



OPEN ACCESS

EDITED BY

Sharon R. Pine,
University of Colorado Anschutz Medical
Campus, United States

REVIEWED BY

Chenbin Liu,
Chinese Academy of Medical Sciences and
Peking Union Medical College, China
Yingjian Yang,
Shenzhen Lanmage Medical Technology Co.,
Ltd, China

*CORRESPONDENCE

Jing Cai

✉ jing.cai@polyu.edu.hk

†PRESENT ADDRESS

Yuanpeng Zhang,
Department of Medical Informatics, Nantong
University, Nantong, Jiangsu, China

†These authors have contributed equally to
this work

RECEIVED 15 July 2024

ACCEPTED 27 October 2025

PUBLISHED 24 November 2025

CITATION

Wang Y, Zwanenburg A, Zhang J, Teng X,
Lam S-K, Cao J, Ma Z, Zhou T, Zhang Y, Ge H
and Cai J (2025) Establishment of a CT-based
radiomic feature robustness databank for
OPC patients via image perturbation in a
multi-institutional study: a practical method
to safeguard model generalizability.
Front. Oncol. 15:1464884.
doi: 10.3389/fonc.2025.1464884

COPYRIGHT

© 2025 Wang, Zwanenburg, Zhang, Teng, Lam,
Cao, Ma, Zhou, Zhang, Ge and Cai. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Establishment of a CT-based radiomic feature robustness databank for OPC patients via image perturbation in a multi-institutional study: a practical method to safeguard model generalizability

Yongqiang Wang^{1,2†}, Alex Zwanenburg^{3,4,5,6†}, Jiang Zhang¹,
Xinzhong Teng¹, Sai-Kit Lam^{7,8}, Jin Cao¹, Zongrui Ma¹, Ta Zhou¹,
Yuanpeng Zhang^{1†}, Hong Ge² and Jing Cai^{1,8,9*}

¹Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, ²Department of Radiation Oncology, The Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, Zhengzhou, China, ³National Center for Tumor Diseases Dresden/University Cancer Center (NCT/UCC), German Cancer Research Center Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany, ⁴Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden (TUD) Dresden University of Technology, Dresden, Germany, ⁵Department of Radiation Physics, Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, ⁶OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden (TUD) Dresden University of Technology, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany, ⁷Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, ⁸Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, ⁹The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

Purpose: To guide the preselection of highly repeatable radiomic features (RFs) in downstream analysis without further analysis its repeatability, a detailed radiomic feature robustness databank (RF-RobustDB) was established via image perturbation.

Methods: Data on 1,274 oropharyngeal carcinoma (OPC) patients who had undergone pretreatment computed tomography (CT) imaging, collected from a public dataset. The original images and corresponding masks underwent systematic perturbations to simulate potential variations encountered during CT image rescanning, including translational shifts, rotational changes, random noise additions, and contour modifications. For each radiomic feature (RF), including unfiltered, wavelet-filtered, and Laplacian-of-Gaussian (LoG)-filtered features, we systematically quantified robustness against these perturbations by intraclass correlation coefficients (ICCs).

Results: Out of 1395 first- and high-order RFs, 470 demonstrated excellent repeatability, i.e., a mean ICC of greater than 0.9. The use of these preselected highly repeatable RFs in model development improved the mean concordance (C) index in two external validation cohorts and reduced the mean C index gap between the training and external validation cohorts. These results demonstrate

that the preselected high repeatable RFs from RF-RobustDB can effectively enhance radiomic model generalizability.

Conclusions: The methodology employed to establish the RF-RobustDB is highly transferable to other tumor sites and different imaging modalities, which will facilitate the creation of RF-RobustDBs to guide the development of universally applicable radiomic models.

KEYWORDS

radiomics, feature repeatability, model generalizability, oropharyngeal carcinoma, progression-free survival

1 Introduction

The role and potential of radiomics in cancer management have been constantly expanding over the past decades, such as the distant metastases prediction of advanced nasopharyngeal carcinoma (NPC) (1), performing risk stratification of oropharyngeal cancer (OPC) (2), breast cancer risk estimation (3), and prediction of treatment response in non-small-cell lung cancer (NSCLC) (4). However, model generalizability remains the prime stumbling block for bend-to-bedside translation of radiomic models. To enhance the generalizability of radiomic models, concerted efforts have been made to enhance repeatability and reproducibility of radiomic features (RFs) for primary model generation (5–7). RF extraction, implemented prior to modeling process (8–11), is crucial for ensuring model reliability and generalizability. Although the Image Biomarker Standardization Initiative (IBSI) provides standardized guidelines for RF extraction (12), RF repeatability and reproducibility remain limited across institutions and imaging protocols (13–15). Consequently, these limitations represent fundamental challenges that need to be addressed before RFs can be effectively incorporated into modeling workflows.

