imaging system for online ART (Lu et al., 2006; Woerner et al., 2017). Previous studies (Nijkamp et al., 2008; Foroudi et al., 2011) have proved that CBCT is helpful in ART for reducing the volume of irradiated healthy tissue and the dose delivered to OAR. In offline ART, CBCTs are used for anatomic change monitoring during the treatment. When needed, a new planning CT is often acquired for plan adaptation to those organ or tumor volume changes. An ideal image dataset for ART should have accurate electron density for dose calculation and high soft tissue contrast resolution for accurate and robust image registrations and/or organ segmentation. For online ART, daily images acquired for treatment alignment are used for adapting the plan to anatomic and tumor changes prior to daily treatment. Unfortunately, online adaptive CBCT is hampered by poor image quality because of scatter artifact and lack of soft-tissue contrast. Furthermore, CBCT image values have poor correlation to electron density which requires post-image processing for correction (van Zijtveld et al., 2007). Poor image quality on CBCT also limits the ability to identify organ boundaries, thus resulting in high inter-observer variability in contour delineation (Lutgendorf-Caucig et al., 2011; Altorjai et al., 2012). Deformable image registration for contours propagation has shown high uncertainties due to poor CBCT image quality (Pukala et al., 2013). Increasing scan settings might improve the image quality and electron density accuracy for CBCT images (Dyer et al., 2019), yet at a cost of increasing imaging dose to patients, which might not be trivial when adding all fractions together.

Recently, deep learning algorithms were proposed to improve CBCT image quality using different network models (Jain, 2008; Xie et al., 2012; Dong et al., 2016). Deep convolutional neural networks (DCNN) can denoise images, reduce blurring, and improve soft tissue contrast resolution (Jain, 2008; Dong et al., 2016). Specifically for those fast-scan-low-dose CBCT scans, a U-NET based DCNN was developed for enhancing image quality for HNC patients, with improved HU accuracy, signal-to-noise ratio, and small anatomical structure preservation (Yuan et al., 2019). Such image quality enhancement should bring clinical benefits specifically for ART, including improved CT-CBCT image registration accuracy, thus improved contour propagation accuracy and better visualization for identifying organs at risk on CBCT images. The present study aimed to evaluate these clinical benefits with the image quality improvements in enhanced CBCT images.

MATERIALS AND METHODS

Patient Data

Forty post-operative HNC patients with a planning CT (pCT) and the first fraction CBCT were retrospectively identified and used for network training. A 2D U-Net shape architecture with 19-layers in 5 depths was specially optimized and trained using a total of 2080 CT and CBCT slice. The network design and architecture were described in the previous study (Yuan et al., 2019). Additional 15 patients with pCT, and replanning CT (rCT) 3–4 weeks into treatment with the same-day CBCT in relation to rCT were selected for DCNN validation. All CBCT scans were acquired with a kV x-ray imaging system mounted on a Synergy® linear accelerator (Elekta AB, Stockholm, Sweden). The CT parameters were set as follows: 512 * 512 matrix size on the axial plane, 1.183 mm * 1.183 mm pixel size, and 3.0 mm thickness. CBCT parameters were set to 270 * 270 matrix size, 1.0 mm * 1.0 mm pixel size, and 3.0 mm thickness. The original CBCT (oCBCT) images were fed into the trained DCNN model to obtain enhanced image quality from CBCT images, namely eCBCT. These images are synthetic CT images created based on the CT-CBCT paired trained DCNN model.

Organs at Risk Selection

For all patients, OARs included: left/right parotid, left/right submandibular gland (SMG), larynx, brainstem, and spinal cord. The reference contours on both pCT and rCT for each patient were manually delineated on the RayStation treatment planning system (Raysearch Laboratory, Stockholm, Sweden) by a radiation oncologist specialized in HNC and confirmed by a senior radiation oncologist. Contours on rCT were directly copied to the corresponding eCBCT and oCBCT through the gray-values based rigid image registration frame as comparison references. To eliminate the potential impact of registration differences between eCBCT and oCBCT images, the eCBCT was first registered to rCT and then the registration result of eCBCT was copied to oCBCT. All organs for delineation were completely covered in the field of CBCT view.

Image Quality Evaluation

The manually segmented OARs on rCT was considered the ground truth for image comparison. Image quality was quantified as the difference of mean pixel values among the region of interests (ROIs) between rCT and CBCT (oCBCT, eCBCT) images, denoted $\mathrm{ROI_m}$. Seven ROIs (left/right parotid, left/right SMG, larynx, brainstem, spinal cord) were used for all patients.

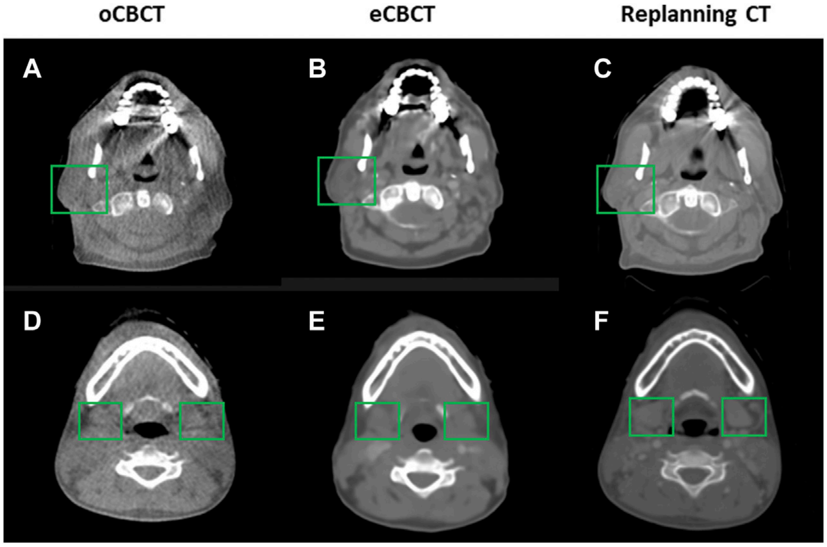


FIGURE 1 | Comparison of image quality for one representative patient. eCBCT has lower image noise and less streak artifacts in the soft tissue region than the oCBCT. eCBCT also has higher image contrast than oCBCT for parotid and submandibular gland areas (see green box).

The definition for signal-to-noise-ratio (SNR) is the ratio of signal power to noise power. The structural similarity index measure (SSIM) is the similarity between two images by comprehensively evaluating different properties such as luminance, contrast, and structure, which is one of human visual system-based metrics. The SNR and the SSIM of CBCTs were measured based on the seven ROIs used in the calculation of spatial non-uniformity for each patient.

$$SNR = 10 \cdot \log_{10}\left[\frac{\sum\sum \left[I_{CT}(x,y)\right]^2}{\sum\sum \left[I_{CT}(x,y) - I_{eCBCT}(x,y)\right]^2}\right]$$

In the formula, I_{CT} represents the CT scan slice and I_{eCBCT} represents the eCBCT scan slice.

$$SSIM = \frac{(2\mu_{eCBCT}\mu_{CT} + C_1)(2\delta_{eCBCT\&CT} + C_2)}{(\mu^2_{eCBCT} + \mu^2_{CT} + C_1)(\delta^2_{eCBCT} + \delta^2_{CT} + C_2)}$$

μ represents the mean value, δ^2 represents the variance, the parameters $C_1 = (k_1 Q)^2$ and $C_2 = (k_2 Q)^2$ are used to stabilize the division with weak denominators, $k_1 = 0.01$ and $k_2 = 0.02$. Q is the dynamic range of the pixel-values.

Contour Accuracy Assessment

For each patient, the CBCT pairs (oCBCT and eCBCT) and the same day rCT were imported into RayStation treatment planning system (TPS). All oCBCTs and eCBCTs were rigid registered based on skull and spine bony anatomy to the pCTs. Subsequently, a deformable image registration was performed between pCT and CBCTs, for organ contour propagation from pCTs to CBCTs image sets (both oCBCT and eCBCT) (Weistrand and Svensson, 2015). The image similarity term measured by correlation coefficient of the anatomically constrained deformation algorithm (ANACONDA) was used for CT/CBCT image comparison/registration. The whole body structure was used to define the registration region. After contour propagation, an experienced HNC radiation oncologist reviewed contours on oCBCT and eCBCT images and made contour modification if necessary. For the same patient, the type of images was not disclosed to the user at the time of contouring to avoid observer bias among different image modalities.

Accuracy of corrected propagated contours on oCBCT and eCBCT images were evaluated against the reference contours on rCTs (Whitfield et al., 2013). Quantitative assessment includes: dice similarity coefficient (DSC), Hausdorff distance (HD), and center of mass (COM) displacement. The DSC was adopted to evaluate the overlap of volumes between two contours. And it is calculated as follows:

$$\mathrm{DSC} = 2 \times \frac{\mathrm{Volume1} \cap \mathrm{Volume2}}{\mathrm{Volume1} + \mathrm{Volume2}}$$

Volume 1 and volume 2 represent the volumes of selected reference contours. A result of 1 means a complete overlap and a result of 0 means no overlap. The HD is to measure the max distance of all the nearest points between contours, define as:

$$\mathrm{HD} = \max\left\{\begin{matrix}\min \mathrm{d}(\mathrm{a}) \\ \mathrm{a} \in \mathrm{A}\end{matrix}, \begin{matrix}\min \mathrm{d}(\mathrm{b}) \\ \mathrm{b} \in \mathrm{B}\end{matrix}\right\}$$

"a" and "b" are points in contours A and B, respectively, where $\min\limits_{a\in A} \mathrm{d}(\mathrm{a})$ is the minimum distance of all points on the contour A to points on the contour B, so as the same definition used for $\min\limits_{b\in B} \mathrm{d}(\mathrm{b})$. While the center of mass displacement (COM) acts as a metric of the overall shift between two contours. It is calculated based on the following equation:

$$\mathrm{COM} = \sqrt[2]{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$$

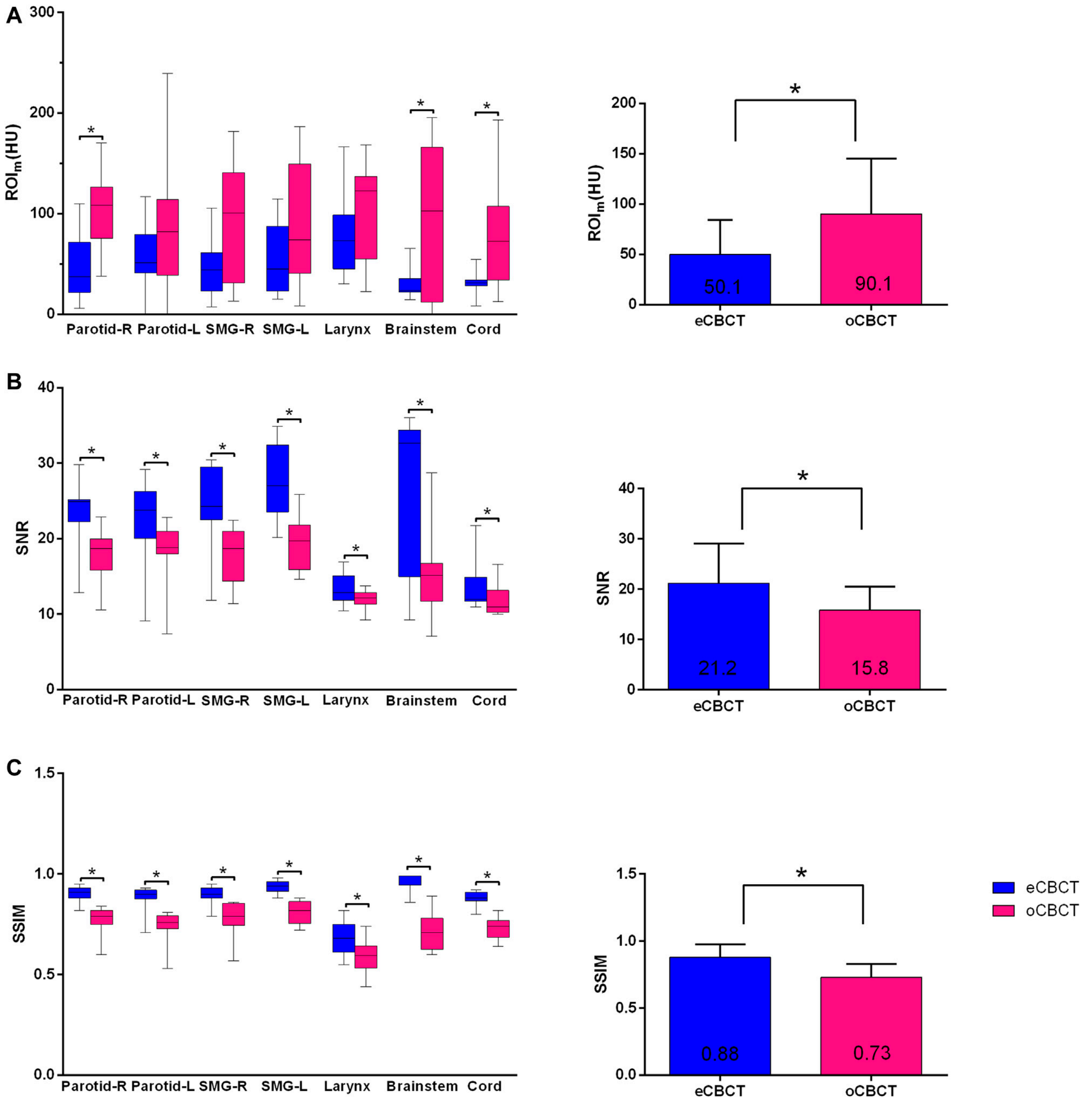


FIGURE 2 | Differences in HU, SNR and SSIM between eCBCT and oCBCT. Box plots on the left side showing the ROI$_m$ (HU) variations **(A)**, signal-to-noise-ratio (SNR) **(B)**, structural similarity index measure (SSIM) **(C)** for parotids, submandibular glands, larynx, brainstem, spinal cord, respectively. The limits of each box represent the 25th and 75th percentiles, the middle black line represents the median, and the upper and lower whiskers represents the highest and lowest values, respectively. The bar graphs on the right side for **(A)–(C)** showing the overall ROI$_m$ (HU), SNR, SSIM variations for all organs, respectively.*Indicates that the *p* value < 0.05, and error bars are standard deviations.

x (1, 2), y (1, 2), z (1,2) are coordinates of the geometric centroid of the contours in comparison (Kumarasiri et al., 2014).

To further evaluate the clinical accessibility of CBCT image quality for manual segmentation, three HNC radiation oncologists visually scored OAR structures on both eCBCT and oCBCT images using a scale 1–3 according to the following criteria: 1) the outline of the structure cannot be identified; 2) the outline of the structure can be identified with moderate difficulty; 3) the image quality is close to CT simulation and the outline of the structure can be clearly identified.

Statistical Analysis

All Statistical analyses were performed in SPSS software version 24.0 (SPSS Inc., Chicago, IL, United States) and GraphPad

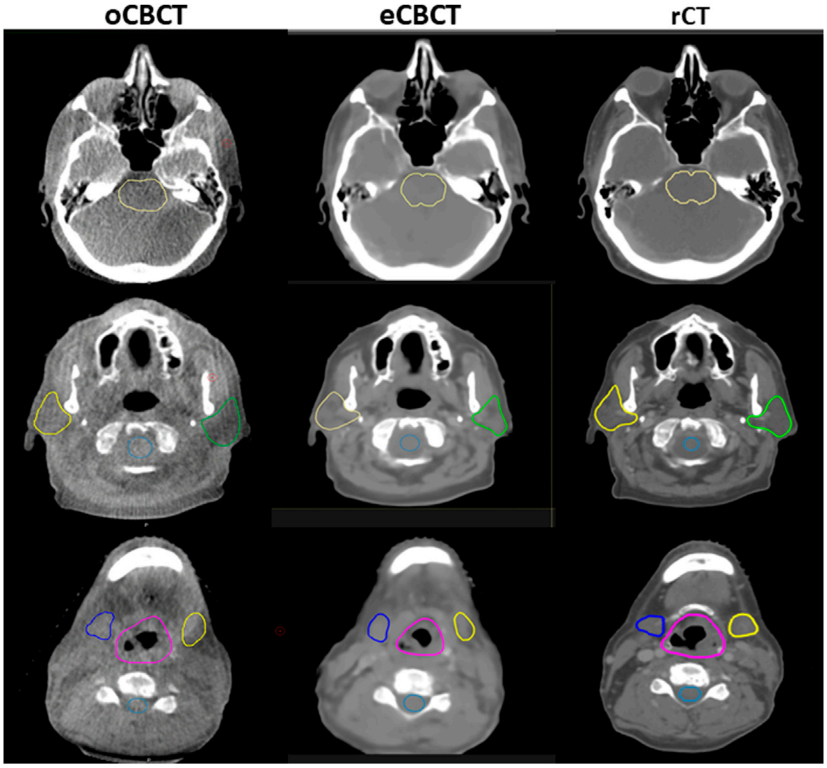


FIGURE 3 | OARs delineated on transverse slices of oCBCT, eCBCT and rCT images for a representative HNC patient. OARs are outlined: brainstem (top, yellow line), parotids [middle, yellow (right) and green (left) lines], spinal cord (middle, light blue line),submandibular glands [bottom, blue (right) and yellow (left) lines], larynx (bottom, purple line).

version 6.0. $p < 0.05$ was considered statistically significant. The Wilcoxon test was used to compare the image quality and the contouring difference between eCBCT and oCBCT.

RESULTS

Figure 1 shows image quality as an example. eCBCT images had lower noise and less streak artifacts in the soft tissue region than oCBCT. eCBCT images also had higher image contrast than oCBCT, particularly in the parotid and submandibular gland regions. A quantitative analysis of image quality for OARs is summarized in **Figure 2**. Seven ROIs were segmented on rCT and the mean pixel values were calculated for each ROI on rCT, oCBCT, and eCBCT images. When compared with rCT, the mean difference in CT values of $\mathrm{ROI_m}$ between rCT and oCBCT were 90 HU, while the difference between rCT and eCBCT reduced to 50 HU. This suggests that the CT values of OARs on eCBCT images more closely match those on rCT than oCBCT. When oCBCT and eCBCT SNR and SSIM were compared, eCBCT was significantly better than oCBCT ($p < 0.05$). This suggests that the DCNN method performs effectively in reducing image noise and improving image quality in eCBCT images, more closely resembling the corresponding rCT images. Metrics of image quality ($\mathrm{ROI_m}$, SNR, and SSIM) were calculated and compared for all OARs on rCT, oCBCT, and eCBCT images. We found that eCBCT showed significant improvement compared to oCBCT for all studied OARs ($p < 0.05$) (**Figure 2**).

Figure 3 shows OAR contours on transverse slices of rCT, oCBCT, and eCBCT images for one representative patient. The mean value of DSC, HD and COM difference for OARs on oCBCT and eCBCT images are shown in **Figure 4**. The average DSC for eCBCT-to-rCT and oCBCT-to-rCT was 0.83 ± 0.06, and 0.70 ± 0.13. The average HD for eCBCT-to-rCT was 0.42 ± 0.13 cm and for oCBCT-to-rCT was 0.72 ± 0.25 cm. The mean COM for eCBCT-to-rCT was 0.28 ± 0.19 cm and for oCBCT-to-rCT was 0.44 ± 0.22 cm eCBCT OARs had a higher DSC than oCBCT for all the structures ($p < 0.05$), except for brainstem. Similarly, the results of HD and COM all showed that OARs delineated on eCBCT were closer to rCT than oCBCT. Statistically, the difference between OARs on eCBCT vs oCBCT for HD and COM were significant for most organs. **Table 1** shows the reported visual scores for OAR identification by three physicians. The scores are higher for all OAR structures on eCBCT vs oCBCT images—particularly for parotid structures. This implies that eCBCT improves ease of manual segmentation compared with oCBCT.

DISCUSSION

The studied DCNN method quantitatively improved CBCT image quality for head and neck patients. The impact of eCBCT image

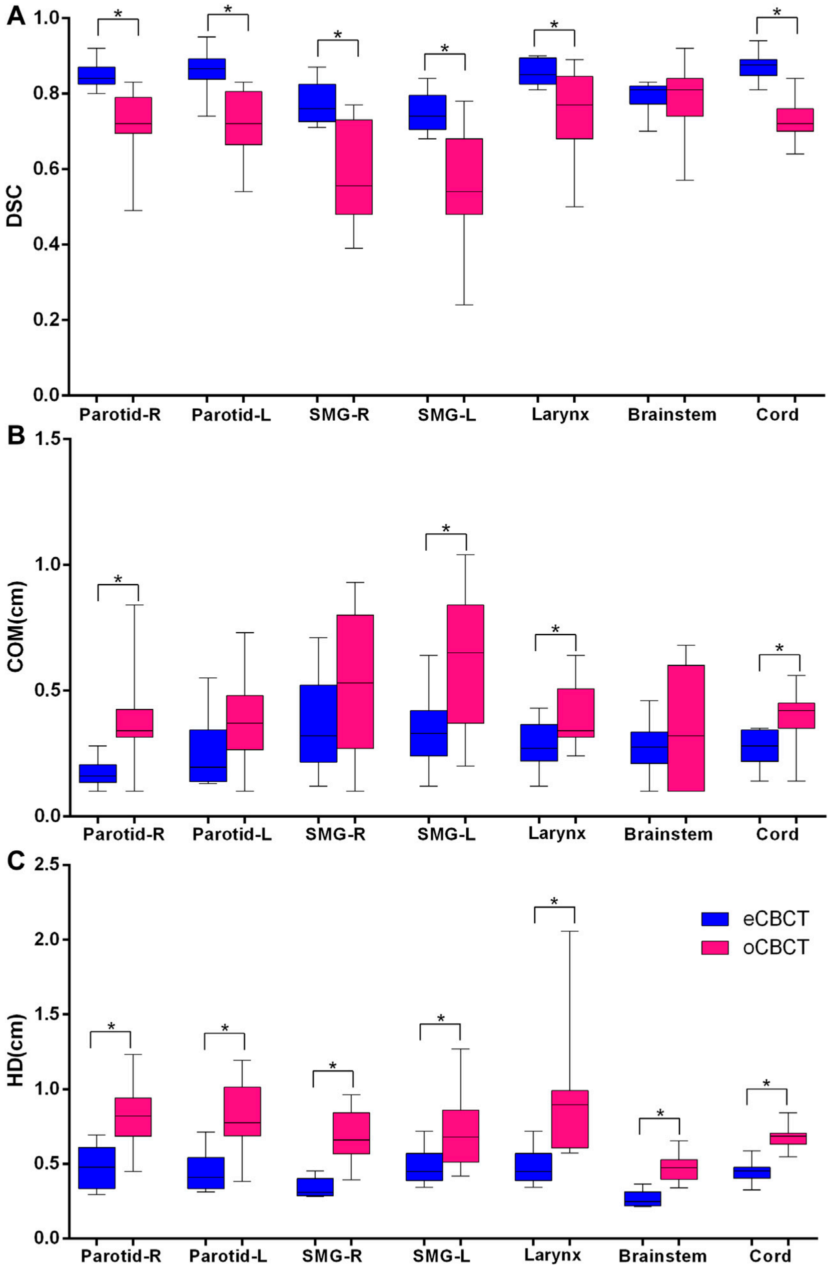


FIGURE 4 | Quantitative assessment of OARs for rCTs vs. oCBCT and eCBCT images. Box plot showing Dice similarity coefficient (DSC) variations **(A)**, Center of mass (COM) displacement **(B)**, Hausdorff distance (HD) variations **(C)** for parotids, submandibular glands, larynx, brainstem, spinal cord, respectively. The limits of each box represent the 25th and 75th percentiles, the middle black line represents the median, and the upper and lower whiskers represents the highest and lowest values, respectively. *Indicates that the *p* value < 0.05.

quality improvements in a clinical context was evaluated. SNR and SSIM of eCBCT both improved compared with those of oCBCT. An overall improvement in image quality also helped users' judgment in identifying OARs and their subsequent contour correction on eCBCT compared with those for oCBCTs.

The inaccurate CBCT Hounsfield units will subsequently compromise dose calculation accuracy (Richter et al., 2008; Usui et al., 2013). Several approaches have been proposed to deal with the shortcomings of CBCT, such as anti-scatter grids and software-based solutions (Letourneau et al., 2007; Stankovic et al., 2017). According to Letourneau et al.'s study (Letourneau et al., 2007), they quantified the magnitude of CBCT image artifacts following the use of an anti-scatter grid and a nonlinear scatter correction. Then the corrected CBCT images were used for online planning and the dosimetric accuracy was satisfied with accepted RT standards. Veiga et al. (2014) indicated that using CT to CBCT deformable image registration provides the tools for calculating "dose of the day" without the need to obtain a new CT. However, they are limited by the time

TABLE 1 | Visual score (mean ± SD) for OAR segmentation ranked by three HNC physicians.

	Parotid-R	Parotid-L	SMG-R	SMG-L	Cord	Larynx	Brainstem
eCBCT	2.3 ± 0.6	2.2 ± 0.5	1.9 ± 0.3	1.9 ± 0.4	1.8 ± 0.5	1.7 ± 0.5	1.3 ± 0.5
oCBCT	1.5 ± 0.3	1.2 ± 0.4	1.1 ± 0.2	1.1 ± 0.3	1.1 ± 0.2	1.3 ± 0.4	1.1 ± 0.3

required to correct the image, and if there are large anatomical changes, these methods will also face problems due to a large challenge to the registration algorithms used in these methods. In our study, we present a fast method for intensity correction for CBCT based on a convolutional neural network. Previously, amongst those using DCNN methods, Kida et al. (2018) showed improved CBCT image quality and noise reduction for 20 prostate cancer patients using a DCNN model. Hansen et al. (2018) presented a proof-of principle of using deep learning techniques for pelvic CBCT correction and dose calculation accuracy, which is superior to conventional methods of mapping image value from the planning CT to CBCT (van Zijtveld et al., 2007), or deforming the planning CT to match a daily CBCT for the dose calculation (Veiga et al., 2015). Original CBCT often suffers from severe scatter contaminations, resulted in significant image value inaccuracy compared to that of CT. In our study, enhanced CBCT images reduced scatter artifacts, improved soft tissue contrast, and improved the HU image values within each OARs.

We compared OAR segmentation on eCBCTs and oCBCTs in reference to rCT, which was acquired on the same day as the CBCT images. Our results indicate that the eCBCTs consistently outperforms oCBCTs in all metrics. The average DSC for parotid glands in eCBCT was more than 0.80. This result is very close to previous studies. According to Zhang et al. (2014), the average DSC for parotid was 0.80 in compressed sensing based CBCT. They also proved that compressed sensing based CBCT can help to improve manual delineation of targets. Although DSC is widely used as a performance metric, it has limitation that the structure volume affects its values. Previous studies (Kumarasiri et al., 2014; Zhang et al., 2018) reported that DSC shows a positive correlation with structure volume, regardless how good the structure overlap is. Therefore, COM and HD were also used as complementary measures to better understand the quality of volume overlaps.

We chose to evaluate DCNN for CBCT image improvement in HNC patients for practical consideration. Due to the complexity of head and neck anatomic structures, and low soft tissue contrast, it is challenging to perform a manual OAR segmentation on the original CBCT. Many had attempted to create a simulated CT from deforming the planning CT to the original CBCT. However, the significant scatter artifacts on CBCT can affect the DIR accuracy. In addition, it was reported (Hou et al., 2011) that deforming contours from CT to CBCT to evaluate anatomic changes or calculate adapted dose during treatment is not reliable or requires significant manual modification. With the current CBCT image quality overall, it seems to be a common clinical practice to obtain propagated contours from the original CT to CBCT after image registration (either rigid or deformable) and correct for any obvious inaccuracy on CBCT. This of course has never been an easy task to users due to the poor quality of CBCT. Thus we included visual scoring as one of the evaluation criteria in this study. Visual score results indicated that physicians felt higher confidence in identifying the outline of those structures on eCBCT, compared to those of oCBCT.

Manual contours defined by experienced physicians were used as the comparison reference. Using manual contours as the “gold standard” is clinically feasible, and many researchers (Li et al., 2016; Zhang et al., 2014) have used this method to evaluate the delineation accuracy. A major limitation of the study is that only a small number of patients’ scans were available for this study. Future study should include more patient data and explore other anatomical regions. Moreover, contouring accuracy of gross tumor volume (GTV) on eCBCT was not studied, due to limited image quality for target delineation on both oCBCT and eCBCT. Therefore, it is worthy of noting that even though the present study has shown significant improvement toward CBCT-based ART, eCBCT image quality still has room for improvement, i.e. on the aspects of target visualization. Yet this study is still valuable for ART, in that eCBCT has improved HU accuracy and can serve for a quick on-line dose verification. The dosimetric deviation can be a trigger for ART, where a regular or high-dose CBCT can be acquired for better image quality should ART is determined necessary. This study presented that DCNN-processed low dose fast scan CBCT images, i.e. eCBCT, have the potential for head and neck adaptive radiotherapy.

CONCLUSION

We validated a DCNN model for improving low-dose-fast-scan CBCT image quality, and enhanced CBCT has the potential to improve delineation accuracy for head and neck patients. These results support that enhanced CBCT has potential for adaptive radiotherapy. In addition, the CBCT image quality may still have room for improvement. Future study includes further improve the performance of the DCNN method, using enhanced CBCT for a direct dose calculation to validate the accuracy by comparing with dose distribution calculated on planning CTs.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

YR and SR conceived and designed the study. WC and YL collected the patient’s information, WC, NY, YL, BD, JQ, SR, and LS performed the experiments. WC and NY were responsible for the data analysis. WC drafted this manuscript. YR, BD, and JQ reviewed and edited the manuscript. JQ, LS, and SB offered constructive suggestions for this study. All authors read and approved the final manuscript.

REFERENCES

Altorjai, G., Fotina, I., Lutgendorf-Caucig, C., Stock, M., Potter, R., Georg, D., et al. (2012). Cone-beam CT-based delineation of stereotactic lung targets: the influence of image modality and target size on interobserver variability. *Int. J. Radiat. Oncol. Biol. Phys.* 82 (2), e265–272. doi:10.1016/j.ijrobp.2011.03.042

Brouwer, C. L., Steenbakkers, R. J., van den Heuvel, E., Duppen, J. C., Navran, A., Bijl, H. P., et al. (2012). 3D Variation in delineation of head and neck organs at risk. *Radiat. Oncol.* 7, 32. doi:10.1186/1748-717X-7-32

Castadot, P., Lee, J. A., Geets, X., and Gregoire, V. (2010). Adaptive radiotherapy of head and neck cancer. *Semin. Radiat. Oncol.* 20 (2), 84–93. doi:10.1016/j.semradonc.2009.11.002

Castelli, J., Simon, A., Louvel, G., Henry, O., Chajon, E., Nassef, M., et al. (2015). Impact of head and neck cancer adaptive radiotherapy to spare the parotid glands and decrease the risk of xerostomia. *Radiat. Oncol.* 10, 6. doi:10.1186/s13014-014-0318-z

Chen, A. M., Daly, M. E., Cui, J., Mathai, M., Benedict, S., and Purdy, J. A. (2014). Clinical outcomes among patients with head and neck cancer treated by intensity-modulated radiotherapy with and without adaptive replanning. *Head Neck.* 36 (11), 1541–1546. doi:10.1002/hed.23477

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (2), 295–307. doi:10.1109/TPAMI.2015.2439281

Dyer, B. A., Nair, C. K., Deardorff, C. E., Wright, C. L., Perks, J. R., and Rao, S. S. (2019). Linear accelerator-based radiotherapy simulation using on-board kilovoltage cone-beam computed tomography for 3-dimensional volumetric planning and rapid treatment in the Palliative setting. *Technol. Canc. Res. Treat.* 18, 623. doi:10.1177/1533033819865623

Foroudi, F., Wong, J., Kron, T., Rolfo, A., Haworth, A., Roxby, P., et al. (2011). Online adaptive radiotherapy for muscle-invasive bladder cancer: results of a pilot study. *Int. J. Radiat. Oncol. Biol. Phys.* 81 (3), 765–771. doi:10.1016/j.ijrobp.2010.06.061

Hansen, D. C., Landry, G., Kamp, F., Li, M., Belka, C., Parodi, K., et al. (2018). ScatterNet: a convolutional neural network for cone-beam CT intensity correction. *Med. Phys.* 45 (11), 4916–4926. doi:10.1002/mp.13175

Hou, J., Guerrero, M., Chen, W., and D'Souza, W. D. (2011). Deformable planning CT to cone-beam CT image registration in head-and-neck cancer. *Med. Phys.* 38 (4), 2088–2094. doi:10.1118/1.3554647

Jain, V. S. H. (2008). Natural image denoising with convolutional networks. *Adv. Neural Inf. Process. Syst.* 64, 769–776.

Kida, S., Nakamoto, T., Nakano, M., Nawa, K., Haga, A., Kotoku, J., et al. (2018). Cone beam computed tomography image quality improvement using a deep convolutional neural network. *Cureus.* 10 (4), e2548. doi:10.7759/cureus.2548

Kumarasiri, A., Siddiqui, F., Liu, C., Yechieli, R., Shah, M., Pradhan, D., et al. (2014). Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting. *Med. Phys.* 41 (12), 121712. doi:10.1118/1.4901409

Letourneau, D., Wong, R., Moseley, D., Sharpe, M. B., Ansell, S., Gospodarowicz, M., et al. (2007). Online planning and delivery technique for radiotherapy of spinal metastases using cone-beam CT: image quality and system performance. *Int. J. Radiat. Oncol. Biol. Phys.* 67 (4), 1229–1237. doi:10.1016/j.ijrobp.2006.09.058

Li, X., Zhang, Y. Y., Shi, Y. H., Zhou, L. H., and Zhen, X. (2016). Evaluation of deformable image registration for contour propagation between CT and cone-beam CT images in adaptive head and neck radiotherapy. *Technol. Health Care.* 24 (Suppl. 2), S747–S755. doi:10.3233/THC-161204

Lim, J. Y., and Leech, M. (2016). Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol.* 55 (7), 799–806. doi:10.3109/0284186X.2016.1173723

Lu, W., Olivera, G. H., Chen, Q., Ruchala, K. J., Haimerl, J., Meeks, S. L., et al. (2006). Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. *Phys. Med. Biol.* 51 (17), 4357–4374. doi:10.1088/0031-9155/51/17/015

Lutgendorf-Caucig, C., Fotina, I., Stock, M., Potter, R., Goldner, G., and Georg, D. (2011). Feasibility of CBCT-based target and normal structure delineation in prostate cancer radiotherapy: multi-observer and image multi-modality study. *Radiother. Oncol.* 98 (2), 154–161. doi:10.1016/j.radonc.2010.11.016

Nelms, B. E., Tome, W. A., Robinson, G., and Wheeler, J. (2012). Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 82 (1), 368–378. doi:10.1016/j.ijrobp.2010.10.019

Nijkamp, J., Pos, F. J., Nuver, T. T., de Jong, R., Remeijer, P., Sonke, J. J., et al. (2008). Adaptive radiotherapy for prostate cancer using kilovoltage cone-beam computed tomography: first clinical results. *Int. J. Radiat. Oncol. Biol. Phys.* 70 (1), 75–82. doi:10.1016/j.ijrobp.2007.05.046

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J. Clin.* 55 (2), 74–108. doi:10.3322/canjclin.55.2.74

Pukala, J., Meeks, S. L., Staton, R. J., Bova, F. J., Manon, R. R., and Langen, K. M. (2013). A virtual phantom library for the quantification of deformable image registration uncertainties in patients with cancers of the head and neck. *Med. Phys.* 40 (11), 111703. doi:10.1118/1.4823467

Richter, A., Hu, Q., Steglich, D., Baier, K., Wilbert, J., Guckenberger, M., et al. (2008). Investigation of the usability of conebeam CT data sets for dose calculation. *Radiat. Oncol.* 3, 42. doi:10.1186/1748-717X-3-42

Schwartz, D. L. (2012). Current progress in adaptive radiation therapy for head and neck cancer. *Curr. Oncol. Rep.* 14 (2), 139–147. doi:10.1007/s11912-012-0221-4

Stankovic, U., Ploeger, L. S., van Herk, M., and Sonke, J. J. (2017). Optimal combination of anti-scatter grids and software correction for CBCT imaging. *Med. Phys.* 44 (9), 4437–4451. doi:10.1002/mp.12385

Usui, K., Ichimaru, Y., Okumura, Y., Murakami, K., Seo, M., Kunieda, E., et al. (2013). Dose calculation with a cone beam CT image in image-guided radiation therapy. *Radiol. Phys. Technol.* 6 (1), 107–114. doi:10.1007/s12194-012-0176-z

van Zijtveld, M., Dirkx, M., and Heijmen, B. (2007). Correction of conebeam CT values using a planning CT for derivation of the "dose of the day. *Radiother. Oncol.* 85 (2), 195–200. doi:10.1016/j.radonc.2007.08.010

Veiga, C., Lourenco, A. M., Mouinuddin, S., van Herk, M., Modat, M., Ourselin, S., et al. (2015). Toward adaptive radiotherapy for head and neck patients: uncertainties in dose warping due to the choice of deformable registration algorithm. *Med. Phys.* 42 (2), 760–769. doi:10.1118/1.4905050

Veiga, C., McClelland, J., Moinuddin, S., Lourenco, A., Ricketts, K., Annkah, J., et al. (2014). Toward adaptive radiotherapy for head and neck patients: feasibility study on using CT-to-CBCT deformable registration for "dose of the day" calculations. *Med. Phys.* 41 (3), 031703. doi:10.1118/1.4864240

Weistrand, O., and Svensson, S. (2015). The ANACONDA algorithm for deformable image registration in radiotherapy. *Med. Phys.* 42 (1), 40–53. doi:10.1118/1.4894702

Whitfield, G. A., Price, P., Price, G. J., and Moore, C. J. (2013). Automated delineation of radiotherapy volumes: are we going in the right direction?. *Br. J. Radiol.* 86 (1021), 20110718. doi:10.1259/bjr.20110718

Woerner, A. J., Choi, M., Harkenrider, M. M., Roeske, J. C., and Surucu, M. (2017). Evaluation of deformable image registration-based contour propagation from planning CT to cone-beam CT. *Technol. Canc. Res. Treat.* 15, 242. doi:10.1177/1533034617697242

Xie, J., Xu, L., and Chen, E. (2012). Image denoising and inpainting with deep neural networks. *Adv. Neural Inf. Process. Syst.* 24, 350–352. doi:10.1364/BOE.8.000679

Yan, H., Cervino, L., Jia, X., and Jiang, S. B. (2012). A comprehensive study on the relationship between the image quality and imaging dose in low-dose cone beam CT. *Phys. Med. Biol.* 57 (7), 2063–2080. doi:10.1088/0031-9155/57/7/2063

Yuan, N., Dyer, B., Rao, S., Chen, Q., Benedict, S., Shang, L., et al. (2019). Convolutional neural network enhancement of fast-scan low-dose cone-beam CT images for head and neck radiotherapy. *Phys. Med. Biol.* 240, 6560. doi:10.1088/1361-6560/ab6240

Zhang, H., Tan, W., and Sonke, J. J. (2014). Effect of compressed sensing reconstruction on target and organ delineation in cone-beam CT of head-and-neck and breast cancer patients. *Radiother. Oncol.* 112 (3), 413–417. doi:10.1016/j.radonc.2014.07.002

Zhang, L., Wang, Z., Shi, C., Long, T., and Xu, X. G. (2018). The impact of robustness of deformable image registration on contour propagation and dose accumulation for head and neck adaptive radiotherapy. *J. Appl. Clin. Med. Phys.* 19 (4), 185–194. doi:10.1002/acm2.12361

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.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/frai.2021.618469

Computed Tomography Radiomics Kinetics as Early Imaging Correlates of Osteoradionecrosis in Oropharyngeal Cancer Patients

Souptik Barua [1,2†]***, Hesham Elhalawani*** [3†]***, Stefania Volpe*** [4,5]***, Karine A. Al Feghali*** [3]***, Pei Yang*** [3]***, Sweet Ping Ng*** [6]***, Baher Elgohari*** [3]***, Robin C. Granberry*** [3]***, Dennis S. Mackin*** [7]***, G. Brandon Gunn*** [3]***, Katherine A. Hutcheson*** [3]***, Mark S. Chambers*** [8]***, Laurence E. Court*** [7]***, Abdallah S. R. Mohamed*** [3]***, Clifton D. Fuller*** [3,7*]***, Stephen Y. Lai*** [9*] ***and Arvind Rao*** [1,2,10*]

[1] Department of Electrical and Computer Engineering, Rice University, Houston, TX, United States, [2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States, [3] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [4] Department of Radiation Oncology, European Institute of Oncology IRCSS, Milan, Italy, [5] Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy, [6] Department of Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia, [7] Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [8] Department of Oncologic Dentistry and Prosthodontics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [9] Department of Head and Neck Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [10] Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, United States

Osteoradionecrosis (ORN) is a major side-effect of radiation therapy in oropharyngeal cancer (OPC) patients. In this study, we demonstrate that early prediction of ORN is possible by analyzing the temporal evolution of mandibular subvolumes receiving radiation. For our analysis, we use computed tomography (CT) scans from 21 OPC patients treated with Intensity Modulated Radiation Therapy (IMRT) with subsequent radiographically-proven $\geq$ grade II ORN, at three different time points: pre-IMRT, 2-months, and 6-months post-IMRT. For each patient, radiomic features were extracted from a mandibular subvolume that developed ORN and a control subvolume that received the same dose but did not develop ORN. We used a Multivariate Functional Principal Component Analysis (MFPCA) approach to characterize the temporal trajectories of these features. The proposed MFPCA model performs the best at classifying ORN vs. Control subvolumes with an area under curve (AUC) = 0.74 [95% confidence interval (C.I.): 0.61–0.90], significantly outperforming existing approaches such as a pre-IMRT features model or a delta model based on changes at intermediate time points, i.e., at 2- and 6-month follow-up. This suggests that temporal trajectories of radiomics features derived from sequential pre- and post-RT CT scans can provide markers that are correlates of RT-induced mandibular injury, and consequently aid in earlier management of ORN.

Keywords: osteoradionecrosis, computed tomography, radiomics, longitudinal, radiotherapy, head and neck cancer, oropharyngeal cancer, functional principal component analysis

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Shivanand Sharanappa Gornale,
Rani Channamma University, India
Sanjay Aneja,
Yale University, United States
Marco Diego Dominietto,
Paul Scherrer Institut
(PSI), Switzerland

****Correspondence:***
Clifton D. Fuller
CDFuller@mdanderson.org
Stephen Y. Lai
SYLai@mdanderson.org
Arvind Rao
ukarvind@umich.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *17 October 2020*
Accepted: *04 March 2021*
Published: *09 April 2021*

Citation:
Barua S, Elhalawani H, Volpe S, Al Feghali KA, Yang P, Ng SP, Elgohari B, Granberry RC, Mackin DS, Gunn GB, Hutcheson KA, Chambers MS, Court LE, Mohamed ASR, Fuller CD, Lai SY and Rao A (2021) Computed Tomography Radiomics Kinetics as Early Imaging Correlates of Osteoradionecrosis in Oropharyngeal Cancer Patients. Front. Artif. Intell. 4:618469. doi: 10.3389/frai.2021.618469

INTRODUCTION

Radiotherapy (RT) is a highly utilized modality in the treatment of head and neck (H&N) cancers with well-established local control and survival benefits (Pan et al., 2016). Advances in radiation delivery techniques from 2-dimensional (2D) and 3-dimensional (3D) techniques to intensity-modulated radiotherapy (IMRT) with the ability to manipulate the beam path to spare normal tissues has significantly improved cure rates and toxicity profile (Allison et al., 2014). Despite that, osteoradionecrosis is a late complication from radiation to the mandibular bone with a serious impact on the quality of life for a growing population of younger surviving head and neck cancer patients (Oh et al., 2004). The incidence of ORN varied between different modalities ranging from 2 to 40% in the conventional era to 0–6% in the IMRT era. Different risk factors were identified to play a role in the development of ORN following radiotherapy treatments (Allison et al., 2013; Zhang et al., 2017). Osteoradionecrosis has a great impact on the patients' quality of life if not detected and managed properly (Tucker et al., 2016; Wong et al., 2017). Diagnosis of ORN mainly relies on clinical and radiological tools such as computed tomography (CT) and magnetic resonance imaging (MRI) with their limited capacity for early detection (Tsien et al., 2014).