Multiple variables influence the repeatability and reproducibility of RFs throughout the imaging process (13–27), such as scanner model (13, 14), scanner type (13, 15), scanning parameters (16), segmentation (17, 18), reconstruction (25), and preprocessing methods (26, 27). However, clear guidelines for selecting highly repeatable RFs in multi-institutional datasets remain unavailable. Test-retest methods pose additional challenges, as they may increase patients' radiation exposure and consume medical resources unnecessarily. Manual re-segmentation further burdens radiologists with additional workload. Although phantom-based studies offer a radiation-free alternative for evaluating RF selection (28), their clinical applicability is limited due to imperfect simulation of human tissues. Given the practical constraints of test-retest studies and manual re-segmentation across institutions, there is an urgent need for a cost-effective, efficient, easily implementable, and clinically transferable RF robustness assessment method. Fortunately, A

software-based image perturbation method proposed by Zwanenburg et al. offers a promising way to simulate the test-retest and re-segmentation process (29). This method simulates patient positioning during imaging, manual segmentation randomization, and varying noise levels of the imaging device. The effectiveness of image perturbation has been demonstrated by improved performance in radiomic models. For example, Teng et al. (30, 31) and Zhang et al. (6) applied image perturbation to select highly reproducible RFs that improved the reliability and generalizability of radiomic models. Moreover, image perturbation has been shown to achieve the same optimal reliability as test-retest imaging for constructing radiomic models (32).

Since the perturbation method demonstrates encouraging/promising capabilities in assessing feature stability, in this study, we aim to establish a reliable RF robustness databank (RF-RobustDB) via perturbation method for guiding the downstream development of radiomic models. Specifically, we included a large-scale of CT images of OPC patients from a total of 7 medical institution. The OPC dataset was obtained from the Cancer Imaging Archive (TCIA) (33). RFs from CT images with and without applications of popular imaging filters were analyzed. The repeatability of the RFs in the RF-RobustDB was quantified by one-way intra-class correlation coefficients (ICCs) (34). We adopted CT dataset for this study mainly owing to its widespread popularity in the cancer management for pre-treatment planning, mid-treatment monitoring, and post-treatment evaluation, as well as the availability of dataset in the community.

Through systemic analysis, the cohort size effects on feature repeatability ensured that the sample size is sufficient to maintain the reliability of RF-RobustDB. Meanwhile, the RF-RobustDB-enhanced selection of highly repeatable RFs significantly improved the generalizability of progression-free survival (PFS) model. These results support the reliability of the comprehensive CT-based RF-RobustDB for OPC, offering a valuable insight into RF repeatability. Moreover, this study provided a comprehensive and generalized methodology for establishing an extensive RF-RobustDB applicable to diverse tissue sites and imaging modalities.

2 Materials and methods

2.1 Patient cohort

This retrospective study analyzed a dataset of pretreatment CT images from 1,418 head-and-neck cancer patients obtained from TCIA (33). The dataset included patients from seven medical institutions: 137 patients from the single-institution HEAD-NECK-RADIOMICS-HN1 (HN1) study (35, 36), 524 patients from the single-institution Radiomic Biomarkers in Oropharyngeal Carcinoma (RBOPC) study (37, 38), 298 patients from four institutions in the Head-Neck-PET-CT (HNPET) study (39, 40), and 459 patients from the single-institution Head and Neck Squamous Cell Carcinoma (HNSCC) study (41–43). To maintain consistency, only OPC patients with primary gross tumor volume (GTV) data were included, resulting in a final cohort of 1,274 OPC patients for establishing the site-specific RF-RobustDB. The baseline characteristics of the selected OPC patient are presented in Table A1. Due to the retrospective nature of this study, informed consent was not required.

2.2 Image perturbation

To simulate the inevitable variabilities in patient setup during image acquisition, a validated image perturbation method was used to mimic patient setup, randomized noise, and manual segmentation diversity. Translational and rotational perturbations were applied to the original (unfiltered) images and tumor masks to mimic patient position. Randomized noise was added to the original images to simulate noise variations during image acquisition. Contour randomization was applied to the tumor mask to mimic variations in manual tumor segmentation.

The image perturbation settings were based on previous studies on repeatability evaluation via image perturbation (6, 29, 31): translation distances were set to 0, 0.4, and 0.8 pixels; rotation angles were set to -20°, 0°, and 20°; noise levels were increased to 0, 1, 2, and 5 times the original noise level; and a three-dimensional

random displacement field was used to deform segmented masks, resulting in randomized contours. For each voxel point, a random field vector component in each dimension is generated from a uniform distribution between -1 and 1. All z-components of the field vectors on the same slice are set to the same value to mimic the uniform inter-slice contour variations resulting from slice-by-slice contouring. The field vectors are then normalized in each dimension by the root mean square. Finally, they are smoothed using a Gaussian filter with a defined sigma value of 10 to ensure continuous changes in the random displacement field and to avoid sharp changes in the deformed contours. Sixty different perturbations were performed to enhance the reliability of our results, as previous studies have suggested that 40 different perturbations are sufficient (6, 30). During each perturbation operation, parameters from the four perturbation modes were randomly combined to simulate the uncertainty in variables during image rescanning.