Fortunately, the recent advances in biomedical imaging were coupled with the rise of radiomics in terms of extracting quantifiable imaging features, possibly of high information yield and subsequent computation of these features kinetics (e.g., delta-radiomics) derived from sequential images (Cacicedo et al., 2016). Paired with machine learning techniques, we hypothesize that radiomic feature kinetics can characterize and distinguish mandibular bone subvolumes at higher risk of developing future ORN. These "temporal virtual digital biopsies" might have the potential to empower earlier intervention and hence improve patients' quality of life.

Consequently, the aims of this study are to:

1. Determine bone radiomic features derived from contrast-enhanced CT (CECT) images that are significantly different between ORN and non-ORN mandibular subvolumes.
2. Develop a predictive radiomic-based signature of ORN based on CECT temporal changes in high-risk mandibular subvolumes
3. Hypothesis generation for future prospective studies.

MATERIALS AND METHODS

Study Population

Following approval from an institutional review board (IRB) at our institution, data for biopsy-proven OPC patients treated between 2002 and 2013 who underwent radiation therapy as a single or multimodality definitive therapy were considered for the current investigation ($n = 83$). This investigation and relevant methodology were performed in compliance with the Health Insurance Portability and Accountability Act (HIPAA) as a retrospective study where the need for informed consent was waived (Freymann et al., 2012). Electronic medical records were scanned for documented diagnosis of mandibular ORN following IMRT in the absence of any prior head and neck re-irradiation along the same lines as a previous ORN study by our team (Mohamed et al., 2017). The aspects of our institutional IMRT approach for oropharyngeal cancer patients were previously reported in detail (Garden et al., 2013). All patients received pre-radiotherapy Dental Oncology service clearance, and, if indicated, prophylactic dental extraction and fluoride trays were prescribed as per standard Head and Neck Service operating procedure (Tsai et al., 2013). Inclusion and exclusion criteria for patients' selection are illustrated in **Figure 1**.

ORN Staging

The severity of ORN was graded I through IV as follows: grade I, i.e., minimal bone exposure requiring conservative management; grade II: minor debridement required; grade III: hyperbaric oxygen therapy (HBOT) received; grade IV: major surgery mandated. This staging system is very comprehensively given its emphasis on response to treatment as a standard to categorize ORN (Tsai et al., 2013). Patients who subsequently suffered from radiographically &/or pathologically proven grade II or worse ORN were included in this study.

CT Acquisition Protocol and Eligibility Criteria

According to our institutional protocol, CECT images were obtained as a prerequisite for pre-treatment diagnostic work-up. Subsequent post-IMRT CECT scans for response evaluation and further surveillance were routinely performed at 2 and 6-month time points and then at regular preset intervals thereafter. Our study revolved about extracting quantitative imaging biomarkers from CECT at pre-IMRT (i.e., baseline), 2-month (post-RT2), and 6-month (post-RT6) post-IMRT, as well as the time instance corresponding to the development of ORN. To that end, CECT scans with available non-reconstructed axial cuts at the aforementioned 4 time points were retrieved. CT slices with evident ORN lesions that were obscured or otherwise affected by visible metal artifacts were not contoured and were not included in the analysis.

All CT scans were attained with a multi-detector row CT scanner. Scan parameters were as follows: slice thickness reconstruction (STR) ranges between 1 and 3 mm, with a median STR of 1 mm, X-ray tube current of 99–584 mA (median: 220 mA) at 120–140 kVp. All images acquired at our institution were composed of 512 × 512 pixels and were acquired following a 90 s delay after intravenous contrast administration. One-hundred and twenty milliliters of contrast were injected at a rate of 3 ml/s. To standardize the image voxel sizes for use in texture feature calculations, all the CT scans were resampled, via a trilinear interpolation voxel resampling filter (Shafiq-ul-Hassan et al., 2017).

Image Segmentation and Registration

We specifically selected CECT scans demonstrating the earliest radiographically evident ORN characteristic lesion(s) as reported by radiologists and further confirmed by physical examination by

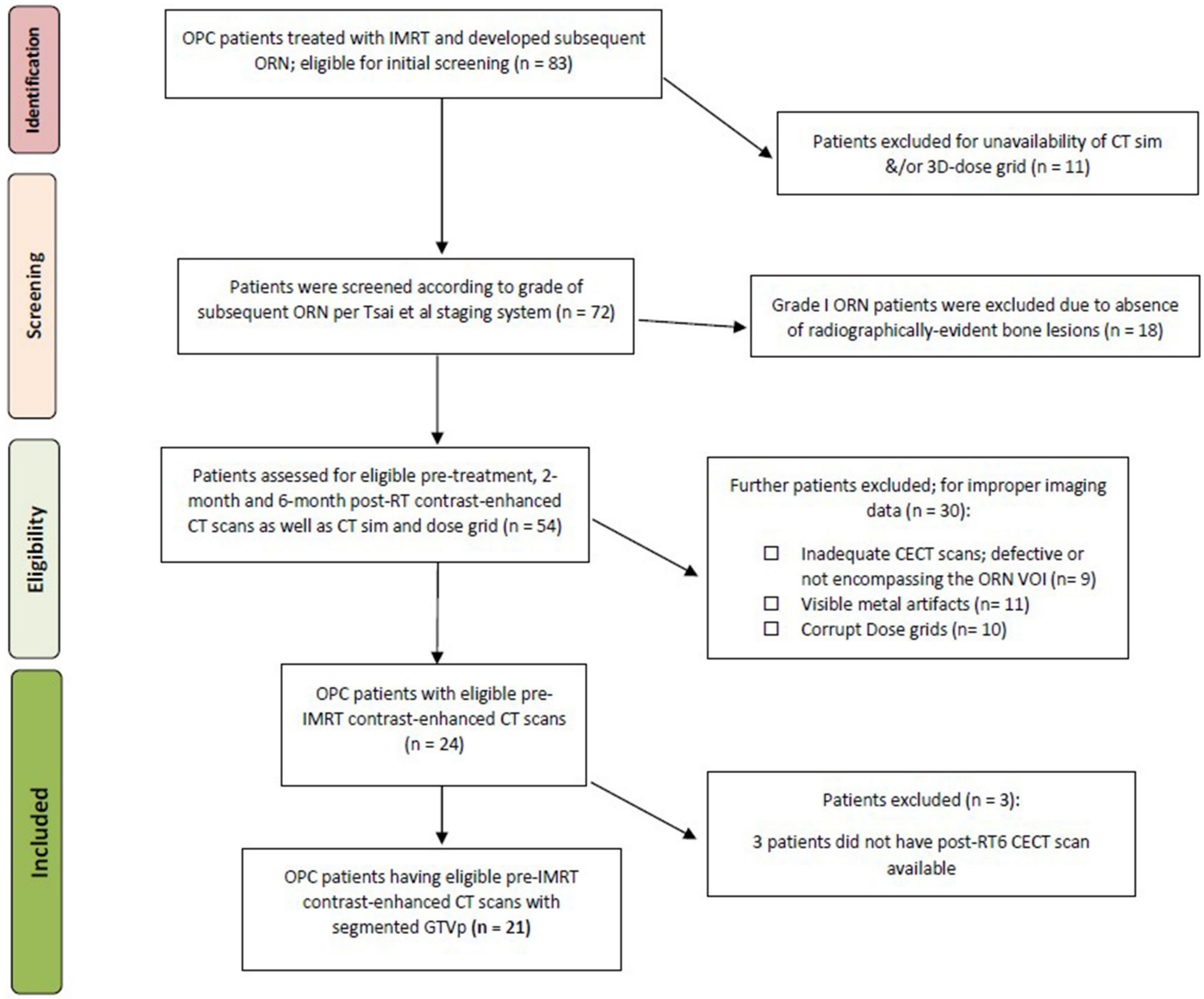

Identification
Screening
Eligibility
Included
OPC patients treated with IMRT and developed subsequent ORN; eligible for initial screening (n = 83)
Patients excluded for unavailability of CT sim &/or 3D-dose grid (n = 11)
Patients were screened according to grade of subsequent ORN per Tsai et al staging system (n = 72)
Grade I ORN patients were excluded due to absence of radiographically-evident bone lesions (n = 18)
Patients assessed for eligible pre-treatment, 2-month and 6-month post-RT contrast-enhanced CT scans as well as CT sim and dose grid (n = 54)
Further patients excluded; for improper imaging data (n = 30):
□ Inadequate CECT scans; defective or not encompassing the ORN VOI (n= 9)
□ Visible metal artifacts (n= 11)
□ Corrupt Dose grids (n= 10)
OPC patients with eligible pre-IMRT contrast-enhanced CT scans (n = 24)
Patients excluded (n = 3):
3 patients did not have post-RT6 CECT scan available
OPC patients having eligible pre-IMRT contrast-enhanced CT scans with segmented GTVp (n = 21)

FIGURE 1 | Patient selection. Flowchart of selection process of patients for this study.

physicians in Head & Neck Surgery as well as in Dental Oncology [ORN CECT].

The original delivered DICOM-RT clinical treatment plans were restored from Pinnacle treatment planning system (Pinnacle, Phillips Medical Systems, Andover, MA) into commercially available image registration software (VelocityAI^{TM} 3.0.1). Diagnostic CECT scans at baseline, post-RT2, post-RT6, and ORN were also imported. Radiographically evident bony lesions were delineated manually by a radiation oncologist (HE) to constitute the ORN volumes of interest (VOIs). Physical exam and other available imaging modalities such as dental-dedicated panoramic X-rays were utilized to guide the segmentation of VOIs.

Planning CT was co-registered with ORN CECT using deformable image registration algorithm of VelocityAI^{TM} 3.0.1. The 3D reconstructed dose grid of RT plan was then overlaid to the ORN CECT. A neighboring radiographically intact mandibular subvolume within the same isodose distribution volume was manually segmented and designated as "Control VOI" at the ORN CECT. Subsequently, baseline, post-RT2, post-RT6 CECT scans were co-registered with ORN CECT using rigid registration algorithms of VelocityAI^{TM} 3.0.1. Both "ORN" & "Control" VOIs were propagated from ORN CECT to other CECT scans at all three prior time points (**Figure 2**).

Radiomics Features Extraction

Computed tomography scans with corresponding contoured VOIs were then extracted in the Digital Imaging and Communications in Medicine format (DICOM), as DICOM-RT and RT-STRUCT files, respectively. These files were then imported into an in-house image biomarker explorer (IBEX) software, built on MATLAB for subsequent radiomics feature extraction (Zhang et al., 2015) along the same lines as previous studies (Elhalawani et al., 2018; Yang et al., 2018).

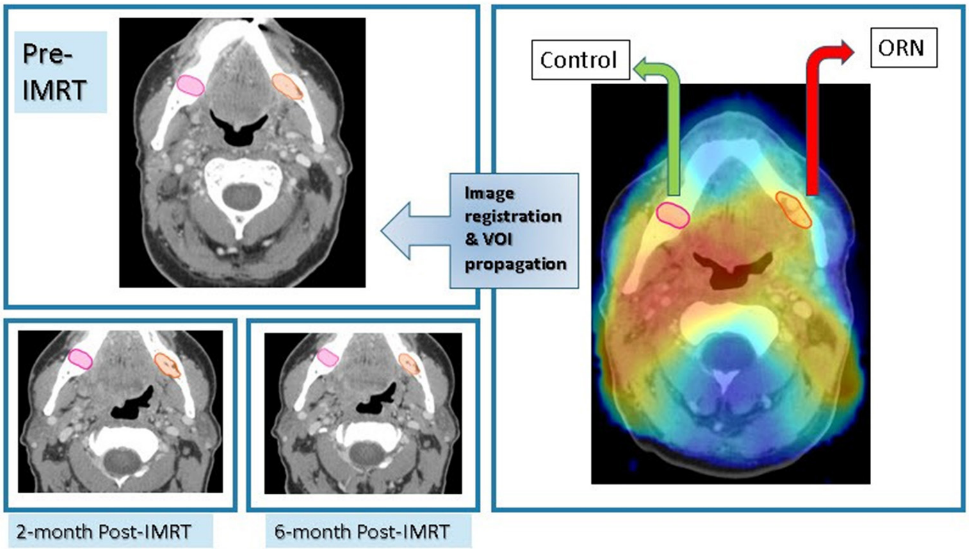


FIGURE 2 | Imaging workflow. Registration of CECT scan at time of diagnosis of ORN to radiation dose grid as well as previous CECT scans at: baseline, 2-month, and 6-month post-RT for each patient with subsequent propagation of ORN & "Control" VOIs.

Radiomic features were derived from two VOIs that correspond to ORN and Control in the 3 prior time points: pre-IMRT, post-RT2, and post-RT6 CECT scans. The number of radiomic features extracted for each VOI summed up to 1,645 individual features. They included a myriad of first- and second-order radiomic features (**Supplementary Table 1**). Second-order radiomic features were calculated in both full 3-dimensional images (3D) as well as 2.5D, i.e., features calculated for each 2-dimensional slice and results were then combined. Other than shape features, a trilinear interpolation voxel resampling filter to 3 mm slice thickness and 1 mm^2 pixel spacing was applied prior to feature extraction to standardize voxel size. First-order feature categories include shape, intensity direct, and intensity histogram. Whereas second-order feature categories encompass: Gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRL) as well as neighborhood intensity difference. For GLCM and GLRL features, calculations from multiple spatial directions were combined to produce one value (Materka and Strzelecki, 1998). For NID, 3 different permutations of neighborhood, i.e., 3, 5, or 7 were employed as in previous projects (Elhalawani et al., 2018,a,b).

Radiomics Features Pre-selection and Reduction

Initially, we worked with radiomic features computed from VOIs corresponding to ORN and Control for 24 patients. The number of radiomic features extracted for each patient is 1,628. Three patients did not have radiomic features computed for the post-RT6 time point and hence completely excluded from subsequent analysis. For these 21 patients, we only kept the radiomic features whose values are available for (i) all 3 time points, and (ii) both in "ORN" and "Control" VOIs. One patient has 2 distinct ORN lesions; accounting for a total number of 43 individual VOIs (22 "ORN" and 21 "Control" VOIs). Thus, we are then left with 1,628 radiomic features from 43 VOIs, i.e., 22 "ORN" and 21 "Control."

Feature reduction by correlation was critical to ensure that the performance of any machine learning algorithm is not degraded because of a high degree of correlation in the features, or multicollinearity (Garg and Tai, 2013). We first compute the Spearman correlation (Landberg et al., 1999; Zar, 2005) for the 1,628 radiomic features at the pre-IMRT time point. We filter out the features whose average correlation level with all the remaining features is greater than a user-defined threshold (Kuhn, 2008). For our data, we used a threshold of 0.5. The threshold was chosen to reasonably balance the dual requirements of multicollinearity reduction and capturing data variation. Following correlation filtering, we reduced the number of features we analyze to 16 features (**Supplementary Table 2**).

First—as a proof of concept—, we sought to establish that radiomics can quantitatively discriminate between ORN and non-ORN mandibular subvolumes. Mann-Whitney test (Mann and Whitney, 1947) was used to identify specific radiomic features that show statistically significant differences between ORN and non-ORN high-risk VOIs.

Functional Principal Component Analysis

We hypothesize that we can predict the risk of ORN by looking at the temporal evolution of radiomic features. A standard way of identifying temporal signatures in time series data is by using functional principal component analysis (FPCA) (Shang, 2014; Aue et al., 2015). FPCA takes multiple time series curves, as an input, and tries to find the underlying shape signatures that optimally can be used to represent all the curves. These shape signatures are called the functional Principal Components (PC). Each time series can now be represented by a weighted combination of each of the PCs. This technique has been used

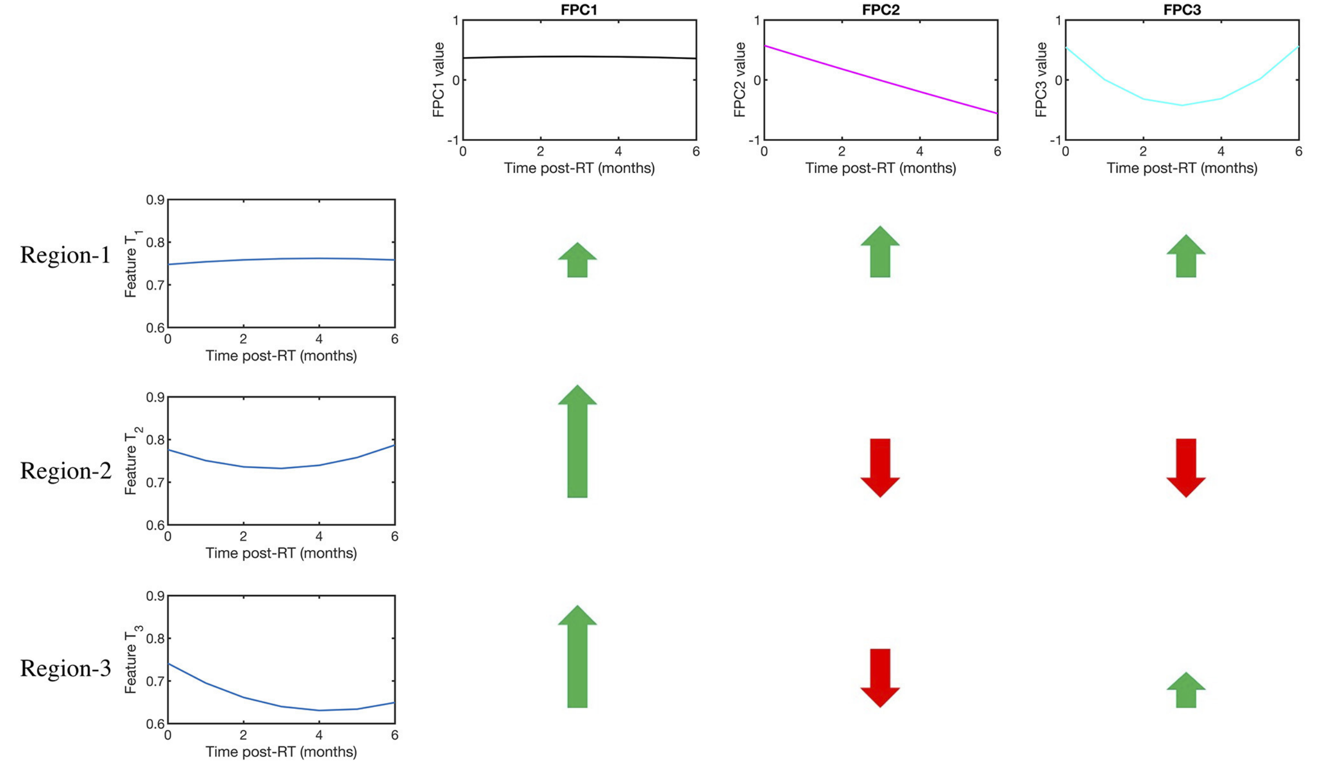


FIGURE 3 | Visual explanation of the FPCA algorithm and its advantages. The first row displays the 3 functional principal components (FPCs). On the left column, the temporal evolution of a Gray Level Co-occurrence Matrix (GLCM)-3D feature is shown for three mandibular regions namely Regions 1,2, and 3. Regions 1 and 2 did not develop ORN, while 3 did. We note that Regions 1,2, and 3 all have similar baseline values, so cannot be distinguished by a model built solely on pre-radiotherapy features. Further, Regions 2 and 3 also have similar change in their values, which a delta radiomics model would see as equivalent scenarios. On the other hand, the difference in the temporal kinetics is efficiently encoded in the 3 FPCs. The color and length of the arrows indicate the sign (+ve or –ve) and magnitude (large or low) of relative contribution made by each FPC in explaining the time series. So, for example, Region 2 and Region 3, which appear alike to a pre-radiotherapy model and a delta radiomics model, can be readily distinguished because of the difference in relative contribution made by the 3rd FPC.

to predict outcomes from sequential data in a wide variety of fields such as remote sensing (Cardot et al., 2003), stock markets (Foutz and Jank, 2010), electroencephalogram (EEG) analysis (Shou et al., 2015), and cancer pathology (Barua et al., 2018). Since our data is multivariate, in that we have a time series for multiple features for the same patient, we can compute the functional PCs for each feature. One way of representation would be to assume each feature is independent, concatenate the PC weights for each feature, and use this concatenated representation as input to a machine learning model. However, since each pair of features is correlated to various degrees, we use a technique called multivariate FPCA (MFPCA), which explicitly accounts for the relationship between the features (Dauxois et al., 1982; Berrendero et al., 2011; Chiou et al., 2014; Happ and Greven, 2018). We utilized the R package MFPCA for our temporal kinetics analysis (Happ and Greven, 2018).

The importance of FPCA is visually explained in **Figure 3**. We display 3 temporal trajectories from our data on the leftmost column. We observe that all 3 sequences T_1, T_2, and T_3, have similar starting points. Further time series T_2 and T_3 have similar end points too. This mimics a significant scenario which we try to address, whereby neither the pre-radiotherapy features, nor the delta features can distinguish between the patients. However, FPCA can distinguish all 3, by accounting for both, the values taken by the time series, and the shape of the trajectory. The top 3 FPCs representing the dataset are shown visually in the top row. The relative contribution of each FPC to each of the time series is shown with arrows, the length of the arrows representing the magnitude, and green and red color indicating the sign (positive and negative, respectively) of the contributions. We can see that the magnitude and sign of the individual contributions from the PCs are quite different, and thus can help distinguish the three-time series.

Training the Random Forest

We used repeated random sampling to produce random forests where validation (Breiman, 2001) ensued where validation of each forest was performed using the left out observations, and the overall accuracy was calculated by averaging the class predictions of each of the forests. The random forest has been shown to be robust to over-fitting and among the most effective of the commonly used classifiers (Breiman, 2001). Each forest used 500 trees, and each split was determined using $\sqrt{p}$ features where p is the number of features. The random forest calculations were performed using the random Forest package for R software (Liaw and Wiener, 2002). To further examine the performance of the model, the ROC curves were plotted and the area under the curve (AUC) was calculated using pROC package for R (Robin et al., 2011).

RESULTS

Patient Information

Twenty-one patients with oropharyngeal cancer (OPC) were identified to have developed ORN after their definitive radiotherapy ± chemotherapy course, either in induction and/or concurrent settings as in **Figure 1**. Eight patients developed grade 2 ORN, whereas 2 patients and 11 patients developed grade 3 and 4 ORN, respectively. The median time to ORN diagnosis was 20.3 months. **Table 1** represents patient demographics, tumor, radiation dose, and ORN disease characteristics.

Radiomics Can Distinguish Between ORN and Non-ORN

An initial set of 1,628 radiomic features were computed for each ORN and Control volume of interest (VOI) obtained from the 21 eligible patients across 3 time points of interest representing baseline (pre-IMRT), 2-month (post-RT2) post-IMRT, and 6-month (post-RT6) post-IMRT. Sixteen radiomics features were ultimately nominated as non-interrelated and consistently available for all three time points. As an initial exploratory step, we computed which of these 16 radiomic features were significantly different between the ORN and Control volumes of interest (VOIs) using a Mann-Whitney test. Furthermore, we also computed if each of these features is larger, or smaller, on average for the ORN VOI compared to the Control VOI. This demonstrates that certain radiomic features differ significantly between ORN and non-ORN regions, motivating us to investigate if their evolution can foretell ORN incidence. The significantly different features and their associated p-values are reported in **Table 2**.

The radiomics features which values are significantly different between the "ORN" and "Control" VOIs at the ORN time point identified using a Mann-Whitney test. The corresponding p-value is reported in the second column. We also report the direction of the difference of means between the ORN and Control VOI feature values in the third column.

Model Construction

We trained random forest models using 500 trees for each of multiple approaches as outlined below: (**Figure 4**)

- **Baseline:** Radiomic features computed on the pre-IMRT CECT scans.
- **Delta (2-month follow-up):** Relative change in the radiomic features from pre-IMRT to post-RT2
- **Delta (6-month follow-up):** Relative change in the radiomic features from pre-IMRT to post-RT6.
- **Temporal Trajectory:** The model built using the proposed multivariate functional principal component analysis

TABLE 1 | Patients, disease, and treatment characteristics.

Characteristics	*N* (%)
SEX	
Male	20 (95.2%)
Female	1 (4.8%)
Age at diagnosis, years: median (range)	61 (57–68)
ETHNICITY	
White or Caucasian	17 (81%)
Hispanic or Latino	2 (9.5%)
African American	2 (9.5%)
SMOKING STATUS	
Current	10 (47.6%)
Former	5 (23.8%)
Never	6 (28.6%)
Smoking pack-years (median; IQR)	10 (0–40.5)
TUMOR LATERALITY	
Right	9 (42.9%)
Left	11 (52.4%)
Midline	1 (4.7%)
OROPHARYNX SUBSITES	
Base of tongue	12 (57.1%)
Tonsil	7 (33.3%)
NOS*	2 (9.6%)
T CATEGORY	
T1	2 (9.5)
T2	10 (47.6%)
T3	5 (23.8%)
T4	4 (19.1%)
N CATEGORY	
N0	2 (9.5%)
N1	0
N2	19 (90.5%)
N3	0 (0)
THERAPEUTIC COMBINATION	
Induction chemotherapy (IC) followed by concurrent chemoradiation	10 (47.6%)
IC followed by radiation alone	1 (4.8%)
CC	10 (47.6%)
VITAL STATUS	
Alive	14 (66.7%)
Dead	7 (33.3%)
Radiation dose (median; IQR) [Gy]	70 (66–70)
Radiation fractions (median; IQR)	33 (30–33)
Onset of post-RT ORN (median; IQR)	20.3 (7.5–95)
ORN LATERALITY (IN RELATION TO PRIMARY TUMOR)	
Ipsilateral	17 (81%)
Contralateral	2 (9.5)
Bilateral	2 (9.5%)
RADIATION DOSE AT THE ORN VOLUME (MEDIAN; IQR) [GY]	
Mean dose	67.9 (59.5–71.2)
Minimum dose	51 (44–59.4)
Maximum dose	68.9 (67.6–73.1)

IQR, inter-quartile range; Gy, Gray; NOS, Not otherwise specified; ORN, osteoradionecrosis.

TABLE 2 | Significantly differing radiomics features between ORN and Control VOIs.

Feature	*p*-value	Mean difference of feature value between ORN and Control feature values
Gray Level Co-occurrence Matrix 25-333-1 InformationMeasureCorr1	0.028	Negative
Gray Level Co-occurrence Matrix 312-4 Cluster Shade	0.034	Positive
Gray Level Co-occurrence Matrix 310-1 Dissimilarity	0.009	Positive
Gray Level Co-occurrence Matrix 38-1InverseDiffMomentNorm	0.0002	Negative
Intensity- Mean	2.43E-7	Negative
Intensity- Local entropy median	4.65E-6	Negative

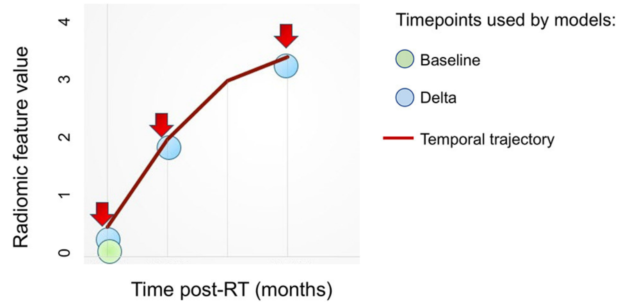


FIGURE 4 | Overview of radiomics features based approaches. Various approaches to integrate radiomics features obtained at multiple (≥1) time points toward building predictive models.

(MFPCA) approach that models the temporal kinetics of the features. Since the time points are not uniformly spaced, we used cubic spline sequence completion to fill in radiomic features at intermediate monthly time points.

- **Baseline + Temporal Trajectory:** We combined the predictions from the baseline model and the temporal trajectory model to give a more robust ORN-risk predictor.

A complete step-by-step guide for the model construction pipeline is presented in **Appendix A1**.

TABLE 3 | A comparison of the Areas under the curves (AUCs) and the 95% confidence intervals for the various approaches.

Method	AUC (95% CI)
Baseline	0.59 (0.41–0.76)
Delta (2-month follow-up)	0.64 (0.46–0.81)
Delta (6-month follow-up)	0.56 (0.39–0.74)
Temporal trajectory	0.74 (0.61–0.90)
Baseline + Temporal trajectory	0.68 (0.53–0.86)

The corresponding areas under the curves (AUCs) and 95% confidence intervals (C.I.) for the prediction of occurrence of ORN "Yes vs. No," in both "ORN" and "Control" VOIs according to the 5 models are depicted in **Table 3** and illustrated in **Figure 5**. We noticed that the baseline features model gives an AUC of 0.59 (95% C.I: 0.41–0.76), while the temporal trajectory gives an AUC of 0.74 (95% C.I: 0.61–0.9). We further built an ensemble model that combines the predictions of the baseline model and the temporal trajectory model, to see if these two models have complementary information that improves performance. We achieved an AUC of 0.68 (95% C.I: 0.53–0.86), likely due to the poor performance of the baseline model which consequently was detrimental to the performance of the combined model. This suggests a more careful approach is needed when choosing pre-IMRT features. Surprisingly, models constructed using percent changes "or delta changes" of the radiomic feature values, performed poorly in predicting ORN incidence with AUCs of 0.64 (95% C.I: 0.46–0.81) and 0.56 (95% C.I: 0.39–0.74) for 2 and 6-month delta changes, respectively. We further observe that the temporal trajectory and combined models have a consistent performance in both low-specificity and high-specificity regimes, in contrast to the delta models which performance is dependent on the regime of choice. This demonstrates that greater reliability is achieved by incorporating the temporal kinetics of the radiomic features. We do note that as a result of the small sample size, the confidence intervals of the models are wide and overlapping. As such, larger validation studies are needed to gauge the true performance of the models.

To enable the use of our temporal trajectory model for the stated aim of ORN prediction, we compute the optimal point on the ROC curve as the point that maximizes the Youden's index (sensitivity+specificity-1) (Youden, 1950). As shown in **Supplementary Figure 1**, the optimal point corresponds to a sensitivity of 0.73 and specificity of 0.62. The optimal threshold for the temporal trajectory model, which represents the cutoff probability value above which a given mandibular region is predicted to be "Control" is found to be 0.54. Thus, if the temporal trajectory model predicts the likelihood of a given region as lower than 0.54, the region is classified as "ORN" and if the probability is higher, the region is classified as "Control." We next generate the confusion matrix for the 43 regions classified using the optimal threshold value; 72.7% of ORN regions and 61.9% of Control regions were correctly classified as shown in **Supplementary Figure 2**.

DISCUSSION

The incidence of head and neck cancer is on the rise, despite reductions in smoking, owing to the recent prevalence of the human papillomavirus (HPV)-associated OPC epidemic (Ang et al., 2010). Forward, it's projected that hundreds of thousands of locally advanced OPC patients worldwide will receive radiation to the head and neck as a definitive treatment modality (Chaturvedi et al., 2011). This rise in RT recipients implies that mandibular bone, which comprises the borders of the oropharynx, will be necessarily irradiated to ensure adequate tumor coverage with subsequently growing incidence of crippling sequelae such as ORN (Gomez et al., 2011).

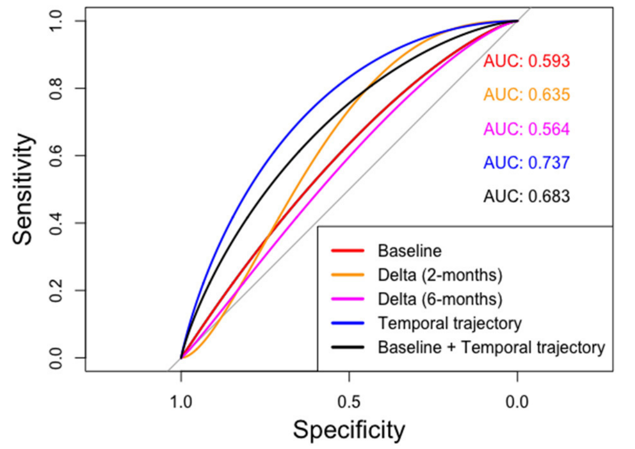


FIGURE 5 | ROC curves computed for various radiomics feature based approaches. The temporal trajectory model using MFPCA (blue) performs better than the other four models: (i) baseline model (red), (ii) delta model after 2-month follow-up (orange), (iii) delta model after 6-month follow-up (purple), and (iv) an ensemble of baseline and temporal trajectory models (black).

Osteoradionecrosis ranges from superficial, slowly progressive bone erosion/devitalization to pathological fracture in a previously irradiated field and may cause significant hardship in the afflicted individual (Mendenhall, 2004; Hamilton et al., 2012). This is particularly apparent when considering devastating lifelong issues with oral hygiene, nutritional inadequacies, and difficulty with speech and resultant preclusion of social interaction (Bonner et al., 2006). Early diagnosis and intervention, whether conservative or surgical, are key for improving outcomes (Ben-David et al., 2007). This essentially applies for grade II ORN, where no consensus has been reached regarding definitive treatment procedures (Oh et al., 2009; Jacobson et al., 2010).

Using CT Radiomics to Identify Mandibular Subvolumes At-Risk of ORN

To date, no imaging modality/clinical nomogram have been shown to precisely foresee the potential risk of developing osteoradionecrosis following IMRT (Allison et al., 2014). Being fully integrated throughout various phases of HNC management, sequential CECT scans via radiomics analytics can provide a plethora of data that can serve as quantifiable surrogates of tissue vitality and vascularity, among others (Wong et al., 2016). To our knowledge, this study is the first to characterize the kinetics of radiomics features of various mandibular subvolumes, before and after exposure to IMRT, to identify subvolumes at high risk ahead of developing ORN. Radiomics features were analyzed longitudinally for quantifying temporal changes in mandibular bone structure in a cohort of OPC patients.

Applying FPCA to Capture Longitudinal Changes in Mandibular Radiomic Features

This has been subsequently integrated into a framework for early prediction of ORN solely based on sequential diagnostic CECT scans. We implemented a Functional Principal Component Analysis (FPCA)-based approach that efficiently models the temporal evolution of radiomic features. The model built using a multivariate FPCA (MFPCA) representation of the entire temporal dataset, predicts the likelihood of ORN development with an AUC = 0.74 (95% C.I 0.61–0.9). We further built an ensemble model that combines the predictions of a baseline model built using pre-IMRT features, and the MFPCA-based model, to leverage information from both baseline feature values and temporal evolution of feature values, which achieved an AUC of 0.68 (0.53–0.86). This emulates the pathophysiology theories that combine pre-irradiation bone condition and RT-induced alterations on tissue, cellular and cytokine levels (Fan et al., 2014). The latter involves: (1) endarteritis and vascular thrombosis with subsequent bone hypoxia and hypocellularity as well as atrophic fibrosis as a consequence of RT-induced activation and dysregulation of fibroblastic activity (Marx, 1983; Jacobson et al., 2010). The fact that the ensemble model does not perform better than the MFPCA-only model suggests the need to choose the pre-IMRT features in a way that is more clinically meaningful than a purely data-driven correlation thresholding approach.

Bone texture analysis has been investigated for years as a potential biomarker of a myriad of structural bone changes related to osteoporosis (Ollivier et al., 2013; Roberts et al., 2013). Interestingly, first-order bone texture features derived from simulation CT scans were correlated to the risk of radiation-induced insufficiency fractures in patients undergoing pelvic radiation (Nardone et al., 2017). Along the same lines, for vascularization status, a previous study by Yin et al. investigated the correlation between angiogenesis (or: new blood vessel formation) in primary renal cell carcinoma and radiomic imaging features from positron-emission tomography (PET) and/or MRI (Yin et al., 2017).

Our study identifies the bone radiomics features which temporal evolution is critical in determining ORN risk. These represent quantifiable imaging biomarkers that capture various intensity and spatial texture dimensions of the aforementioned RT-related bone environment changes in the irradiated field. Most of the discriminating features belong to: “Neighborhood intensity difference” (NID) and “Gray level co-occurrence matrix” (GLCM) categories. The GLCM is a matrix that expresses how combinations of discretized gray levels of neighboring pixels, or voxels in a 3D volume, are distributed along one of the image directions. Generally, the neighborhood for GLCM is a 26-connected neighborhood in 3D and an 8-connected neighborhood in 2D (Liang et al., 2016). The “NID 2.5D Texture strength” quantifies how uniform a texture is, i.e., complex textures are non-uniform and rapid changes in gray levels are common (Amadasun and King, 1989). GLCM3 Cluster shade is a measure of the skewness or asymmetry of the matrix and is believed to be a more objective uniformity metric (Unser, 1986). On the other hand, GLCM3 Contrast gauges gray level variations in the volume of interest, i.e., the difference between the highest and the lowest values of a continuous set of pixels (Haralick et al., 1973). GLCM3 Correlation is a measure of texture smoothness, where higher values denote regions with similar gray-levels (Yang et al., 2012). Nonetheless, it is unclear how these radiomic features are linked to well-known physiological underpinnings of ORN evolution. A future validation study including biological imaging is warranted to investigate the link between these radiomic features and physiological properties.

We have seen that there is significant information regarding ORN progression in the first 6 months after radiotherapy that can be robustly correlated to risk of ORN. Functional principal component analysis is an efficient statistical algorithm to capture the temporal evolution of the mandible landscape. Competing techniques such as pre-radiotherapy only models and delta radiomics models do not encapsulate how different features evolve with time. The FPCA efficiently encodes the temporal kinetics of the features into its functional principal components (FPCs). The radiomics data can now be compactly represented by only a small set of numbers but can still capture its time-varying properties.

Furthermore, we implement a multivariate FPCA (MFPCA) that accounts for the correlations that exist between various radiomics features. MFPCA distills a large set of features to a few specific ones that encompass most of the data variation. This makes our prediction model more likely to generalize to new, unseen data (Happ and Greven, 2018). We observe from the receiver operating characteristic curves that the temporal trajectory model performs consistently better than the other models in both the high- and low-specificity or false positive regions. This demonstrates the reliability of using temporal kinetics, for example, compared to a delta model, which we observed to have vastly different performance depending on the specificity value. The combined prediction model does not improve over the temporal trajectory only model, possibly because of the extremely poor performance of the baseline model. However, the combined model also performs consistently in both the low- false positive or high false positive regimes. We envisage that with a more careful choice of features, the baseline model can be improved, which will significantly improve the performance of the combined model. We note however that while the average performance of the MFPCA model is at least 10 percentage points better than either the baseline or the two delta models, the respective confidence intervals are overlapping across models. As such, larger validation studies are needed to find out the true predictive ability of each of the models investigated.

The preliminary feature filtering step was performed by setting an upper limit of 0.5 on average correlation value for a given radiomic feature. Meaning, if a given radiomic feature correlated with all other features more than 0.5 on average, it was dropped from our feature set. The choice of value was made to whittle the number of features down from a mammoth 1,628 to a more manageable 16 given the small sample size of our cohort. The reduction of features is necessary to compute robust functional principal components as well as reduce the possibility of overfitting by the random forest models. We also found that reducing the number of features further led to a drop in the model performance, which suggests loss of information crucial to prediction performance.

Our study accounted for the fact that artifacts from metal dental fillings are known to encumber target delineation and subsequent radiomics analysis (Leijenaar et al., 2015; Block et al., 2018). For this purpose, the presence of visible dental artifacts effect anywhere in the slices that encompassed "ORN" or "Control" VOIs at any time point precluded the integration of this scan and hence the patient's data as an input to the model.

Study Limitations

The fact that we excluded these patients with metal dental fillings, combined with the low event rate of ORN in the IMRT era, as well as the fact that we excluded patients with grade I ORN with no radiographically-evident bone lesions to delineate, contributed to the low sample size; hence limiting the generalizability of the resulting model. The small sample size limited us to apply automatically generated radiomics features instead of engineering features that are explicit surrogates for early vascular injuries of the mandible. Sub-group analysis based on variables such as T-stage, radiation dose, and chemotherapy usage were infeasible because smaller sample sizes within each group reduces the robustness of the functional principal components computed and hence the statistical value of any subsequent sub-group analysis. Another limitation of this study is the conceivable uncertainties introduced from varied acquisition parameters or incongruence among various scanners, or even between different models from the same vendor (Mackin et al., 2015). Most patients had their scan performed at our center along the same acquisition parameters. Moreover, we have applied a pre-processing trilinear interpolation aiming at standardizing voxel size to reduce or eliminate relevant variability in radiomics features (Mackin et al., 2017). The results also suggested that the performance changed rapidly when we changed the number of features, which suggests the need for a more careful feature-filtering algorithm. Designating a "Control" VOI that share the same image, time point, and deposited radiation dose with the "ORN" VOI is an approach we have used and would recommend for future multi-institutional radiomics studies. However, it should be noted that our model was trained on a homogenous, carefully selected set of patients with OPC where mandibles received similarly high doses of radiation; hence limiting model generalization to varying clinical scenarios.

Future Directions

Not far from longitudinal imaging studies, our team previously showed that Dynamic Contrast-Enhanced (DCE-MRI) can provide biomarkers that are physiological correlates of acute mandibular vascular injury and recovery temporal kinetics (Joint and Neck Radiotherapy, 2016). This has further motivated a National Institute of Dental and Craniofacial Research (NIDCR)-funded prospective trial that explores the correlation between DCE-MRI derived spatiotemporal parameter maps following external beam radiation therapy (EBRT) and subsequent development of ORN (ClinicalTrials.gov S., 2020). Upon accrual completion, CT scans from this study will be used for re-training and externally validating our model. This could potentially optimize model generalization since patients will display more diverse and representative head and neck cancer sites, radiation doses, and other clinical variables. Our results may prompt the investigation of DCE-MRI-derived radiomics analytics and subsequent integration into the overall predictive model; thus, providing more physiologically and biologically cognizant data inputs for the machine learning techniques tested.

Furthermore, the availability of larger cohorts will provide potential avenues for model validation and generalization over the whole mandible in patients with ORN vs. healthy controls. Specifically, a larger cohort would make it possible to examine the performance of our FPCA-based model across T-stage, radiation dose, and chemotherapy usage, providing additional insights into the impact of these variables on ORN development. In future validation studies, we plan to enroll more patients with more evenly distributed variable levels. A proposed application would be engineering radiomics features that are explicit surrogates for osteoclastic dysregulation and subsequent fibro-atrophic bone changes, and maybe monitoring the response to common therapeutic maneuvers, such as pentoxifylline.

CONCLUSION

Radiomics analysis allows for quantification of changes in RT-related bone structure from diagnostic imaging modalities with subsequent integration of serially derived radiomics features into an ORN probability computational tool. Computationally, FPCA efficiently encodes the temporal kinetics of a given radiomic feature. The MFPCA then compactly combines the temporal information from FPCA from multiple radiomic features.

In summary, we hope this study calls professionals' attention to non-traditional inputs (radiomics), dimensions (temporal kinetics), and innovative statistical approaches (MFPCA) to improve interpretation and integration of imaging biomarkers into RT toxicities prediction and mitigation. In this work, we have thus provided an end-to-end framework for predicting the risk of RT-related ORN based entirely on radiomic features.