2.3 RF extraction

Image pre-processing and RF extraction were conducted in accordance with the IBSI guidelines (12). Before RF extraction, all images were resampled to a $1 \times 1 \times 1 \text{ mm}^3$ resolution, and re-segmentation was performed to limit pixel values between -150 and 180 HU, effectively excluding non-tumor tissue (such as air and bone) within the volume of interest (31). As a previous study suggested, using the fixed bin number between 8 and 128 discretize images can reduce the infinite possible number of intensity values to a finite set and image noise (44). Hence, a fixed bin number of 30 was used for image discretization in this study. RF extraction was performed using PyRadiomics v2.2 (45) in Python v3.7. Shape-based features, first-order features, and high-order features from the gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), gray-level dependence matrix (GLDM), and neighboring gray-tone difference matrix (NGTDM) were extracted from the GTVs in original, Laplacian-of-Gaussian (LoG)-filtered (with sigma values of 1, 2, 3, 4, 5, and 6 mm) and Coiflet-1 wavelet-filtered images.

Fourteen shape-based features were extracted from each tumor mask, and 93 first-order and high-order features were extracted from each of the unfiltered, LoG-filtered, and wavelet-filtered images. Following image perturbation, we additionally extracted corresponding feature sets from all perturbed image variants. Finally, the RF-RobustDB contained 1,316 unfiltered, LoG-filtered, and wavelet-filtered RFs, and 78,960 perturbed features were extracted for ICC analysis.

2.4 RF repeatability assessment

Since the feature is extracted from different perturbed mask region, the assignment of perturbation parameters is independent to patients. Therefore, the robustness of each RF was quantified in

TABLE A1 Baseline patient characteristics of the dataset in different cohorts.

Data cohort	Sex	Median age	Overall stage
HNSCC	Male: 395	57(28-87)	I-IV
	Female:64		
HN1	Male: 67	60(44-80)	I-IVb
	Female:21		
RBOPC	Male: 423	60(33-89)	I-IVb
	Female:101		
HNPET	Male: 151	63(34-90)	I-IVb
	Female:52		

terms of a one-way, random, absolute-agreement ICC, which was calculated using Equation 1, as follows (34).

$$ICC(1, 1) = \frac{MS_n - MS_W}{MS_n + (k+1)MS_W} \quad (1)$$

where MS_n is the mean square for different patients, MS_W is the mean square for residual sources of variance, and k is the number of perturbation times plus one for the unperturbed image. As recommended by a previous study (34), features with an $ICC < 0.5$ were regarded as having poor repeatability, ICC between 0.50 and 0.75 were regarded as having moderate repeatability, ICC between 0.75 and 0.90 were regarded as having good repeatability, and $ICC > 0.9$ were regarded as having excellent repeatability.

2.5 Establishment of the RF-RobustDB

To determine the reliable patient sample size (n) required for constructing the RF-RobustDB, the mean ICC values as a function of patient numbers were systematically analyzed. The methodology

was implemented as follows: Starting with 10 patients, the sample size was incrementally increased by 10 patients up to 100, followed by 100-patient increments thereafter. For each specific patient-number subgroup, 10 rounds of random resampling were performed from the oropharyngeal carcinoma (OPC) datasets to calculate the corresponding mean ICC values. The six radiomic features (RFs) demonstrating the greatest variations in mean ICC values across different sample sizes were selected to illustrate the sample size dependency of ICC metrics. Based on this analysis, a patient number that showed a stabilized trend in mean ICC s was identified and ultimately used for establishing the RF-RobustDB.

2.6 RF selection and PFS model development

To demonstrate the efficacy of the RF-RobustDB, six PFS models were constructed using RFs selected through different ICC thresholds: (1) $ICC > 0.9$, (2) $ICC > 0.85$, (3) $ICC > 0.8$, (4) $ICC > 0.75$, (5) $ICC > 0.5$, and (6) non-preselected RFs. PFS events were

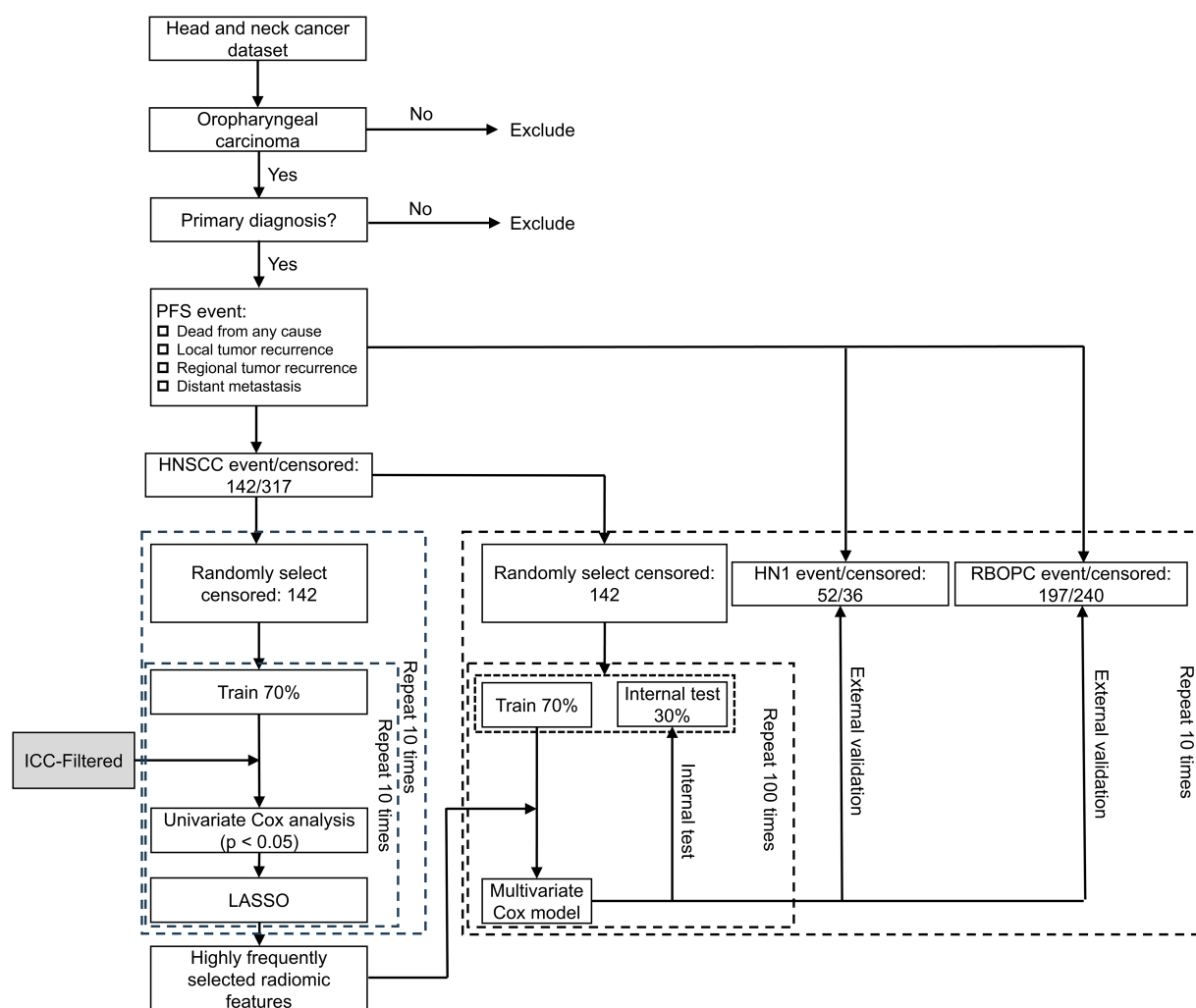


FIGURE 1

Workflow used for building progression-free survival (PFS) models based on CT images of oropharyngeal carcinoma patients. ICC = intraclass correlation coefficient; LASSO = least absolute shrinkage and selection operator; RF = radiomic feature; HNSCC = Head and Neck Squamous Cell Carcinoma; HN1 = HEAD-NECK-RADIOMICS-HN1; RBOPC = Radiomic Biomarkers in Oropharyngeal Carcinoma.

defined as local/regional recurrence, distant metastasis, or death from any cause. Figure 1 demonstrate the complete feature selection and modeling workflow. For the feature selection procedure, the highly reliable RFs derived from unperturbed images were initially chosen based on their robustness, as defined by the mean ICC. Subsequently, univariate Cox analysis was utilized to identify the RFs associated with PFS events within the pre-selected RFs in the training group. RFs with a *p*-value of less than 0.05 were considered significant. Finally, the least absolute shrinkage and selection operator algorithm was employed to select RFs with non-zero coefficients in the training group. The training groups were randomly bootstrapped 10 times from the 10 resampled balanced HNSCC dataset. The features that appeared frequently were selected to construct the PFS models.

To determine the optimal number of RFs for modeling, the relationship between the feature number and model performance was systematically investigated in the HNSCC dataset (Figure A1). The results revealed that the model constructed using non-preselected RFs exhibited optimal performance in the internal testing group when the feature number reached five (Figure A1(A)). Specifically, the mean C index exhibited the highest value in the testing cohorts, indicating the superior predictive ability of the model at this feature threshold. Moreover, using a greater number of RFs resulted in larger mean C index gaps between the training and internal testing groups. Similar trends of optimal performance were observed in the other five experiments (Figure A1(B)-(F)). Therefore, the top five RFs that most frequently appeared in all experiments were ultimately selected to ensure that the feature number would not introduce conflicts into the final results. The selected RFs for each experiment are listed in Table A2.

Multivariate Cox regression was employed to model the survival risks for PFS in the HNSCC dataset. The performance of the developed PFS models was evaluated by concordance (C) index for the training, internal testing, and external validation (HN1, RBOPC) cohorts. The HNPET dataset was excluded from external validation due to insufficient follow-up data on local/regional recurrence and distant metastasis. To assess the robustness of the models, the mean C index and its 95% confidence intervals were calculated in 100 bootstrap experiments on the 10 randomly resampled balanced datasets.

2.7 Model generalizability assessment

To assess model generalizability across external validation (EV) cohorts, a generalizability index (G) that quantifies the absolute difference in C index values between training and EV groups. The G index is defined by Equation 2:

$$G = \sum_m^M \frac{|C_{Train} - C_{EV_m}|}{M} \quad (2)$$

where C_{Train} represents the C index of the trained model, C_{EV_m} is the C index for the m^{th} EV cohort, and M is the total number of EV cohorts. Lower G-index values indicate superior model generalizability, reflecting smaller performance discrepancies

between training and validation datasets. The mean G index and corresponding 95% confidence intervals were calculated from 1,000 cross validation models to comprehensively evaluate the distribution of model generalizability performance.