DATA AVAILABILITY STATEMENT

Clinical dataset is not available as it includes personal health identifiers (PHI). It is possible for de-identified data to be made available upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by MD Anderson IRB RCR-03-800. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All authors performed substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, drafting the work or revising it critically for important intellectual content, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. SB and HE: manuscript writing, conceived study idea, and designed workflow. SB designed the temporal kinetics software framework to filter radiomics features and compute FPCA, implemented the random forest models to compare pre-IMRT, delta, and FPCA models, and conducted statistical analyses. HE: direct oversight of image segmentation/image post-processing, supervision of DICOM-RT analytic workflows and clinical data collection workflows. SV, KA, PY, SN, and BE: electronic medical record screening, data extraction, image segmentation/registration, and clinical data collection. RG: Informatics software support. MC: Database construction, clinical/oncologic oropharynx database curation and oversight, conceptual feedback, and support. KH, and GG: data provision, patient case extraction, supervisory support, editorial oversight, programmatic oversight under Stiefel Program activities. DM and LC: development support for radiomics workflow and curation of the radiomics-based image features. AM: supervision of image segmentation/image post-processing and clinical data collection, and manuscript editing. SL and CF: co-corresponding authors, primary investigators, conceived, coordinated, and directed all study activities, responsible for data collection, project integrity, manuscript content and editorial oversight and correspondence, direct oversight of trainee personnel. AR: co-corresponding author, primary investigator, conceived, coordinated, and directed all study activities, supervised SB: toward building the FPCA and random forest models, responsible for project integrity, manuscript content, editorial oversight and correspondence, analytic support, and conceptual advice regarding model construction. Preliminary analyses and portions of this data were presented as a poster at the 2018 American Society of Radiation Oncology (ASTRO) Multidisciplinary Head and Neck Cancer Symposium, February 15–17, 2018, Scottsdale, AZ, USA.

FUNDING

SB and AR were supported by CCSG Bioinformatics Shared Resource P30 CA016672, CPRIT RP170719, CPRIT RP150578, NCI 1R37CA214955-01A1, a gift from Agilent technologies, and a Research Scholar Grant from the American Cancer Society (RSG-16-005-01). AR was a member of Voxel Analytics, LLC, reports receiving other commercial research support from Agilent Technologies. HE was supported in part by the philanthropic donations from the Family of Paul W. Beach to GG, HE, and CF receive funding and project-relevant salary support from NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007-10). This research was supported by the Andrew Sabin Family Foundation; CF was a Sabin Family Foundation Fellow. CF, SL, and AM receive funding and project-relevant salary support from the National Institutes of Health (NIH), including: National Institute for Dental and Craniofacial Research Award (1R01DE025248-01/R56DE025248-01). CF, SL, KH, and AM receive funding and project-relevant salary support from NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148-01). CF and AM receive funding and project-relevant salary support from National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825-01). CF receive funding and project-relevant salary support from NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672) and National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787). CF has received direct industry grant support and travel funding from Elekta AB. AM was funded for this work by The University of Texas MD Anderson Cancer Center-Oropharynx Cancer Program generously supported by Mr. and Mrs. Charles W. Stiefel.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.medrxiv.org/content/10.1101/2020.10.09.20208827v1 (Barua et al., 2020). The authors would like to acknowledge the funding support they received, which are detailed in the funding statement. We would also like to thank the assistance of M.D Anderson's scientific publishing team in greatly enhancing the readability of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.618469/full#supplementary-material

Supplementary Figure 1 | ROC curve for the FPCA-based temporal trajectory model with optimal operating point. The optimal operating point of the ROC curve is identified by maximizing the Youden's index. The corresponding sensitivity and specificity values for the optimal operating point are shown.

Supplementary Figure 2 | Confusion matrix for the ORN prediction task at the optimal operating point. The confusion matrix showing the classification performance by the FPCA-based temporal trajectory model at the optimal operating point of the ROC curve.

Supplementary Table 1 | Computed tomography- derived intensity histogram, shape and texture analysis features set.

Supplementary Table 2 | Filtered radiomic feature set. The final set of 16 features chosen after correlation filtering, which are used for building baseline, delta, and FPCA models.

Appendix A1 | Step-by-step implementation guide for the model construction pipeline.

REFERENCES

Allison, R. R., Patel, R. M., and McLawhorn, R. A. (2014). Radiation oncology: physics advances that minimize morbidity. *Fut Oncol.* 10, 2329–2344. doi: 10.2217/fon.14.176

Allison, R. R., Sibata, C., and Patel, R. (2013). Future radiation therapy: photons, protons and particles. *Fut Oncol.* 9, 493–504. doi: 10.2217/fon.13.13

Amadasun, M., and King, R. (1989). Textural features corresponding to textural properties. *IEEE Transact. Syst. Man Cybernet.* 19, 1264–1274. doi: 10.1109/21.44046

Ang, K. K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D. I., Nguyen-Tân, P. F., et al. (2010). Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* 363, 24–35. doi: 10.1056/NEJMoa0912217

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *J Am Statist Associat.* 110, 378–392. doi: 10.1080/01621459.2014.909317

Barua, S., Elhalawani, H., Volpe, S., Feghali, K. A., Yang, P., Ng, S. P., et al. (2020). Discovering early imaging biomarkers of osteoradionecrosis in oropharyngeal cancer by characterization of temporal changes in computed tomography mandibular radiomic features. *medRxiv.* 2020:2020.10.09.20208827. doi: 10.1101/2020.10.09.20208827

Barua, S., Solis, L., Parra, E. R., Uraoka, N., Jiang, M., Wang, H., et al. (2018). A functional spatial analysis platform for discovery of immunological interactions predictive of low-grade to high-grade transition of pancreatic intraductal papillary mucinous neoplasms. *Cancer Informat.* 17:1176935118782880. doi: 10.1177/1176935118782880

Ben-David, M. A., Diamante, M., Radawski, J. D., Vineberg, K. A., Stroup, C., Murdoch-Kinch, C. A., et al. (2007). Lack of osteoradionecrosis of the mandible after intensity-modulated radiotherapy for head and neck cancer: likely contributions of both dental care and improved dose distributions. *Int J Radiation Oncol Biol Phys.* 68, 396–402. doi: 10.1016/j.ijrobp.2006.11.059

Berrendero, J. R., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computat. Statist. Data Analysis.* 55, 2619–2634. doi: 10.1016/j.csda.2011.03.011

Block, A. M., Cozzi, F., Patel, R., Surucu, M., Hurst, N., Jr., et al. (2018). Radiomics in head and neck radiation therapy: impact of metal artifact reduction. *Int. J. Radiation Oncol. Biol. Phys.* 99:E640. doi: 10.1016/j.ijrobp.2017.06.2146

Bonner, J. A., Harari, P. M., Giralt, J., Azarnia, N., Shin, D. M., Cohen, R. B., et al. (2006). Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* 354, 567–578. doi: 10.1056/NEJMoa053422

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cacicedo, J., Navarro, A., Del Hoyo, O., Gomez-Iturriaga, A., Alongi, F., Medina, J. A., et al. (2016). Role of fluorine-18 fluorodeoxyglucose PET/CT in head and neck oncology: the point of view of the radiation oncologist. *Br J Radiol.* 89:20160217. doi: 10.1259/bjr.20160217

Cardot, H., Faivre, R., and Goulard, M. (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *J. Appl. Statistics* 30, 1185–1199. doi: 10.1080/0266476032000107187

Chaturvedi, A. K., Engels, E. A., Pfeiffer, R. M., Hernandez, B. Y., Xiao, W., Kim, E., et al. (2011). Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J. Clin. Oncol.* 29, 4294–4301. doi: 10.1200/JCO.2011.36.4596

Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica* 24, 1571–1596. doi: 10.5705/ss.2013.305

ClinicalTrials.gov S. (2020). ClinicalTrials.gov [Internet] Bethesda (MD): National Library of Medicine (US). 2000 Feb 29 -. Identifier NCT03145077, Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Diagnosing Osteoradionecrosis in Patients With Head and Neck Cancer That Is Primary, Has Come Back, or Has Spread to Other Places in the Body (2020). Available online at: https://clinicaltrials.gov/ct2/show/NCT03145077?term=stephen$+$laianddraw=2andrank=1 (accessed Feb 21, 2021).

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Analysis* 12, 136–154. doi: 10.1016/0047-259X(82)90088-4

Elhalawani, H., Kanwar, A., Mohamed, A. S. R., White, A., Zafereo, J., Wong, A., et al. (2018). Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. *Sci. Rep*. 8:1524. doi: 10.1038/s41598-017-14687-0

Elhalawani, H., Mohamed, A. S. R., Kanwar, A., Dursteler, A., Rock, C. D., Eraj, S. E., et al. (2018b). EP-2121: serial parotid gland radiomic-based model predicts post-radiation xerostomia in oropharyngeal cancer. *Radiother. Oncol*. 127, S1167–S8. doi: 10.1016/S0167-8140(18)32430-7

Elhalawani, H. E., Mohamed, A. S. R., Volpe, S., Yang, P., Campbell, S., Granberry, R., et al. (2018a). PO-0991: serial tumor radiomic features predict response of head and neck cancer treated with Radiotherapy. *Radiother. Oncol*. 127:S551. doi: 10.1016/S0167-8140(18)31301-X

Fan, H., Kim, S. M., Cho, Y. J., Eo, M. Y., Lee, S. K., and Woo, K. M. (2014). New approach for the treatment of osteoradionecrosis with pentoxifylline and tocopherol. *Biomater. Res*. 18:13. doi: 10.1186/2055-7124-18-13

Foutz, N. Z., and Jank, W. (2010). Research note-prerelease demand forecasting for motion pictures using functional shape analysis of virtual stock markets. *Market. Sci*. 29, 568–579. doi: 10.1287/mksc.1090.0542

Freymann, J. B., Kirby, J. S., Perry, J. H., Clunie, D. A., and Jaffe, C. C. (2012). Image data sharing for biomedical research—meeting HIPAA requirements for de-identification. *J. Digital Imaging* 25, 14–24. doi: 10.1007/s10278-011-9422-x

Garden, A. S., Dong, L., Morrison, W. H., Stugis, E. M., Glisson, B. S., Frank, S. J., et al. (2013). Patterns of disease recurrence following treatment of oropharyngeal cancer with intensity modulated radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys*. 85, 941–947. doi: 10.1016/j.ijrobp.2012.08.004

Garg, A., and Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *Int. J. Modell. Identificat. Control* 18, 295–312. doi: 10.1504/IJMIC.2013.053535

Gomez, D. R., Estilo, C. L., Wolden, S. L., Zelefsky, M. J., Kraus, D. H., Wong, R. J., et al. (2011). Correlation of osteoradionecrosis and dental events with dosimetric parameters in intensity-modulated radiation therapy for head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys*. 81:e207–e13. doi: 10.1016/j.ijrobp.2011.02.003

Hamilton, J. D., Lai, S. Y., and Ginsberg, L. E. (2012). Superimposed infection in mandibular osteoradionecrosis: diagnosis and outcomes. *J. Computer Assisted Tomogr*. 36, 725–731. doi: 10.1097/RCT.0b013e3182702f09

Happ, C., and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (Dimensional) domains. *J. Am. Statistical Associat*. 113, 649–659. doi: 10.1080/01621459.2016.1273115

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transact. Syst. Man Cybernet*. SMC-3, 610-21. doi: 10.1109/TSMC.1973.4309314

Jacobson, A. S., Buchbinder, D., Hu, K., and Urken, M. L. (2010). Paradigm shifts in the management of osteoradionecrosis of the mandible. *Oral Oncol*. 46, 795–801. doi: 10.1016/j.oraloncology.2010.08.007

Joint, H., and Neck Radiotherapy, M. R. (2016). Dynamic contrast-enhanced MRI detects acute radiotherapy-induced alterations in mandibular microvasculature: prospective assessment of imaging biomarkers of normal tissue injury. *Sci. Rep*. 6:29864. doi: 10.1038/srep29864

Kuhn, M. (2008). Building predictive models in R using the caret *Package*. 2008. 28,26. doi: 10.18637/jss.v028.i05

Landberg, T., Chavaudra, J., Dobbs, J., Gerard, J. P., Hanks, G., Horiot, J. C., et al. (1999). Report 62. *J. Int. Commission Radiat. Units Measurements* os32, NP-NP. doi: 10.1093/jicru/os32.1.Report62

Leijenaar, R. T. H., Carvalho, S., Hoebers, F. J. P., Aerts, H. J. W. L., van Elmpt, W. J. C., Huang, S. H., et al. (2015). External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncolog*. 54, 1423–1429. doi: 10.3109/0284186X.2015.1061214

Liang, C., Huang, Y., He, L., Chen, X., Ma, Z., Dong, D., et al. (2016). The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget*. 7, 31401–31412. doi: 10.18632/oncotarget.8919

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News*. 2, 18–22.

Mackin, D., Fave, X., Zhang, L., Fried, D., Yang, J., Taylor, B., et al. (2015). Measuring CT scanner variability of radiomics features. *Invest. Radiol*. 50, 757–765. doi: 10.1097/RLI.0000000000000180

Mackin, D., Fave, X., Zhang, L., Yang, J., Jones, A. K., Ng, C. S., et al. (2017). Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE* 12:e0178524. doi: 10.1371/journal.pone.0178524

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist*. 18, 50–60. doi: 10.1214/aoms/1177730491

Marx, R. E. (1983). Osteoradionecrosis: a new concept of its pathophysiology. *J Oral Maxillofac Surg*. 41:283–8. doi: 10.1016/0278-2391(83)90294-X

Materka, A., and Strzelecki, M. (1998). *Texture Analysis Methods —A Review*. Institute of Electronics, Technical University of Lodz, Brussels.

Mendenhall, W. M. (2004). Mandibular osteoradionecrosis. *J. Clin. Oncol*. 22, 4867–4868. doi: 10.1200/JCO.2004.09.959

Mohamed, A. S. R., Hobbs, B. P., Hutcheson, K. A., Murri, M. S., Garg, N., Song, J., et al. (2017). Dose-volume correlates of mandibular osteoradionecrosis in Oropharynx cancer patients receiving intensity-modulated radiotherapy: results from a case-matched comparison. *Radiother. Oncol*. 124, 232–239. doi: 10.1016/j.radonc.2017.06.026

Nardone, V., Tini, P., Carbone, S. F., Grassi, A., Biondi, M., Sebaste, L., et al. (2017). Bone texture analysis using CT-simulation scans to individuate risk parameters for radiation-induced insufficiency fractures. *Osteoporosis Int*. 28, 1915–1923. doi: 10.1007/s00198-017-3968-5

Oh, H.-K., Chambers, M. S., Garden, A. S., Wong, P.-F., and Martin, J. W. (2004). Risk of osteoradionecrosis after extraction of impacted third molars in irradiated head and neck cancer patients. *J. Oral Maxillofacial Surg*. 62, 139–144. doi: 10.1016/j.joms.2003.08.009

Oh, H.-K., Chambers, M. S., Martin, J. W., Lim, H.-J., and Park, H.-J. (2009). Osteoradionecrosis of the mandible: treatment outcomes and factors influencing the progress of osteoradionecrosis. *J. Oral and Maxillofacial Surg*. 67, 1378–1386. doi: 10.1016/j.joms.2009.02.008

Ollivier, M., Le Corroller, T., Blanc, G., Parratte, S., Champsaur, P., Chabrand, P., et al. (2013). Radiographic bone texture analysis is correlated with 3D microarchitecture in the femoral head, and improves the estimation of the femoral neck fracture risk when combined with bone mineral density. *Eur. J. Radiol*. 82, 1494–1498. doi: 10.1016/j.ejrad.2013.04.042

Pan, H. Y., Haffty, B. G., Falit, B. P., Buchholz, T. A., Wilson, L. D., Hahn, S. M., et al. (2016). Supply and demand for radiation oncology in the United States: updated projections for 2015 to 2025. *Int. J. Radiation Oncol. Biol. Phys*. 96, 493–500. doi: 10.1016/j.ijrobp.2016.02.064

Roberts, M. G., Graham, J., and Devlin, H. (2013). Image texture in dental panoramic radiographs as a potential biomarker of osteoporosis. *IEEE Transact. Bio-Medical Eng*. 60, 2384–2392. doi: 10.1109/TBME.2013.2256908

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Shafiq-ul-Hassan, M., Zhang, G. G., Latifi, K., Ullah, G., Hunt, D. C., Balagurunathan, Y., et al. (2017). Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys*. 44, 1050–1062. doi: 10.1002/mp.12123

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Adv Stat Anal*. 98, 121–142. doi: 10.1007/s10182-013-0213-1

Shou, H., Zipunnikov, V., Crainiceanu, C.M., and Greven, S. (2015). Structured functional principal component analysis. *Biometrics* 71, 247–257. doi: 10.1111/biom.12236

Tsai, C. J., Hofstede, T. M., Sturgis, E. M., Garden, A. S., Lindberg, M. E., Wei, Q., et al. (2013). Osteoradionecrosis and radiation dose to the mandible in patients with oropharyngeal cancer. *Int. J. Radiat. Oncol. Biol. Phys*. 85, 415–420. doi: 10.1016/j.ijrobp.2012.05.032

Tsien, C., Cao, Y., and Chenevert, T. (2014). Clinical applications for diffusion magnetic resonance imaging in radiotherapy. *Semin. Radiat. Oncol*. 24, 218–226. doi: 10.1016/j.semradonc.2014.02.004

Tucker, J. R., Xu, L., Sturgis, E. M., Mohamed, A. S. R., Hofstede, T. M., Chambers, M. S., et al. (2016). Osteoradionecrosis in patients with salivary gland malignancies. *Oral Oncol*. 57, 1–5. doi: 10.1016/j.oraloncology.2016.03.006

Unser, M. (1986). Sum and difference histograms for texture classification. *IEEE Transact. Pattern Anal. Machine Intelligence* PAMI-8, 118–125. doi: 10.1109/TPAMI.1986.4767760

Wong, A. J., Kanwar, A., Mohamed, A. S., and Fuller, C. D. (2016). Radiomics in head and neck cancer: from exploration to application. *Translat. Cancer Res*. 5, 371–382. doi: 10.21037/tcr.2016.07.18

Wong, A. T. T., Lai, S. Y., Gunn, G. B., Beadle, B. M., Fuller, C. D., Barrow, M. P., et al. (2017). Symptom burden and dysphagia associated with osteoradionecrosis in long-term oropharynx cancer survivors: a cohort analysis. *Oral Oncol*. 66, 75–80. doi: 10.1016/j.oraloncology.2017.01.006

Yang, P., Mackin, D., Chen, C., Mohamed, A. S. R., Elhalawani, H., Shi, Y., et al. (2018). Discrimination of epstein-barr virus status in NPC Using CT-derived radiomics features: linking imaging phenotypes to tumor biology. *Int. J. Radiat. Oncol. Biol. Phys*. 100:1361. doi: 10.1016/j.ijrobp.2017.12.142

Yang, X., Tridandapani, S., Beitler, J. J., Yu, D. S., Yoshida, E. J., Curran, W. J., et al. (2012). Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: an *in vivo* study of late toxicity. *Med. Phys*. 39, 5732–5739. doi: 10.1118/1.4747526

Yin, Q., Hung, S.-C., Wang, L., Lin, W., Fielding, J. R., Rathmell, W. K., et al. (2017). Associations between tumor vascularity, vascular endothelial growth factor expression and PET/MRI radiomic signatures in primary clear-cell–renal-cell-carcinoma: proof-of-concept study. *Sci. Rep*. 7:43356. doi: 10.1038/srep43356

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*. 3, 32–35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Zar, J. H. (2005). "Spearman rank correlation," in *Encyclopedia of Biostatistics*, eds P. Armitage and T. Colton. doi: 10.1002/0470011815.b2a15150

Zhang, L., Fried, D. V., Fave, X. J., Hunter, L. A., Yang, J., and Court, L. E. (2015). IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. *Med. Phys*. 42, 1341–1353. doi: 10.1118/1.4908210

Zhang, W., Zhang, X., Yang, P., Blanchard, P., Garden, A. S., Gunn, B., et al. (2017). Intensity-modulated proton therapy and osteoradionecrosis in oropharyngeal cancer. *Radiother. Oncol*. 123, 401–405. doi: 10.1016/j.radonc.2017.05.006

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.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 23 April 2021
doi: 10.3389/frai.2021.624038

A Novel Machine Learning Model for Dose Prediction in Prostate Volumetric Modulated Arc Therapy Using Output Initialization and Optimization Priorities

*P. James Jensen, Jiahan Zhang, Bridget F. Koontz and Q. Jackie Wu**

Department of Radiation Oncology, Duke Cancer Institute, Durham, NC, United States

Treatment planning for prostate volumetric modulated arc therapy (VMAT) can take 5–30 min per plan to optimize and calculate, limiting the number of plan options that can be explored before the final plan decision. Inspired by the speed and accuracy of modern machine learning models, such as residual networks, we hypothesized that it was possible to use a machine learning model to bypass the time-intensive dose optimization and dose calculation steps, arriving directly at an estimate of the resulting dose distribution for use in multi-criteria optimization (MCO). In this study, we present a novel machine learning model for predicting the dose distribution for a given patient with a given set of optimization priorities. Our model innovates upon the existing machine learning techniques by utilizing optimization priorities and our understanding of dose map shapes to initialize the dose distribution before dose refinement via a voxel-wise residual network. Each block of the residual network individually updates the initialized dose map before passing to the next block. Our model also utilizes contiguous and atrous patch sampling to effectively increase the receptive fields of each layer in the residual network, decreasing its number of layers, increasing model prediction and training speed, and discouraging overfitting without compromising on the accuracy. For analysis, 100 prostate VMAT cases were used to train and test the model. The model was evaluated by the training and testing errors produced by 50 iterations of 10-fold cross-validation, with 100 cases randomly shuffled into the subsets at each iteration. The error of the model is modest for this data, with average dose map root-mean-square errors (RMSEs) of 2.38 ± 0.47% of prescription dose overall patients and all optimization priority combinations in the patient testing sets. The model was also evaluated at iteratively smaller training set sizes, suggesting that the model requires between 60 and 90 patients for optimal performance. This model may be used for quickly estimating the Pareto set of feasible dose objectives, which may directly accelerate the treatment planning process and indirectly improve final plan quality by allowing more time for plan refinement.

Keywords: dose prediction, multi-criterial optimization, treatment planning, prostate VMAT, machine learning, artificial intelligence, residual neural networks

Edited by:
Thomas Hartung,
Johns Hopkins University,
United States

Reviewed by:
Himanshu Arora,
University of Miami, United States
Fred Wright,
North Carolina State University,
United States

***Correspondence:**
Q. Jackie Wu
jackie.wu@duke.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *30 October 2020*
Accepted: *18 March 2021*
Published: *23 April 2021*

Citation:
Jensen PJ, Zhang J, Koontz BF and Wu QJ (2021) A Novel Machine Learning Model for Dose Prediction in Prostate Volumetric Modulated Arc Therapy Using Output Initialization and Optimization Priorities. Front. Artif. Intell. 4:624038. doi: 10.3389/frai.2021.624038

INTRODUCTION

Volumetric modulated arc therapy (VMAT) is a cancer treatment option that can effectively irradiate a target while minimizing the nearby healthy tissue irradiation in relatively short delivery times (Otto, 2008; Teoh et al., 2011). Dual-arc VMAT has been shown to be an effective treatment technique for prostate cancer (Guckenberger et al., 2009; Zhang et al., 2010). VMAT treatment planning relies on inverse planning techniques that perform dose optimization and dose calculation to create a deliverable treatment plan. To employ existing single-function minimization algorithms, VMAT optimization techniques typically scalarize the dose objectives into a weighted sum to use as the optimization loss function, with the weights (priorities) decided by a treatment planner. Dose objective scalarization allows the treatment planner to create and evaluate several plans by providing multiple priority combinations to the optimizer to create a subjectively optimal treatment plan. This problem can be formulated as a multi-criteria optimization (MCO) problem, in which the treatment planner has to learn about the set of feasible plan doses which cannot be strictly improved, which is historically named, the Pareto surface (Hwang and Masud, 1979; Miettinen, 1999). MCO has been studied extensively and many methods for exactly sampling the Pareto surface have been implemented for radiation therapy treatment planning systems (Craft et al., 2007; Monz et al., 2008; Bokrantz and Forsgren, 2013).

However, these contemporary MCO methods ultimately require the generation of many treatment plans to sample the Pareto surface. In this framework, the treatment planner samples the Pareto surface and linearly interpolates the sampled plans to infer the feasible ranges of dose trade-offs. However, VMAT treatment planning using current commercial treatment planning systems can take 5–30 min per plan to optimize and calculate, so that the exact methods for sampling the Pareto surface can take a longer time to run. This time cost reduces the remaining amount of time that the planner has for manual plan refinement and also limits the precision of the surface sampling, decreasing the accuracy of any subsequent surface interpolations and limiting the understanding of the planner with regard to feasible dose trade-offs. All these factors combine to reduce the quality of the final treatment plan.

The primary goal of this study is to present a method for quickly estimating the dose distribution for a given set of optimization priorities. This method would be able to quickly and accurately estimate the Pareto surface for a given patient and indirectly improve the quality of the final plan by allowing the treatment planner more time for plan refinement.

In recent years, machine learning has seen success in image classification and processing tasks, due to the ability of modern convolutional neural network variants, such as residual networks (ResNets) and U-Nets, to quickly detect and manipulate learned image patterns (Simonyan and Zisserman, 2014; Ronneberger et al., 2015; He et al., 2016). Inspired by the speed and accuracy of these results, we hypothesized that it was possible to use a similar model to bypass the time-intensive dose optimization and dose calculation steps in treatment planning, arriving directly at the resulting dose distribution and computing the relevant dose objectives. Such a model would greatly benefit the treatment planning system (TPS), as it would provide a way to quickly estimate the dose distributions of many treatment plans to infer the Pareto surface of a given patient for feasible dose objectives.

In this study, we present a novel machine learning model for predicting the TPS-simulated dose distribution for a given patient. Similar models have previously been implemented which are more directly drawn from the U-Net architecture (Babier et al., 2019; Nguyen et al., 2019), but these models have undergone only modest modification for the specific task of dose prediction. The primary motivation behind our model is to use our understanding of the general shape of dose distributions to remove much of the non-linearity of the dose prediction problem and decreasing the difficulty of subsequent network predictions. Our model takes the optimization priorities of the treatment plan, which were taken into account during dose prediction, and infers feasible dose distributions across a range of optimization priority combinations, allowing for indirect Pareto surface inference. This model is also relatively fast (0.05 s per plan), and it is capable of sampling the entire Pareto surface much faster than commercial dose optimization and dose calculation engines.

METHODS

Patient Cohort and Treatment Planning Technique

Hundred prostate cancer patients were retrospectively included in this study. The data of each patient consisted of an abdominal CT scan and contours of their planning target volume (PTV), the bladder, the rectum, the left femoral head, and the right femoral head. After anonymization, patient datasets were imported to a commercial treatment planning system for dose optimization and dose calculation. The PTV dose prescription was set to 70 Gy in 29 fractions, as is the current standard for clinical practice at our institution. During treatment planning, each plan included two concentric, coplanar VMAT beams centered on the PTV, with field sizes set to encompass the PTV during a 358-degree beam rotation. Beam collimators were set at 15° and 345° to reduce the effect of collimator leaf gap overlap. During optimization, priorities were placed on the PTV homogeneity index (HI = D2%–D98%), bladder D25%, and rectum D25%. These objectives were chosen to represent the dimensions of trade-off during treatment planning, since the primary goals of prostate VMAT are uniform PTV coverage, bladder sparing, and rectum sparing. These objectives had different optimization priority combinations for each plan to sample the Pareto surface of dose trade-offs. After optimization, plans were normalized such that PTV D95% equaled 100% of the dose prescription of the target. Fixed constraints for each plan optimization included PTV D93% < 101% to reduce the dose-shifting effect

Abbreviations: VMAT, volumetric modulated arc therapy; MCO, multi-criteria optimization; TPS, treatment planning system; PTV, planning target volume; IMRT, Intensity-modulated radiation therapy; HI, homogeneity index; ResNet, residual network; RMSE, root-mean-square error; APD, average projected distance; ANPD, average nearest point distance; ICRU, International Commission on Radiation Units and Measurements; AAPM, the American Association of Physicists in Medicine; FC, fully connected; L-ReLU, leaky rectified linear unit; SS, scaled softsign.

of plan normalization, as well as D0.01cc < 65% for both femoral heads in accordance with the standard practice of our institution for normal critical structure constraints. Variable constraints included PTV HI < 10%, the bladder D25% < 30% of prescription, and the rectum D25%. For each patient, the Pareto surface was sampled by optimizing and calculating 25 plans that follow. Each plan had a different optimization priority combination and therefore sampled a different location on the Pareto surface. Bounding points on the surface were chosen through manual plan optimization such that the bounding points represented clinically feasible plans. Subsequent points on the surface were created using linear combinations of the objective priorities of the bounding points; this ensured that all interior points also represented clinically feasible plans on the Pareto surface. Beamlet fluence optimization and dose calculation were performed with the commercial treatment planning system. After each plan was calculated, the corresponding dose map, critical structure maps, and optimization priority combination were exported for use during model training and evaluation.

Dose Prediction Model Architecture

An overview of the architecture of the dose prediction is depicted in **Figure 1**. The inputs of the model are the objective priorities and structure maps of the PTV, the bladder, and the rectum, resized to slices of 128 × 128 voxels to increase model efficiency. These structure maps are binary image-domain representations of the corresponding structures, indicating for each pixel whether that pixel is inside the contour of the structure. These structure maps have been scaled by the objective priorities of their corresponding structure for each plan. This is a straightforward way to incorporate objective trade-off priorities without complicating the architecture of the model.

Dose Initialization

The model begins by creating an initial dose distribution *via* an inverse fit of inter-slice and intra-slice PTV distance maps on a voxel-wise basis. The functional form of the initialized dose fit is as follows:

$$D_i = \left[1 + a_1 * ISD_1^{a_2} \; + c * ISD_2^{a_3}\right]^{-1}$$

where ISD_1 refers to the inter-slice distance from the voxel to the nearest PTV location within the slice of the voxel, ISD_2 refers to the intra-slice distance from the voxel to the nearest PTV location at the row and column of the voxel, and a_1, a_2, and a_3 are variables that need to be fitted. The purpose of this initialization is to allow the subsequent neural network to predict the shift between the initialization and the TPS-simulated dose distribution rather than the dose distribution itself. We hypothesize that these shifts are more likely linear than the dose distribution itself and therefore more easily learned.

Patch Extraction

The model proceeds by extracting three sets of 9 × 9 transverse patches from all structure maps and the initialized dose map at each voxel. Each set of patches has a different atrous rate, which is the number of voxels skipped between the sampled voxels. The first patches have an atrous rate of 1, i.e., they do not skip any voxel and are contiguous. These patches allow the model to infer local structure information near the pixels on which they are centered. For the second patches, the structure maps are smoothed by convolution with a uniform 3 × 3 kernel, and the patches are extracted with an atrous rate of 3. Similarly, the third patches are extracted with an atrous rate of 10 from the structure maps after smoothing by a 10 × 10 kernel. The smoothing convolutions are performed to make each voxel within the atrous patches contain structure information from the nearby voxels that the atrous sampling skips.

The idea of atrous convolution (also called, dilated convolution) was originally presented by Yu and Koltun (2015). The motivation for including multiple patches with different atrous rates is to capture the features of the input data at both coarse and fine levels. This removes the need for traditional downsampling and upsampling layers in the network. For the patches with atrous rates >1, the combination of average smoothing and atrous sampling essentially increases the receptive field size per layer of the model, so that the model can infer the effect of critical structures at both large and short distances without significantly increasing the amount of memory or model parameters required. It is to be noted that the model architect can choose the number of patches and the atrous rates of every patch, and a similar model with atrous rates near 1, 3, and 10 will produce results similar to the result of this model. For this model, the atrous rates were chosen based on the nature of the input data. The patches with an atrous rate of 1 captured every fine detail, the patches with an atrous rate of 10 spanned most of the images and captured every coarser detail, and the patches with an atrous rate of 3 reflected the more intermediate features. The patches are then cast into 81-element vectors per voxel, and the vectors and optimization priorities are all concatenated voxel-wise to serve as input for the residual network.

Residual Network

The model then uses the patch vectors as inputs for a neural network, which is inspired by the recently developed ResNet (He et al., 2016). This network is used to determine an update to the dose initialization rather than computing the dose from scratch. The natural choice for the construction of this intermediate network was the residual network (ResNet), because the residual blocks of ResNets were originally designed with a similar concept in mind. As explained by He et al., residual blocks tend to perform much better than the conventional network blocks at a higher depth when the effects of the input features resemble linear residuals. This happens partially because the residual formalism makes the gradients to be less susceptible to vanishing or exploding, improving the convergence (He et al., 2016). Unpublished internal testing confirmed that the performance of our model degrades when it replaces the residual blocks with standard convolutional or fully connected (FC) blocks. Moreover, residual blocks have been shown to make the performance of the model less dependent on the number of blocks included, which reduces the need for fine-tuning the number of blocks in the model.

Our network consists of a series of six residual blocks that sequentially update the initialized dose map. Each residual block, depicted in **Figure 2**, consists of three FC layers. The first two

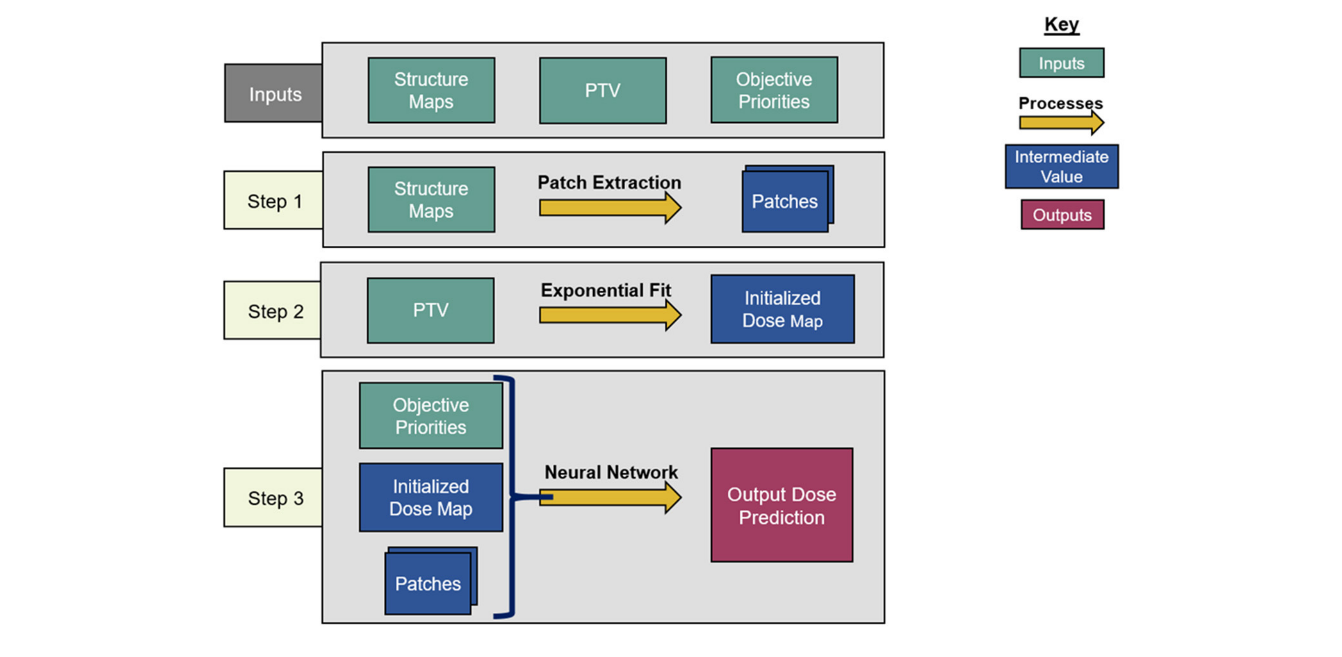


FIGURE 1 | Overview of the dose prediction model architecture.

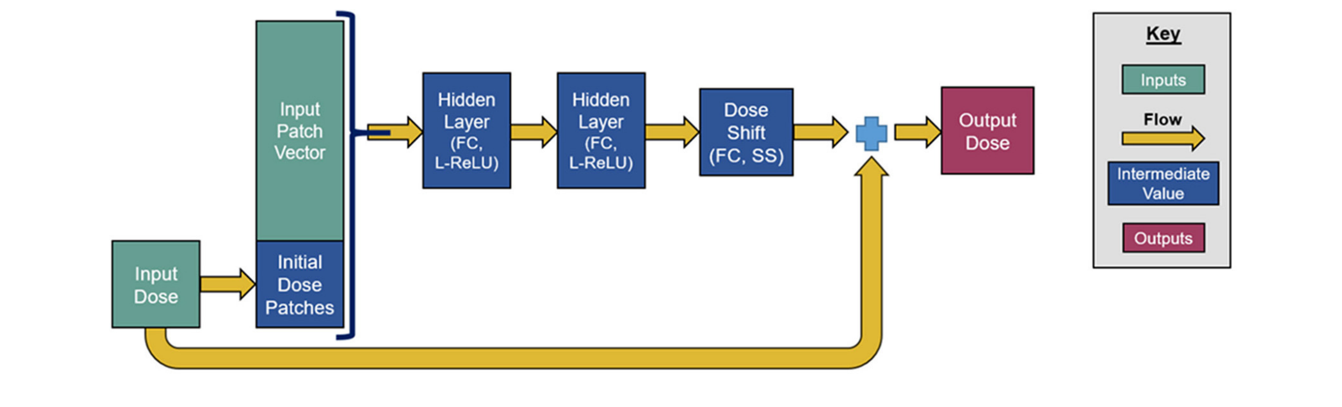


FIGURE 2 | Graphical depiction of a residual block within the neural network.

layers have 100 output units and leaky rectified linear unit (L-ReLU) activations are defined as follows;

$$L - ReLU\ (x) = x \text{ when } x > 0 \text{ and } L - ReLU\ (x) = 0.2x \text{ when } x \leq 0.$$

These first two layers extract quasi-linear features from the patch vectors. The last layer has a single output and scaled softsign (SS) activation, defined as, $SS(x) = 0.3x/(1 + |x|)$. The purpose of this last activation function is to take the quasi-linear combinations from the previous layer and map them to a suitable dose shift with a limited range. Since each residual block changes the initialized dose map, the dose map patches need to be reextracted after each update. The number of residual blocks, layers per block, and output units per layer were chosen somewhat subjectively, and we anticipated that the accuracy achieved by this neural network can be achieved through similar network designs and hyperparameter tunings.

Model Training

The training loss function was the root-mean-square error (RMSE) between the predicted dose map and TPS-simulated dose map, restricted to voxels within the body contour and restricted to slices containing at least one critical structure. Dose initialization variables were fit according to the RMSE between the initialized dose map and the TPS-simulated dose, and these variables were trained before the residual network variables. Gradients for the loss function were estimated using batches of training data, with each batch containing several slices approximately equal to the typical number of slices that a patient would have. Slices in the batches were sampled diagonally, such that the batch slices were located at different levels within different patients. This sampling means that each batch contains slices from most patients in the training set at most slice positions, such that each batch is a good representation of the entire cohort. Therefore, the gradients computed from the batches were in close approximations to the gradients of the loss function applied to the entire cohort, improving the optimization of convergence and stability. The model was trained using the Adam optimization algorithm, which was designed for stochastic gradient-based optimization (Kingma and Ba, 2014). Kingma and Ba recommend specific hyperparameters, including step size $\alpha = 0.001$, decay hyperparameters $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and error epsilon $\epsilon = 10^{-8}$, all of which are used in the training of our model. The Adam optimizer is particularly appropriate here because the batch gradient computations are stochastic. The trainable parameters in each layer were initialized using the Glorot uniform initializer, which initializes variables by sampling randomly from a uniform distribution bounded by $\pm\sqrt{6\,/(number\ of\ inputs + number\ of\ outputs)}$ (Glorot and Bengio, 2010). The Glorot uniform initializer was designed to model the inherent variance of rectified linear unit activation functions, similar to the activation functions used in our residual network. All aspects of the model, including optimization and evaluation, were implemented using the Tensorflow machine learning platform with an NVIDIA Quadro M4000. Optimization proceeded for 2,000 iterations before termination.

Predicted Pareto Surface Evaluation

Pareto surfaces are generated from the model by passing several optimization priority combinations as inputs and evaluating the relevant dose-volume metrics from the resulting dose maps. For analysis, this study compared the Pareto surfaces of the clinical and predicted dose maps using the same optimization priority combinations for both surfaces, allowing for direct comparison of matched plans which should produce the same dose maps and objective metrics. However, when evaluating the accuracy of a predicted Pareto surface, we were more interested in the entire surface as a connected set rather than a few points which sample the surface. Although we can use the sampled points to interpolate the Pareto surface, distances between the sampled points do not necessarily represent the distances between points which are interpolated from the sampled points (Jensen et al., 2020). For this reason, we believe that it is insufficient to simply evaluate the RMSEs between the sampled points in Pareto space as a metric for the closeness of the represented predicted Pareto surface to the TPS-simulated Pareto surface. To our knowledge, no previous publications on dose prediction for radiation therapy have directly evaluated the distance between the Pareto surfaces generated by their models.

To overcome this insufficiency, we tested three metrics in addition to RMSE between the matched points in Pareto space. The first additional metric is the Hausdorff distance, mathematically defined between the two sets A and B are as follows:

$$d_H(A,B) = max\left\{\sup_{x\epsilon A}\inf_{y\epsilon B}\left|x-y\right|, \sup_{y\epsilon B}\inf_{x\epsilon A}\left|x-y\right|\right\}$$

where A and B, *in this case*, represent the vertices (sampled points) of each Pareto surface. One benefit of the Hausdorff distance is that it is sensitive to outliers so that the Hausdorff distance between the sets of Pareto surface vertices should be similar to the Hausdorff distance between the actual Pareto surfaces as interpolated sets. However, this sensitivity to outliers causes Hausdorff distances to represent the maximum error rather than the average error more strongly. In the context of machine learning, this is a drawback because the outliers which influence the Hausdorff distance can fluctuate because of the random initial conditions of the model.

The second additional metric is the average projected distance (APD) in Pareto space, which addresses some of the insufficiencies of Pareto space RMSE and Hausdorff distance. A more abstract discussion has been published about the properties of the APD and why this metric is superior to the RMSE in Pareto space (Jensen et al., 2020). The APD examines the vector displacements between matched points between two sets and then averages the displacements when projected along the direction of the average vector displacement. The APD between two sets, A and B is mathematically defined as follows:

$$APD(A,B) = E\left[\left(x_i - y_i\right)\cdot E\left[x_i - y_i\right]\right] / \left|E\left[x_i - y_i\right]\right|$$

where the $\left(x_i, y_i\right)$ symbols enumerate the matched pairs of points between sets, A and B (in our case, the TPS-simulated and predicted Pareto surfaces), and E refers to taking an average over these matched pairs for one patient. The primary motivation behind the APD as a Pareto space metric is depicted in **Figure 3**, in which we can see that overall Pareto surface interpolation accuracy is not affected by the pointwise error components along the respective Pareto surfaces. APDs first remove these error components, so we expect the APD to better measure the closeness of the interpolated Pareto surfaces compared to the RMSE or HD.