3 Results

3.1 Patient-number dependence analysis

Figure 2 depicts the relationship between the number of patients and the mean ICCs of six RFs, which were selected as the top six RFs exhibiting the most significant variance changes with varying patient numbers. As the number of patients increased, the mean ICCs of the *Firstorder_Maximum* and *GLCM_ClusterTendency* features initially increased and then stabilized. In contrast, the mean ICCs of the remaining four selected features first fluctuated before eventually stabilizing. All six features tended to stabilize once the patient count reached 200. This stabilization trend of the mean ICCs demonstrates that the patient sample size used in this study was sufficient for constructing the RF-RobustDB. Specifically, 800 resampled patients were subjected to 100 iterations of resampling to compute the mean ICCs and their corresponding 95% confidence intervals.

3.2 Establishment and validation of the RF-RobustDB

3.2.1 The repeatability of shape-based RFs

Figure 3 presents a comprehensive evaluation of the ICCs for shape-based RFs across multiple simulated test conditions. Notably, all shape-based features demonstrated consistently high repeatability, with mean ICC values surpassing the 0.9 threshold (range: 0.955-0.999) across various perturbation scenarios. This robust performance supports the clinical applicability of shape-based RFs, as their measurements consistently reflect the tumor's shape characteristics and are less influenced by other clinical factors.

3.2.2 The repeatability of first-order and textural RFs

Figure 4 presents the ICCs of the first-order and textural RFs. The left panel of Figure 4 displays the mean ICCs and their 95% confidence intervals for each unfiltered RF. The right panel of Figure 4 gives the mean ICCs of the unfiltered, wavelet-filtered and LoG-filtered RFs. Collectively, Figures 3 and 4 establish the RF-RobustDB. Detailed information on the mean ICCs and their 95% confidence intervals can be found in the Appendix.

3.2.3 RF-RobustDB help improving the generalizability of PFS models

Figure 5A, B illustrates the selection of highly reproducible radiomic features (RFs) using the established RF-RobustDB. Figure 5A quantifies the absolute counts, while Figure 5B presents the relative proportions of these reproducible RFs across six feature

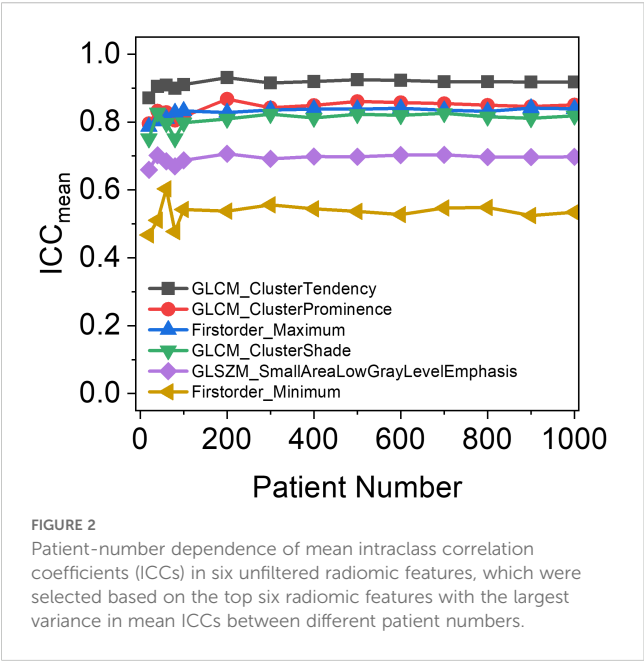
TABLE A2 Final selected radiomic features for each PFS model.

ICC threshold	Feature name
ICC>0	log-sigma-2-mm-3D_glszm_LowGrayLevelZoneEmphasis_30_binCount
	log-sigma-5-mm-3D_glszm_LowGrayLevelZoneEmphasis_30_binCount
	log-sigma-6-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis_30_binCount
	wavelet-LLH_firstorder_Skewness_30_binCount
	log-sigma-3-mm-3D_glszm_ZoneEntropy_30_binCount
ICC>0.5	wavelet-HHH_firstorder_Energy_30_binCount
	wavelet-LHH_firstorder_Energy_30_binCount
	log-sigma-5-mm-3D_firstorder_Energy_30_binCount
	log-sigma-4-mm-3D_firstorder_Energy_30_binCount
	wavelet-LLH_firstorder_Energy_30_binCount
ICC>0.75	wavelet-LHH_glrlm_GrayLevelNonUniformityNormalized_30_binCount
	wavelet-LHH_glcml_Idn_30_binCount
	log-sigma-6-mm-3D_glcml_ClusterShade_30_binCount
	original_glcml_Correlation_30_binCount
	wavelet-LLL_glcml_JointAverage_30_binCount
ICC>0.8	wavelet-LHH_glcml_Idn_30_binCount
	log-sigma-6-mm-3D_glcml_ClusterShade_30_binCount
	log-sigma-5-mm-3D_glszm_ZoneEntropy_30_binCount
	original_glcml_Correlation_30_binCount
	log-sigma-4-mm-3D_glszm_SizeZoneNonUniformity_30_binCount
ICC>0.85	wavelet-LHH_gldm_SmallDependenceEmphasis_30_binCount
	wavelet-LLL_glrlm_ShortRunHighGrayLevelEmphasis_30_binCount
	wavelet-LLL_firstorder_Range_30_binCount
	log-sigma-2-mm-3D_firstorder_90Percentile_30_binCount
	log-sigma-6-mm-3D_glrlm_RunLengthNonUniformity_30_binCount
ICC>0.9	log-sigma-1-mm-3D_firstorder_Maximum_30_binCount
	log-sigma-2-mm-3D_glszm_SizeZoneNonUniformity_30_binCount
	original_glcml_Correlation_30_binCount
	log-sigma-3-mm-3D_glszm_SizeZoneNonUniformity_30_binCount
	log-sigma-6-mm-3D_glcml_DifferenceVariance_30_binCount

classes: first-order, GLCM, GLRLM, GLSZM, GLDM, and NGTDM. Among 1,395 first-order and high-order RFs analyzed, 470 features (33.7%) demonstrated excellent repeatability, defined by mean ICCs > 0.9. Notably, unfiltered and LoG-filtered features exhibited significantly higher repeatability rates compared to wavelet-filtered features.

Using the established RF-RobustDB, six PFS models were systematically evaluated by recommended RFs at varying mean ICC thresholds. The C index in the training and internal testing groups first decreased and then increased as the ICC increased, as the red and black curve shown in Figure 5C. However, it is

noteworthy that the EV cohorts (HN1 and RBOPC) demonstrated superior discriminative performance for models constructed using ICC preselected RFs, as shown by the pink and blue lines in Figure 5C. In addition, there was a smaller C index gap between the training and EV cohorts in the mean ICC preselected RF groups than in the non-preselected RF group, as shown by the G index trend in Figure 5D. The smallest mean G index was obtained in the RF groups preselected with a mean ICC of 0.5, and the largest mean G index was obtained in the non-preselected RFs groups. By taking into account the large C index and small G index of the RFs preselected with mean ICCs, it was found that the PFS models



incorporating preselected RFs demonstrated significantly higher generalizability compared to those using non-preselected RFs.

4 Discussion

4.1 Result analysis

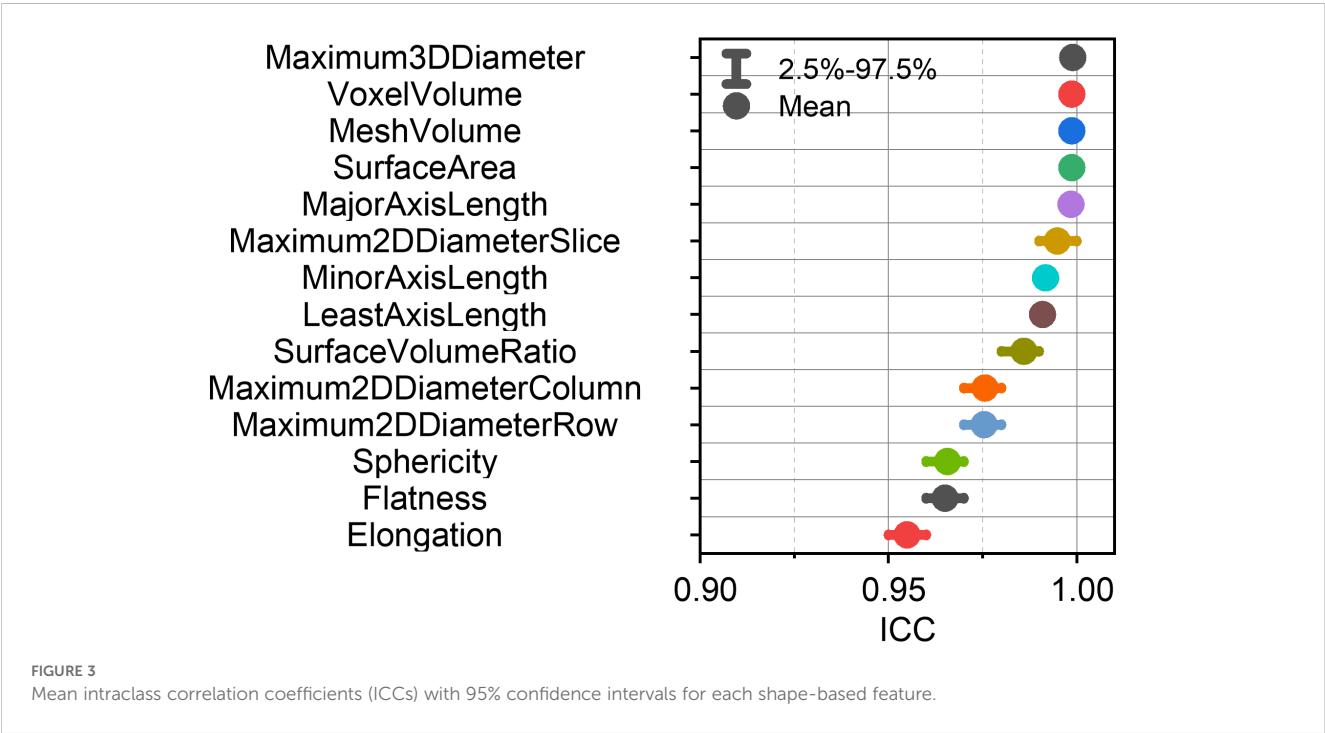
A robust RF-RobustDB of pretreatment CT-derived RFs in OPC patients was established through image perturbation. This database effectively guides the preselection of repeatable RFs and

enhances the generalizability of multi-cohort PFS studies. A new G index was introduced to quantitatively evaluate the generalizability of the constructed PFS models. The methodology developed in this study can be easily extended to other anatomical sites and imaging modalities, providing a feasible solution for establishing standardized RF-RobustDBs to comprehensively assess RF repeatability across various clinical scenarios.