The third metric is the average nearest point distance (ANPD) in the Pareto space, which supersamples the simplicial complex representations of the Pareto surfaces and averages the distance from each sampled point of one surface to the simplicial complex of the other surface. The ANPD between the two sets, A and B is mathematically defined as follows:

$$ANPD(A,B) = avg\left\{\underset{y\epsilon S(B)}{avg}\ \underset{x\epsilon S(A)}{inf}\left|\left|x-y\right|\right|_2,\ \underset{x\epsilon S(A)}{avg}\ \underset{y\epsilon S(B)}{inf}\left|\left|x-y\right|\right|_2\right\}$$

where the sets A and B represent the vertices of each Pareto surface and $S(A)$ and $S(B)$ represent the simplicial complexes spanned by the vertices of A and B, respectively. Here, each Pareto surface vertex corresponds to one dose distribution generated by different optimization priorities. The primary motivation behind the ANPD as a Pareto space metric is that it reflects the individual distances from each point on one surface to the other surface, which we imagine to be the "true" distance between that point and the surface. Like the APD, the ANPD has already been discussed according to its properties and comparison to the RMSE in another publication (Jensen et al., 2020). For this publication, all these metrics are presented because a consensus about the optimal metric has not yet been established.

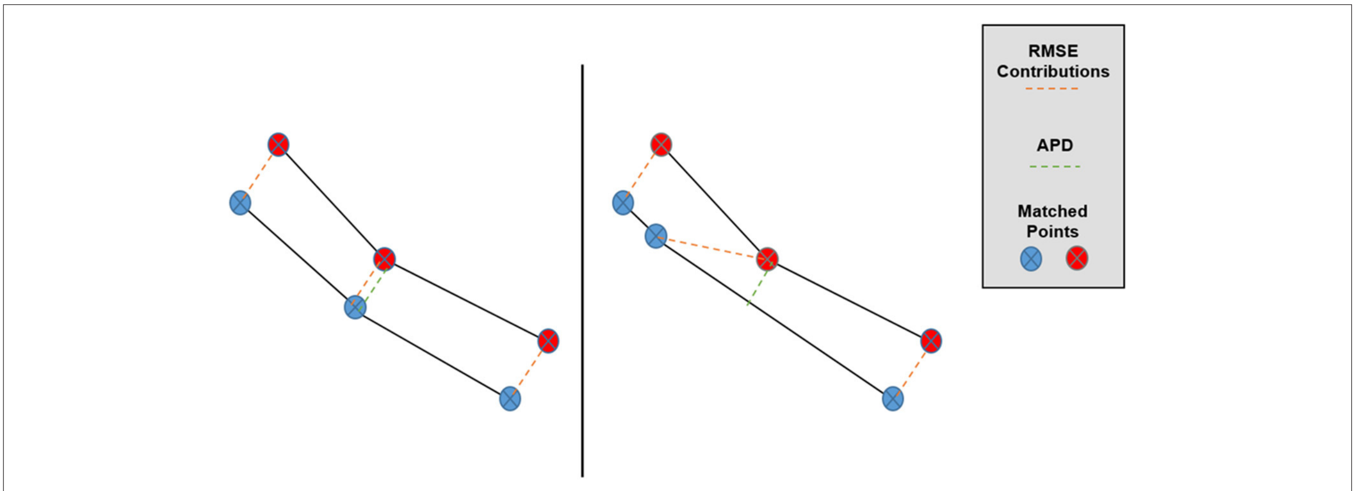


FIGURE 3 | Graphical depiction of the effect of matched point error along the Pareto surfaces on RMSE and APD. Despite the lower surfaces being very similar, the central matched pair has a much larger contribution to Pareto RMSE in the right case while the APD remains approximately the same.

Model Evaluation

When evaluating this model, a single instance of training and testing the model is insufficient because the performance of the model depends on the specific training set and testing set. To counteract the randomness associated with choosing a training set and testing set, the following evaluation scheme was used. After the model was designed and developed, it was evaluated using a 10-fold cross-validation repeated 50 times. In this evaluation, one repetition of 10-fold cross-validation involves randomly partitioning the patient dataset into ten 10-patient subsets and training the model 10 times, with each training set using a different subset for testing and the rest of the subsets for evaluation. In one repetition of 10-fold cross-validation, each patient appears in training sets exactly nine times, and each patient appears in the testing sets exactly once. Therefore, the 10-fold cross-validation partially negates the effect of randomly assigning patients into training and testing sets. In this evaluation, 10-fold cross-validation was repeated 50 times, with the patient dataset partitioned into different subsets for each repetition. By repeating the cross-validation many times, the randomness associated with the random selection of the training and testing sets is reduced further. Overall, this evaluation involved training and testing the model 500 times, which is 10 training/testing pairs for each of the 50 cross-validation repetitions. The results from all 500 model validations were aggregated by the training set and the testing set using the metrics described in the sections above.

To test the performance of the model with smaller training datasets, another set of cross-validations was performed using different ratios of training data to testing data. These cross-validations evaluated the performance of the model with training-to-testing data set ratios of 90%:10%, 80%:20%, 70%:30%, 60%:40%, 50%:50%, 40%:60%, 30%:70%, 20%:80%, and 10%:90%. For example, the second of these cross-validations used 80 patients to train each model and 20 patients to test each model. For this cross-validation, the patients were grouped into ten 10-patient subsets, enumerated #1, #2, #3, #4, #5, #6, #7, #8, #9, and #10. The first model validation in the cross-validation was trained on subsets #1-8 and tested on subsets #9 and #10; the second validation was trained on subsets #2-9 and tested on subsets #10 and #1; the third validation was trained on subsets #3-10 and tested on subsets #1 and #2; and so on. In this way, each of the cross-validations at smaller training-to-testing ratios evaluated the model ten times. This is not strictly a 10-fold cross-validation, but it is a cross-validation because every patient appears in the same number of training subsets and the same number of testing subsets. For comparison, the cross-validation with an 80%:20% ratio evaluates 10 trainings of the model, while normal 5-fold cross-validation evaluates 5 trainings of the model. The utility of this approach can be seen clearly for the 50%:50% case, where the corresponding 2-fold cross-validation only involves training the model twice. Clearly, this is not thorough enough, but the other extremity of testing every possible 50%:50% partition of the data is not feasible due to time constraints. This approach of cycling through the ten subsets strikes a compromise between thoroughness and efficiency when evaluating the model at smaller training set sizes. For each cross-validation, the performance of the model on the training and testing sets was aggregated to evaluate the performance of the model while reducing the randomness associated with grouping the patients into training and testing subsets.

RESULTS

Direct Dose Map Evaluation

Figure 4 shows the dose map RMSE for the aggregated training and testing sets during a randomly selected model instance training. After training, the mean dose map RMSEs were 2.44 ± 0.89% and 2.42 ± 0.47% for the training and testing set dose predictions, respectively, across all cross-validations. These errors demonstrate that the model can achieve good prediction accuracy on a voxel-by-voxel basis. The difference between the

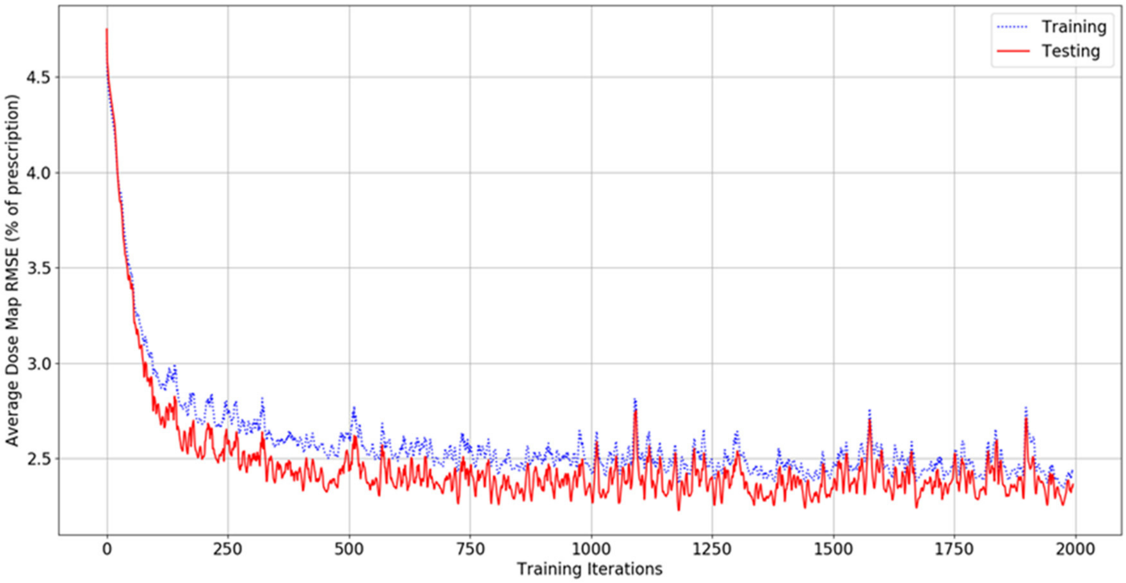


FIGURE 4 | Graph of dose map root-mean-square error for the training set (blue, dashed) and testing set (red, solid) as a function of the number of iterations during model training.

training and testing dose map RMSEs is less than their standard deviations, suggesting that the performance of the model is similar for both the training and testing datasets. The dose map RMSEs due to the initialization fit alone were 5.12 ± 0.55%, and 5.54 ± 1.30% for the training and testing sets, respectively. This indicates that the residual network makes a measurable improvement to the dose initialization and that the model successfully learns after the dose initialization. Note that these values differ from the general dose map RMSE of the model at 0 iterations into training because the residual parameters and effects of the network on prediction are non-zero and initialized randomly. For comparison, the International Commission on Radiation Units and Measurements (ICRU) and Task Group 142 of the American Association of Physicists in Medicine (AAPM) have stated that a 5% maximum dosimetric uncertainty is appropriate for standard intensity-modulated radiation therapy (IMRT) treatments (ICRU, 1976; Klein et al., 2009). Therefore, these dose map RMSEs are comparable to the maximum error permitted in treatment delivery.

Figures 5, **6** show side-by-side comparisons between the effect of prioritizing PTV HI or prioritize rectum $D_{25\%}$ in a dose map prediction and its corresponding TPS simulation. Visually, we can see that the dose map predictions are jagged compared to their respective TPS-simulated dose maps, specifically around the 30% isodose line. We expect this to be the case because the neural architecture of the network does not explicitly promote local smoothness in the dose distribution predictions. However, real dose distributions tend to be smooth and continuous, so that the artificial jaggedness in the prediction of our model is a drawback reflecting the artificial nature of the model. Note that the jaggedness makes the isodose lines look dissimilar, but the general location of the isodose lines corresponds much more strongly to voxel-by-voxel error than the precise shape of the isodose lines. Additionally, we see that the region of the largest isodose displacement is the low dose region anterior to the PTV. Note that this region is not near the PTV or the surrounding critical structures, so that the inaccuracy in this region has a small impact on the predicted PTV/OAR doses.

Figures 7, **8** show the performance of the model on training and testing sets as a function of the ratios of training set data to testing set data, so that they show the effect of decreasing the number of patients used to train the model. **Figure 7** shows that the training set errors decrease as training set size decreases. On the other hand, **Figure 8** shows that the testing set errors increase as training set size decreases. From these results, we can see that the performance of the model degrades slightly as the amount of training data shrinks because it increases overfitting to the training data. It is difficult to conclude from these figures exactly how much data is needed to properly fit to the data, but that number is likely between 60 and 90 patients based on the figures (though this number might not generalize to other treatment sites). **Figures 7**, **8** also suggest that the spread of errors becomes larger with smaller training data sizes, indicating that the performance of the model is more random with smaller training data sizes.

Model Evaluation Time

Total dose prediction requires an average of 1.26 s to evaluate an entire 25-plan Pareto surface for one patient, or just 0.05 s per plan. This is significantly faster than current commercial dose optimization and dose calculation engines, which can take ∼5–30 min per plan. Due to this speed, we anticipate that this model can be used in real-time without needing to interpolate plan doses from a set of previously predicted doses.

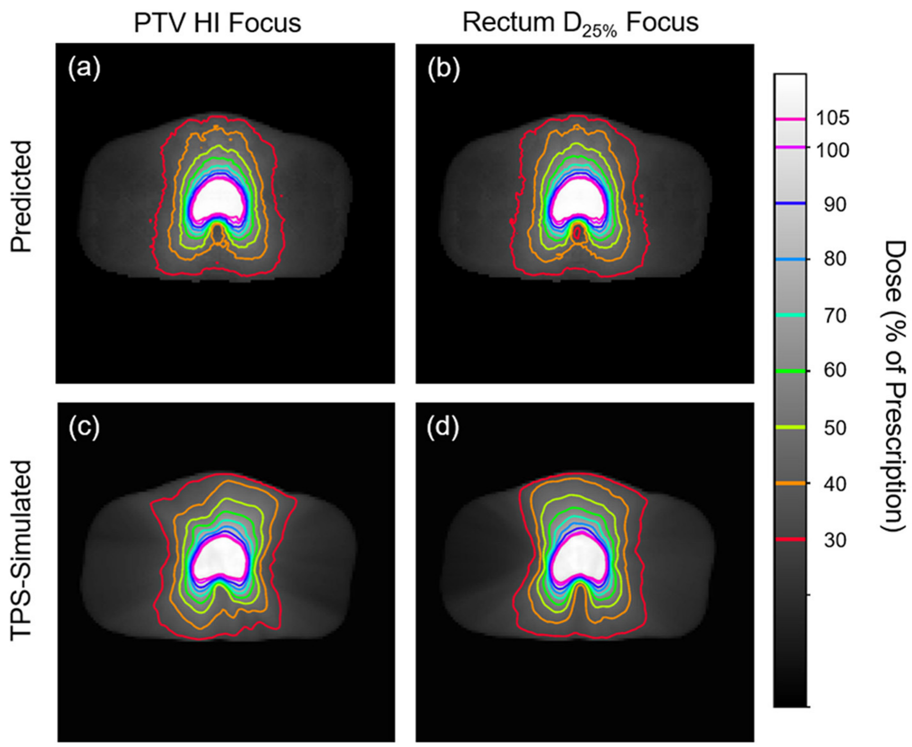


FIGURE 5 | Side-by-side comparisons between the effect of prioritizing PTV HI **(a,c)** or prioritizing rectum $D_{25\%}$ **(b,d)** in a dose map prediction **(a,b)** and its corresponding TPS-simulated dose map **(c,d)**. Transverse slices are taken from the center of the PTV, and the patient was randomly sampled from the testing dataset.

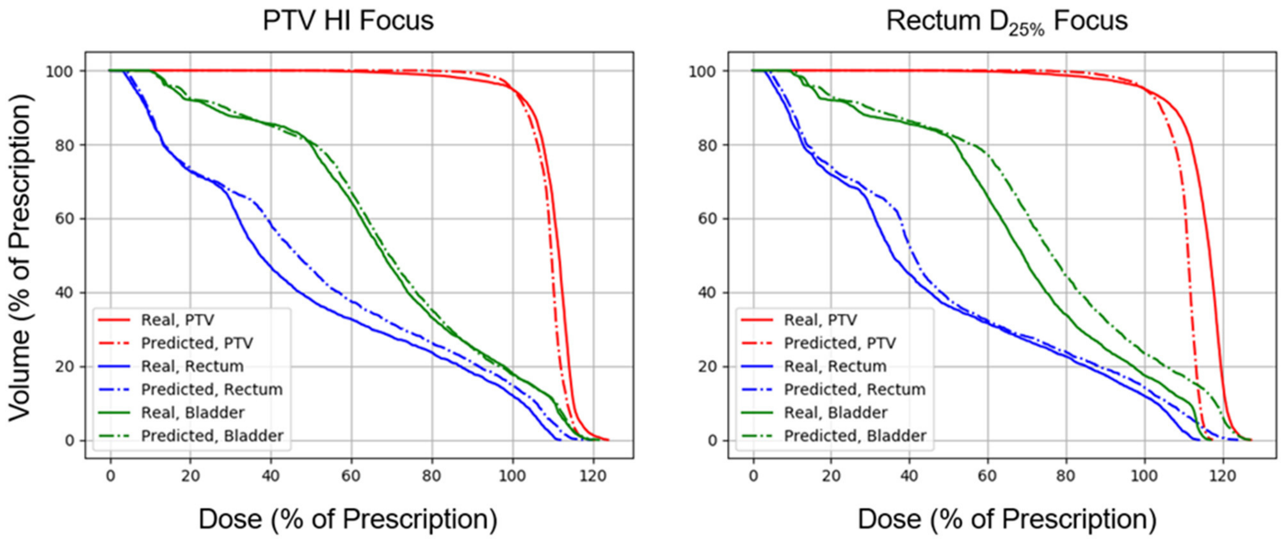


FIGURE 6 | Side-by-side comparisons between the effect of prioritizing PTV HI or prioritizing rectum $D_{25\%}$ in a DVH prediction and its corresponding TPS-simulated DVH.

Predicted Pareto Surface Evaluation

The mean Pareto space RMSEs were 10.33 ± 3.57% and 10.11 ± 4.61% for the training and testing sets, respectively, when aggregated over the fifty splits of 10-fold cross-validation. These errors indicate that the training and testing set dose predictions have similar distances to their corresponding TPS-simulated doses in objective space. This contrasts the dose map RMSEs for the training and testing set, which

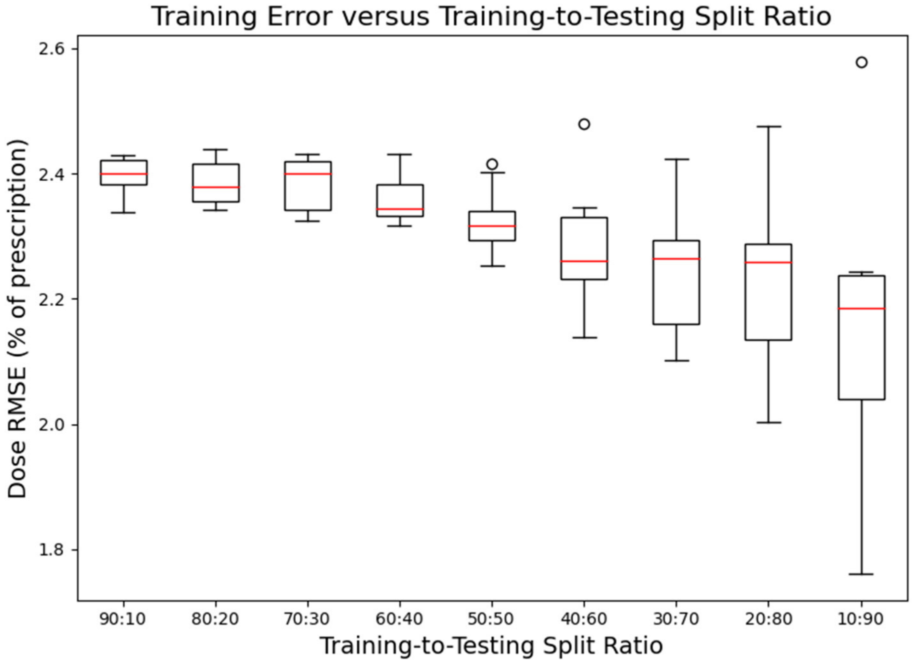


FIGURE 7 | Box-and-whisker plots representing the training set errors of the model as a function of decreasing training-to-testing set split ratios. Each box-and-whisker plot represents the aggregate errors from one split of 10-fold cross-validation.

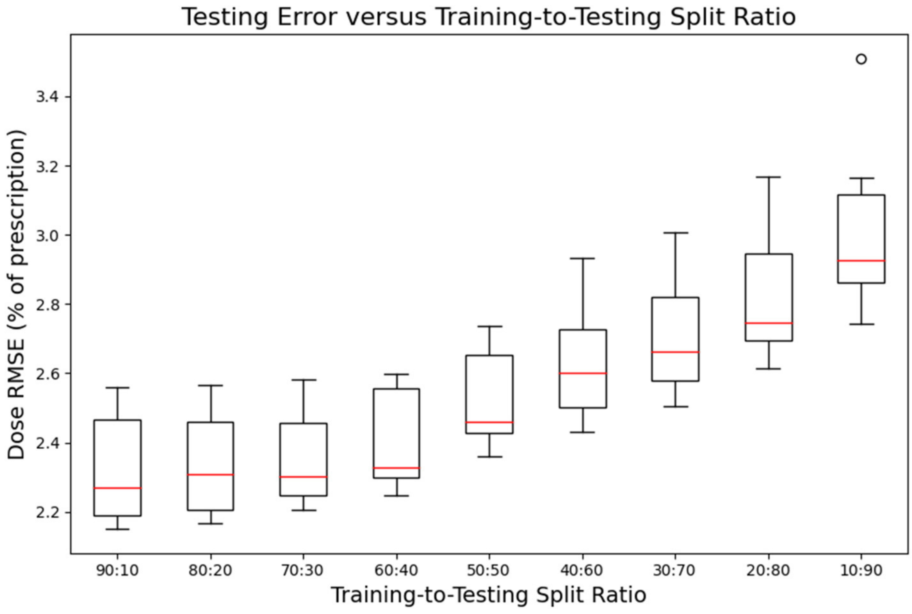


FIGURE 8 | Box-and-whisker plots representing the testing set errors of the model as a function of decreasing training-to-testing set split ratios. Each box-and-whisker plot represents the aggregate errors from one split of 10-fold cross-validation.

were more dissimilar than the Pareto space RMSEs. Note that the Pareto space RMSE combines the errors across the objectives *via* accumulation rather than averaging, so we expected these numbers to be significantly larger than dose map RMSE.

The mean Pareto space Hausdorff distances were 14.98 ± 5.91% and 14.79 ± 5.77% for the training and testing set dose predictions, respectively, when aggregated over the fifty splits of 10-fold cross-validation. These errors are notably larger and have more variance than the corresponding Pareto RMSEs. However, Hausdorff distances, in general, are more sensitive to outliers than set averaged RMSEs, so we expect this increased magnitude and variance. We see that the training and testing set Hausdorff distances are also similar, indicating that the errors of our model primarily occur at low-dose regions away from the PTV and at critical structures.

The mean Pareto space APDs were 10.17 ± 3.52% and 9.81 ± 4.74% for the training and testing set dose predictions, respectively, when aggregated over the fifty splits of 10-fold cross-validation. These results confirm that the model fitting is not significant, as the training and testing sets had comparable projected distances. As expected, the Pareto APDs are slightly lower than the Pareto RMSEs and Pareto space Hausdorff distance.

The mean Pareto space ANPDs were 8.44 ± 3.29% and 8.85 ± 4.21% for the training and testing set dose predictions, respectively, when aggregated over the fifty splits of 10-fold cross-validation. The ANPD results demonstrate that the performance of the model in Pareto space is similar for both training and testing set predictions. As expected, the Pareto ANPDs are lower than the three other distance metrics because the minimal distance between Pareto surface interpolations tends to be lower than the distance between their vertices.

DISCUSSION

In this work, we have presented a novel machine learning dose prediction model which takes optimization objective priorities into account, allowing for indirect Pareto surface estimation. Our results indicate that the model can predict doses with good accuracy, as the predicted dose map RMSEs have few percentages of their corresponding TPS-simulated doses. These dose map RMSEs are less than the maximum error tolerance proposed by the ICRU and AAPM TG 142, suggesting that our predictions may be appropriate for clinical dose distribution estimation. Moreover, the model produces just a dose distribution without actually creating a plan, so the model requires a final real plan optimization and dose calculation which will correct these dose map prediction errors prior to treatment delivery. This means that the error in the results of our model only affects treatment planning and not treatment delivery. The evaluated Pareto surface metrics indicate that these dose map predictions make reasonable translations in Pareto space. Our results also indicate that the overfitting of the model to training data dose map RMSE is modest because the training and testing errors are similar.

The prediction speed of our model is particularly encouraging. By predicting each plan in ∼0.05 s, our model may be used for real-time treatment planning without needing to interpolate between previously sampled points, allowing the treatment planner to very quickly estimate the doses produced by a given optimization priority combination. This indirectly gives the planner more time to plan per patient, which may improve the quality of the final plan. Moreover, our model only requires patient anatomy and optimization priorities, so it can generate samples from the Pareto surface automatically. This is potentially useful for large-scale automatic theoretical dosimetric investigations of new treatment planning paradigms, such as testing the effects of pushing a dose limit past its historical value or determining the feasibility of treating new structures. More research is needed to investigate these possibilities.

We believe that the speed, accuracy, and proper fitting of our model are due to the design of the model. The implementation of a dose initialization combined with a residual neural network is a novel proposal that appears to model the dose prediction process well. Also, the combination of contiguous and atrous patches during contour processing increases the effective receptive field size of each layer in the ResNet. Achieving a similar effective field-of-view in a more traditional convolutional neural network would involve either increasing the size of each convolution kernel or adding many more layers to the network. However, both of these options involve more model parameters, have increased computational requirements, and are more prone to overfitting. The patch extraction process of our model innovates by incorporating local and global information within each layer without increasing computational requirements or promoting overfitting.

Despite its potential advantages, our model has some limitations which hinder its accuracy and utility. The dose initialization of our model assumes an isotropic inverse exponential decay of dose as a function of inter-slice and intra-slice distances from the PTV. Although this assumption is only appropriate for VMAT plans which involve beam arcs wrapping nearly 360° around the patient, it is likely that other forms of dose initialization exist which are appropriate for IMRT or VMAT with significantly fewer than 360° per arc. Additionally, the model required several hyperparameters (i.e., 6 residual blocks in the neural network, 100 output units for the first two layers in each block, atrous rates of 1, 3, and 10 in patch sampling, etc.), and it is not immediately clear how to determine the optimal values for these hyperparameters aside from trial and error. However, we expect that slight adjustments from our chosen values for the hyperparameters should not significantly change the model performance. Finally, since the output of the model is a dose distribution without an actual plan optimization or dose calculation, the model can only be used to determine the subjectively optimal optimization priorities, which then need to be used in a real plan optimization and dose calculation to actually create a deliverable plan.

This study itself also has several shortcomings that make it difficult to be certain of the performance and generalizability of the model. Due to time constraints and the size of the dataset, it was not feasible to compute the gamma index passing rates of the plans predicted by our model. Gamma index analysis could be useful for confirming the quality of our results, and future research should seek to include this data. However, gamma indices are more generous than their dose difference thresholds (typically 3%, which is higher than our model's performance of 2.42%), so we anticipate that the gamma index passing rates of our data would be quite high. Also, gamma passing rates have been shown to increase in the presence of random noise, so we anticipate that the slight noisiness of our data makes gamma passing rates less useful.

It is also difficult to confirm whether these results would extend to other treatment sites. The dataset in this study is likely large enough to sufficiently represent the population of prostate VMAT treatment cases because of the similarity between our training and testing set errors. This coincides with our expectations because the relevant anatomical structures of prostate cases all tend to be somewhat similar. However, it is not immediately clear how this result is generalizable with other training sets, which may experimentally find that they require more or less training data. Also, this study does not include treatment planning data from other treatment sites, so it is difficult to determine whether this model would generalize well to model another treatment site. Further research needs to be done to test this model on other treatment sites. In particular, the current structure of the model is not built to process the data from multiple treatment sites concurrently. However, it is feasible to modify the structure of this model to learn from multiple treatment sites by incorporating the structure maps alongside their DVH constraints. Further research needs to be done to test these claims.

We have implemented several metrics for evaluating the error between a predicted Pareto surface and its corresponding TPS-simulated Pareto surface. Our metrics reported similar values around 8–15% of dose prescription for both training and testing sets. Again, it is worth noting that these metrics accumulate the errors from each dimension rather than averaging them, which is why these surface metrics are significantly larger than the dose map RMSE of 2–3%. Most of these metrics have an inherent limitation in that they measure errors from the matched pairs of plans which sample their respective surfaces rather than measure errors from the surfaces themselves. Of the metrics presented, we hypothesize that the ANPD is the most appropriate of these metrics due to its use of point-by-point nearest distances between the surfaces, which likely reflects the actual distance between the Pareto surfaces. However, a more theoretical investigation is required to justify the ANPD here as the appropriate metric. Our results show that these metrics are significantly different from each other, which provides evidence that there exists an optimal metric to represent the distance between Pareto surfaces. Also, to our knowledge, no other body of research has applied Pareto space metrics to evaluate the Pareto surfaces of radiation therapy dose predictions. This prevents us from comparing our Pareto space results with previous dose prediction research. To account for this, we have included all these metrics for ease of comparison with future research.

CONCLUSION

We have presented a novel machine learning dose prediction model which takes optimization objective priorities into account. The error of the model is modest when applied to our prostate VMAT cases, with average dose map RMSEs of 2.42 ± 0.47% overall patients and all optimization priority combinations in the patient testing set. This model may be used for quickly estimating the Pareto set of feasible dose objectives, which may directly accelerate the treatment planning process and indirectly improve the final plan quality by allowing more time for plan refinement. Future research needs to be done to determine the generalizability of this model to other treatment sites and datasets.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

PJ was responsible for conducting the bulk of the research and manuscript writing. JZ provided feedback and suggestions for the scientific work to the primary author during research. BK provided clinical feedback regarding the data produced in this research. QW oversaw the research process. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by NIH Grant R01CA201212.

REFERENCES

Babier, A., Mahmood, R., McNiven, A. L., Diamant, A., and Chan, T. C. Y. (2019). Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med. Phys.* 47, 297–306. doi: 10.1002/mp.13896

Bokrantz, R., and Forsgren, A. (2013). An Algorithm for approximating convex pareto surfaces based on dual techniques. *INFORMS J. Comput.* 25, 377–393. doi: 10.1287/ijoc.1120.0508

Craft, D., Halabi, T., Shih, H. A., and Bortfeld, T. (2007). An approach for practical multiobjective IMRT treatment planning. *Int. J. Radiat. Oncol. Biol. Phys.* 69, 1600–1607. doi: 10.1016/j.ijrobp.2007.08.019

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res. Proc. Track* 9, 249–256.

Guckenberger, M., Richter, A., Krieger, T., Wilbert, J., Baier, K., and Flentje, M. (2009). Is a single arc sufficient in volumetric-modulated arc therapy (VMAT) for complex-shaped target volumes? *Radiother. Oncol.* 93, 259–265. doi: 10.1016/j.radonc.2009.08.015

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. doi: 10.1109/CVPR.2016.90

Hwang, C. L., and Masud, A. S. M. (1979). *Multiple Objective Decision Making, Methods and Applications: A State-of-the-Art Survey*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-45511-7

ICRU (1976). Determination of absorbed dose in a patient irradiated by beams of X or gamma rays in radiotherapy procedures. *Med. Phys.* 4:461. doi: 10.1118/1.594356

Jensen, P. J., Zhang, J., and Wu, Q. J. (2020). Technical note: interpolated pareto surface similarity metrics for multi-criteria optimization in radiation therapy. *Med. Phys.* 47, 6450–6457. doi: 10.1002/mp.14541

Kingma, D., and Ba, J. (2014). "Adam: a method for stochastic optimization, in *International Conference on Learning Representations*.

Klein, E. E., Hanley, J., Bayouth, J., Yin, F.-F., Simon, W., Dresser, S., et al. (2009). Task group 142 report: quality assurance of medical accelerators. *Med. Phys.* 36, 4197–4212. doi: 10.1118/1.3190392

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. Boston, MA: Springer. doi: 10.1007/978-1-4615-5563-6

Monz, M., Küfer, K.-H., Bortfeld, T., and Thieke, C. (2008). Pareto navigation - algorithmic foundation of interactive multi-criteria IMRT planning. *Phys. Med. Biol*. 53, 985–998. doi: 10.1088/0031-9155/53/4/011

Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., et al. (2019). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci. Rep*. 9:1076. doi: 10.1038/s41598-018-37741-x

Otto, K. (2008). Volumetric modulated arc therapy: IMRT in a single gantry arc. *Med. Phys*. 35, 310–317. doi: 10.1118/1.2818738

Ronneberger, O., Fischer, P., and Brox, T. U. (2015). *Net: Convolutional Networks for Biomedical Image Segmentation*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-24574-4_28

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv[Preprint].arXiv:1409.1556.*

Teoh, M., Clark, C. H., Wood, K., Whitaker, S., and Nisbet, A. (2011). Volumetric modulated arc therapy: a review of current literature and clinical use in practice. *Br. J. Radiol*. 84, 967–996. doi: 10.1259/bjr/22373346

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv:arXiv[Preprint].*1511.07122.

Zhang, P., Happersett, L., Hunt, M., Jackson, A., Zelefsky, M., and Mageras, G. (2010). Volumetric modulated arc therapy: planning and evaluation for prostate cancer cases. *Int. J. Radiat. Oncol. Biol. Phys*. 76, 1456–1462. doi: 10.1016/j.ijrobp.2009.03.033

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.

Evaluating the Effectiveness of Personalized Medicine With Software

Adam Kapelner[1]*, Justin Bleich[2], Alina Levine[1], Zachary D. Cohen[3], Robert J. DeRubeis[3] and Richard Berk[2]

[1]Department of Mathematics, Queens College, CUNY, Queens, NY, United States, [2]Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA, United States, [3]Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

We present methodological advances in understanding the effectiveness of personalized medicine models and supply easy-to-use open-source software. Personalized medicine involves the systematic use of individual patient characteristics to determine which treatment option is most likely to result in a better average outcome for the patient. Why is personalized medicine not done more in practice? One of many reasons is because practitioners do not have any easy way to holistically evaluate whether their personalization procedure does better than the standard of care, termed *improvement*. Our software, “Personalized Treatment Evaluator” (the R package PTE), provides inference for improvement out-of-sample in many clinical scenarios. We also extend current methodology by allowing evaluation of improvement in the case where the endpoint is binary or survival. In the software, the practitioner inputs 1) data from a single-stage randomized trial with one continuous, incidence or survival endpoint and 2) an educated guess of a functional form of a model for the endpoint constructed from domain knowledge. The bootstrap is then employed on data unseen during model fitting to provide confidence intervals for the improvement for the average future patient (assuming future patients are similar to the patients in the trial). One may also test against a null scenario where the hypothesized personalization are not more useful than a standard of care. We demonstrate our method’s promise on simulated data as well as on data from a randomized comparative trial investigating two treatments for depression.

Keywords: personalized medicine, inference, bootstrap, treatment regimes, randomized comparative trial, statistical software

OPEN ACCESS

Edited by:
Jake Y. Chen,
University of Alabama at Birmingham,
United States

Reviewed by:
Sunyoung Jang,
Princeton Radiation Oncology Center,
United States
Himanshu Arora,
University of Miami, United States

****Correspondence:***
Adam Kapelner
kapelner@qc.cuny.edu

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Big Data

Received: *14 June 2020*
Accepted: *03 February 2021*
Published: *18 May 2021*

Citation:
Kapelner A, Bleich J, Levine A,
Cohen ZD, DeRubeis RJ and Berk R
(2021) Evaluating the Effectiveness of
Personalized Medicine With Software.
Front. Big Data 4:572532.
doi: 10.3389/fdata.2021.572532

1 INTRODUCTION

Personalized medicine, sometimes called “precision medicine” or “stratified medicine” (Smith, 2012), is a medical paradigm offering the possibility for improving the health of individuals by judiciously treating individuals based on his or her heterogeneous prognostic or genomic information (Zhao and Zeng, 2013). These approaches have been described under the umbrella of “P4 medicine,” a systems approach that combines predictive, personalized, preventive and participatory features to improve patient outcomes (Weston and Hood, 2004; Hood and Friend, 2011). And the interest in such personalization is exploding.

Fundamentally, personalized medicine is a statistical problem. However, much recent statistical research has focused on how to best estimate *dynamic treatment regimes* or *adaptive interventions* (Collins et al., 2004; Chakraborty and Murphy, 2014). These are essentially strategies that vary treatments administered over time as more is learned about how particular patients respond to one or more interventions. Elaborate models are often proposed that purport to estimate optimal dynamic treatment regimes from multi-stage experiments (Murphy, 2005b) as well as the more difficult situation of inference in observational studies.

Thus, the extant work, at least in the field of statistics, is highly theoretical. There is a dearth of software that can answer two fundamental questions practitioners will need answered before they can personalize future patients' treatments: 1) How much better is this personalization model expected to perform when compared to my previous "naive" strategy for allocating treatments? 2) How confident can I be in this estimate? Can I reject a null hypothesis that it will perform no better than the standard of care? Chakraborty and Moodie (2013), page 168 believe that "more targeted research is warranted" on these important questions; and the goal of our paper is to provide a framework and usable software that fills in part of this gap.

Personalized medicine is a broad paradigm encompassing many real-world situations. The setting we focus on herein is a common one: using previous randomized comparative/controlled trial (RCT) data to be able to make better decisions for future patients. We consider RCTs with two treatment options (two-arm), with one endpoint measure (also called the "outcome" or "response" which can be continuous, binary or survival) and where the researchers also collected a variety of patient characteristics to be used for personalization. The practitioner also has an idea of a model of the response (usually a simple first-order interaction model). Although this setting is simplistic, our software can then answer the two critical questions listed above.

Our advances are modest but important for practitioners. 1) Practitioners now have easy-to-use software that automates the testing of their personalization models. 2) We introduce a more intuitive metric that gauges how well the personalization is performing: "improvement" vs. a baseline strategy. 3) Our estimates and hypothesis tests of improvement are all cross-validated, making the estimates honest even when the data at hand was overfit and thereby generalizable to future patients. This external validity is only possible if future patients can be thought to come from the same population as the clinical trial, a generalization that is debated but beyond the scope of our work. 4) We have extended this well-established methodology to the setting of binary and survival endpoints, the most common endpoints in clinical trials.

The paper proceeds as follows. In **Section 2**, we review the modern personalized medicine literature and locate our method and its limitations within. **Section 3** describes our methods and its limitations in depth, by describing the conceptual framework emphasizing our methodological advances. We then carefully specify the data and model inputs, define the improvement metric, and illustrate a strategy for providing practitioners with estimates and inference. **Section 4** applies our methods to 1) a simple simulated dataset in which the response model is known, 2) a more complicated dataset characterized by an unknown response model and 3) a real data set from a published clinical trial investigating two treatments for major depressive disorder. **Section 5** demonstrates the software for all three types of endpoints: continuous, binary and survival. **Section 6** concludes and offers future directions of which there are many.

2 BACKGROUND

Consider an individual seeking one of two treatments, neither of which is known to be superior for all individuals. "What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?" (Paul, 1967).[1] Sometimes practitioners will select a treatment based informally on personal experience. Other times, practitioners may choose the treatment that their clinic or peers recommend. If the practitioner happens to be current on the research literature and there happens to be a published RCT whose results have clear clinical implications, the study's superior treatment (on average) may be chosen.

Each of these approaches can sometimes lead to improved outcomes, but each also can be badly flawed. For example, in a variety of clinical settings, "craft lore" has been demonstrated to perform poorly, especially when compared to very simple statistical models (Dawes, 1979). It follows that each of these "business-as-usual" *treatment allocation procedures* can in principle be improved if there are patient characteristics available which are related to how well an intervention performs.

Personalized medicine via the use of patient characteristics is by no means a novel idea. As noted as early as 1865, "the response of the average patient to therapy is not necessarily the response of the patient being treated" (translated by Bernard, 1957). There is now a substantial literature addressing numerous aspects of personalized medicine. The field is quite fragmented: there is literature on treatment-covariate interactions, locating subgroups of patients as well as literature on personalized treatment effects estimation. However, a focus on inference is rare in the literature and available software for inference is negligible.

Byar (1985) provides an early review of work involving treatment-covariate interactions. Byar and Corle (1977) investigates tests for treatment-covariate interactions in survival models and discusses methods for treatment recommendations based on covariate patterns. Shuster and van Eys (1983) considers two treatments and proposes a linear model composed of a treatment effect, a prognostic factor, and their interaction. Using this model, the authors create confidence intervals to determine for which values of the prognostic factor one of two treatments is superior.

[1] Note that this problem is encountered in fields outside of just medicine. For example, finding the movie that will elicit the most enjoyment to the individual (Zhou et al., 2008) or assessing whether a certain unemployed individual can benefit from job training (LaLonde, 1986). Although the methods discussed herein can be applied more generally, we will employ examples and the vocabulary from the medical field.

Many researchers also became interested in discovering "qualitative interactions," which are interactions that create a subset of patients for which one treatment is superior and another subset for which the alternative treatment is superior. Gail and Simon (1985) develop a likelihood ratio test for qualitative interactions which was further extended by Pan and Wolfe (1997) and Silvapulle (2001). For more information and an alternative approach, see Foster (2013).

Much of the early work in detecting these interactions required a prior specification of subgroups. This can present significant difficulties in the presence of high dimensionality or complicated associations. More recent approaches such as Su et al. (2009) and Dusseldorp and Van Mechelen (2014) favor recursive partitioning trees that discover important nonlinearities and interactions. Dusseldorp et al. (2016) introduce an R package called QUINT that outputs binary trees that separate participants into subgroups. Shen and Cai (2016) propose a kernel machine score test to identify interactions and the test has more power than the classic Wald test when the predictor effects are non-linear and when there is a large number of predictors. Berger et al. (2014) discuss a method for creating prior subgroup probabilities and provides a Bayesian method for uncovering interactions and identifying subgroups. Berk et al. (2020) uses trees to both locate subgroups and to provides valid inference on the local heterogeneous treatment effects. They do so by exhaustively enumerating every tree and every node, running every possible test and then providing valid post-selection inference (Berk et al., 2013b).

In our method, we make use of RCT data. Thus, it is important to remember that "clinical trials are typically not powered to examine subgroup effects or interaction effects, which are closely related to personalization … even if an optimal personalized medicine rule can provide substantial gains, it may be difficult to estimate this rule with few subjects" (Rubin and van der Laan, 2012). This is why a major bulk of the literature does not focus on finding covariate-treatment interactions or locating subgroups of individuals, but instead on the entire model itself (called the "regime") and that entire model is then used to sort individuals. Holistic statements can then be made on the basis of this entire sorting procedure. We turn to some of this literature now.

Zhang et al. (2012a) consider the context of treatment regime estimation in the presence of model misspecification when there is a single-point treatment decision. By applying a doubly robust augmented inverse probability weighted estimator that under the right circumstances can adjust for confounding and by considering a restricted set of policies, their approach can help protect against misspecification of either the propensity score model or the regression model for patient outcome. Brinkley et al. (2010) develop a regression-based framework of a dichotomous response for personalized treatment regimes within the rubric of "attributable risk." They propose developing optimal treatment regimes that minimize the probability of a poor outcome, and then consider the positive consequences, or "attributable benefit," of their regime. They also develop asymptotically valid inference for a parameter similar to improvement with business-as-usual as the random (see our **Section 3.3**), an idea we extend. Within the literature, their work is the closest conceptually to ours. Gunter et al. (2011b) develop a stepwise approach to variable selection and in Gunter et al. (2011a) compare it to stepwise regression. Rather than using a traditional sum-of-squares metric, the authors' method compares the estimated mean response, or "value," of the optimal policy for the models considered, a concept we make use of in **Section 3**. Imai and Ratkovic (2013) use a modified Support Vector Machine with LASSO constraints to select the variables useful in an optimal regime when the response is binary. van der Laan and Luedtke (2015) uses a loss-based super-learning approach with cross-validation.

Also important within the area of treatment regime estimation, but not explored in this paper, is the estimation of dynamic treatment regimes (DTRs). DTRs constitute a set of decision rules, estimated from many experimental and longitudinal intervals. Each regime is intended to produce the highest mean response over that time interval. Naturally, the focus is on optimal DTRs—the decision rules which provide the highest mean response. Murphy (2003) and Robins (2004) develop two influential approaches based on regret functions and nested mean models respectively. Moodie et al. (2007) discuss the relationship between the two while Moodie and Richardson (2009) and Chakraborty et al. (2010) present approaches for mitigating biases (the latter also fixes biases in model parameter estimation stemming from their non-regularity in SMART trials). Robins et al. (2008) focus on using observational data and optimizing the time for administering the stages—the "when to start"—within the DTR. Orellana et al. (2010) develop a different approach for estimating optimal DTRs based on marginal structural mean models. Henderson et al. (2010) develop optimal DTR estimation using regret functions and also focus on diagnostics and model checking. Barrett et al. (2014) develop a doubly robust extension of this approach for use in observational data. Laber et al. (2014) demonstrate the application of set-valued DTRs that allow balancing of multiple possible outcomes, such as relieving symptoms or minimizing patient side effects. Their approach produces a subset of recommended treatments rather than a single treatment. Also, McKeague and Qian (2014) estimate treatment regimes from functional predictors in RCTs to incorporate biosignatures such as brain scans or mass spectrometry.

Many of the procedures developed for estimating DTRs have roots in reinforcement learning. Two widely used methods are Q-learning Murphy (2005a) and A-learning (see Schulte et al., 2014 for an overview of these concepts). One well-noted difficulty with Q-learning and A-learning are their susceptibility to model misspecification. Consequently, researchers have begun to focus on "robust" methods for DTR estimation. Zhang et al. (2013) extend the doubly robust augmented inverse probability weighted method (Zhang et al., 2012a) by considering multiple binary treatment stages.

Many of the methods mentioned above can be extended to censored survival data. Zhao et al. (2015) describe a computationally efficient method for estimating a treatment regimes that maximizes mean survival time by extending the weighted learning inverse probability method. This method is doubly robust; it is protected from model misspecification if either the censoring model or the survival model is correct.

Additionally, methods for DTR estimation can be extended. Goldberg and Kosorok (2012) extend Q-learning with inverse-probability-of-censoring weighting to find the optimal treatment plan for individual patients, and the method allows for flexibility in the number of treatment stages.

It has been tempting, when creating these treatment regime models, to directly employ them to predict the differential response of individuals among different treatments. This is called in the literature "heterogeneous treatment effects models" or "individualized treatment rules" and there is quite a lot of interest in it.

Surprisingly, methods designed for accurate estimation of an overall conditional mean of the response may not perform well when the goal is to estimate these individualized treatment rules. Qian and Murphy (2011) propose a two-step approach to estimating "individualized treatment rules" based on single-stage randomized trials using ℓ_1-penalized regression while Lu et al. (2013) use quadratic loss which facilitates variable selection. Rolling and Yang (2014) develop a new form of cross-validation which chooses between different heterogeneous treatment models.

One current area of research in heterogeneous effect estimation is the development of algorithms that can be used to create finer and more accurate partitions. Kallus (2017) presents three methods for the case of observational data: greedily partitioning data to find optimal trees, bootstrap aggregating to create a "personalization forest" a la Random Forests, and using the tree method coupled with mixed integer programming to find the optimal tree. Lamont et al. (2018) build on the prior methods of parametric multiple imputation and recursive partitioning to estimate heterogeneous treatment effects and compare the performance of both methods. This estimation can be extended to censored data. Henderson et al. (2020) discuss the implementation of Bayesian Additive Regression Trees for estimating heterogeneous effects which can be used for continuous, binary and censored data. Ma et al. (2019) proposes a Bayesian predictive method that integrates multiple sources of biomarkers.

One major drawback of many of the approaches in the literature reviewed is their significant difficulty evaluating estimator performance. Put another way, given the complexity of the estimation procedures, statistical inference is very challenging. Many of the approaches require that the proposed model be correct. There are numerous applications in the biomedical sciences for which this assumption is neither credible nor testable in practice. For example, Evans and Relling (2004) consider pharmacogenomics, and argue that as our understanding of the genetic influences on individual variation in drug response and side-effects improves, there will be increased opportunity to incorporate genetic moderators to enhance personalized treatment. But we will ever truly understand the endpoint model well enough to properly specify it? Further, other biomarkers (e.g. neuroimaging) of treatment response have begun to emerge, and the integration of these diverse moderators will require flexible approaches that are robust to model misspecification (McGrath et al., 2013). How will the models of today incorporate important relationships that can be anticipated but have yet to be identified? Further, many proposed methods employ non-parametric models that use the data to decide which internal parameters to fit and then in turn estimates these internal parameters. Thus a form of model selection that introduces difficult inferential complications (see Berk et al., 2013b).

At the very least, therefore, there should be an alternative inferential framework for evaluating treatment regimes that do not require correct model specification (and thereby obviating the need for model checking and diagnostics) nor knowledge of unmeasured characteristics (see discussion in Henderson et al., 2010) accompanied by easy-to-use software. This is the modest goal herein.

3 METHODS

This work seeks to be didactic and thus carefully explains the extant methodology and framework (**Section 3.1**), data inputs (**Section 3.2**), estimation (**Section 3.4**) and a procedure for inference (**Section 3.5**). For those familiar with this literature, these sections can be skipped. Our methodological contributions then are to 1) employ out-of-sample validation to (**Section 3.4**) specifically to 2) *improvement*, the metric defined as the difference in how patients allocated via personalized model fare in their outcome and a patients allocated via business-as-usual model fare in their outcome (**Section 3.3**) and 3) to extend this validation methodology to binary and survival endpoints (**Sections 3.4.2** and **3.4.3**). **Table 1** serves as a guide to the main notation used in this work, organized by section.

3.1 Conceptual Framework

We imagine a set of random variables having a joint probability distribution that can be also be viewed as a population from which data could be randomly and independently realized. The population can also be imagined as all potential observations that could be realized from the joint probability distribution. Either conception is consistent with our setup.

A researcher chooses one of the random variables to be the response Y which could be continuous, binary or survival (with a corresponding variable that records its censoring, explained later). We assume without loss of generality that a higher-valued outcome is better for all individuals. Then, one or more of the other random variables are covariates $\boldsymbol{X} \in \mathcal{X}$. At the moment, we do not distinguish between observed and unobserved covariates but we will later. There is then a conditional distribution $\mathbb{P}(Y|\boldsymbol{X})$ whose conditional expectation function $\mathbb{E}[Y|\boldsymbol{X}]$ constitutes the population response surface. No functional forms are imposed on the conditional expectation function and thus it is allowed to be nonlinear with interactions among the covariates which is the case in artificial intelligence procedures (machine learning).

All potential observations are hypothetical study subjects. Each can be exposed to a random treatment denoted $A \in \{0, 1\}$ where zero codes for the first experimental condition also equivalently referred to as T_1 (which may be considered the "control" or "comparison" condition) and one codes for another experimental condition equivalently referred to as T_2. We make the standard assumption of no interference

TABLE 1 | A compendium of the main notation in our methodology by section.

Notation	Description
Framework (**Section 3.1**)	
Y	The random variable (r.v.) for the outcomes for the subjects
X, $\mathcal{X}$	The r. v. for the observed measurements for the subjects, its support
A	The r. v. for the treatment
T_1, T_2	The two treatments, shorthand for their codes, zero and one
d or $d[f]$	The decision function; this function maps observed measurements to treatment
V or $V[d]$	The value of the decision function, the average outcome over all patients if this decision is used to allocate treatment
d^*	The unknown optimal decision function i.e. the one with highest V
d_0	A naive, baseline, business-as-usual or null decision function
μ_{I_0}	The unknown improvement of d over d_0; it is the difference of values
The RCT data (**Section 3.2.1**)	
n	The number of subjects in the randomized comparative trial (RCT)
p	The number of measurements assessed on each subject
x_i	The vector of p measurements for the ith subject
$x_{i,j}$	The jth measurement for the ith subject
$\boldsymbol{X}$	The $n \times p$ matrix of all measurements for all subjects
$\boldsymbol{A}$	The vector of treatments for all the n subjects
$\boldsymbol{y}$	The vector of outcomes for all the n subjects
The response model (**Section 3.2.2**)	
f	The function that relates the p measurements and A to the response
U_i	The r. v. for the unknown covariates for subject i
$\xi(x_i, A_i, U_i)$	The function that computes misspecification in the response
$\mathcal{E}_i$	The r. v. for irreducible noise for the ith subject
β_j	The linear coefficient for the jth measurement when $A = 0$
γ_j	The additional linear coefficient for the jth measurement when $A = 1$
Out of sample estimation and validation (**Section 3.4**)	
$\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}}$	The subset of the data used to create the fit of f
$\boldsymbol{X}_{\text{test}}, \boldsymbol{y}_{\text{test}}$	The subset of the data used to validate the fit of f
$\widehat{f}$, $\widehat{d}$, $\widehat{V}$, $\widehat{I}_0$	The finite-sample estimates of f, d, V, μ_{I_0}
$\overline{y}_{set}$	The arithmetic average of $\{y_i : i \in set\}$
$\widehat{\beta}_j$, $\widehat{\gamma}_j$	The finite-sample estimates of β_j, γ_j
Inference (**Section 3.5**)	
B	The number of bootstrap samples
$\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}$	A sample of the rows of $\boldsymbol{X}, \boldsymbol{y}$ with replacement
$\tilde{I}_{0,b}$	The bth estimate of μ_{I_0} in the bootstrap
α	The size of the hypothesis test
Personalization of future subjects' treatments (**Section 3.6**)	
x_*	A future subject (not part of the RCT)

between study subjects, which means that the outcome for any given subject is unaffected by the interventions to which other subjects are randomly assigned (Cox, 1958) and outcomes under either condition can vary over subjects (Rosenbaum, 2002, Section 2.5.1). In short, we employ the conventional Neyman-Rubin approach (Rubin, 1974) but treat all the data as randomly realized (Berk et al., 2013a).

A standard estimation target in RCTs is the population average treatment effect (PATE), defined here as $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$, the difference between the population expectations. That is, the PATE is defined as the difference in mean outcome were all subjects exposed to T_2 or alternatively were all exposed to T_1. In a randomized controlled trial, the PATE is synonymous with the overall efficacy of the treatment of interest and it is almost invariably the goal of the trial (Zhao and Zeng, 2013).

For personalization, we want to make use of any association between Y and X. For the hypothetical study subjects, there is a conditional population response surface $\mathbb{E}[Y|X, A = 1]$ and another conditional population response surface $\mathbb{E}[Y|X, A = 0]$, a key objective being to exploit the difference in these response surfaces for better treatment allocation. The typical approach is to create a deterministic *individualized treatment decision rule* d that takes an individual's covariates and maps them to a treatment. We seek $d : \mathcal{X} \rightarrow \{0, 1\}$ based on knowledge of the differing conditional population response surfaces. The rule is sometimes called an *allocation procedure* because it determines which treatment to allocate based on measurements made on the individual. To compare different allocation procedures, our metric is the expectation of the outcome Y using the allocation procedure d averaged over all subjects $\mathcal{X}$. Following the notation of Qian and Murphy (2011), we denote this expectation as the value of the decision rule

$$V[d] := \mathbb{E}^{d}_{X,A}[Y] \triangleq \int_{X} \left(\sum_{a \in \{0,1\}} \left(\int_{\mathbb{R}} y f_{Y|X,A}(y,x,a) dy \right) \mathbb{1}_{a=d(x)} \right) f_X(x) \mathrm{d}x.$$

Although the integral expression appears complicated, when unpacked it is merely an expectation of the response averaged over $\mathcal{X}$, the space of all patients characteristics. When averaging over $\mathcal{X}$, different treatments will be recommended based on the rule, i.e. $a = d(x)$, and that in turn will modify the density of the response, $f_{Y|X}$. Put another way, $V[d]$ is the mean patient outcome when personalizing each patient's treatment.

We have considered all covariates to be random variables because we envision *future* patients for whom an appropriate treatment is required. Ideally, their covariate values are realized from the same joint distribution as the covariate values for the study subjects, an assumption that is debated and discussed in the concluding section.

In addition, we do not intend to rely on estimates of the two population response surfaces. As a practical matter, we will make do with a population response surface approximation for each. No assumptions are made about the nature of these approximations and in particular, how well or poorly either population approximation corresponds to the true conditional response surfaces.

Recall that much of the recent literature has been focused on finding the optimal rule, $d^* \triangleq \text{argmax}_d\{V[d]\}$. Although this is an admirable ideal (as in Qian and Murphy 2011), our goals here are more modest. We envision an imperfect rule d far from d^*, and we wish to gauge its performance relative to the performance of another rule d_0, where the "naught" denotes a business-as-usual allocation procedure, sometimes called "standard of care". Thus, we define the population value improvement μ_{I_0} as the value of d minus the value of d_0,

$$\mu_{I_0} \triangleq V[d] - V[d_0] = \mathbb{E}^{d}_{X,A}[Y] - \mathbb{E}^{d_0}_{X,A}[Y], \tag{1}$$

which is sometimes called "benefit" in the literature. Since our convention is that higher response values are better, we seek large, positive improvements that translate to better average performance (as measured by the response). Note that this is a natural measure when Y is continuous. When Y is incidence or survival, we redefine μ_{I_0} (see **Sections 3.4.2** and **3.4.3**).

The metric μ_{I_0} is not standard in the literature but we strongly believe it to be the natural metric for personalization following Kallus (2017). There are many other such metrics beyond value and improvement. For example, Ma et al. (2019) uses three 1) the expected number of subjects misassigned to their optimal treatment, 2) exepected gain or loss in treatment utility and 3) the expected proportion where the model correctly predicted the response which is useful only in the binary response case which we address later.

3.2 Our Framework's Required Inputs

Our method depends on two inputs 1) access to RCT data and 2) either a prespecified parametric model $f(x, A; \theta)$ for the population approximation of the true response surfaces or an explicit d function. If we prespecified f, we then use the RCT data to estimate parameters of the model $\hat{\theta}$, and the estimates are embedded in the estimated model, $\hat{f}$. This model estimate permits us, in turn, to construct an estimated decision rule $\hat{d}$ and an estimate of the improved outcomes future subjects will experience (explained later in **Section 3.4**). We assume that the model f is specified before looking at the data. "Data snooping" (running our method, checking the p-value, changing the model f and running again) fosters overfitting and can introduce serious estimation bias, invalidating our confidence intervals and statistical tests (Berk et al., 2013b).

3.2.1 The RCT Data

Our procedure strictly requires RCT data to ensure there is a causal effect of the heterogeneous parameters. Much of the research discussed in the background (**Section 2**) applies in the case of observational data. We realize this limits the scope of our proposal. The RCT data must come from an experiment undertaken to estimate the PATE for treatments T_1 and T_2 for a diagnosis of a disease of interest. T_1 and T_2 are the same treatments one would offer to future subjects with the same diagnosis.

There are n subjects each with p covariates which are denoted for the ith subject as $\boldsymbol{x_i} \triangleq [x_{i1}, x_{i2}, \ldots, x_{ip}]$. Because these covariates will be used to construct a decision rule applied with future patients in clinical settings, the $\boldsymbol{x_i}$'s in the RCT data must be the same covariates measured for new subjects. Thus, such characteristics such as the site of treatment (in a multi-center trial) or the identification of the medical practitioner who treated each subject or hindsight-only variables are not included.

We assume the outcome measure of interest y_i is assessed once per subject. Aggregating all covariate vectors, binary allocations and responses rowwise, we denote the full RCT data as the column-bound matrix $[\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{y}]$. In practice, missing data can be imputed (in both the RCT data and the future data), but herein we assume complete data.

We will be drawing inference to a patient population beyond those who participated in the experiment. Formally, new subjects must be sampled from that same population as were the subjects in the RCT. In the absence of explicit probability sampling, the case would need to be made that the model can generalize. This requires subject-matter expertise and knowledge of how the study subjects were recruited.

3.2.2 The Model for the Response Based on Observed Measurements

The decision rule d is a function of x through f and is defined as

$$d[f(x)] \triangleq \underset{A\in\{0,1\}}{\text{argmax}}\, f(x, A) = \mathbb{1}_{f(x,1)-f(x,0)}. \tag{2}$$

As in Berk et al. (2014), we assume the model f provided by the practitioner to be an approximation using the available information,

$$Y_i = \underbrace{f(X_i, A_i) + \xi(X_i, U_i, A_i)}_{\mathbb{E}[Y_i|X_i,U_i,A_i]} + \varepsilon_i, \tag{3}$$

where f differs from the true response expectation by a term dependent on U, the unobserved information. The last term ε_i is the irreducible noise around the true conditional expectations and is taken to be independent and identically distributed, mean-centered and uncorrelated with the covariates. Even in the absence of ε_i, f will always differ from the true conditional

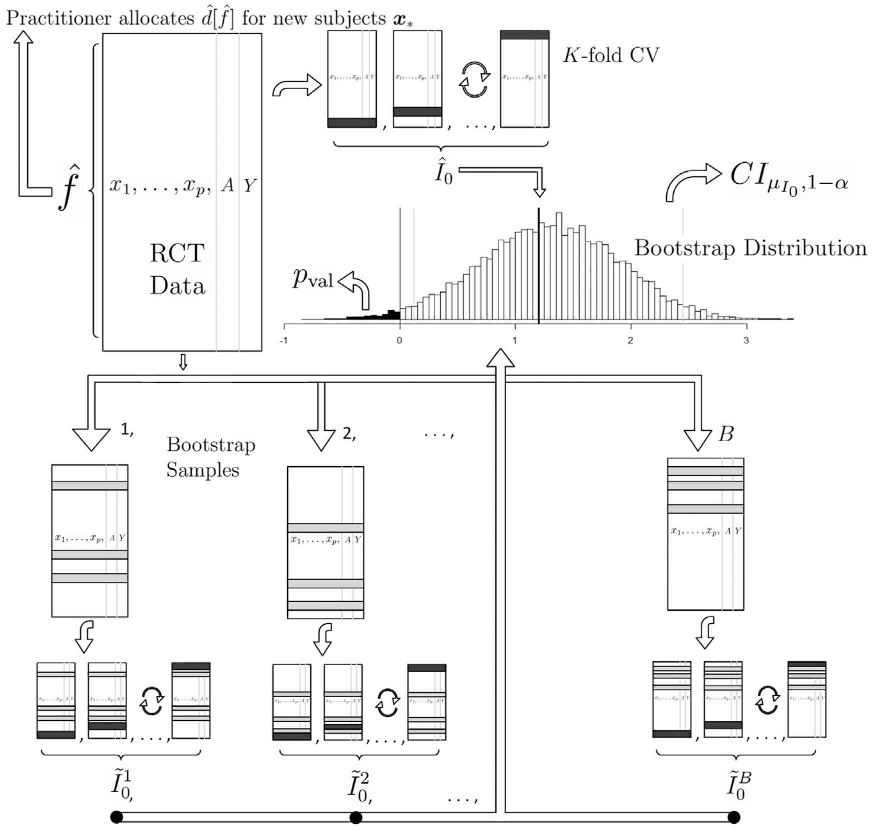


FIGURE 1 | A graphical illustration of (1) our proposed method for estimation and (2) our proposed method for inference on the population mean improvement of an allocation procedure and (3) our proposed future allocation procedure (top left of the illustration). To compute the best estimate of the improvement $\hat{I}_0$, the RCT data goes through the K-fold cross validation procedure of **Section 3.4** (depicted in the top center). The black slices of the data frame represent the test data. To draw inference, we employ the non-parametric bootstrap procedure of **Section 3.5** by sampling the RCT data with replacement and repeating the K-fold CV to produce $\tilde{I}_0^1, \tilde{I}_0^2, \ldots, \tilde{I}_0^B$ (bottom). The gray slices of the data frame represent the duplicate rows in the original data due to sampling with replacement. The confidence interval and significance of $H_0 : \mu_{I_0} \leq 0$ is computed from the bootstrap distribution (middle center). Finally, the practitioner receives $\hat{f}$ which is built with the complete RCT data (top left).

expectation function by $\xi_i(X_i, U_i, A_i)$, which represents model misspecification (Box and Draper, 1987, Chapter 13).

We wish only to determine whether an estimate of f is *useful* for improving treatment allocation for future patients (that are similar to the patients in the RCT) and do not expect to recover the optimal allocation rule d^* which requires the unseen U. Further, we do not concern ourselves with substantive interpretations associated with any of the p covariates, a goal of future research. Thus, our method is robust to model misspecification by construction.

What could f look like in practice? Assume a continuous response (binary and survival are discussed later) and consider the conventional linear regression model with first order interactions. Much of the literature we reviewed in **Section 2** favored this class of models. We specify a linear model containing a subset of the covariates used as main effects and a possibly differing subset of the covariates to be employed as first order interactions with the treatment indicator, $\{x_{1'}, \ldots, x_{p'}\} \subset \{x_1, \ldots, x_p\}$, selected using domain knowledge:

$$f(x_{i1}, A_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + A_i\Big(\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'}\Big). \quad (4)$$

These interactions induce heterogeneous effects between T_1 and T_2 for a subject x in a very interpretable way: $d[f(x)] = 1$ when $\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'} > 0$ and 0 otherwise. The γ's are the critical component of the model if there are systematic patient-specific differences between the interventions. Thereby, d varies over different points in $\mathcal{X}$ space. Note that rules derived from this type of conventional model also have the added bonus as being interpretable as a best linear approximation of the true relationship.

We stress that our models are *not* required to be of this form, but we introduce them here mostly for familiarity and pedagogical simplicity. There are times when this linear model will perform terribly even if $\{x_{1'}, \ldots, x_{p'}\}$ are the correct moderating variables. For a non-linear example, see Zhao and Zeng (2013), **Figure 1**, right. Although this model is the default implementation, the user can specify any model desired in the software. This will be discussed in **Section 5**.

We also stress that although the theory for estimating linear models' coefficients (as well as those for logistic regression and Weibull regression) is well-developed, we are not interested in inference for these coefficients in this work as our goal is only estimation and inference for overall usefulness of the personalization scheme, i.e. the unknown parameter μ_{I_0}. This will become clear in the next few sections.

3.3 Other Allocation Procedures

Although d_0 can be any allocation rule, for the purposes of the paper, we examine only two "business-as-usual" allocation procedures (others are discussed as extensions in **Section 6**). The first we call **random** denoting the allocation where the patient receives T_1 or T_2 with a fair coin flip, probability 50%. This serves as a baseline or "straw man" but nevertheless an important standard—the personalization model should be able to provide better patient outcomes than a completely random allocation.

The second business-as-usual procedure we call **best**. This procedure gives all patients the better of the two treatments as determined by the comparison of the sample average for all subjects who received T_1 denoted $\bar{y}_{T_1}$ and the sample average of all subjects who received T_2 denoted $\bar{y}_{T_2}$. This is used as the default in many frameworks for example Kang et al. (2014). We consider beating this procedure the gold standard in proof that the personalization truly works as practitioners most often employ the current best known treatment. However, some consider this comparison conservative (Brinkley et al., 2010, Section 7) and the next section will describe why it is statistically conservative as we lose sample size when demanding this comparison. Due to this conservativeness, barring conclusive evidence that either T_1 or T_2 is superior, the random procedure should be the standard of comparison. This case is not infrequent in RCTs which feature negative comparison results, the case in our clinical trial example of **Section 4.3**.

3.4 Estimating the Improvement Scores

3.4.1 For a Continuous Response

How well do unseen subjects with treatments allocated by d do on average compared to the same unseen subjects with treatments allocated by d_0? We start by computing the estimated improvement score, a sample statistic given by

$$\hat{I}_0 \triangleq \hat{V}[\hat{d}] - \hat{V}[\hat{d}_0], \tag{5}$$

where $\hat{d}$ is an estimate of the rule d derived from the population response surface approximation, $\hat{V}$ is an estimate of its corresponding value V and $\hat{I}_0$ is an estimate of the resulting population improvement μ_{I_0} (**Eq. 1**). The $\hat{d}_0$ notation indicates that sometimes the competitor d_0 may have to be estimated from the data as well. For example, the allocation procedure **best** must be calculated by using the sample average of the responses for both T_1 and T_2 in the data.

TABLE 2 | The elements of y_{test} cross-tabulated by their administered treatment A_i and our model's estimate of the better treatment $\hat{d}(\boldsymbol{x}_i)$.

	$\hat{d}(\boldsymbol{x}_i) = 0$	$\hat{d}(\boldsymbol{x}_i) = 1$
$A_i = 0$	P	Q
$A_i = 1$	R	S

In order to properly estimate μ_{I_0}, we use cross-validation (Hastie et al., 2013, Chapter 7.10). We split the RCT data into two disjoint subsets: training data with n_{train} of the original n observations $[\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}}]$ and testing data with the remaining $n_{\text{test}} = n - n_{\text{train}}$ observations $[\boldsymbol{X}_{\text{test}}, \boldsymbol{y}_{\text{test}}]$. Then $\hat{f}_{\text{train}}$ can be fit using the training data to construct $\hat{d}$ via **Eq. 2**. Performance of $\hat{d}$ as calculated by **Equation 5**, is then evaluated on the test data. Hastie et al. (2013) explain that a single train-test split yields an estimate of the "performance" of the procedure on future individuals conditional on $[\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}}]$, the "past". Thus, the $\hat{I}_0$ statistic defined in **Eq. 5** computed using $[\boldsymbol{X}_{\text{test}}, \boldsymbol{y}_{\text{test}}]$ can provide an honest assessment of improvement (i.e. immune to overfitting in $\hat{f}$) who are allocated using our proposed methodology compared to a baseline business-as-usual allocation strategy (Faraway, 2016). This can be thought of as employing a replicated trial, often required in drug development programs, which separates rule construction (in-sample) from rule validation (out-of-sample) as recommended by Rubin and van der Laan (2012). Note that this comes at a cost of more sample variability (as now our estimate will be based on the test subset with a sample size much smaller than n). Our framework and software is the first to provide user-friendly out-of-sample validation for the overall utility of personalized medicine models as a native feature.

Given the estimates $\hat{d}$ and $\hat{d}_0$, the question remains of how to explicitly compute $\hat{V}$ for subjects we have not yet seen in order to estimate $\hat{I}_0$. That is, we are trying to estimate the expectation of an allocation procedure over covariate space $\mathcal{X}$.

Recall that in the test data, our allocation prediction $\hat{d}(\boldsymbol{x}_\text{i})$ is the binary recommendation of T_1 or T_2 for each $\boldsymbol{x}_{\text{test},i}$. If we recommended the treatment that the subject actually was allocated in the RCT, i.e. $\hat{d}(\boldsymbol{x}+_i) = A_i$, we consider that subject to be "lucky". We define lucky in the sense that by the flip of the coin, the subject was randomly allocated to the treatment that our model-based allocation procedure estimates to be the better of the two treatments.

The average of the lucky subjects' responses should estimate the average of the response of new subjects who are allocated to their treatments based on our procedure d and this is the estimate of $\hat{V}[\hat{d}]$ we are seeking. Because the x's in the test data are assumed to be sampled randomly from population covariates, this sample average estimates the expectation over $\mathcal{X}$, i.e. $\mathbb{E}^d_{X,A}[Y]$ conditional on the training set. In order to make this concept more clear, it is convenient to consider **Table 2**, a 2 × 2 matrix which houses the sorted entries of the out-of-sample y_{test} based on the predictions, the $\hat{d}(\boldsymbol{x}_i)$'s.

The diagonal entries of sets P and S contain the "lucky" subjects. The notation $\bar{y}_{.}$ indicates the sample average among the elements of y_{test} specified in the subscript located in the cells of the table.

How do we compute $\hat{V}[\hat{d}_0]$, the business-as-usual procedure? For **random**, we simply average all of the y_{test} responses; for **best**, we average the y_{test} responses for the treatment group that has a larger sample average. Thus, the sample statistics of **Eq. 5** can be written as

$$\hat{I}_{\text{random}} \triangleq \bar{y}_{P \cup S} - \bar{y}_{\text{test}}, \tag{6}$$

$$\hat{I}_{\text{best}} \triangleq \bar{y}_{P \cup S} - \begin{cases} \bar{y}_{P \cup Q}, & \text{when} \quad \bar{y}_{P \cup Q} \geq \bar{y}_{R \cup S}, \\ \bar{y}_{R \cup S}, & \text{when} \quad \bar{y}_{P \cup Q} < \bar{y}_{R \cup S}. \end{cases} \tag{7}$$

Note that the plug-in estimate of value $\hat{V}[\hat{d}] = \bar{y}_{P \cup S}$ is traditional in the personalized medicine literature. For example, in Kallus (2017), Corollary 3 it is written as

$$\hat{V}\left[\hat{d}\right] := \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{\hat{d}(x_i) = A_i}}{\sum_{i=1}^{n} \mathbb{1}_{\hat{d}(x_i) = A_i}}. \tag{8}$$

There is one more conceptual point. Recall that the value estimates $\hat{V}[\cdot]$ are conditional on the training set. This means they do not estimate the unconditional $\mathbb{E}_{\mathbf{X},A}^{d}[Y]$. To address this, Hastie et al. (2013), Chapter 7 recommend that the same procedure be performed across many different mutually exclusive and collectively exhaustive splits of the full data. This procedure of building many models is called "K-fold cross-validation" (CV) and its purpose is to integrate out the effect of a single training set to result in the unconditional estimate of generalization. This is "an alternative approach … [that] for simplicity … [was not] consider [ed] … further" in the previous investigation of Chakraborty et al. (2014), page 5.

In practice, how large should the training and test splits be? Depending on the size of the test set relative to the training set, CV can trade bias for variance when estimating an out-of-sample metric. Small training sets and large test sets give more biased estimates since the training set is built with less data than the n observations given. However, large test sets have lower variance estimates since they are composed of many examples. There is no consensus in the literature about the optimal training-test split size (Hastie et al., 2013, page 242) but 10-fold CV is a common choice employed in many statistical applications and provides for a relatively fast algorithm. In the limit, n models can be created by leaving each observation out, as done in DeRubeis et al. (2014). In our software, we default to 10-fold cross validation but allow for user customization.

This estimation procedure outlined above is graphically illustrated in the top of **Figure 1**. We now extend this methodology to binary and survival endpoints in the next two sections.

3.4.2 For a Binary Response

In the binary case, we let V be the expected probability of the positive outcome under d just like in Kang et al. (2014). We consider three improvement metrics 1) the probability difference, 2) the risk ratio and 3) the odds ratio:

$$\text{(a)} \quad \mu_{I_0} \triangleq V[d] - V[d_0],$$

$$\text{(b)} \quad \mu_{I_0} \triangleq \frac{V[d]}{V[d_0]},$$

$$\text{(c)} \quad \mu_{I_0} \triangleq \frac{V[d] / (1 - V[d])}{V[d_0] / (1 - V[d_0])}.$$

and the estimate of all three (the probability difference, the risk ratio and the odds ratio) is found by placing hats on each term in the definitions above (all V's, d's and d_0's).

Following the example in the previous section we employ the analogous model, a logistic linear model with first order treatment interactions where the model f now denotes the probability of the positive outcome $y = 1$,

$$\begin{aligned} f(\boldsymbol{x}_{i1}, A_i) = \text{logit}\big(&\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &+ A_i(\gamma_0 + \gamma_{1'} x_{1'} + \ldots + \gamma_{p'} x_{p'})\big). \end{aligned} \tag{9}$$

This model, fit via maximum likelihood numerically (Agresti, 2018), is the default in our software implementation. Here, higher probabilities of success imply higher logit values so that algebraically we have the same form of the decision rule estimate, $\hat{d}[\hat{f}(\boldsymbol{x})] = 1$ when $\hat{\gamma}_0 + \hat{\gamma}_{1'} x_{1'} + \cdots + \hat{\gamma}_{p'} x_{p'} > 0$.

If the risk ratio or odds ratio improvement metrics are desired, **Eqs. 6**, **7** are modified accordingly but otherwise estimation is then carried out the same as in the previous section.

3.4.3 For a Survival Response

Survival responses differ in two substantive ways from continuous responses: 1) they are always non-negative and 2) some values are "censored" which means it appropriates the value of the last known measurement but it is certain that the true value is greater. The responses y are coupled with this censoring information $\boldsymbol{c}$, a binary vector of length n where the convention is to let $c_i = 0$ to indicate that y_i is censored and thus set equal to its last known value.

To obtain $\hat{d}$, we require a survival model. For example purposes here we will assume the exponential regression model (the exponentiation enforces the positivity of the response values) with the usual first order treatment interactions,

$$\begin{aligned} f(x_{i1}, A_i) = \exp\big(&\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &+ A_i(\gamma_0 + \gamma_{1'} x_{1'} + \cdots + \gamma_{p'} x_{p'})\big). \end{aligned} \tag{10}$$

Under the exponential model, the convention is that the noise term ε is multiplicative instead of additive, $Y_i = f(x_{i1}, A_i)\varepsilon_i$. Note that at this step, a fully parametric model is needed; the non-parametric Kaplan-Meier or the semi-parametric Cox proportion hazard model are insufficient as we need a means of explicitly estimating $\mathbb{E}[Y \mid X, A]$ for all values of X and both values of A.

Moreso than for continuous and incidence endpoints, parameter estimation is dependent on the choice of error distribution. Following Hosmer and Lemeshow (1999), a flexible model is to let $\ln(\varepsilon_1), \ldots, \ln(\varepsilon_n) \overset{\text{iid}}{\sim} \text{Gumbel}(0, \sigma^2)$, implying the popular Weibull model for survival (and the default in our software). As was the case previously, the user is free to choose whatever model they wish. The β_j's, γ_j's and the nuisance scale parameter σ^2 are fit using maximum likelihood taking care to ensure the correct contributions of censored and uncensored values. Similar to the case of logistic regression, the likelihood function does not have a closed form solution and must be approximated numerically.

Some algebra demonstrates that the estimated decision rule is the same as those above, i.e. $\widehat{d}[\widehat{f}(x)] = 1$ when $\widehat{\gamma}_0 + \widehat{\gamma}_{1'}x_{1'} + \cdots + \widehat{\gamma}_{p'}x_{p'} > 0$. In other words, the subject is given the treatment that yields the longest expected survival.

Subjects are then sorted in cells like **Table 2** but care is taken to keep the corresponding c_i values together with their paired y_i values following Yakovlev et al. (1994). At this point, we need to specify analogous computations to **Eqs. 6**, **7** that are sensitive to the fact that many y_i values are censored (The sample averages $\bar{y}$ obviously cannot be employed here because it ignores this censoring).

Of course we can reemploy a new Weibull model and define improvement as we did earlier as the difference in expectations (**Eq. 1**). However, there are no more covariates needed at this step as all subjects have been sorted based on $\widehat{d}(x)$. Thus, there is no reason to require a parametric model that may be arbitrarily wrong.

For our default implementation, we have chosen to employ the difference of the Kaplan-Meier median survival statistics here because we intuitively feel that a non-parametric estimate makes the most sense. Once again, the user is free to employ whatever they feel is most appropriate in their context. Given this default, please note that the improvement measure of **Eq. 1** is no longer defined as the difference in survival expectations, but now the difference in survival medians. This makes our framework slightly different in the case of survival endpoints.

3.5 Inference for the Population Improvement Parameter

Regardless of the type of endpoint, the $\widehat{I}_0$ estimates are drawn from an elaborate estimator whose sampling distribution is not available in closed form. We can employ the nonparametric bootstrap to obtain an asymptotic estimate of its sampling variability, which can be used to construct confidence intervals and testing procedures (Efron and Tibshirani, 1994).

In the context of our proposed methodology, the bootstrap procedure works as follows for the target of inference μ_{I_0}. We take a sample with replacement from the RCT data of size n denoted with tildes: $[\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}]$. Using the 10-fold CV procedure described at the end of **Section 3.4**, we create an estimate $\tilde{I}_0$. We repeat the resampling of the RCT data and the recomputation of $\tilde{I}_0$ B times where B is selected for resolution of the confidence interval and significance level of the test. In practice we found $B = 3000$ to be sufficient, so we leave this as the default in our software implementation. Because the n's of usual RCTs are small, and the bootstrap is embarrassingly parallelizable, this is not an undue computational burden.

In this application, the bootstrap approximates the sampling of many RCT datasets. Each $\tilde{I}$ that is computed corresponds to one out-of-sample improvement estimate for a particular RCT dataset drawn from the population of RCT datasets. We stress again that the frequentist confidence intervals and tests that we develop for the improvement measure do *not* constitute inference for a new individual's improvement, it is inference for the average improvement for future subjects vs. **random** allocation, μ_{I_0}.

To create a $1-\alpha$ level confidence interval, first sort the $\{\tilde{I}_{0,1}, \ldots, \tilde{I}_{0,B}\}$ by value, and then report the values corresponding to the empirical $\alpha/2$ and $1-\alpha/2$ percentiles. This is called the "percentile method." "Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its "rationale is somewhat obscure" (Rice, 1994, page 272). There are other ways to generate asymptotically valid confidence intervals using bootstrap samples but there is debate about which has the best finite sample properties. We have also implemented the "basic method" (Davison and Hinkley, 1997, page 194) and the bias-corrected "BC_a method" of Efron (1987) that DiCiccio and Efron (1996) claim performs an order of magnitude better in accuracy than the percentile method. Implementing other confidence interval methods for the bootstrap may be useful future work.

If a higher response is better for the subject, we set $H_0 : \mu_{I_0} \leq 0$ and $H_a : \mu_{I_0} > 0$. Thus, we wish to reject the null hypothesis that our allocation procedure is at most as useful as a naive business-as-usual procedure. To obtain an asymptotic p value based on the percentile method, we tally the number of bootstrap sample $\tilde{I}$ estimates below 0 and divide by B. This bootstrap procedure is graphically illustrated in the bottom half of **Figure 1** and the bootstrap confidence interval and p value computation is illustrated in the center. Note that for incidence outcomes where the improvement is defined as the risk ratio or odds ratio, we use $H_0 : \mu_{I_0} \leq 1$ and $H_a : \mu_{I_0} > 1$ and tally the number of $\tilde{I}$ estimates below 1.

We would like to stress once again that we are not testing for qualitative interactions—the ability of a covariate to "flip" the optimal treatment for subjects. Tests for such interactions would be hypothesis tests on the γ parameters of **Eqs. 4**, **9**, **10**, which assume model structures that are not even required for our procedure. Qualitative interactions are controversial due to model dependence and entire tests have been developed to investigate their significance. In the beginning of **Section 2** we commented that most RCTs are not even powered to investigate these interactions. "Even if an optimal personalized medicine rule [based on such interactions] can provide substantial gains it may be difficult to estimate this rule with few subjects" (Rubin and van der Laan, 2012). The bootstrap test (and our approach at large) looks at the holistic picture of the personalization scheme without focus on individual covariate-treatment interaction effects to determine if the personalization scheme in totality is useful, conceptually akin to the omnibus F-test in an OLS regression.

3.5.1 Concerns With Using the Bootstrap for This Inference

There is some concern in the personalized medicine literature about the use of the bootstrap to provide inference. First, the estimator for V is a non-smooth functional of the data which may result in an inconsistent bootstrap estimator (Shao, 1994). The non-smoothness is due to the indicator function in **Eq. 8** being non-differentiable, similar to the example found in (Horowitz, 2001, Chapter 52, Section 4.3.1). However, “the value of a fixed [response model] (i.e., one that is not data-driven) does not suffer from these issues and has been addressed by numerous authors” (Chakraborty and Murphy, 2014). Since our $\widehat{V}$ is constructed out-of-sample, it is merely a difference of sample averages of the hold-out response values that are considered pre-sorted according to a fixed rule.[2] This setup does not come without a substantial cost. Estimation of the improvement score out-of-sample means the effective sample size of our estimate is small and our power commensurately suffers. One can also implement the double bootstrap (see e.g. the comparisons in Chakraborty et al., 2010) herein and that is forthcoming in our software (see **Section 5**).

There is an additional concern. Some bootstrap samples produce null sets for the “lucky subjects” (i.e. $P \cup S = \varnothing$ of **Table 2** or equivalently, all values of the indicator in **Eq. 8** are zero). These are safe to ignore as we are only interested in the distribution of estimates conditional on feasibility of estimation. Empirically, we have noticed that as long as $n > 20$, there are less than 1% of bootstrap samples that exhibit this behavior. Either way, we print out this percentage when using the PTE package and large percentages warn the user that inference is suspect.

3.6 Future Subjects

The implementation of this procedure for future patients is straightforward. Using the RCT data, estimate f to arrive at $\hat{f}$. When a new individual, whose covariates are denoted $\boldsymbol{x}_*$, enters a clinic, our estimated decision rule is calculated by predicting the response under both treatments, then allocating the treatment which corresponds to the better outcome, i.e. $\widehat{d}(\boldsymbol{x}_*)$. This final step is graphically illustrated in the top left of **Figure 1**.

It is important to note that $\widehat{d}(\boldsymbol{x}_*)$ is built with RCT data where treatment was allocated randomly and without regard to the subject covariates. In the example of the first order linear model with treatment interactions, the γ parameters have a causal interpretation—conditional causation based on the values of the moderating covariates. Thus $\widehat{d}(\boldsymbol{x}_*)$ reflects a treatment allocation that causes the response to be higher (or lower). We reiterate that this would not be possible with observational data which would suffer from elaborate confounding relationships between the treatment and subject covariates (see discussion in **Sections 2** and **6.1**).

4 DATA EXAMPLES

We present two simulations in **Sections 4.1** and **4.2** that serve only as illustrations that our methodology both works as purported but degrades in the case of pertinent information that goes missing. We then demonstrate a real clinical setting in **Section 4.3**.

4.1 Simulation With Correct Regression Model

Consider a simulated RCT dataset with one covariate x where the true response function is known:

$$Y = \beta_0 + \beta_1 X + A\left(\gamma_0 + \gamma_1 X\right) + \varepsilon, \tag{11}$$

where ε is mean-centered. We employ $f(x, A)$ as the true response function, $\mathbb{E}[Y|X, A]$. Thus, $d = d^*$, the “optimal” rule in the sense that a practitioner can make optimal allocation decisions (modulo noise) using $d(x) = \mathbb{1}_{\gamma_0+\gamma_1 x>0}$. Consider d_0 to be the **random** allocation procedure (see **Section 3.3**). Note that within the improvement score definition (**Eq. 1**), the notation $\mathbb{E}_X^d[Y]$ is an expectation over the noise $\mathcal{E}$ and the joint distribution of X, A. After taking the expectation over noise, the improvement under the model of **Eq. 11** becomes

$$\begin{aligned}\mu_{I_0} &= \mathbb{E}_X\left[\beta_0 + \beta_1 X + \mathbb{1}_{\gamma_0+\gamma_1 X>0}\left(\gamma_0 + \gamma_1 X\right)\right] - \mathbb{E}_X\left[\beta_0 + \beta_1 X + 0.5\left(\gamma_0 + \gamma_1 X\right)\right] \\ &= \mathbb{E}_X\left[\left(\mathbb{1}_{\gamma_0+\gamma_1 x>0} - 0.5\right)\left(\gamma_0 + \gamma_1 X\right)\right] \\ &= \gamma_0\left(\mathbb{P}\left(\gamma_0 + \gamma_1 X > 0\right) - 0.5\right) + \gamma_1\left(\mathbb{E}_X\left[X\mathbb{1}_{\gamma_0+\gamma_1 x>0}\right] - 0.5\mathbb{E}_X[X]\right).\end{aligned}$$

We further assume $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and we arrive at

$$\mu_{I_0} = \left(\gamma_0 + \gamma_1\mu_X\right)\left(0.5 - \Phi\left(-\frac{\gamma_0}{\gamma_1}\right)\right) + \gamma_1\frac{\sigma_X}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_X^2}\left(-\frac{\gamma_0}{\gamma_1} - \mu_X\right)^2\right).$$

We simulate under a simple scenario to clearly highlight features of our methodology. If $\mu_X = 0, \sigma_X^2 = 1$ and $\gamma_0 = 0$, neither treatment T_1 or T_2 is on average better. However, if $x > 0$, then treatment T_2 is better in expectation by $\gamma_1 \times x$ and analogously if $x < 0$, T_1 is better by $-\gamma_1 \times x$. We then set $\gamma_1 = \sqrt{2\pi}$ to arrive at the round number $\mu_{I_0} = 1$. We set $\beta_0 = 1$ and $\beta_1 = -1$ and let $\mathcal{E}_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. We let the treatment allocation vector A be a random block permutation of size n, balanced between T_1 and T_2. Since there is no PATE, the random and best d_0 procedures (see **Section 3.3**) are the same in value. We then vary $n \in \{100, 200, 500, 1000\}$ to assess convergence for both d_0 procedures and display the results in **Figure 2**.