Our analysis of mean ICC dependence on patient cohort size (Figure 2) revealed that the six selected unfiltered RFs showing the highest variance across different sample sizes achieved stabilization when the patient number exceeded 200. This suggests that a minimum of 200 patients provides sufficient data for reliable ICC-based assessment of RF robustness, confirming that the sample size in this study (1,274) ensured the reliability of the RF-RobustDB. The RF-RobustDB evaluation demonstrated significant differences in feature repeatability between filtering methods: only 11.7% of wavelet-filtered RFs exhibited excellent repeatability ($ICC > 0.9$), compared to 56.3% of Laplacian-of-Gaussian (LoG)-filtered RFs. This substantial disparity (44.6%) establishes the superior robustness of LoG-filtered features. The low repeatability of wavelet-filtered RFs likely stems from the characteristics of wavelet filtering, image resampling strategies, and perturbation settings (6). Therefore, radiomic model construction requires more stringent selection criteria for wavelet-filtered RFs compared to their LoG-filtered counterparts.

4.2 RF-RobustDB reliability analysis

The clinical utility of the RF-RobustDB was evaluated through external validation using two independent cohorts derived from separate institutions. Importantly, these validation cohorts were



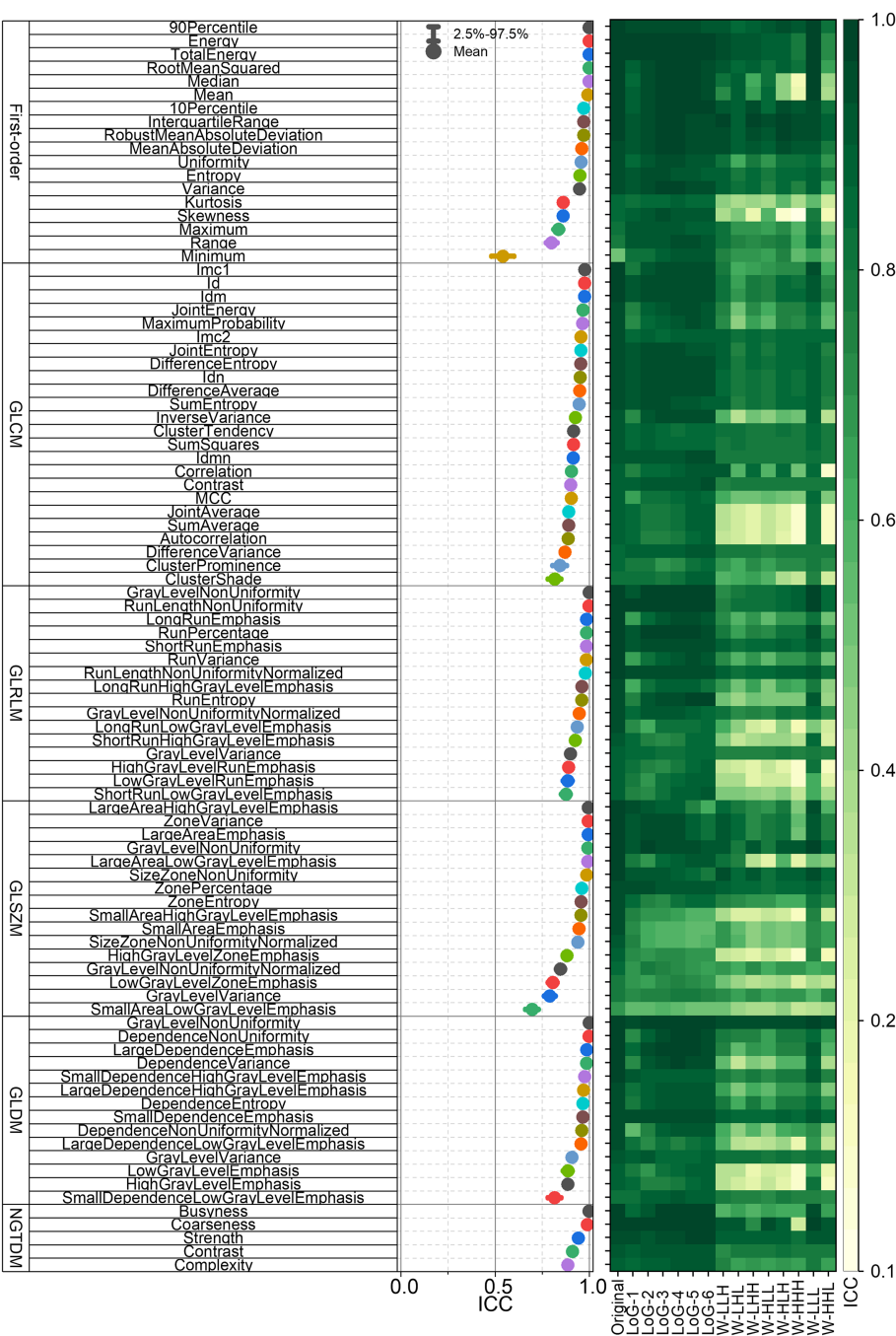


FIGURE 4 Intraclass correlation coefficients (ICCs) of first-order and textural radiomic features. GLCM = gray-level co-occurrence matrix; GLRLM = gray-level run-length matrix; GLSZM = gray-level size-zone matrix; GLDM = gray-level dependence matrix; NGTDM = neighboring gray-tone difference matrix.