Convergence to $\mu_{I_0} = 1$ is observed clearly for both procedures but convergence for d_0 best is slower than d_0 rand. This is due to the $\widehat{V}$ being computed with fewer samples: $\bar{y}_{\text{test}}$, which uses all of the available data, vs. $\bar{y}_{P\cup Q}$ or $\bar{y}_{R\cup S}$, which uses only half the available data on average (see **Eqs. 6**, **7**). Also note that upon visual inspection, our bootstrap distributions seem to be normal. Our intuition is that non-normality in this distribution when using the software package warns the user that the inference is suspect.

In this section we assumed knowledge of f and thereby had access to an optimal rule. In the next section we explore convergence when we do not know f but pick an approximate model yielding a non-optimal rule that would still provide clinical utility.

[2] Note also that we do not have the additional non-smoothness created by Q-learning during the maximization step (Chakraborty et al., 2010, Section 2.4).

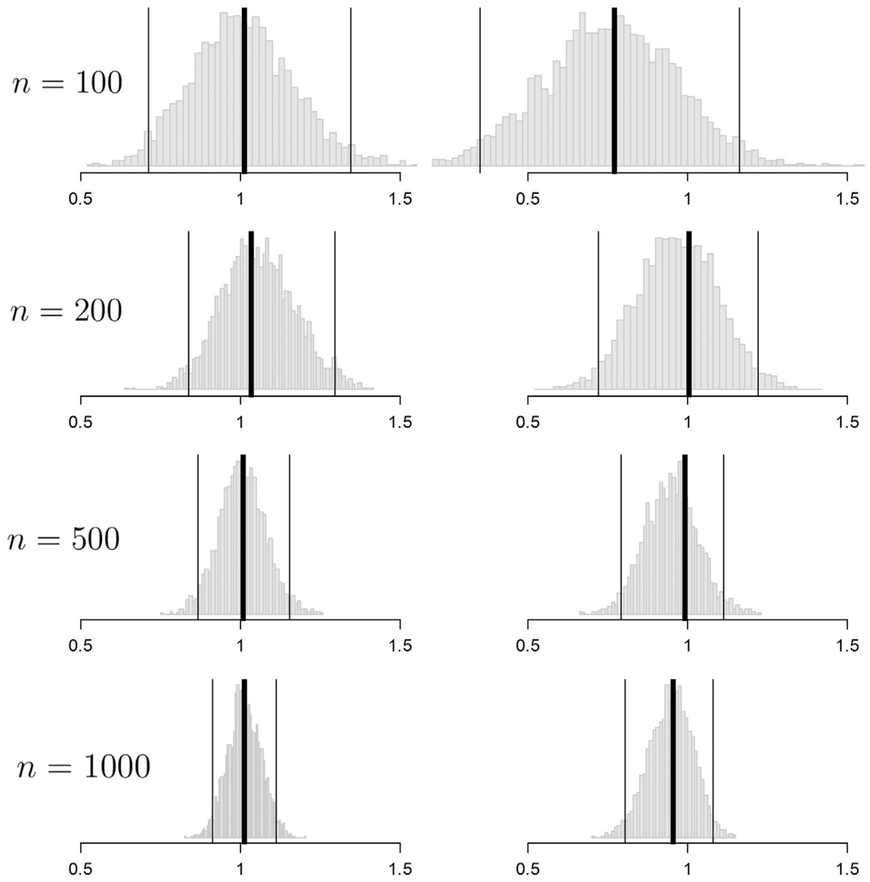


FIGURE 2 | Histograms of the bootstrap samples of the out-of-sample improvement measures for d_0 **random** (left column) and d_0 **best** (right column) for the response model of **Eq. 11** for different values of n. $\hat{I}_0$ is illustrated with a thick black line. The CI$_{\mu_{I_0},95\%}$ computed by the percentile method is illustrated by thin black lines.

4.2 Simulation With an Approximate Regression Model

Consider RCT data with a continuous endpoint where the true response model is

$$Y = \beta_0 + \beta_1 X + \beta_2 U + A\left(\gamma_0 + \gamma_1 X^3 + \gamma_2 U\right) + \varepsilon, \tag{12}$$

where X denotes a covariate recorded in the RCT and U denotes a covariate that is not included in the RCT dataset. The optimal allocation rule d^* is one when $\gamma_0 + \gamma_1 X^3 + \gamma_2 U > 0$ and 0 otherwise. The practitioner, however, does not have access to the information contained in U, the unobserved covariate, and has no way to ascertain the exact relationship between X and the treatment. Consider the default model that is an approximation of the true population response surface,

$$f(X, A) = \beta_0 + \beta_1 X + A\left(\gamma_0 + \gamma_1 X\right), \tag{13}$$

which is different from the true response model due to (a) the misspecification of X (linear instead of cubic) and (b) the absence of covariate U (see **Eq. 3**). This is a realistic scenario; even with infinite data, d^* cannot be located because of both ignorance of the true model form and unmeasured subject characteristics.

To simulate, we set the X's, U's and $\mathcal{E}$'s to be standard normal variables and then set $\beta_0 = 1, \beta_1 = -1, \beta_2 = 0.5, \gamma_0 = 0, \gamma_1 = 1$ and $\gamma_2 = -3$. The X_i's and the U_i's are deliberately made independent of one another so that the observed covariates cannot compensate for the unobserved covariates, making the comparison between the improvement under d^* and d more stark. To find the improvement when the true model's d^* is used to allocate, we simulate under **Eq. 12** and obtain $\mu_{I_0}^* \approx 1.65$ and analogously, to find the improvement under the approximation model's d, we simulate under **Eq. 13** and obtain $\mu_{I_0} \approx 0.79$. Further simulation shows that not observing U is responsible for 85% of this observed drop in improvement performance and employing the linear X in place of the non-linear X^3 is responsible for the remaining 15%. Since $\gamma_0 = 0$ and the seen and unseen covariates are mean-centered, there is no PATE and thus these simulated improvements apply to both the cases where d_0 is **random** and d_0 is **best**.

Figure 3 demonstrates results for $n = \{100, 200, 500, 1000\}$ analogous to **Figure 2**. We observe that the bootstrap confidence intervals contain μ_{I_0} but not $\mu_{I_0}^*$. This is expected; we are not allocating using an estimate of d^*, only an estimate of d.

Convergence toward $\mu_{I_0} = 0.79$ is observed clearly for both procedures and once again the convergence is slower for the best

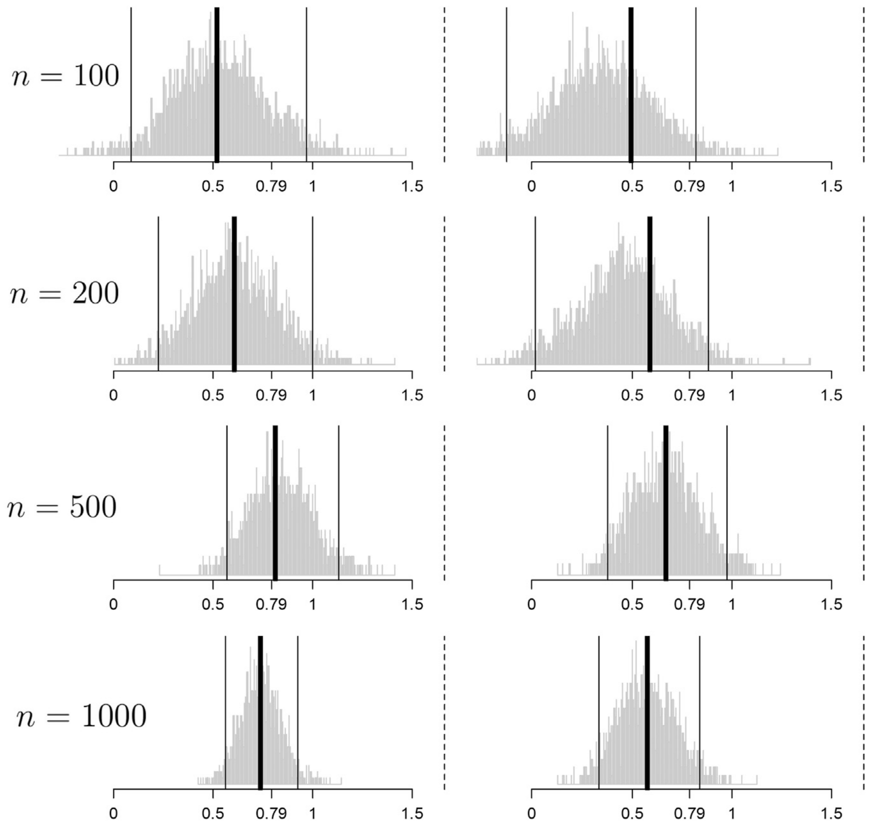


FIGURE 3 | Histograms of the bootstrap samples of the cross-validated improvement measures for d_0 **random** (left column) and d_0 **best** (right column) for the response model of **Eq. 12** for different values of n. $\hat{I}_0$ is illustrated with a thick black line. The $\text{CI}_{\mu_{I_0},95\%}$ computed via the percentile method is illustrated by thin black lines. The true population improvement $\mu_{I_0}^*$ given the optimal rule d^* is illustrated with a dotted black line.

procedure for the same reasons outlined in **Section 4.1**. Note that the coverage illustrated here is far from $\mu_{I_0}^*$, the improvement using the optimal allocation rule. Kallus (2017) presents a coefficient of personalization metric similar to R^2 where a value of 100% represents perfect personalization and 0% represents standard of care. Here, we would fall far short of the 100%.

The point of this section is to illustrate what happens in the real world: the response model is unknown and important measurements are missing and thus any personalized medicine model falls far short of optimal. However, the effort can still yield an improvement that can be clinically significant and useful in practice.

There are many cases where our procedure will not find signal yielding improvement in patient outcomes either because it does not exist or we are underpowered to detect it. For example 1) in cases where there are many variables that are important and a small sample size. Clinical trials are not usually powered to find even single interaction effects, let alone many. The small sample size diminishes power to find the effect, similar to any statistical test. 2) If the true heterogeneity in the functional response cannot be approximated by a linear function. For instance, parabolic or sine functions cannot be represented whatsoever by best fit lines.

In the next section, we use our procedure in RCT data from a real clinical trial where both these limitations apply. The strategy is to approximate the response function using a reasonable model f built from domain knowledge and the variables at hand and hope to find demonstrate a positive, clinically meaningful μ_{I_0} knowing full well it will be much smaller than $\mu_{I_0}^*$.

4.3 Clinical Trial Demonstration

We consider an illustrative example in psychiatry, a field where personalization, called "*precision psychiatry*, promises to be even more transformative than in other fields of medicine" (Fernandes et al., 2017). However, evaluation of predictive models for precision psychiatry has been exceedingly rare, with a recent systematic review identifying 584 studies in which prediction models had been developed, with only 10.4% and 4.6% having conducted proper internal and external validation, respectively (Salazar de Pablo et al., 2020).

We will demonstrate our method (that provides this proper validation) on the randomized comparative trial data of DeRubeis et al. (2005). In this depression study, there were two treatments with very different purported mechanisms of action: cognitive behavioral therapy (T_1) and paroxetine, an antidepressant medication (T_2). After omitting patients who dropped out

TABLE 3 | Baseline characteristics of the subjects in the clinical trial example for the moderating variables employed in our personalization model. These statistics differ slightly from those found in the table of DeRubeis et al. (2005, page 412) as here they are tabulated for subjects only after dropout (n = 154).

Variable	Sample Average or Proportion
Age	40.3 ± 11.3
Chronicity	55.1%
Life stressors	6.6 ± 4.8
Personality disorder	48.1%
Unemployed	14.9%
Married	37.6%

there were $n = 154$ subjects with 28 baseline characteristics measured. Although this dataset did not have explicit patient-level identifying information, using these 28 characteristics could potentially identify some of the patients. Note that this study was funded and begun before clinical trials registration and thus it does not have a clinical trial registration number.

The primary outcome measure y is the continuous Hamilton Rating Scale for Depression (HRSD), a composite score of depression symptoms where lower means less depressed, assessed by a clinician after 16 weeks of treatment. A simple t test revealed that there was no statistically significant HRSD difference between the cognitive behavioral therapy and paroxetine arms, a well-supported finding in the depression literature. Despite the seeming lack of a population mean difference among the two treatments, practitioner intuition and a host of studies suggest that the covariates collected can be used to build a principled personalized model with a significant negative μ_{I_0}. The lack of a difference also suggests that the **random** d_0 is an appropriate baseline comparison. If there were to be a clinically and statistically significant average difference in treatment outcomes between T_1 and T_2, then the **best** d_0 would be appropriate. In this latter case, we have found in our experience (in analyses of other datasets) that proving a personalization improvement is elusive as it is difficult to beat **best**. Even though the **best** d_0 is inappropriate in our context, we still provide its results in this section for illustrative purposes.

We now must specify a model, f. For the purpose of illustration, we employ a linear model with first-order interactions with the treatment (as in **Eq. 4**). Which of the 28 variables should be included in the model? Clinical experience and theory should suggest both prognostic (main effect) and moderator (treatment interaction) variables (Cohen and DeRubeis, 2018). We should not use variables selected using methods performed on this RCT data such as the variables found in DeRubeis et al. (2014), **Table 3**. Such a procedure would constitute data snooping and it will invalidate the inference provided by our method. The degree of invalidation is not currently known and is much needed to be researched.

Of the characteristics measured in this RCT data, previous researchers have found significant treatment moderation in age and chronicity (Cuijpers et al., 2012), early life trauma (Nemeroff et al., 2003) (which we approximate using a life stressor metric), presence of personality disorder (Bagby et al., 2008), employment status and marital status (Fournier et al., 2009) but almost remarkably baseline severity of the depression does not moderate (Weitz et al., 2015) and baseline severity is frequently the most important covariate in response models (e.g. Kapelner and Krieger, 2020, **Figure 1B**). We include these $p = 6$ variables as moderators and as main effects[3] and statistics of their baseline characteristics are found in **Table 3**.

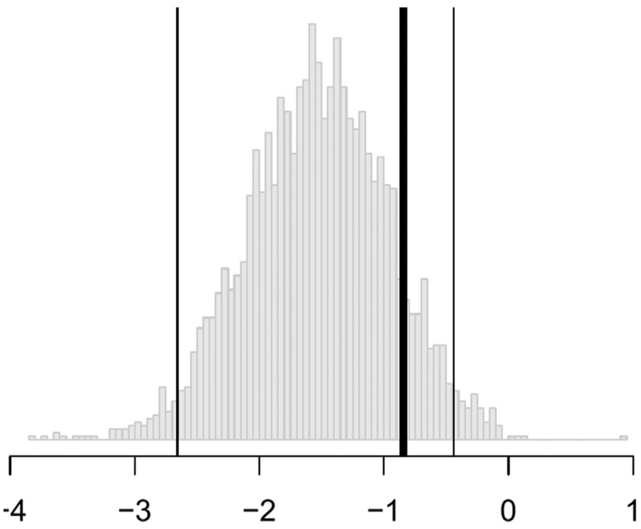


FIGURE 4 | Histograms of the bootstrap samples of $\tilde{I}_{\text{Rand}}$ i.e. for the randomd_0 business-as-usual allocation procedure. The thick black line is the best estimate of $\hat{I}_0$, the thin black lines are the confidence interval computed via the percentile method. More negative values are "better" as improvement is defined as lowering the HSRD composite score corresponding to a patient being less depressed.

The output from 3,000 bootstrap samples are shown in **Figure 4**. From these results, we anticipate that a new subject allocated using our personalization model will be less depressed on average by 0.84 HRSD units with a 95% confidence interval of [0.441, 2.657] compared to that same subject being allocated randomly to cognitive behavioral therapy or paroxetine. We can easily reject the null hypothesis that personalized allocation over **random** is no better for the new subject (p value = 0.001).

In short, the results are statistically significant, but the estimated improvement may not be clinically significant. According to the criterion set out by the National Institute for Health and Care Excellence, three points on the HRSD is considered clinically important. Nevertheless, this personalization scheme can be implemented in practice with new patients for a modest improvement in patient outcome at little cost.

[3] Although this is standard linear modeling practice, it is not absolutely essential in our methodology, where our goal is neither inference for the variables nor prediction of the endpoint.

5 THE PTE PACKAGE

5.1 Estimation and Inference for Continuous Outcomes

The package comes with two example datasets. The first is the continuous data example. Below we load the library and the data whose required form is 1) a vector of length n for the responses, $\boldsymbol{y}$ and 2) a matrix $\boldsymbol{X}$ of dimension $n \times (p + 1)$ where one column is named "treatment" and the other p are appropriate names of the covariates.

```
R> library(PTE); library(dplyr)
R> data(continuous_example)
R> X = continuous_example$X
R> y = continuous_example$y
R> continuous_example$X %>% sample_n(5)
# A tibble: 5 x 6
  treatment     x1     x2     x3     x4         x5
      <dbl> <fctr> <fctr> <fctr> <fctr>      <dbl>
1         1     NO    OFF    YES MEDIUM  1.3009448
2         0    YES    OFF    YES MEDIUM -0.5483983
3         0     NO    OFF    YES    LOW  0.3762733
4         1     NO    OFF    YES MEDIUM -1.1648459
5         1    YES     ON    YES   HIGH -0.8566221
> round(head(continuous_example$y), 3)
[1] -0.746 -1.359  0.020  0.632 -0.823 -2.508
```

The endpoint $\boldsymbol{y}$ is continuous and the RCT data has a binary treatment vector appropriately named (this is required) and five covariates, four of which are factors and one is continuous. We can run the estimation for the improvement score detailed in **Section 3.4.1** and the inference of **Section 3.5** by running the following code:

```
R> options(mc.cores = 4)
R> pte_results = PTE_bootstrap_inference(X, y,
     B = 1000, run_bca_bootstrap = TRUE)
R> plot(pte_results)
```

Here, 1,000 bootstrap samples were run on four cores in parallel to minimize runtime. The f model defaults to a linear model where all variables included are interacted with the treatment and fit with least squares. Below are the results.

```
R> pte_results
    I_random observed_est = 0.077,  pctile p-val = 0.021,
      95% CI's: basic = [-0.237, 0.143], pctile = [0.011, 0.391],
        BCa = [-0.085, 0.145],
    I_best observed_est = 0.065,  pctile p-val = 0.089,
      95% CI's: basic = [-0.198, 0.2], pctile = [-0.071, 0.328],
        BCa = [-0.146, 0.198]
```

Note how the three bootstrap methods are different from another. The percentile method barely includes the actual observed statistic for the random comparison (see discussion in **Section 3.5**). The software also plots the $\tilde{I}$'s in a histogram (unshown).

To demonstrate the flexibility of the software, consider the case where the user wishes to use x_1, x_2, x_3, x_4 as main effects and x_5 as the sole treatment moderator. And further, the user wishes to estimate the model parameters using the ridge penalty instead of OLS. Note that this is an elaborate model that would be difficult to justify in practice and it is only shown here as an illustration of the customizability of the software. Below is the code used to test this approach to personalization.

```
R> library(glmnet)
R> pte_results = PTE_bootstrap_inference(X, y, B = 1000,
  personalized_model_build_function = function(Xytrain){
    Xytrain_mm = model.matrix(˜ . - y + x5 * treatment, Xytrain)
    cv.glmnet(Xytrain_mm, Xytrain[, ncol(Xytrain)], alpha = 0)
  },
  predict_function = function(mod, Xyleftout){
    Xyleftout$censored = NULL
    Xyleftout_mm = model.matrix(˜ . + x5 * treatment, Xyleftout)
    predict(mod, Xyleftout_mm)
  })
```

Here, the user passes in a custom function that builds the ridge model to the argument `personalized_model_build_function`. The specification for ridge employed here uses the package `glmnet` (Friedman et al., 2010) that picks the optimal ridge penalty hyperparameter automatically. Unfortunately, there is added complexity: the `glmnet` package does not accept formula objects and thus model matrices are generated both upon model construction and during prediction. This is the reason why a custom function is also passed in via the argument `predict_function` which wraps the default `glmnet` predict function by passing in the model matrix.

5.2 Estimation and Inference for Binary Outcomes

In order to demonstrate our software for the incidence outcome, we use the previous data but threshold its response arbitrarily at its 75th percentile to create a mock binary response (for illustration purposes only).

```
R> y = ifelse(y > quantile(y, 0.75), 1, 0)
```

We then fit a linear logistic model using all variables as fixed effects and interaction effects with the treatment. As discussed in **Section 3.4.2**, there are three improvement metrics for incidence outcomes. The default is the odds ratio. The following code fits the model and performs the inference.

```
R> pte_results = PTE_bootstrap_inference(X, y, B = 1000,
     regression_type = "incidence", run_bca_bootstrap = TRUE)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Note that the response type incidence has to be explicitly made known otherwise the default would be to assume the endpoint is continuous and perform regression. Below are the results.

```
R> pte_results
    I_random observed_est = 1.155,  pctile p-val = 0.104,
      95% CI's: basic = [0.317, 1.475], pctile = [0.836, 1.994],
        BCa = [0.497, 1.503],
    I_best observed_est = 1.04,  pctile p-val = 0.323,
      95% CI's: basic = [0.329, 1.431], pctile = [0.649, 1.751],
        BCa = [0.52, 1.528]
```

The p value is automatically calculated for $H_0 : \mu_{I_0} < 1$ (i.e. the odds of improvement is better in d_0 than d). Other tests can be specified by changing the `H_0_mu_equals` argument. Here, the test failed to reject H_0. Information is lost when a continuous metric is coerced to be binary. If the user wished to define improvement via the risk ratio (or straight probability difference), an argument would be added to the above, `incidence_metric = "risk_ratio"` (or `"probability_difference"`).

5.3 Estimation and Inference for Survival Outcomes

Our package also comes with a mock RCT dataset with a survival outcome. In addition to the required input data $\boldsymbol{y}$, $\boldsymbol{X}$ described in **Section 5.1**, we now also require a binary vector $\boldsymbol{c}$ also of length n where a value $c_i = 1$ denotes that the ith subject's y_i is a censored value. Below, we load the data.

```
R> data(survival_example)
R> X = survival_example$X
R> y = survival_example$y
R> censored = survival_example$censored
```

There are four covariates, one factor and three continuous. We can run the estimation for the improvement score detailed in **Section 3.4.3** and inference for the true improvement by running the following code.

```
R> pte_results = PTE_bootstrap_inference(X, y, censored = censored,
  B = 1000, run_bca_bootstrap = TRUE, regression_type = "survival")
```

The syntax is the same as the above two examples except here we pass in the binary $\boldsymbol{c}$ vector separately and declare that the endpoint type is survival. Again by default all covariates are included as main effects and interactions with the treatment in a linear Weibull model.

In the default implementation for the survival outcome, improvement is defined as median survival difference of personalization vs. standard of care. The median difference can be changed via the user passing in a new function with the `difference_function` argument. The median difference results are below.