exclusively utilized for feature selection and model training phases, thereby maintaining the integrity of the validation process. Comparative analysis revealed that models incorporating RF-RobustDB-preselected features demonstrated superior performance in external validation, as evidenced by the higher mean concordance indices compared to models using non-preselected features and a reduction in the performance gap between training and validation cohorts. Zhang and colleagues improved the generalizability of a disease-free survival model for head and neck cancer by pre-selecting highly reproducible RFs using the perturbation method (6). The study by Gong et al. also provides compelling evidence supporting the critical importance of

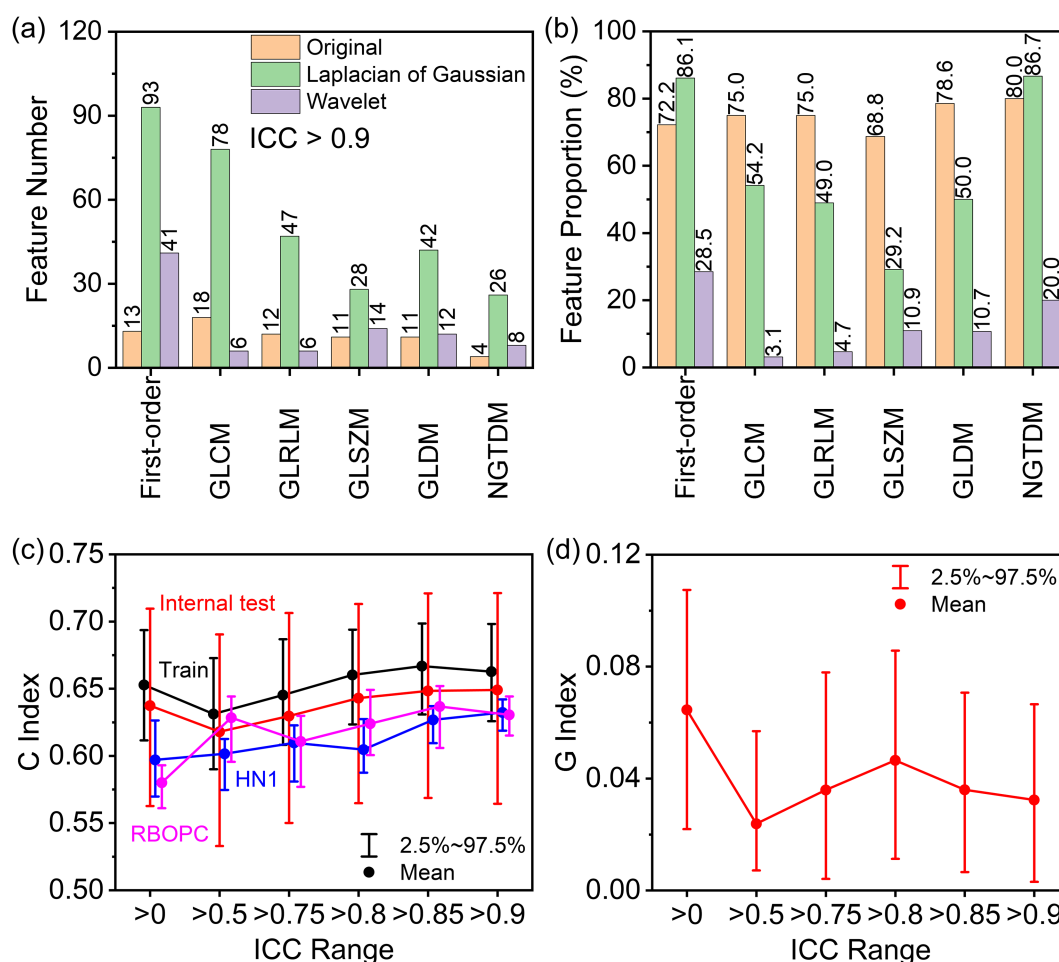


FIGURE 5

(A) Numbers and (B) proportions of excellent repeatable radiomic features in first-order, gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), gray-level dependence matrix (GLDM), and neighboring gray-tone difference matrix (NGTDM) features selected based on mean ICCs. (C) Performance of progression-free survival (PFS) models built using preselected highly repeatable radiomic features with different mean intraclass correlation coefficient (ICC) thresholds in the training (Train), internal testing (Test), and two external validation (HN1 and RBOPC) cohorts. (D) Generalizability index (G) of the PFS models