```
R> pte_results
    I_random observed_est = 0.148,  p val = 0.017,
      95% CI's: basic = [0.011, 0.296], pctile = [0, 0.285],
        BCa = [-0.001, 0.274],
    I_best observed_est = -0.041,  p val = 0.669,
      95% CI's: basic = [-0.134, 0.082], pctile = [-0.164, 0.052],
        BCa = [-0.195, 0.013]
```

It seems that the personalized medicine model increases median survival by 0.148 vs. d_0 being the **random** allocation of the two treatments. If survival was measured in years (the typical unit), this would be about 2 months. However, it cannot beat the d_0 being the **best** of the two treatments. Remember, this is a much more difficult improvement metric to estimate as we are really comparing two cells in **Table 2** to another two cells, one of which is shared. Thus the sample size is low and power suffers. This difficulty is further compounded in the survival case because censored observations add little information.

6 DISCUSSION

We have provided a methodology to test the effectiveness of personalized medicine models. Our approach combines RCT data with a statistical model f of the response for estimating improved outcomes under different treatment allocation protocols. Using the non-parametric bootstrap and cross-validation, we are able to provide confidence bounds for the improvement and hypothesis tests for whether the personalization performs better compared to a business-as-usual procedure. We demonstrate the method's performance on simulated data and on data from a clinical trial on depression. We also present our statistical methods in an open source software package in `R` named `PTE` which is available on `CRAN`. Our package can be used to evaluate personalization models generally e.g. in heart disease, cancer, etc. and even outside of medicine e.g. in Economics and Sociology.

6.1 Limitations and Future Directions

Our method and corresponding software have been developed for a particular kind of RCT design. The RCT must have two arms and one endpoint (continuous, incidence or survival). An extension to more than two treatment arms is trivial as **Eq. 2** is already defined generally. Implementing extensions to longitudinal or panel data are simple within the scope described herein. And extending the methodology to count endpoints would also be simple.

Although we agree that a "once and for all" treatment strategy [may be] suboptimal due to its inflexibility" (Zhao et al., 2015), this one-stage treatment situation is still common in the literature and the real world and this is the setting we chose to research. We consider an extended implementation for dynamic treatment regimes on multi-stage experiments fruitful future work. Consider being provided with RCT data from sequential multiple assignment randomized trials ("SMARTs," Murphy, 2005b) and an a priori response model f. The estimate of $\widehat{V}[\widehat{d}]$ (**Eq. 5**) can be updated for a SMART with k stages (Chakraborty and Murphy, 2014) where our **Table 2** is a summary for only a single stage. In a SMART with k stages, the matrix becomes a hypercube of dimension k. Thus, the average of diagonal entries in the multi-dimensional matrix is the generalization of the estimate of $\widehat{V}[\widehat{d}]$ found in **Eq. 6**. Many of the models for dynamic treatment regimes found in Chakraborty and Moodie (2013) can then be incorporated into our methodology as d, and we may be able to provide many of these models with valid statistical inference. Other statistics computed from this multi-dimensional matrix may be generalized as well.

Our choices of d_0 explored herein were limited to the **random** or the **best** procedures (see **Section 3.3**). There may be other business-as-usual allocation procedures to use here that make for more realistic baseline comparisons. For instance, one can modify **best** to only use the better treatment if a two-sample t-test rejects at prespecified Type I error level and otherwise default to **random**. One can further set d_0 to be a regression model or a physician's decision tree model and then use our framework to pit two personalized medicine models against each other.

It might also be useful to consider how to extend our methodology to observational data. The literature reviewed in **Section 2** generally does not require RCT data but "only" a model that accurately captures selection into treatments e.g. if "the [electronic medical record] contained all the patient information used by a doctor to prescribe treatment up to the vagaries and idiosyncrasies of individual doctors or hospitals" (Kallus, 2017, Section 1). This may be a very demanding requirement in practice. In this paper, we do not even require valid estimates of the true population response surface. In an observational study one would need that selection model to be correct and/or a correct model of the way in which subjects and treatments were paired (Freedman and Berk, 2008). Although assuming one has a model that captures selection, it would be fairly straightforward to update the estimators of **Section 3.4** to inverse weight by the probability of treatment condition (the "IPWE") making inference possible for observational data (Zhang et al., 2012b; Chakraborty and Murphy, 2014; Kallus, 2017).

Another extension would be to drop the requirement of specifying the model f whose specification is a tremendous constraint in practice: what if the practitioner cannot construct a suitable f using domain knowledge and past research? It is tempting to use a machine learning model that will both specify the structure of f and provide parameter estimates within e.g. Kallus's personalization forests (Kallus, 2017) or convolutional neural networks (LeCun and Bengio, 1998) if the raw subject-level data had images. We believe the bootstrap of **Section 3.5** will withstand such a machination but are awaiting a rigorous proof. Is there a solution in the interim? As suggested as early as Cox (1975), we can always pre-split the data in two where the first piece can be used to specify f and the second piece can be injected into our procedure. The cost is less data for estimation and thus, less power available to prove that the personalization is effective.

If we do not split, all the data is to be used and there are three scenarios that pose different technical problems. Under one scenario, a researcher is able to specify a suite of possible models before looking at the data. The full suite can be viewed as comprising a single procedure for which nonparametric bootstrap procedures may in principle provide simultaneous confidence intervals (Buja and Rolke, 2014). Under the other two scenarios, models are developed inductively from the data. This problem is more acute for instance in Davies (2015) where high-dimensional genomic data is incorporated for personalization (e.g. where there are many more SNPs than patients in the RCT). If it is possible to specify exactly how the model search is undertaken (e.g., using the lasso), some forms of statistical inference may be feasible. This is currently an active research area; for instance, Lockhart et al. (2014) and Lee et al. (2016) develop a significance test for the lasso and there is even some evidence to suggest that the double-peeking is not as problematic as the community has assumed (Zhao et al., 2020).

Our method's generalizability to future patients is also in question as our validation was done within the patients of a RCT. The population of future patients is likely not the same as the population of patients in the RCT. Future patients will likely have wider distributions of the p covariates as typical RCTs feature strict inclusion criteria sometimes targeting high risk patients for higher outcome event rates. A good discussion of these issues is found in Rosenberger and Lachin (2016), Chapter 6. The practitioner will have to draw on experience and employ their best judgment to decide if the estimates our methodology provides will generalize.

And of course, the method herein only evaluates if a personalization scheme works on average over an entire population. "Personalized medicine" eponymously refers to personalization for an individual. Ironically, that is not the goal herein, but we do acknowledge that estimates and inference at an individual level coupled to valid inference for the improvement score is much-needed. This is not without difficulty as clinical trials are typically not powered to examine subgroup effects. A particularly alarming observation is made by Cuijpers et al. (2012), page 7, "if we want to have sufficient statistical power to find clinically relevant differential effect sizes of 0.5, we would need …. about 23,000 patients".

DATA AVAILABILITY STATEMENT

The datasets presented in this article in **Section 4.3** is not readily available because the clinical data cannot be anonymized sufficiently according to the IRB guidelines of the institutions where the study was performed. For more information, contact the authors of the study.

AUTHOR CONTRIBUTIONS

All authors were responsible for development of methodology and drafting the manuscript. Kapelner and Bleich did the data analysis of **Section 4**.

FUNDING

Adam Kapelner acknowledges support for this project provided by a PSC-CUNY Award, jointly funded by The Professional Staff Congress and The City University of New York. ZC and RD acknowledge support from MQ: Transforming mental health (MQDS16/72).

ACKNOWLEDGMENTS

We would like to thank our three reviewers and the editor whose input greatly improved this work. We would like to thank Abba Krieger, Larry Brown, Bibhas Chakraborty, Andreas Buja, Ed George, Dan McCarthy, Linda Zhao and Wei-Yin Loh for helpful discussions and comments. Earlier versions of this manuscript were released as a preprint on arXiv (Kapelner et al., 2014).

REFERENCES

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Bagby, R. M., Quilty, L. C., Segal, Z. V., McBride, C. C., Kennedy, S. H., and Costa, P. T. (2008). Personality and Differential Treatment Response in Major Depression: a Randomized Controlled Trial Comparing Cognitive-Behavioural Therapy and Pharmacotherapy. *Can. J. Psychiatry* 53, 361–370. doi:10.1177/070674370805300605

Barrett, J. K., Henderson, R., and Rosthøj, S. (2014). Doubly Robust Estimation of Optimal Dynamic Treatment Regimes. *Stat. Biosci.* 6, 244–260. doi:10.1007/s12561-013-9097-6

Berger, J. O., Wang, X., and Shen, L. (2014). A Bayesian Approach to Subgroup Identification. *J. Biopharm. Stat.* 24, 110–129. doi:10.1080/10543406.2013.856026

Berk, R. A., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013b). Valid Post-selection Inference. *Ann. Stat.* 41, 802–837. doi:10.1214/12-aos1077

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., et al. (2014). Misspecified Mean Function Regression. *Sociological Methods Res.* 43, 422–451. doi:10.1177/0049124114526375

Berk, R., Olson, M., Buja, A., and Ouss, A. (2020). Using Recursive Partitioning to Find and Estimate Heterogenous Treatment Effects in Randomized Clinical Trials. *J. Exp. Criminol.*, 1–20. doi:10.1007/s11292-019-09410-0

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013a). Covariance Adjustments for the Analysis of Randomized Field Experiments. *Eval. Rev.* 37, 170–196. doi:10.1177/0193841x13513025

Bernard, C. (1957). *An Introduction to the Study of Experimental Medicine*. New York, NY: Dover Publications.

Box, G. E. P., and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York, NY: Wiley.

Brinkley, J., Tsiatis, A., and Anstrom, K. J. (2010). A Generalized Estimator of the Attributable Benefit of an Optimal Treatment Regime. *Biometrics* 66, 512–522. doi:10.1111/j.1541-0420.2009.01282.x

Buja, A., and Rolke, W. (2014). *Calibration for Simultaneity: (Re)sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data*. University of Pennsylvania working paper.

Byar, D. P. (1985). Assessing Apparent Treatment-Covariate Interactions in Randomized Clinical Trials. *Statist. Med.* 4, 255–263. doi:10.1002/sim.4780040304

Byar, D. P., and Corle, D. K. (1977). Selecting Optimal Treatment in Clinical Trials Using Covariate Information. *J. chronic Dis.* 30, 445–459. doi:10.1016/0021-9681(77)90037-6

Chakraborty, B., Laber, E. B., and Zhao, Y.-Q. (2014). Inference about the Expected Performance of a Data-Driven Dynamic Treatment Regime. *Clin. Trials* 11, 408–417. doi:10.1177/1740774514537727

Chakraborty, B., and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*. New York, NY: Springer.

Chakraborty, B., and Murphy, S. A. (2014). Dynamic Treatment Regimes. *Annu. Rev. Stat. Appl.* 1, 447–464. doi:10.1146/annurev-statistics-022513-115553

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes. *Stat. Methods Med. Res.* 19, 317–343. doi:10.1177/0962280209105013

Cohen, Z. D., and DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annu. Rev. Clin. Psychol.* 14, 209–236. doi:10.1146/annurev-clinpsy-050817-084746

Collins, L. M., Murphy, S. A., and Bierman, K. L. (2004). A Conceptual Framework for Adaptive Preventive Interventions. *Prev. Sci.* 5, 185–196. doi:10.1023/b:prev.0000037641.26017.00

Cox, D. R. (1975). A Note on Data-Splitting for the Evaluation of Significance Levels. *Biometrika* 62, 441–444. doi:10.1093/biomet/62.2.441

Cox, D. R. (1958). *Planning of Experiments*. New York, NY: Wiley.

Cuijpers, P., Reynolds, C. F., III, Donker, T., Li, J., Andersson, G., and Beekman, A. (2012). Personalized Treatment of Adult Depression: Medication, Psychotherapy, or Both? a Systematic Review. *Depress. Anxiety* 29, 855–864. doi:10.1002/da.21985

Davies, K. (2015). *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine*. New York, NY: Simon & Schuster.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. 1st Edn. Cambridge, UK: Cambridge University Press.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *Am. Psychol.* 34, 571–582. doi:10.1037/0003-066x.34.7.571

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L., and Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individual Treatment Recommendations. *A. Demonstration. PLoS One* 9, e83875. doi:10.1371/journal.pone.0083875

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive Therapy vs Medications in the Treatment of Moderate to Severe Depression. *Arch. Gen. Psychiatry* 62, 409–416. doi:10.1001/archpsyc.62.4.409

DiCiccio, T. J., and Efron, B. (1996). Bootstrap Confidence Intervals. *Stat. Sci.* 11, 189–212. doi:10.1214/ss/1032280214

Dusseldorp, E., Doove, L., and Mechelen, I. v. (2016). QUINT: An R Package for the Identification of Subgroups of Clients Who Differ in Which Treatment Alternative Is Best for Them. *Behav. Res.* 48, 650–663. doi:10.3758/s13428-015-0594-z

Dusseldorp, E., and Van Mechelen, I. (2014). Qualitative Interaction Trees: a Tool to Identify Qualitative Treatment-Subgroup Interactions. *Statist. Med.* 33, 219–237. doi:10.1002/sim.5933

Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* 82, 171–185. doi:10.1080/01621459.1987.10478410

Efron, B., and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Evans, W. E., and Relling, M. V. (2004). Moving towards Individualized Medicine with Pharmacogenomics. *Nature* 429, 464–468. doi:10.1038/nature02626

Faraway, J. J. (2016). Does Data Splitting Improve Prediction?. *Stat. Comput.* 26, 49–60. doi:10.1007/s11222-014-9522-9

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The New Field of "precision Psychiatry". *BMC Med.* 15, 1–7. doi:10.1186/s12916-017-0849-x

Foster, J. C. (2013). Subgroup Identification and Variable Selection from Randomized Clinical Trial Data. PhD thesis. Ann Arbor, MI: The University of Michigan.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., and Gallop, R. (2009). Prediction of Response to Medication and Cognitive Therapy in the Treatment of Moderate to Severe Depression. *J. consulting Clin. Psychol.* 77, 775–787. doi:10.1037/a0015401

Freedman, D. A., and Berk, R. A. (2008). Weighting Regressions by Propensity Scores. *Eval. Rev.* 32, 392–409. doi:10.1177/0193841x08317586

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Gail, M., and Simon, R. (1985). Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics* 41, 361–372. doi:10.2307/2530862

Goldberg, Y., and Kosorok, M. R. (2012). Q-learning with Censored Data. *Ann. Stat.* 40, 529. doi:10.1214/12-aos968

Gunter, L., Chernick, M., and Sun, J. (2011a). A Simple Method for Variable Selection in Regression with Respect to Treatment Selection. *Pakistan J. Stat. Operations Res.* 7, 363–380. doi:10.18187/pjsor.v7i2-sp.311

Gunter, L., Zhu, J., and Murphy, S. (2011b). Variable Selection for Qualitative Interactions in Personalized Medicine while Controlling the Family-wise Error Rate. *J. Biopharm. Stat.* 21, 1063–1078. doi:10.1080/10543406.2011.608052

Hastie, T., Tibshirani, R., and Friedman, J. H. (2013). *The Elements of Statistical Learning*. 10th Edn. Berlin, Germany: Springer Science.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models. *Biostatistics* 21, 50–68. doi:10.1093/biostatistics/kxy028

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for Optimal Dynamic Treatment Regimes. *Biometrics* 66, 1192–1201. doi:10.1111/j.1541-0420.2009.01368.x

Hood, L., and Friend, S. H. (2011). Predictive, Personalized, Preventive, Participatory (P4) Cancer Medicine. *Nat. Rev. Clin. Oncol.* 8, 184–187. doi:10.1038/nrclinonc.2010.227

Horowitz, J. L. (2001). *Chapter 52 - the Bootstrap (Elsevier), of Handbook of Econometrics*, Vol. 5, 3159–3228. doi:10.1016/s1573-4412(01)05005-xThe Bootstrap

Hosmer, D. W., and Lemeshow, S. (1999). *Applied Survival Analysis: Time-To-Event*. Hoboken, NJ: Wiley.

Imai, K., and Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Ann. Appl. Stat.* 7, 443–470. doi:10.1214/12-aoas593

Kallus, N. (2017). Recursive Partitioning for Personalization Using Observational Data. *Proc. 34th Int. Conf. on Machine Learning* 70, 1789–1798.

Kang, C., Janes, H., and Huang, Y. (2014). Combining Biomarkers to Optimize Patient Treatment Recommendations. *Biom* 70, 695–707. doi:10.1111/biom.12191

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J., and Berk, R. (2014). Inference for the Effectiveness of Personalized Medicine with Software. Available at: http://arxiv.org/abs/1404.7844.

Kapelner, A., and Krieger, A. (2020). A Matching Procedure for Sequential Experiments that Iteratively Learns Which Covariates Improve Power. Available at: http://arxiv.org/abs/2010.05980.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014). Set-valued Dynamic Treatment Regimes for Competing Outcomes. *Biom* 70, 53–61. doi:10.1111/biom.12132

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *Am. Econ. Rev.* 76, 604–620.

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., et al. (2018). Identification of Predicted Individual Treatment Effects in Randomized Clinical Trials. *Stat. Methods Med. Res.* 27, 142–157. doi:10.1177/0962280215623981

LeCun, Y., and Bengio, Y. (1998). "Convolutional Networks for Images, Speech, and Time Series," in *The Handbook of Brain Theory and Neural Networks*. Editor M. A. Arbib (Cambridge: MIT Press), 255–258.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact Post-selection Inference, with Application to the Lasso. *Ann. Stat.* 44, 907–927. doi:10.1214/15-aos1371

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A Significance Test for the Lasso. *Ann. Stat.* 42, 413–468. doi:10.1214/13-aos1175

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable Selection for Optimal Treatment Decision. *Stat. Methods Med. Res.* 22, 493–504. doi:10.1177/0962280211428383

Ma, J., Stingo, F. C., and Hobbs, B. P. (2019). Bayesian Personalized Treatment Selection Strategies that Integrate Predictive with Prognostic Determinants. *Biometrical J.* 61, 902–917. doi:10.1002/bimj.201700323

McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., et al. (2013). Toward a Neuroimaging Treatment Selection Biomarker for Major Depressive Disorder. *JAMA psychiatry* 70, 821–829. doi:10.1001/jamapsychiatry.2013.143

McKeague, I. W., and Qian, M. (2014). Estimation of Treatment Policies Based on Functional Predictors. *Stat. Sin* 24, 1461–1485. doi:10.5705/ss.2012.196

Moodie, E. E. M., and Richardson, T. S. (2009). Estimating Optimal Dynamic Regimes: Correcting Bias under the Null: [Optimal Dynamic Regimes: Bias Correction]. *Scand. Stat. Theor. Appl* 37, 126–146. doi:10.1111/j.1467-9469.2009.00661.x

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007). Demystifying Optimal Dynamic Treatment Regimes. *Biometrics* 63, 447–455. doi:10.1111/j.1541-0420.2006.00686.x

Murphy, S. A. (2005a). A Generalization Error for Q-Learning. *J. Mach Learn. Res.* 6, 1073–1097.

Murphy, S. A. (2005b). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statist. Med.* 24, 1455–1481. doi:10.1002/sim.2022

Murphy, S. A. (2003). Optimal Dynamic Treatment Regimes. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 65, 331–355. doi:10.1111/1467-9868.00389

Nemeroff, C. B., Heim, C. M., Thase, M. E., Klein, D. N., Rush, A. J., Schatzberg, A. F., et al. (2003). Differential Responses to Psychotherapy versus Pharmacotherapy in Patients with Chronic Forms of Major Depression and Childhood Trauma. *Proc. Natl. Acad. Sci.* 100, 14293–14296. doi:10.1073/pnas.2336126100

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content. *The Int. J. biostatistics* 6, 1–47. doi:10.2202/1557-4679.1200

Pan, G., and Wolfe, D. A. (1997). Test for Qualitative Interaction of Clinical Significance. *Statist. Med.* 16, 1645–1652. doi:10.1002/(sici)1097-0258(19970730)16:14<1645::aid-sim596>3.0.co;2-g

Paul, G. L. (1967). Strategy of Outcome Research in Psychotherapy. *J. consulting Psychol.* 31, 109–118. doi:10.1037/h0024436

Qian, M., and Murphy, S. A. (2011). Performance Guarantees for Individualized Treatment Rules. *Ann. Stat.* 39, 1180–1210. doi:10.1214/10-aos864

Rice, J. A. (1994). *Mathematical Statistics and Data Analysis*. 2nd Edn. Belmont, CA: Duxbury Press.

Robins, J. M. (2004). "Optimal Structural Nested Models for Optimal Sequential Decisions," in Proceedings of the Second Seattle Symposium on Biostatistics. (Springer), 189–326. doi:10.1007/978-1-4419-9076-1_11

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and Extrapolation of Optimal Treatment and Testing Strategies. *Statist. Med.* 27, 4678–4721. doi:10.1002/sim.3301

Rolling, C. A., and Yang, Y. (2014). Model Selection for Estimating Treatment Effects. *J. R. Stat. Soc. B* 76, 749–769. doi:10.1111/rssb.12043

Rosenbaum, P. R. (2002). *Observational Studies*. New York, NY: Springer.

Rosenberger, W. F., and Lachin, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*. 2nd Edn. Hoboken, NJ: John Wiley & Sons.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J. Educ. Psychol.* 66, 688–701. doi:10.1037/h0037350

Rubin, D. B., and van der Laan, M. J. (2012). Statistical Issues and Limitations in Personalized Medicine Research with Clinical Trials. *Int. J. biostatistics* 8, 1–20. doi:10.1515/1557-4679.1423

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., et al. (2020). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophrenia Bulletin* 47, 284–297. doi:10.1093/schbul/sbaa120

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Stat. Sci. a Rev. J. Inst. Math. Stat.* 29, 640–661. doi:10.1214/13-sts450

Shao, J. (1994). Bootstrap Sample Size in Nonregular Cases. *Proc. Amer. Math. Soc.* 122, 1251. doi:10.1090/s0002-9939-1994-1227529-8

Shen, Y., and Cai, T. (2016). Identifying Predictive Markers for Personalized Treatment Selection. *Biom* 72, 1017–1025. doi:10.1111/biom.12511

Shuster, J., and van Eys, J. (1983). Interaction between Prognostic Factors and Treatment. *Controlled Clin. trials* 4, 209–214. doi:10.1016/0197-2456(83)90004-1

Silvapulle, M. J. (2001). Tests against Qualitative Interaction: Exact Critical Values and Robust Tests. *Biometrics* 57, 1157–1165. doi:10.1111/j.0006-341x.2001.01157.x

Smith, R. (2012). British Medical Journal Group Blogs. *Stratified, Personalised or Precision Medicine.* Available at: https://blogs.bmj.com/bmj/2012/10/15/richard-smith-stratified-personalised-or-precision-medicine/. Online. (Accessed July 01, 2021).

Su, X., Tsai, C. L., and Wang, H. (2009). Subgroup Analysis via Recursive Partitioning. *J. Machine Learn. Res.* 10, 141–158.

van der Laan, M. J., and Luedtke, A. R. (2015). Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *J. causal inference* 3, 61–95. doi:10.1515/jci-2013-0022

Weitz, E. S., Hollon, S. D., Twisk, J., Van Straten, A., Huibers, M. J. H., David, D., et al. (2015). Baseline Depression Severity as Moderator of Depression Outcomes Between Cognitive Behavioral Therapy vs Pharmacotherapy. *JAMA psychiatry* 72, 1102–1109. doi:10.1001/jamapsychiatry.2015.1516

Weston, A. D., and Hood, L. (2004). Systems Biology, Proteomics, and the Future of Health Care: toward Predictive, Preventative, and Personalized Medicine. *J. Proteome Res.* 3, 179–196. doi:10.1021/pr0499693

Yakovlev, A. Y., Goot, R. E., and Osipova, T. T. (1994). The Choice of Cancer Treatment Based on Covariate Information. *Statist. Med.* 13, 1575–1581. doi:10.1002/sim.4780131508

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat* 1, 103–114. doi:10.1002/sta.411

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* 68, 1010–1018. doi:10.1111/j.1541-0420.2012.01763.x

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions. *Biometrika* 100, 681–694. doi:10.1093/biomet/ast014

Zhao, S., Witten, D., and Shojaie, A. (2020). Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *arXiv.*

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *J. Am. Stat. Assoc.* 110, 583–598. doi:10.1080/01621459.2014.937488

Zhao, Y., and Zeng, D. (2013). Recent Development on Statistical Methods for Personalized Medicine Discovery. *Front. Med.* 7, 102–110. doi:10.1007/s11684-013-0245-7

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). Large-scale Parallel Collaborative Filtering for the Netflix Prize. *Lecture Notes Comput. Sci.* 5034, 337–348. doi:10.1007/978-3-540-68880-8_32

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.

frontiers
in Artificial Intelligence

MINI REVIEW
published: 07 June 2021
doi: 10.3389/frai.2021.627369

Predicting Cervical Cancer Outcomes: Statistics, Images, and Machine Learning

Wei Luo *

Department of Radiation Medicine, University of Kentucky, Lexington, KY, United States

Cervical cancer is a very common and severe disease in women worldwide. Accurate prediction of its clinical outcomes will help adjust or optimize the treatment of cervical cancer and benefit the patients. Statistical models, various types of medical images, and machine learning have been used for outcome prediction and obtained promising results. Compared to conventional statistical models, machine learning has demonstrated advantages in dealing with the complexity in large-scale data and discovering prognostic factors. It has great potential in clinical application and improving cervical cancer management. However, the limitations of prediction studies and prediction models including simplification, insufficient data, overfitting and lack of interpretability, indicate that more work is needed to make clinical outcome prediction more accurate, more reliable, and more practical for clinical use.

Keywords: cervical cancer, clinical outcome prediction, statistical model, machine learning, medical image, radiomics

OPEN ACCESS

Edited by:
Bertram Müller-Myhsok,
Max Planck Institute of Psychiatry (MPI), Germany

Reviewed by:
Shivanand Sharanappa Gornale,
Rani Channamma University, India
Lanchun Lu,
The Ohio State University, United States

****Correspondence:***
Wei Luo
wei.luo@uky.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *09 November 2020*
Accepted: *25 May 2021*
Published: *07 June 2021*

Citation:
Luo W (2021) Predicting Cervical Cancer Outcomes: Statistics, Images, and Machine Learning. Front. Artif. Intell. 4:627369. doi: 10.3389/frai.2021.627369

INTRODUCTION

Cancer treatment is one of the most complicated and challenging tasks in medicine. Although cancer survival rate has been significantly improved for the last decades with the introduction of new drugs, technologies and techniques, there are still uncertainties on the effect of those advances on clinical outcomes. The information of clinical outcomes is critical for the evaluation of treatment effectiveness and optimization of treatment strategies. Clinical outcomes usually are not available until enough clinical data have been accumulated following up a large number of patients for long periods. To know clinical outcomes more quickly so that treatment can be improved or adjusted timely, accurate prediction of clinical outcomes is expected. There are two approaches used for clinical outcome prediction. One is to use radiobiological models including tumor control probability (TCP) model, normal tissue complication probability (NTCP) model, and equivalent uniform dose (EUD). The other is to build statistical models utilizing all the information that is relevant to disease prognosis such as demographics, laboratory tests, images, and dosimetry, to find the relationship between those factors and clinical outcomes. The more data is used, the more accurate the prediction would be. In this regard, artificial intelligence especially machine learning (ML) has a great capacity to process huge and complex data and thus has been used in many areas including medicine. Recently, ML has been introduced into radiation oncology to predict clinical outcomes (Kang et al., 2015; Luo et al., 2020).

Cervical cancer is the third most common cancer and a leading cause to death for women worldwide (Ferlay et al., 2019; Rebecca, 2020). It is one of a few cancers that were first treated with radiation therapy successfully (Mazeron and Gerbaulet, 1998). The treatment of cervical cancer is also one of the most complex and challenging cancer management tasks and may involve all three

TABLE 1 | The reported actual clinical outcomes.

Modality	Study	5-year survival rate			
		I	II	III	IV
RT	Joslin et al. (2001)	94.5	62.6	37.3	
	Kim et al. (1988)	83.2	68.9	30.9	27
	Landoni et al. (1997)	84			
Surgery	Brunschwig. (1968)	77.4	51.6		
	Landoni et al. (1997)	88			
RT + surgery	Landoni et al. (1997)	78			
Chemoradiotherapy	Eifel et al. (2004)	81.8		62.6	

cancer treatment modalities (surgery, chemotherapy, and radiation therapy (RT)) and all radiation therapy techniques (external beam radiation therapy (EBRT), intracavitary/ interstitial brachytherapy (BT), high dose rate (HDR)/low dose rate (LDR) brachytherapy, and permanent seed implant). This paper does not intend to provide a comprehensive review of cervical cancer outcome predictions, but mainly focuses on the prediction results with different methods, the efficacy and limitations of prediction associated with radiation therapy.

Reported Clinical Outcomes

Actual clinical outcomes are directly derived from the results obtained following up patients. Numerus studies have revealed cervical cancer survival rates for different International Federation of Gynecology and Obstetrics (FIGO) stages (I–IV) and different treatment techniques. In the United States, the 5-year survival rates of cervical cancer patients ranged from 17 to 92% with the all-stage rate of 66% according to the American Cancer Society (American Cancer Society, 2020). Surgery, chemotherapy, and radiation therapy are the treatment options for cervical cancer. The clinical outcomes are associated with treatment modalities and FIGO stages. The actual 5-year survival rates have been reported and are summarized in **Table 1** (Brunschwig, 1968; Kim et al., 1988; Landoni et al., 1997; Joslin et al., 2001; Eifel et al., 2004). Severe complications were also reported for 9% of patients with radiation therapy alone (Podczaski et al., 1990) and 20% of patients with chemoradiotherapy (Small et al., 2011). Those reported results were summaries of previous clinical data, but not predictions of clinical outcomes. Mathematical models can establish quantitative relationship between disease-related factors and outcomes and thus predict clinical outcomes based on identified prognostic factors or predictors.

Outcome Prediction Using Conventional Statistical Models

Statistical models have been commonly used to analyze clinical results and also for cervical cancer outcome prediction. To make accurate and meaningful predictions, identifying predictors is critical. The linear regression model was introduced to analyze the correlation between the mRNA expression of Homeobox (HOX) genes in cervical cancer and overall survival. It was found that high HOX expression significantly reduced the overall survival in a cohort of 308 cervical cancer patients and the difference in 15-years survival rate between high and low expression was up to around 25% (Eoh et al., 2017). The Cox proportional hazards regression model (CPHR) uses hazard ratio to distinguish different groups and evaluates the relative importance of predictors. Tumor diameter has been identified as an important predictor based on CPHR (Landoni et al., 1997). A retrospective study reviewed the hospital records of 4,490 patients with stage IB, IIA, or IIB cervical cancer at a single institution, and found that the disease-specific survival (DSS) rate and pelvic disease control (PDC) rate had strong correlations with tumor diameter, FIGO stage, histological subtype, and clinical node status. Overall, the 5-year DSS for tumor diameter ≤4, 4.1–6, and >6 cm, was 85, 69, and 52%, respectively; for stages I, IIA, and IIB disease DSS was 80, 68, and 59%, respectively, and the PDC rates were 90, 87, and 82%, respectively (Eifel et al., 2009).

Outcome Prediction Using Image Analysis

Radiation therapy heavily relies on medical imaging. Various three-dimensional (3D) imaging techniques such as computerized tomography (CT), nuclear magnetic resonance imaging (MRI) and positron emission tomography (PET) have been widely used for cervical cancer diagnosis and treatment. Those images may also contain the information about clinical outcomes. By analyzing the F-18 fluorodeoxyglucose (FDG) pretreatment images of 248 cervical cancer patients staged from IA2 to IVB and using CPHR, a study reported that the maximal standardized uptake value (SUVmax) that quantifies cervical tumor uptake of FDG is associated with treatment response and prognosis in cervical cancer patients and gave better outcome prediction than lymph node status, stage, or tumor volume (Kidd et al., 2007). The results showed that the overall survival rate at 5 years was 95% for patients with an SUVmax ≤5.2, 70% for patients with an SUVmax from >5.2 to ≤13.3, and 44% for patients with an SUVmax >13.3.

Recently, radiomics has been introduced as a powerful tool to extract huge and complex image features from PET/CT and MRI images for prediction of cervical cancer clinical outcome. It was reported that radiomics features could contribute to prognoses in cervical cancer (Lucia et al., 2018). Using CPHR, two textural features, Grey Level Non Uniformity gray-level run-length matrix (GLRLM) in PET and Entropy gray-level co-occurrence matrix GLCM in ADC maps from DWI MRI, were identified as independent prognostic factors. They were significantly stronger correlated with prognoses than clinical parameters, with an accuracy of 94% for predicting recurrence and 100% for predicting lack of loco-regional control compared with ~50–60% accuracy with clinical parameters. It was also found that the high gray-level run emphasis (HGRE) derived from GLRLM and used to measure high SUV distribution can serve as a predictor (Chen et al., 2018a). This study included 142 cervical cancer patients who had took 18F-FDG PET/CT for pretreatment staging and treated with EBT and intracavitary brachytherapy as well as concurrent chemotherapy. The binary logistic regression model was used to identify the independent prognostic factors among all the radiomic features and predict clinical outcomes. The log-rank test and CPHR analysis were performed to examine the effects of explanatory variables on outcome endpoints including overall survival, progression-free

TABLE 2 | The results from machine learning.

Algorithm	Study	Accuracy	Sensitivity	Specificity	AUC	End point
PNN	Obrzut et al. (2017)	0.9	1.0			5 year-survival
	Obrzut et al. (2019)		0.9	0.7		10-Survival
Network in network	Shen et al. (2019)	89.0	71.0	93.0		Recurrence
		87.0	77.0	90.0		Metastasis
CNN	Zhen et al. (2017)		72.0	59.0	0.700	Rectal toxicity
SVM	Chen et al. (2018b)		87.8	79.9	0.910	Rectal toxicity
SVM	Tian et al. (2019)		97.1	88.5	0.904	Radiation-induced fistula

survival, distant metastasis-free survival, and pelvic relapse-free survival. The results showed that the value of HGRE >3.68 or <3.68 were associated with significant different progression-free survival and pelvic relapse-free survival. Thus, HGRE was identified as an important factor in predicting chemoradiotherapy outcomes.

Outcome Prediction Using Machine Learning

To the author's knowledge, ML was first used to predict overall survival for 134 cervical cancer patients in 2002, using an artificial neural network model (ANN) including 11 prognostic factors (age, performance status, hemoglobin, total protein in serum, FIGO stage, histological type, histological grading at 30 Gy, histological grading at 40 Gy, histological grading at the end of therapy, cytological grading at 30 Gy, cytological grading at 40 Gy, cytological grading at the end of therapy) (Ochi et al., 2002). The predicted survival result was able to achieve an area under the receiver operating characteristic (ROC) curve (AUC) of 0.7782. A more recent study included 102 patients with cervical cancer staged as IA2-IIB, selected 23 demographic and tumor-related parameters, and collected perioperative data of each patient (Obrzut et al., 2017). The study predicted the 5-year survival rate using six machine learning methods: the probabilistic neural network (PNN), multilayer perceptron network (MLP), gene expression programming classifier (GEP), support vector machines algorithm (SVM), radial basis function neural network (RBFNN) and k-Means algorithm. Compared with other models, PNN provided the best prediction with an accuracy of 0.892 and sensitivity of 0.975. PNN was further used to predict the 10-year survival for the same cohort and also achieved high predictability (Obrzut et al., 2019).

Deep-learning (DL) has also been introduced for outcome prediction. A neural network model was implemented to predict survival utilizing clinicolaboratory variables among recurrent cervical cancer patients (Matsuo et al., 2017). The study tried to find among 13 clinicolaboratory variables the predictors for life expectancy in 157 recurrent cervical cancer patients. Those variables included age, body habitus change, pain score, blood pressure, and heart rate, white blood cell, hemoglobin, platelet, bicarbonate, blood urea nitrogen, creatinine, and albumin. The results showed that the 3-month survival decrease was associated with older age, decreasing albumin level, decreasing body mass index, increasing pain score, decreasing systolic blood pressure, decreasing white blood cell count, increasing platelet counts, and decreasing hemoglobin levels. This study group further predicted survival rate for 768 cervical cancer patients using the same DL model with 40 features that included patient demographics, vital signs, laboratory test results, tumor characteristics, and treatment types (Matsuo et al., 2019). They showed that the results of DL were better than that of CPHR.

In a recent study, a DL model called network in network was developed to predict treatment failures including local relapse and distant metastasis based on the analysis of the PET/CT images (Shen et al., 2019). The prediction of local relapse and distant metastasis obtained reasonable accuracy, sensitivity, and specificity. (**Table 2**) Four groups of radiomic features were also calculated, but none of the radiomic features was able to predict distant metastasis in this study.

ML is also able to predict treatment complications. A retrospective study applied the convolutional neural network (CNN) algorithm to analyze rectum dose distribution and predict rectum complications (Zhen et al., 2017). The study included 42 cervical cancer patients treated with EBRT combined with BT. The results showed that the texture features derived from the rectum surface dose map can generate better predictive performance than the volume parameters $D_{0.1/1/2cc}$ that are prescribed for dose constrains, in terms of sensitivity, specificity and AUC. The same research group applied the SVM algorithm to predict rectal toxicity for the same patient cohort and also achieved higher sensitivity, specificity, and AUC when compared with $D_{0.1/1/2cc}$ (**Table 2**) (Chen et al., 2018b).

The radiation-induced fistula is a concern for treating advanced gynecological (GYN) malignancies using radiation therapy. Another SVM model was developed to predict the risk of fistula formation caused by radiation therapy (Tian et al., 2019). The study included 35 gynecological cancer patients treated with interstitial BT. The model used the features of mixed data types that might be correlated to fistula formation, and included patient demographics, patient health status, tumor characteristics, additional invasive procedures, and dosimetric parameters. The predicted outcomes achieved a high prediction accuracy as shown in **Table 2**.

DISCUSSION

Accurate prediction of clinical outcomes would guide treatment to focus on specific prognostic factors and optimize the treatment scheme for each patient. The prediction of cervical cancer outcomes is one of the most challenging tasks as the management of cervical cancer involves the most complicated cancer treatment strategies. The studies reviewed in this paper have utilized models to discover many new prognostic factors such as tumor diameter, histological subtype, FDG SUVmax, radiomic features, and clinicolaboritory variables, and establish the relationships between those factors and clinical outcomes. Therefore, clinical outcomes can be accurately predicted. But the accuracy of prediction is related to models and algorithms. Several models performed very well in the studies. For example, CHPR predicted the 5-year survival rates 80% (I), 68% (IIA), and 59% (IIB) (Eifel et al., 2009), which were comparable to the reported results of 83.2% (I) and 68.9% (II) (Kim et al., 1988). Also, several DL models gave high accuracy predictions (**Table 2**). Such promising results have indicated that model-based outcome prediction has great potential for clinical applications.

The models used for prediction can be categorized into conventional statistical models and ML models. Conventional statistical models include the linear regression, the logistic regression, and CPHR. CPHR is one of most commonly used models for outcome prediction. It models relative hazards treating all the relevant factors proportionally. It can determine which factor is the most influential. But the proposed proportionality or linearity may not be valid because many prognostic factors are not linear and interact with each other. Thus the performance of prediction may not be ideal. In contrast, ML is able to deal with complex and non-linear relations in the data. Especially, it is able to learn feature representations automatically from raw data without direct feature engineering. Overall, ML outperformed statistical models in cervical cancer outcome prediction (Matsuo et al., 2017; Luo et al., 2019; Matsuo et al., 2019; Tian et al., 2019).

However, there was also evidence that the superiority of ML in outcome prediction is not always supported (Christodoulou et al., 2019). In addition, the ML models and algorithms have their own limitations, notably, overfitting (Zhen et al., 2017), and lack of interpretability (Luo et al., 2019; Luo et al., 2020). Overfitting would undermine predictive performance. Lack of interpretability would hinder the use of ML. ML works like a "black box" due to the complex algorithms. It is not easy to understand how it works and the predicted outcomes are not easy to understand as well. For instance, some predictors such as Albumin level were identified as significant prognostic factors by CPHR, but not by the DL (Matsuo et al., 2017). Thus, the prediction using ML may not be as convincing or well accepted as that using conventional models that are explicitly formulated. Furthermore, it is difficult to catch bugs or errors if they occur. Development of independent validation methods may help resolve this issue.

It should also be realized that the studies reviewed in this paper have limitations as well. First of all, most prediction studies did not have enough data, which would reduce the accuracy of the predictions. Secondly, most studies did not distinguish between treatment modalities and techniques. The treatment of cervical cancer involves almost all available cancer treatment modalities and techniques. Each modality and technique play specific roles and has different contributions to clinical outcomes. For example, LDR brachytherapy led to the 4-year disease-free survivals of 87, 66, and 28% for FIGO stages I, II, and III, respectively, (Coia et al., 1990), while HDR was able to achieve the 5-year survival of 94.4, 62, and 37.2%, for state I, II, III, respectively (Utley et al., 1984). Thus, the impact of different techniques on the outcomes should be determined separately and weighted in the prediction models. More attention should be paid to brachytherapy as brachytherapy is a major and complex treatment modality for cervical cancer. Especially, brachytherapy is sensitive to radiobiological effect. Radiobiological effect such as, dose-rate effect should be included in prediction models. Finally, most studies were limited to a single institution and small number of patients, and the results may have bias and significant uncertainties. The predicted outcomes are expected to be comparable to the actual outcomes independently derived from clinical trials or actual patient records.

CONCLUSION

The prediction of cervical cancer outcomes utilizing statistical models, images, and ML has produced promising results. Particularly, ML has capacity to process complex and non-linear relations in large-scale data, discover new prognostic factors, and perform predictions. It has great potential in clinical applications. However, more work is needed to make ML practical and reliable for clinical use. Future studies may include development of new methods and algorithms to minimize the effect of data scarcity, differentiating treatment modalities and techniques in prediction and evaluating individual contributions to clinical outcomes, and independent validation of machine learning algorithms.

AUTHOR CONTRIBUTIONS

The author is fully responsible for the design of the study and writing of the paper.

REFERENCES

American Cancer Society (2020). *Cancer Facts and Figures*. New York, NY: Oxford University Press.

Brunschwig, A. (1968). The Surgical Treatment of Cancer of the Cervix: Stage I and II. *Am. J. Roentgenology* 102, 147–151. doi:10.2214/ajr.102.1.147

Chen, J., Chen, H., Zhong, Z., et al. (2018). Investigating Rectal Toxicity Associated Dosimetric Features with Deformable Accumulated Rectal Surface Dose Maps for Cervical Cancer Radiotherapy. *Radiat. Oncol.* 13, 125. doi:10.1186/s13014-018-1068-0

Chen, S.-W., Shen, W.-C., Hsieh, T.-C., Liang, J.-A., Hung, Y.-C., Yeh, L.-S., et al. (2018). Textural Features of Cervical Cancers on FDG-PET/CT Associate with Survival and Local Relapse in Patients Treated with Definitive Chemoradiotherapy. *Sci. Rep.* 8, 11859. doi:10.1038/s41598-018-30336-6

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A Systematic Review Shows No Performance Benefit of

Machine Learning over Logistic Regression for Clinical Prediction Models. *J. Clin. Epidemiol.* 110, 12–22. doi:10.1016/j.jclinepi.2019.02.004

Coia, L., Won, M., Lanciano, R., Marcial, V. A., Martz, K., and Hanks, G. (1990). The Patterns of Care Outcome Study for Cancer of the Uterine Cervix. Results of the Second National Practice Survey. *Cancer* 66, 2451–2456. doi:10.1002/1097-0142(19901215)66:12<2451::aid-cncr2820661202>3.0.co;2-5

Eifel, P. J., Jhingran, A., Levenback, C. F., and Tucker, S. (2009). Predictive Value of a Proposed Subclassification of Stages I and II Cervical Cancer Based on Clinical Tumor Diameter. *Int. J. Gynecol. Cancer* 19, 2–7. doi:10.1111/igc.0b013e318197f185

Eifel, P. J., Winter, K., Morris, M., Levenback, C., Grigsby, P. W., Cooper, J., et al. (2004). Pelvic Irradiation with Concurrent Chemotherapy versus Pelvic and Para-Aortic Irradiation for High-Risk Cervical Cancer: An Update of Radiation Therapy Oncology Group Trial (RTOG) 90-01. *Jco* 22, 872–880. doi:10.1200/jco.2004.07.197

Eoh, K. J., Kim, H. J., Lee, J.-Y., Nam, E. J., Kim, S., Kim, S. W., et al. (2017). Upregulation of Homeobox Gene Is Correlated with Poor Survival Outcomes in Cervical Cancer. *Oncotarget* 8, 84396–84402. doi:10.18632/oncotarget.21041

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., et al. (2019). Estimating the Global Cancer Incidence and Mortality in 2018: GLOBOCAN Sources and Methods. *Int. J. Cancer* 144, 1941–1953. doi:10.1002/ijc.31937

Joslin, C. (2001). "High Dose Rate Brachytherapy for Treating Cervix Cancer," in *Principles and Practice of Brachytherapy Using after Loading Systems*. Editors C. Joslin, A. Flynn, and E. Hall (New York: Arnold).

Kang, J., Schwartz, R., Flickinger, J., et al. (2015). Machine Learning Approaches for Predicting Radiation Therapy Outcomes: a Clinician's Perspective. *Int. J. Radiat. Oncol Biol Phys* 93 (No. 5), 1127–1135. doi:10.1016/j.ijrobp.2015.07.2286

Kidd, E. A., Siegel, B. A., Dehdashti, F., and Grigsby, P. W. (2007). The Standardized Uptake Value for F-18 Fluorodeoxyglucose Is a Sensitive Predictive Biomarker for Cervical Cancer Treatment Response and Survival. *Cancer* 110, 1738–1744. doi:10.1002/cncr.22974

Kim, R. Y., Trotti, A., Wu, C. J., et al. (1988). "Results of Radiation Therapy Alone in the Treatment of Carcinoma of the Uterine Cervix," in *Radiological Society of North America 74th Scientific Assembly and Annual Meeting, Anon; 395* (Oak Brook, IL: Radiological Society of North America Inc), 176.

Landoni, F., Maneo, A., Colombo, A., Placa, F., Milani, R., Perego, P., et al. (1997). Randomised Study of Radical Surgery versus Radiotherapy for Stage Ib-IIa Cervical Cancer. *The Lancet* 350, 535–540. doi:10.1016/s0140-6736(97)02250-2

Lucia, F., Visvikis, D., Desseroit, M.-C., Miranda, O., Malhaire, J.-P., Robin, P., et al. (2018). Prediction of Outcome Using Pretreatment 18F-FDG PET/CT and MRI Radiomics in Locally Advanced Cervical Cancer Treated with Chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* 45, 768–786. doi:10.1007/s00259-017-3898-7

Luo, Y., Chen, S., and Valdes, G. (2020). Machine Learning for Radiation Outcome Modeling and Prediction. *Med. Phys.* 47 (5), e178–184. doi:10.1002/mp.13570

Luo, Y., Tseng, H.-H., Cui, S., Wei, L., Ten Haken, R. K., and El Naqa, I. (2019). Balancing Accuracy and Interpretability of Machine Learning Approaches for Radiation Treatment Outcomes Modeling. *BJR Open* 1, 20190021. doi:10.1259/bjro.20190021

Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R. S., Takiuchi, T., Liu, Y., et al. (2019). Survival Outcome Prediction in Cervical Cancer: Cox Models vs Deep-Learning Model. *Am. J. Obstet. Gynecol.* 220 (1-14), 381–e14. doi:10.1016/j.ajog.2018.12.030

Matsuo, K., Purushotham, S., Moeini, A., Li, G., Machida, H., Liu, Y., et al. (2017). A Pilot Study in Using Deep Learning to Predict Limited Life Expectancy in Women with Recurrent Cervical Cancer. *Am. J. Obstet. Gynecol.* 217, 703–705. doi:10.1016/j.ajog.2017.08.012

Mazeron, J. J., and Gerbaulet, A. (1998). The Centenary of Discovery of Radium. *Radiother. Oncol.* 49, 205–216. doi:10.1016/s0167-8140(98)00143-1

Obrzut, B., Kusy, M., Semczuk, A., et al. (2017). Prediction of 5–year Overall Survival in Cervical Cancer Patients Treated with Radical Hysterectomy Using Computational Intelligence Methods. *BMC Cancer* 17, 840. doi:10.1186/s12885-017-3806-3

Obrzut, B., Kusy, M., Semczuk, A., Obrzut, M., and Kluska, J. (2019). Prediction of 10-year Overall Survival in Patients with Operable Cervical Cancer Using a Probabilistic Neural Network. *J. Cancer* 10 (18), 4189–4195. doi:10.7150/jca.33945

Ochi, T., Murase, K., Fujii, T., Kawamura, M., and Ikezoe, J. (2002). Survival Prediction Using Artificial Neural Networks in Patients with Uterine Cervical Cancer Treated by Radiation Therapy Alone. *Int. J. Clin. Oncol.* 7, 0294–0300. doi:10.1007/s101470200043

Podczaski, E., Stryker, J. A., Kaminski, P., Ndubisi, B., Larson, J., Degeest, K., et al. (1990). Extended-field Radiation Therapy for Carcinoma of the Cervix. *Cancer* 66, 251–258. doi:10.1002/1097-0142(19900715)66:2<251::aid-cncr2820660210>3.0.co;2-e

Rebecca, L. (2020). Siegel, Cancer Statistics. *Ca Cancer J. Clin.* 70, 1. doi:10.3322/caac.21590

Shen, W.-C., Chen, S.-W., Wu, K.-C., Hsieh, T.-C., Liang, J.-A., Hung, Y.-C., et al. (2019). Prediction of Local Relapse and Distant Metastasis in Patients with Definitive Chemoradiotherapy-Treated Cervical Cancer by Deep Learning from [18F]-Fluorodeoxyglucose Positron Emission Tomography/computed Tomography. *Eur. Radiol.* 29, 6741–6749. doi:10.1007/s00330-019-06265-x

Small, W., Jr, Winter, K., Levenback, C., Iyer, R., Hymes, S. R., Jhingran, A., et al. (2011). Extended-field Irradiation and Intracavitary Brachytherapy Combined with Cisplatin and Amifostine for Cervical Cancer with Positive Para-Aortic or High Common Iliac Lymph Nodes: Results of Arm II of Radiation Therapy Oncology Group (RTOG) 0116. *Int. J. Gynecol. Cancer* 21, 1266–1275. doi:10.1097/IGC.0b013e31822c2769

Tian, Z., Yen, A., Zhou, Z., Shen, C., Albuquerque, K., and Hrycushko, B. (2019). A Machine-Learning-Based Prediction Model of Fistula Formation after Interstitial Brachytherapy for Locally Advanced Gynecological Malignancies. *Brachytherapy* 18, 530–538. doi:10.1016/j.brachy.2019.04.004

Utley, J. F., von Essen, C. F., Horn, R. A., and Moeller, J. H. (1984). High-dose-rate Afterloading Brachytherapy in Carcinoma of the Uterine Cervix. *Int. J. Radiat. Oncology*Biology*Physics* 10, 2259–2263. doi:10.1016/0360-3016(84)90231-1

Zhen, X., Chen, J., Zhong, Z., Hrycushko, B., Zhou, L., Jiang, S., et al. (2017). Deep Convolutional Neural Network with Transfer Learning for Rectum Toxicity Prediction in Cervical Cancer Radiotherapy: a Feasibility Study. *Phys. Med. Biol.* 62, 8246–8263. doi:10.1088/1361-6560/aa8d09

Conflict of Interest: The 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.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 30 November 2021
doi: 10.3389/frai.2021.582928

A Modified AUC for Training Convolutional Neural Networks: Taking Confidence Into Account

Khashayar Namdar[1,2]*, Masoom A. Haider[3,4] and Farzad Khalvati[1,2]

[1]Department of Medical Imaging, University of Toronto, Toronto, ON, Canada, [2]The Hospital for Sick Children (SickKids), Toronto, ON, Canada, [3]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada, [4]Sunnybrook Research Institute, Toronto, ON, Canada

Receiver operating characteristic (ROC) curve is an informative tool in binary classification and Area Under ROC Curve (AUC) is a popular metric for reporting performance of binary classifiers. In this paper, first we present a comprehensive review of ROC curve and AUC metric. Next, we propose a modified version of AUC that takes confidence of the model into account and at the same time, incorporates AUC into Binary Cross Entropy (BCE) loss used for training a Convolutional neural Network for classification tasks. We demonstrate this on three datasets: MNIST, prostate MRI, and brain MRI. Furthermore, we have published GenuineAI, a new python library, which provides the functions for conventional AUC and the proposed modified AUC along with metrics including sensitivity, specificity, recall, precision, and F1 for each point of the ROC curve.

Keywords: AUC, ROC, CNN, binary classification, loss function

OPEN ACCESS

Edited by:
Jake Y. Chen,
University of Alabama at Birmingham,
United States

Reviewed by:
Akram Mohammed,
University of Tennessee Health
Science Center (UTHSC),
United States
Rubul Kumar Bania,
North-Eastern Hill Uiveristy, Tura
Campus, India

****Correspondence:***
Khashayar Namdar
ernest.namdar@utoronto.ca

Specialty section:
This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Artificial Intelligence

Received: *13 July 2020*
Accepted: *30 September 2021*
Published: *30 November 2021*

Citation:
Namdar K, Haider MA and Khalvati F
(2021) A Modified AUC for Training
Convolutional Neural Networks: Taking
Confidence Into Account.
Front. Artif. Intell. 4:582928.
doi: 10.3389/frai.2021.582928

INTRODUCTION

Classification is an important task in different fields, including Engineering, Social Science, and Medical Science. To evaluate quality of classification, a metric is needed. Accuracy, precision, and F1 score are three popular examples. However, there are other metrics that are more accepted in specific fields. For example, sensitivity and specificity are widely used in Medical Science.

For binary classification, Receiver Operating Characteristic (ROC) curve incorporates different evaluation metrics. The Area Under ROC Curve (AUC) is a widespread metric, especially in Medical Science (Sulam et al., 2017). In engineering, AUC has been used to evaluate the classification models since the early 1990s (Burke et al., 1992), and AUC research has continued ever since. Kottas *et al.* proposed a method to report confidence intervals for AUC (Kottas et al., 2014). Yu *et al.* proposed a modified AUC which is customized for gene ranking (Yu et al., 2018). Yu also proposed another version of AUC for penalizing regression models used for gene selection with high dimensional data (Yu and Park, 2014). Rosenfeld *et al.* used AUC as a loss function and demonstrated AUC-based training lead to better generalization (Rosenfeld et al., 2014). Their research, however, is not in the context of Neural Networks (NN); instead, they use Support Vector Machines (SVM). Therefore, their method does not address the challenges we address in this paper, including taking confidence of the model into account in calculating AUC and thus, making it a better metric for training neural networks. Zhao et al. proposed an algorithm for AUC maximization in online learning (Zhao et al., 2011). A stochastic approach for the same task was introduced by Ying et al. (2016). Cortes and Mohri studied correlation of AUC, as it is optimized, and error rate (Cortes and Mohri, 2004). Their research showed that minimizing the error rate may not result in maximizing AUC. Ghanbari and

Scheinberg directly optimized error rate and AUC of the classifiers; however, their approach only applies to linear classifiers (Ghanbari and Scheinberg, 2018).

This paper explains in detail the meaning of AUC, how reliable it is, under which circumstances it should be used, and its limitations. It also proposes a novel approach to eliminate these limitations. Our primary focus is on deep learning and Convolutional Neural Networks (CNNs), which differentiates our work from the previous work in the literature. We propose confidence-incorporated AUC (cAUC) as a modified AUC which directly correlates to Cross-Entropy Loss function and thus, helps to stop CNN training at a more optimum point in terms of confidence. This is not possible with conventional AUC, as not only the minimum of Binary Cross-Entropy loss function may not correlate with the maximum of AUC, but also AUC does not take the confidence of the model into account. We have also published a new library called GeuineAI[1], which contains our modified AUC and conventional AUC with more features in comparison to the existing standard python libraries.

REVISITING THE CONCEPT OF AUC

In supervised binary classification, each datapoint has a label. Conformed with standards of Machine Learning, labels are either 0/1 or 01/10 or sometimes +1/-1 and the model's (classifier's) outputs are usually probabilities. In the case of cancer detection, for example, input data may be CT or MRI images. Cancerous cases will be images labeled with 1 (positive) and normal (healthy) images will have 0 (negative) as their labels. The model returns a probability for each image. In the ideal scenario, the model's output will be 1 for cancerous images and 0 for normal ones.

Four possible outcomes of binary classification are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From **Table 1**, it can be inferred that TX means Truly predicted as X and FX means Falsely predicted as X.

Defined as the total number of correct predictions out of total cases, Accuracy is calculated by **Equation (1)**.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

As it can be seen, accuracy is only concerned about correct versus wrong predictions. In many situations, especially in Medical Science, this is not enough. The consequences of misclassifying a normal case as cancerous and considering a cancerous case as normal are way different. The first one is referred to FP, also known as Type I error, whereas the second one is a FN or Type II error. True Positive Rate (TPR) and False Positive Rate (FPR) are two criterions which distinguish the error types.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

TABLE 1 | Possible outcomes of binary classification.

	Actual value	Predicted value
TN	0	0
FP	0	1
FN	1	0
TP	1	1

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

TPR is also known as sensitivity and refers to the ratio of correct predictions to total within actual positives. FPR is the ratio of wrong predictions within actual negatives. FPR is related to specificity by **Eq. 4**, which is used frequently in Medical Science.

$$FPR = 1 - specificity \tag{4}$$

As mentioned before, predicted value should be binary, but output of the model is probability. Thresholding is how probabilities are converted to predicted values. As an example, if the output is 0.6 and the threshold is 0.5, predicted value is 1.

$$y = \begin{cases} 0\, if\ p \leq t \\ 1,\ otherwise \end{cases} \tag{5}$$

y in **Eq. 5** is the predicted value, p is the output of the model, which is a probability, and t is the threshold. Depending on t, TPR and FPR will be different. ROC is the curve formed by plotting TPR versus FPR for all possible thresholds and AUC is the area under that curve.

In the following, we take an example-based approach to highlight the fundamentals of AUC.

Example 1: **Table 2** contains the simplest possible example. It should be followed from left to right. y^d refers to the desired value which is the same as the label (ground truth).

It can be seen from **Table 2** that actual positives and actual negatives are necessary to draw an ROC curve. Although it may seem trivial, lack of one category in one batch leads to NaN in training of Machine Learning (ML) models. Furthermore, if the batch size is equal to one, the batch AUC is always NaN. Consequently, for any NN to be directly trained with a modified AUC, or for any code where AUC is calculated within each batch, batch size of one cannot be used. Furthermore, the sampler should be customized in a way to return samples from both classes in each batch.

Example 2: **Table 3** contains an example of classifying one positive and one negative cases and **Figure 1** shows the

TABLE 2 | Example 1.

$y^d = 1$	$t < 0.5$	$y = 1$	$TP = 1$ $TN = 0$	$TPR = \frac{TP}{TP+FN} = \frac{1}{1+0} = 1$
$p = 0.5$			$FP = 0$ $FN = 0$	$FPR = \frac{FP}{FP+TN} = \frac{0}{0+0} = NaN$

[1] https://pypi.org/project/GenuineAI/

TABLE 3 | Example 2.

$y_1^d = 1$ $y_2^d = 0$	$t < 0.4$	$y_1 = 1$ $y_2 = 1$	$TP = 1$ $TN = 0$ $FP = 1$ $FN = 0$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{1}{1+0} = 1$
$p_1 = 0.4$ $p_2 = 0.6$	$0.4 \le t < 0.6$	$y_1 = 0$ $y_2 = 1$	$TP = 0$ $TN = 0$ $FP = 1$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $FPR = \frac{1}{1+0} = 1$
	$0.6 \le t$	$y_1 = 0$ $y_2 = 0$	$TP = 0$ $TN = 1$ $FP = 0$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $TPR = \frac{0}{0+1} = 0$

TABLE 4 | Example 3.

$y_1^d = 1$ $y_2^d = 0$	$t < 0.4$	$y_1 = 1$ $y_2 = 1$	$TP = 1$ $TN = 0$ $FP = 1$ $FN = 0$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{1}{1+0} = 1$
$p_1 = 0.4$ $p_2 = 0.6$	$0.4 \le t < 0.6$	$y_1 = 0$ $y_2 = 1$	$TP = 1$ $TN = 1$ $FP = 0$ $FN = 0$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{0}{0+1} = 0$
	$0.6 \le t$	$y_1 = 0$ $y_2 = 0$	$TP = 0$ $TN = 1$ $FP = 0$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $FPR = \frac{0}{0+1} = 0$

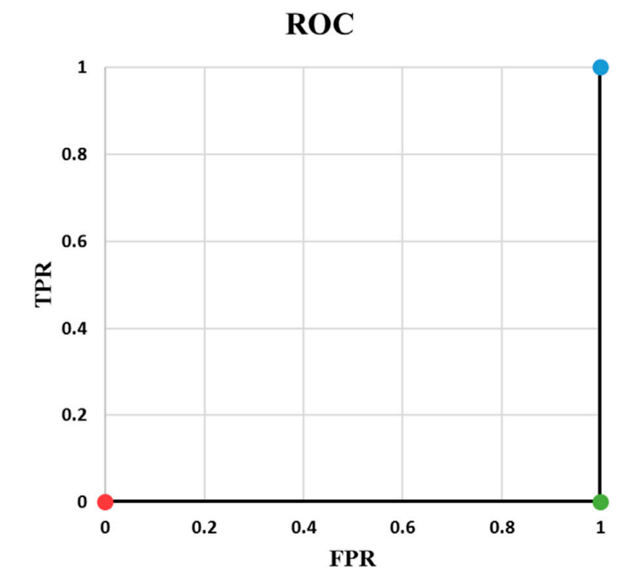


FIGURE 1 | ROC of Example 2.

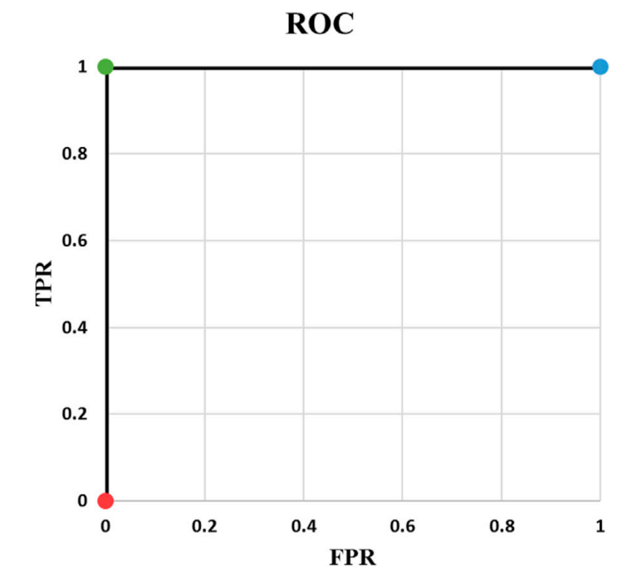


FIGURE 2 | ROC of Example 3.

corresponding ROC curve. There are important points in this example. ROC curves always start from (0,0) and always end at (1,1). The reason is that if threshold is 0, all predicted values are 1. They will be either TP or FP. Therefore, both TPR and FPR are 1. On the other hand, if threshold is 1, everything is predicted as negative. In this case, predictions are all TN or FN. Consequently, TPR and FPR will be both zero. Two things must be taken into account when writing a ML code: $t = 0$, and $t = 1$ should be treated separately and t should be iterated backward if going from (0, 0) to (1, 1) is desired. Backward iteration necessity comes from the fact that the highest t corresponds to the lowest TPR and FPR. Exceptions of $t = 0$ and $t = 1$ are needed for rare cases when the output of the model is exactly 0 or 1.

Example 3: Our third example is complement of Example 2. As it is indicated in **Table 4**, output probability for the positive case (y_1^d) is higher. Under these conditions, AUC is equal to 1, as depicted in **Figure 2**. In other words, ideal situation for classification of one positive and one negative example in terms of AUC is when output probability of the positive case is higher.

It should be noted that if the two probabilities were slightly different, e.g., $p_1 = 0.501$ and $p_2 = 0.499$, AUC would be 1. The separation of probabilities does not have to be at 0.5. $p_1 = 0.0002$ and $p_2 = 0.0001$ would still result in AUC = 1. This leads to an important issue which is confidence. It turns out AUC does not take into account the confidence of the model.

Example 4: In the fourth example (**Table 5**), the output probabilities are the same for the two samples. This leads to AUC of 0.50. This example shows that whenever all output probabilities are equal, AUC is 0.50 and ROC is a straight line from (0, 0) to (1, 1) (**Figure 3**). This is true for all different values of N where N is batch size or number of samples.

Example 5: In example 5, N is equal to 3 and there are 4 points in the ROC curve (**Figures 4**, **5**). The reason for this phenomenon is

TABLE 5 | Example 4.

$y_1^d = 1$ $y_2^d = 0$	$t < p$	$y_1 = 1$ $y_2 = 1$	$TP = 1$ $TN = 0$ $FP = 1$ $FN = 0$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{1}{1+0} = 1$
$p_1 = p_2 = p$	$p \leq t$	$y_1 = 0$ $y_2 = 0$	$TP = 0$ $TN = 1$ $FP = 0$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $FPR = \frac{1}{1+0} = 1$

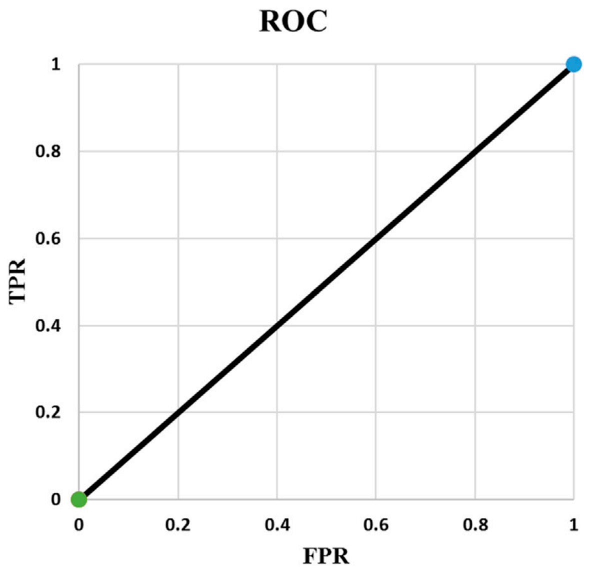


FIGURE 3 | ROC of Example 4.

TABLE 6 | Example 5.

$y_1^d = 1$ $y_2^d = 0$ $y_3^d = 1$	$t < 0.4$	$y_1 = 1$ $y_2 = 1$ $y_3 = 1$	$TP = 1$ $TN = 0$ $FP = 2$ $FN = 0$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{2}{2+0} = 1$
	$0.4 \leq t < 0.45$	$y_1 = 0$ $y_2 = 1$ $y_3 = 1$	$TP = 0$ $TN = 0$ $FP = 1$ $FN = 1$	$TPR = \frac{1}{1+0} = 1$ $FPR = \frac{1}{1+1} = 0.5$
$p_1 = 0.4$ $p_2 = 0.55$ $p_2 = 0.55$	$0.45 \leq t < 0.55$	$y_1 = 0$ $y_2 = 1$ $y_3 = 0$	$TP = 0$ $TN = 1$ $FP = 1$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $FPR = \frac{1}{1+1} = 0.5$
	$0.55 \leq t$	$y_1 = 0$ $y_2 = 0$ $y_3 = 0$	$TP = 0$ $TN = 2$ $FP = 0$ $FN = 1$	$TPR = \frac{0}{0+1} = 0$ $FPR = \frac{0}{0+2} = 0$

effective threshold boundaries. As it can be seen in **Table 6**, up to t = 0.4, no value of t changes the model's predictions. It turns out that those effective boundaries are defined by predicted probabilities. It should now be highlighted, in Examples 2 and 3, N was 2 and there were 3 points on the ROC curve. In the general form, for N predictions, there will be N+1 points on the ROC curve. For each pair of predictions with equal probabilities, one point is omitted. The extreme case is when all output probabilities are equal. In this case, there will be two points on the ROC curve and AUC is 0.5 (Example 4).

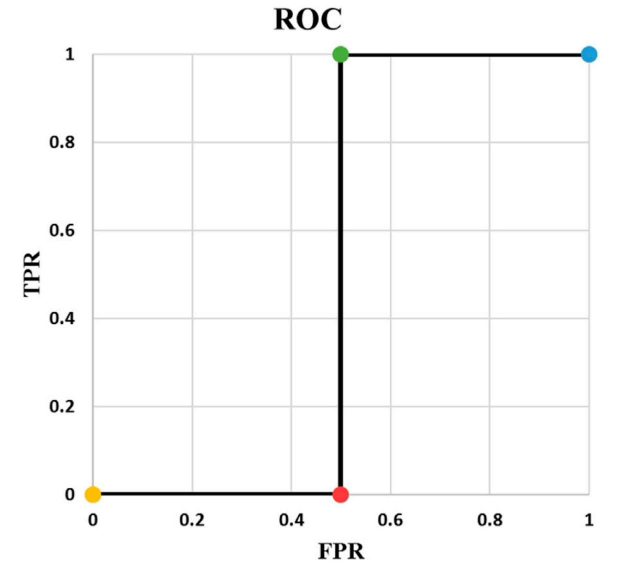


FIGURE 4 | ROC of Example 5.

METHODS

Inspired by the previous examples, we will now investigate some characteristics of ROC and AUC. We will demonstrate how misclassification of a single data point can decrease AUC, and what extreme scenarios of misclassification look like. We will then provide an example to show a higher AUC does not necessarily correspond to better classification. The section is concluded with introducing cAUC, our proposed modified AUC, and mathematical support for its correlation to Binary Cross Entropy (BCE).

A result of having *N+1* points on the ROC curve is that *N+1* different effective values can be assigned to threshold *t*. In other words, while infinite values for *t* can be selected, selecting more than *N+1* values for *t* would not help to achieve more accurate AUC or "smoother" ROC curve. Even if calculations are precise, the efficiency will be degraded because if t values are not selected from different effective intervals, they will result in the same point on ROC. In Example 3, *t* = 0, 0.1, 0.2, 0.3, or any other value below 0.4 will result in (1, 1) on ROC. Furthermore, because continuous variables have to be discretized, selecting fixed step size to increase *t* may result in inaccuracy. It happens almost certainly if two probabilities are highly close to each other and the fixed step is not small enough to land between them. Usually high values of N create

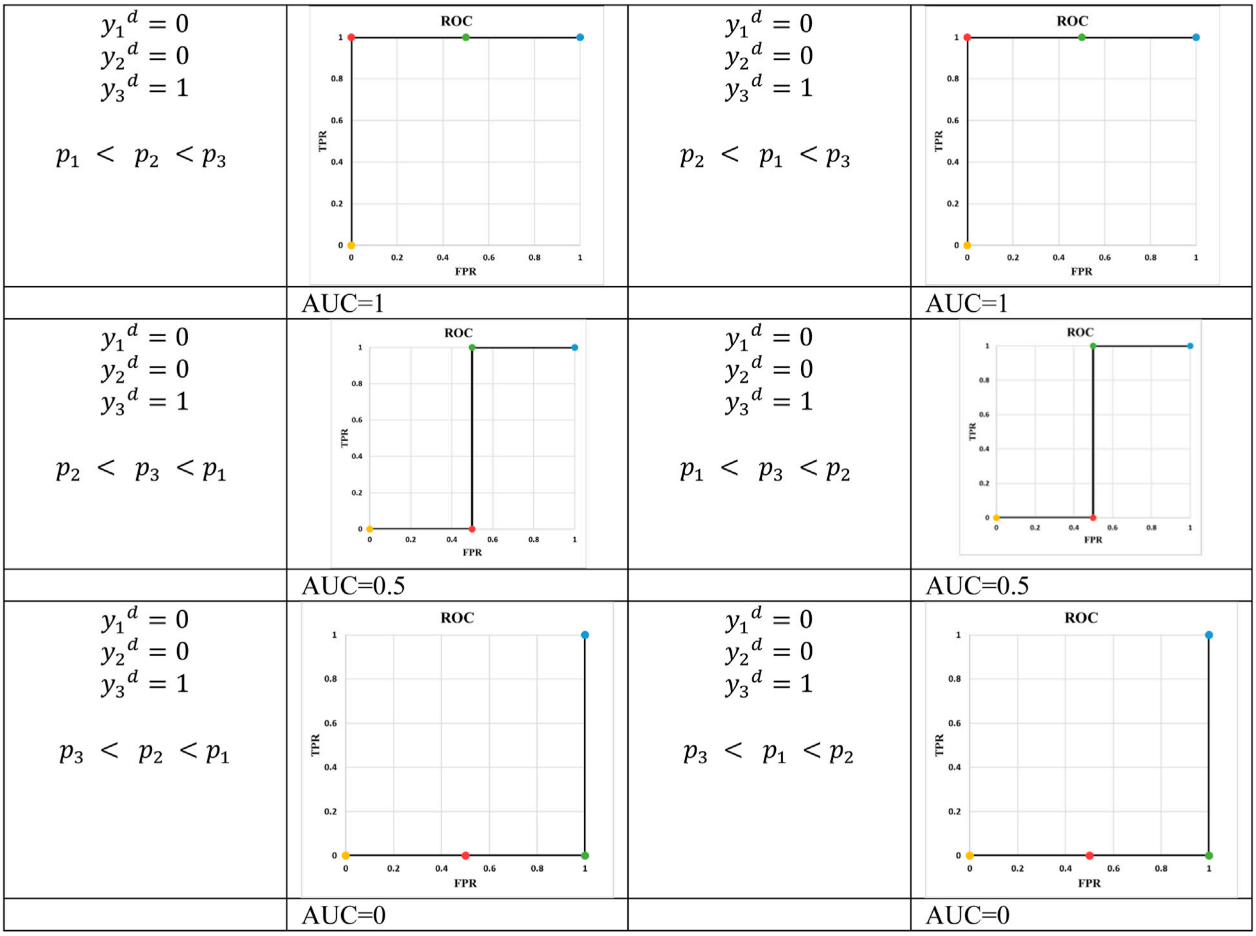


FIGURE 5 | ROC curves for N = 3, two actual negative and an actual positive.

such circumstances. Therefore, having a method for selecting optimal threshold is crucial. Changing value of t is effective if and only if it affects predictions. Assuming probabilities are sorted, any value of t between p_i and p_{i+1} does not change predictions. Supported by the same rational, the optimal values of t we suggest is given by **(6)**. An optimum set, based on the rule of having $N+1$ points in ROC, has to have $N+1$ members. However, our proposed set has $N+2$ elements. If **Eq. 5** is conformed, 1 can be removed from the set. Nevertheless, adding 0 and 1 to the set is a safe approach for avoiding programming errors.

$$t \in \{0,\ p_i\ ,\ 1\}\,, i = 1, 2, \ldots, N \tag{6}$$

Figure 5 depicts all possible outcomes (except special cases of equal probabilities). It seems ROC is always staircase looking, except for the situations where a pair of predicted probabilities are equal. Thus, using trapezoid integration is the best and most accurate technique to calculate AUC. Furthermore, **Figure 5** demonstrates order of predicted probabilities plays a key role in amount of AUC. If there is at least one threshold t where the probabilities of all actual positives and negatives are above and below it, respectively, then the AUC is equal to 1. Although the mathematical proof needs more fundamentals, there is one key support: selecting t at the boundary of positive and negative data points results in a perfect classification corresponding to (0, 1) on ROC.

TABLE 7 | A group of realizations with N = 3, AN = 2, and AP = 1.

	–		t
Sorted Actual Values	0	0	1
Predicted Probabilities	p-ε	p-ε	p
–	TN	TN	TP

$$AUC = 1\, if\ \exists t \left| \left\{ \forall p_i, y_i^d \in AP\ \ t < p_i\ and\ \forall p_j, y_i^d \in AN\ \ p_j \le t \right\} \right. \tag{7}$$

Where AP and AN are actual positives and actual negatives, respectively.

To be able to separate positive and negative datapoints in a way that probabilities of positive cases are higher, we introduce ε. In **Table 7**, ε is a positive real number which

TABLE 8 | A group of realizations with N = 8, AN = 4, AP = 4, and AUC = 1.

	–	–	–		t	–	–	–
Sorted Actual Values	0	0	0	0	1	1	1	1
Sorted Probabilities	p-ε-3δ	p-ε-2δ	p-ε-δ	p-ε	p	p+δ	p+2δ	p+3δ
–	TN	TN	TN	TN	TP	TP	TP	TP

TABLE 9 | A group of realizations with N = 8, AN = 4, AP = 4, and AUC = 0.

	–	–	–		t	–	–	–
Sorted actual values	0	0	0	0	1	1	1	1
Sorted probabilities	p-3δ	p-2δ	p-δ	p	p-ε	p-ε+δ	p-ε+2δ	p-ε+3δ
–	FP	FP	FP	FP	FN	FN	FN	FN

TABLE 10 | A group of realizations with N = 8, AN = 4, AP = 4, and 0 < AUC<1.

(a)	–	–	–	–		t	–	–	–	
	sorted Actual values	0	0	0	0	1	1	1	1	
	sorted Probabilities	p-ε-2δ	p-ε-δ	p-ε	p	p	p+δ	p+2δ	p+3δ	
		–	TN	TN	TN	🟥		TP	TP	TP
(b)	–	–	–	t	–	–	–	–		
	sorted Actual values	0	0	0	0	1	1	1	1	
	sorted Probabilities	p-ε-2δ	p-ε-δ	p-ε	p	p	p+δ	p+2δ	p+3δ	
		–	TN	TN	TN	FP	TP	TP	TP	TP
(c)	–	–	–	–	–	t	–	–		
	sorted Actual values	0	0	0	0	1	1	1	1	
	sorted Probabilities	p-ε-2δ	p-ε-δ	p-ε	p	p	p+δ	p+2δ	p+3δ	
	–	TN	TN	TN	TN	FN	TP	TP	TP	

is less than or equal to p. This ensures p-ε is zero or positive and implies that p-ε is less than p. For example, if p is 0.8, ε can be in the range of 0–0.8. It also explains why **(7)** is true. For any $t\ \in [p-\varepsilon,\ p)$, conditions of **(7)** are met and the AUC is equal to 1. In this case, (0,0), (0,1), and (1,1) are points of ROC.

In **Table 7**, probabilities of all actual negatives are equal (p-ε). To be able to sort probabilities within each class of datapoints, δ is introduced. **Table 8** extends **Table 7** scenario to more general cases where probabilities are not necessarily equal. In this case, δ can be considered as a random noise which is a non-negative real number. It helps to simulate predicted probabilities better. With δ, the predicted probabilities do not follow a distinct pattern of having a fixed distance.

Table 9 shows the other extreme. When there is threshold t such that probabilities of all actual positives and negatives are below and above it, respectively, then the AUC is zero.

$$AUC = 0\,if\ \exists\, t\ \Big|\ \left\{\forall p_i, y_i^d \in AP\ \ p_i \le t\ and\ \forall p_j, y_j^d \in AN\ \ t < p_j\right\} \tag{8}$$

Table 10 depicts all remaining possible scenarios where AUC is greater than zero (0 < AUC). **Table 10** gives the big picture. For $t\ \in [p-\varepsilon,\ p)$, there will be one FP in predicted values (**Table 10**), which means TPR is 1 and FPR is positive. For $t\ \in [p,\ p+\delta)$, there will be one FN in predicted values (**Table 10**), which means FPR = 0 and TPR less than one. In other words, in the ROC curve, (**Tables 10(b),(c)**) correspond to points (d_1,1) and (0, d_2), respectively, where d_1 and d_2 are positive real numbers (**Figure 6**). Obviously, this causes a reduction in AUC as much as the area of a triangle. (0, d_2), (d_1,1), and (0, 1) are vertices of the triangle. FN contributes to TPR whereas FP is part of FPR. Therefore, d_1 is influenced by FP and d_2, is a function of FN. Because they both play a role in the triangle's area, it can be concluded that the AUC does not discriminate between FP and FN. All it does is scaling the

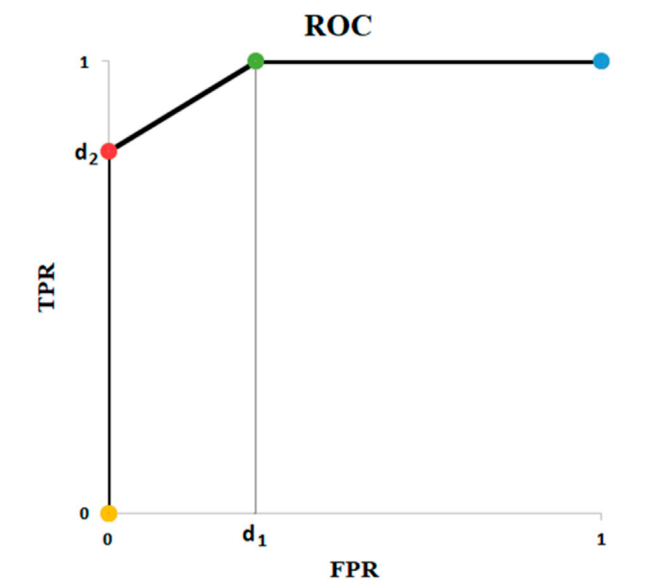


FIGURE 6 | ROC of Example of **Table 10**.

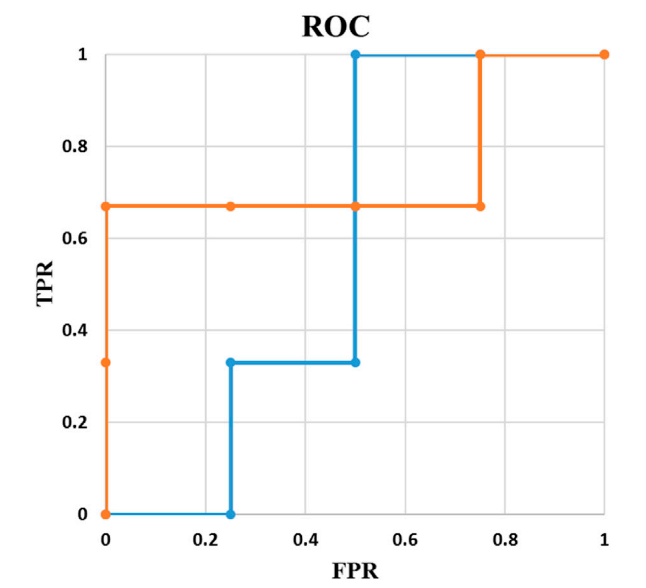


FIGURE 7 | ROC curves for two different models with N = 7.

importance with respect to degree of imbalance. In other words, AUC equalizes importance of positive and negative cases as if the number of *AP*s and *AN*s were the same. In this perspective, ROC has a built-in normalizer mechanism. However, in real world, that may not be desired. In most cancer detection situations, for example, importance of a positive case massively outweighs that of a negative case.

The fact that the AUC does not discriminate between FP and FN implies that what should be used as a criterion when training a model is ROC curve itself and not the AUC. Hence, in order to translate probabilities to predictions, one specific $t \in [0, 1]$ is needed.

In medical science (e.g., cancer detection), instead of AUC value, the clinical value of a classification method is usually studied in terms of TPR or FPR. For example, for a desired TPR, using the ROC curve, the point with lowest FPR is selected. From there, the desired threshold is derived, and the classification is performed. Thus, to evaluate the performance, confusion matrix is the most informative way of reporting where a model with a lower AUC may be preferred when the specific TPR/FPR are considered. One possible example is illustrated in **Figure 7**.

In **Figure 7**, AUC of the orange line and the blue line are 0.75 and 0.58, respectively. Although the orange line has a higher AUC, if the acceptable sensitivity is set at 1, the blue line corresponds to the best model. In other words, to be able to identify every single positive example, with the orange line we will misclassify 75% of our negative examples compared with 50% of misclassification by the blue one.

Proposed AUC With Confidence

We call a model confident if it returns probabilities near 1 for all positive cases and probabilities near 0 for all negative examples. In previous section, it was demonstrated that AUC does not provide the confidence of the classification model under study. In other words, whether the predicted probabilities are close to each other or not does not affect the AUC value. As a result, a classification model that is able to separate the positive and negative cases by a small margin (e.g., 5%), has the same AUC as the one that separates the positive and negative cases by a large margin (e.g., 25%). Risk assessment in Medical Science and regression in Statistics are cases where having large margins may not be the target. However, in the context of classification, the margin is a key point. The whole idea of Support Vector Machines (SVM) is formed around large margin classification (Parikh and Shah, 2016). The ultimate effect of Cross Entropy (CE) loss function on NNs is imposing separation between predicted probability of positive and negative examples (Zhang and Sabuncu, 2018).

To address this issue, we propose a modified AUC (cAUC), which provides a confidence measure for the classification model. To do so, we introduce two coefficients, α and β.

$$\alpha = \max(p_i) - \min(p_j) \Big| \{p_i \in AP,\ p_j \in AN\} \tag{9}$$

$$\beta = \min(p_i) - \max(p_j) \Big| \{p_i \in AP,\ p_j \in AN\} \tag{10}$$

$$cAUC = e^{(\alpha-1)} e^{(\beta-1)} AUC \tag{11}$$

The idea behind **Eq. 11** is the smaller the range between the probabilities of the two classes, the lower the AUC will be and vice versa. If the range is the maximum possible value (which is 1), the AUC remains unchanged. Otherwise, it is decreased.

In the following, we show that our cAUC local maximums correspond to BCE local minimums. Intuitively, BCE is minimized when the probabilities created by the model are close to 1 for APs and near 0 for ANs. This translates to the concept of confidence we discussed above. Mathematically, BCE is explained through **Eq. 12**. Using the same separation approach, we have used so far, BCE can be rewritten for APs and ANs as

TABLE 11 | Comparison of AUC and the proposed AUC for a random case.

Real values	Sorted probabilities	Parameters
1	0.803258838	α = 0.80325884 – 0.27759354 = 0.5256653 β = 0.30374599 – 0.69960646 = –0.39586047 *AUC* = 0.6666666666666666 *cAUC* = 0.1027290563696407 —
0	0.517853202	
1	0.639592674	
1	0.303745995	
0	0.699606458	
0	0.318090495	
0	0.277593543	
1	0.421482502	
1	0.556011119	
1	0.548716153	

Eq. 13. From **Eq. 13**, it can be concluded ideal BCE loss is resulted under conditions of **Eq. 14**.

$$BCE = \frac{-1}{N}\sum_{i=1}^{N} y_i^d \log(p_i) + \left(1 - y_i^d\right)\log\left(1 - p_i\right) \tag{12}$$

$$BCE = \frac{-1}{N}\left[\sum_{i=1}^{N}\begin{cases}\log(p_i)\big|\, y_i^d \in AP\\ 0,\ otherwise\end{cases} + \sum_{j=1}^{N}\begin{cases}\log\left(1 - p_j\right)\big|\, y_j^d \in AN\\ 0,\ otherwise\end{cases}\right] \tag{13}$$

$$BCE = 0\, if\ \forall y_i^d \in AP,\ p_i = 1\ and\ \forall y_i^d \in AN,\ p_i = 0 \tag{14}$$

If conditions of **Eq. 14** are met, from **Eq. 7** it can be inferred AUC is equal to 1 because for any threshold between 0 and 1, all datapoints are correctly classified. In this case **Eq. 9**, **10** result in $\alpha = \beta = 1$. Ultimately, our definition of cAUC, **Eq. 11**, returns $cAUC = 1$. Therefore, the ideal cases of cAUC and BCE correspond to each other. Through a similar procedure, it can be proved their worst cases (cAUC = 0 and BCE $\rightarrow \infty$) correspond too. In the transition between the two extremes, BCE and confidence-related part of cAUC (the exponential coefficients) have a monotonic behavior.

We proved that if AUC is equal to 1, the probability of positive and negative examples can be close to each other and thus, leading to high BCE. Therefore, a high AUC does not necessarily mean low BCE. Thus, instead of AUC, we propose monitoring cAUC, which in global optimums is guaranteed to result in ideal BCE and AUC, and in local optimums has higher potential for stopping the training when the model is confident, not overfit, and achieves a high AUC.

RESULTS

We will evaluate our confidence-incorporated AUC (cAUC) on 4 different scenarios: random predictions, a customized dataset based on MNIST (LeCun and Cortes, 2010), our proprietary Prostate Cancer (PCa) dataset, and a dataset based on BraTS19 (Menze et al., 2015; Bakas et al., 2017; Bakas, 2018). Our PCa dataset of Diffusion-weighted MRI is described in our previous research (Yoo et al., 2019). The CNN architectures and the utilized settings are similar to our shallow models used in other research projects (Hao et al., 2020). Nonetheless, the details are provided in **Supplementary Appendix A**. Given the fact that AUC is not differentiable, to train the network we used BCE. The only essential point which should be covered is input channels of our CNN for MNIST classification. Because MNIST is a single channel dataset, we revised the network to be compatible with it.

cAUC vs AUC on Random Data

To test the proposed AUC, in an N = 10 simulation, real values and predicted probabilities were generated randomly using U [0, 1] as **Table 11**. In case of arbitrary classification, expected value of AUC is 0.5. The goal here is to calculate expected values of cAUC for such conditions. Another point for the presented values in **Table 11** is to highlight importance of sample size. With the widespread use of AI in Medical Science, researchers must care about sample sizes. Our experiment shows AUC = 0.66 is not hard to achieve through chance when N is not high enough.

Simulations with N = 100 and 10,000 trials show expected value of AUC is 0.50 and expected value of the revised AUC is 0.07. Intuitively, AUC = 0.5 happens when everything is by chance. We showed one example is when output of the model is constant. In other words, when variance of the output vector is zero. In this case, coefficients α and β also are zero in limit [according to **(9)** and **(10)**]. Therefore, cAUC will be $0.5{*}e^{(-1)}e^{(-1)}$ which is 0.07.

cAUC vs. AUC on an MNIST-Based Dataset

MNIST is a well-known dataset of handwritten digits, including 60,000 train and 10,000 test images (LeCun and Cortes, 2010). It includes single channel, 28 × 28 pixel, normalized images. The 10 different digits form classes of data in MNIST, by default. Because our ultimate goal was

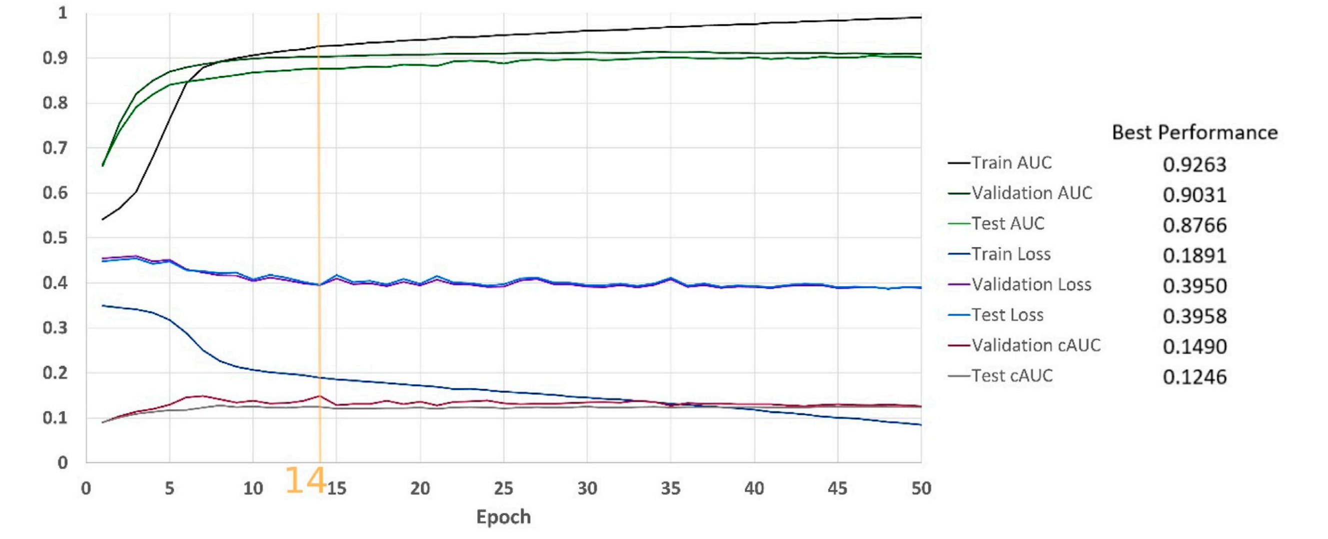


FIGURE 8 | Classification results on the MINIST-based dataset.

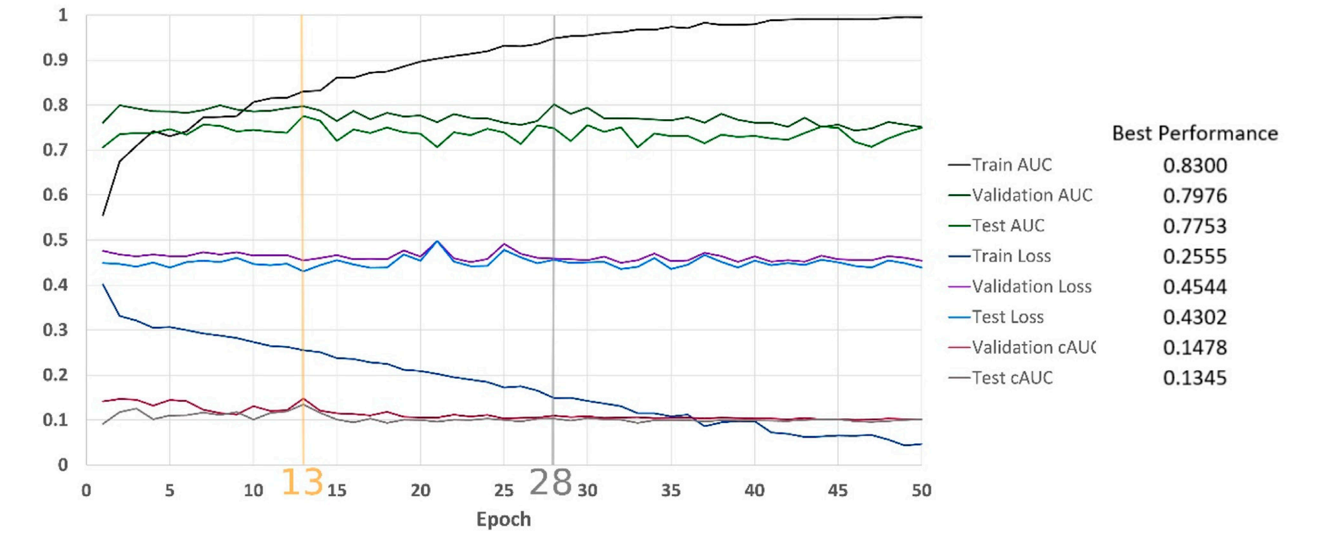


FIGURE 9 | Classification results on the PCa dataset.

Medical applications, we marked examples of 7 as positive and all other digits as negative to create our imbalanced binary MNIST-based dataset. Our train set included the first 5,000 examples of training cohort of MNIST and our validation set was 1,500 examples (indices: 45,000–46,500) of it. Our test set was built from the first 1,000 examples of MNIST test. This was done to ensure our dataset size is reasonable in comparison to Medical ones. To make our data noisy, as it is always seen in Medical datasets, we added uniform random noise to each pixel. For that end, we first scaled MNIST examples in order to have each pixel values in the range of [0, 1]. Then we added 5 times of a random image to it and scaled the result back to [0, 1] as stated in **Eq. 15**

$$image = \frac{\frac{MNIST\ image}{255} + 5 * numpy.random.random\left((28, 28)\right)}{6} \quad (15)$$

Figure 8 shows results of the classification over 50 epochs of training. In each epoch, average BCE loss, AUC, and cAUC for training, validation, and test cohorts are calculated. This procedure is maintained until the last epoch and then the monitored values are plotted.

cAUC vs. AUC on a Proprietary PCa Dataset

Figure 9 depicts the results of classification over our institutional review board approved PCa dataset, which

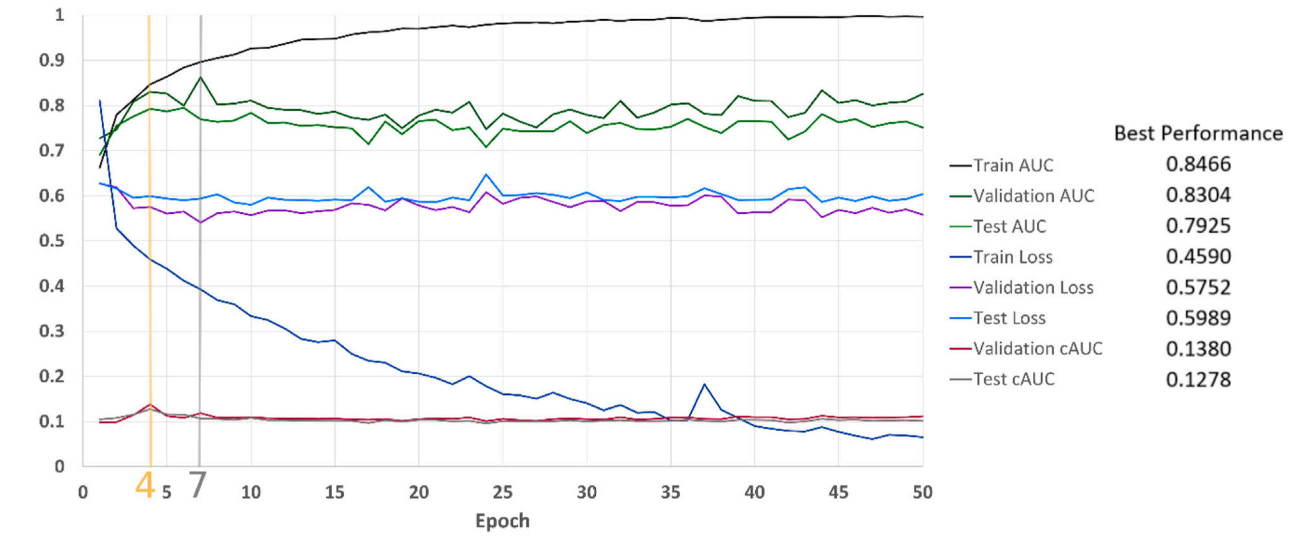


FIGURE 10 | Classification results on our BraTS-based dataset.

included Diffusion-weighted MRI images of 414 prostate cancer patients (5,706 2D slices). The dataset was divided into training (217 patients, 2,955 slices), validation (102 patients, 1,417 slices), and test sets (95 patients, 1,334 slices). Label for each slice was generated based on the targeted biopsy results where a clinically significant prostate cancer (Gleason score>6) was considered a positive label. The golden vertical line is where cAUC guides us to stop and the grey vertical line is where we would stop if AUC was used.

cAUC vs. AUC on a BraTS-Based Dataset

We used the BraTS19 dataset, with the same setting as our previous research (Hao et al., 2021). The dataset contains 335 patients of which 259 patients were diagnosed with high-grade glioma (HGG) and 76 patients had low-grade glioma (LGG). For each patient, we stacked three MRI sequences, which are T1-weighted, post–contrast-enhanced T1-weighted (T1C), and T2-weighted (T2) volumes. With the help of BraTS segmentations, we randomly extracted 20 slices per patient with the tumor region in axial plane. Our training dataset contained 203 patients, which corresponds to 2,927 slices (1,377 LGG and 1,550 HGG examples). 66 patients were included in the validation set (970 slices, 450 LGG and 520 HGG examples). Another 66 patients formed our test set (970 slices, 450 LGG and 520 HGG examples). LGG slices were labeled as 0 and HGGs were assigned to be 1. The images were resized to 224 × 224 pixels. **Figure 10** illustrates the results of classification over the dataset. cAUC directs the model to stop at epoch number 4 whereas both AUC and BCE would lead to the seventh epoch.

DISCUSSION

In this research, we first highlighted several important ROC and AUC characteristics. We demonstrated that to draw ROC curve, both actual positives and actual negatives are needed. Threshold equal to 1 corresponds to (0,0) in the ROC curve and $t = 0$ appears as (1,1). If a function is to calculate TPR, FPR or other metrics, it should iterate backward on the t values. The AUC is not concerned about confidence of the model. Regardless of N, if all the predictions are the same ($p_1 = p_2 = \ldots = p_N$), AUC will be 0.5 and the ROC curve will be a straight line from (0,0) to (1,1). Selecting more thresholds does not result in a smoother ROC or more accurate AUC. Thresholds must be selected from the set of the predicted probabilities plus 0 and 1. The order of predicted probabilities is correlated to the ROC shape and has a major impact on AUC. If there is at least a threshold where the probabilities of all actual positives and all actual negatives are above and below it, respectively, then the AUC is equal to 1. Conversely, the AUC will be 0 for the opposite case. The AUC does not differentiate FP from FN. All it does is scaling actual positive and actual negatives in a way that they have equal contributions to AUC. Therefore, the ROC curve should be used as the criterion and not AUC, if FP and FN have different weights. Because the final goal is classification, what is important is the performance of the model at a specific threshold. Therefore, there may be cases where a model with a lower AUC performs better at one threshold. The right approach is finding the optimum threshold from ROC and reporting the confusion matrix at that threshold.

The core of our research was the amendment of AUC in terms of margins. To add confidence to the optimized model, AUC needs to be refined. Using two coefficients, a revised AUC was proposed. Through simulations and mathematics, we showed the revised AUC reflects confidence of the model.

Unlike AUC, through experiments on MNIST, our PCa, and BraTS dataset, we demonstrated that local maximums in the proposed modified AUC correspond to local minimums of cross-entropy loss function. It was shown that selecting the best model based on cAUC is computationally efficient, mathematically reasonable, and it results in avoiding overfitting.

The conventional approach for when to stop training a CNN to achieve the highest AUC is to monitor the AUC while the model is being trained with a loss function such as BCE, and save the model whenever AUC breaks the previous highest score. However, when BCE is set to be used as the loss function, the hypothesis is that the best model has the lowest loss and therefore, the minimum loss is what the model is trained for. Hence, choosing the best model based on the highest AUC is not well rationalized and may not lead to the optimum point.

Our proposed metric inherits several limitations of the standard AUC and ROC but does not add any additional restrictions. Similar to AUC, cAUC is not differentiable and cannot be directly used as a loss function for training any NN. Additionally, calculating cAUC for a batch of data, especially if the batch size is small, will not help because it will be a measure of ranking in a small sample of the dataset. Similar to the standard AUC, cAUC does not give more importance to the positive examples.

CONCLUSION

Our results demonstrate the proposed cAUC is a better metric to choose the best performing model. On our MNIST-based dataset, when training a CNN, it results in stopping earlier which is computationally desirable. Moreover, it has landed in a less overfitting-prone area. Our results on the prostate MRI dataset are particularly interesting. With standard AUC we would stop training the CNN model at a suboptimal point with regards to BCE. With our proposed cAUC, we are able to stop at an optimal point where the training model gives the highest AUC. Our BraTS dataset experiments demonstrate cAUC can indicate optimum points that neither AUC nor BCE would direct the model towards them.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Three datasets have been used for the research. The MNIST and the BraTS datasets are publicly available. The prostate dataset analyzed in this research is available from the corresponding author on reasonable request pending the approval of the institution(s) and trial/study investigators who contributed to the dataset. Requests to access these datasets should be directed to mahaider@radfiler.com.

AUTHOR CONTRIBUTIONS

KN and FK contributed to the design of the concept and implementation of the algorithms. MH contributed in collecting and reviewing the data. All authors contributed to the writing and reviewing of the manuscript. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.582928/full#supplementary-material

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. *Sci. Data* 4, 170117. doi:10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge.

Bottou, L., Curtis, F. E., and Nocedal, J. (2016). Optimization Methods for Large-Scale Machine Learning. *SIAM Review* 60 (2), 223–311. doi:10.1137/16M1080173

Burke, H. B., Ph, D., Rosen, D. B., Ph, D., and Goodman, P. H. (1992). *Comparing Artificial Neural Nrworks to Other Statistical Memods for Medical Outcome Prediction*, 2213–2216.

Cortes, C., and Mohri, M. (2004). AUC Optimization vs. Error Rate Minimization. *Adv. Neural Inf. Process. Syst.*.

Ghanbari, H., and Scheinberg, K. (2018). Directly and Efficiently Optimizing Prediction Error and AUC of Linear Classifiers. [Online]. Available: http://arxiv.org/abs/1802.02535.

Hao, R., Namdar, K., Liu, L., Haider, M. A., and Khalvati, F. A. (2020). A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks. *J. Digit Imaging* 34 (4), 862–876. doi:10.1007/s10278-021-00478-7

Hao, R., Namdar, K., Liu, L., and Khalvati, F. (2021). A Transfer Learning-Based Active Learning Framework for Brain Tumor Classification. *Front. Artif. Intell.* 4, 61. doi:10.3389/frai.2021.635766

Kottas, M., Kuss, O., and Zapf, A. (2014). A Modified Wald Interval for the Area under the ROC Curve (AUC) in Diagnostic Case-Control Studies. *BMC Med. Res. Methodol.* 14 (1), 1–9. doi:10.1186/1471-2288-14-26

LeCun, Y., and Cortes, C. (2010). {MNIST} Handwritten Digit Database. [Online]. Available: http://yann.lecun.com/exdb/mnist/.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 34 (10), 1993–2024. doi:10.1109/TMI.2014.2377694

Parikh, K. S., and Shah, T. P. (2016). Support Vector Machine - A Large Margin Classifier to Diagnose Skin Illnesses. *Proced. Tech.* 23, 369–375. doi:10.1016/j.protcy.2016.03.039

Rosenfeld, N., Meshi, O., Tarlow, D., and Globerson, A. (2014). Learning Structured Models with the AUC Loss and its Generalizations. *J. Mach. Learn. Res.* 33, 841–849.

Sulam, J., Ben-Ari, R., and Kisilev, P. (2017). *"Maximizing AUC with Deep Learning for Classification of Imbalanced Mammogram Datasets," Eurographics Work.* Bremen, Germany: Vis. Comput. Biol. Med.. [Online]. Available: https://www.cs.bgu.ac.il/~rba/Papers/MaximizingAUC_MG.pdf.

Ying, Y., Wen, L., and Lyu, S. (2016). Stochastic Online AUC Maximization. *Adv. Neural Inf. Process. Syst. No. Nips*, 451–459.

Yoo, S., Gujrathi, I., Haider, M., and Khalvati, F. (2019). Prostate Cancer Detection Using Deep Convolutional Neural Networks. *Nat. Sci. Rep.*. doi:10.1038/s41598-019-55972-4

Yu, W., Chang, Y.-C. I., and Park, E. (2018). Applying a Modified AUC to Gene Ranking. *Csam* 25 (3), 307–319. doi:10.29220/CSAM.2018.25.3.307

Yu, W., and Park, T. (2014). AucPR: AucPR: An AUC-Based Approach Using Penalized Regression for Disease Prediction with High-Dimensional Omics Data. *BMC Genomics* 15 (Suppl. 10), 1–12. doi:10.1186/1471-2164-15-S10-S1

Zhang, Z., and Sabuncu, M. R. (2018).Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *Adv. Neural Inf. Process. Syst.*, 8778–8788.

Zhao, P., Hoi, S. C. H., Jin, R., and Yang, T. (2011). Online AUC Maximization. *Proc. 28th Int. Conf. Mach. Learn. ICML*, 233–240.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership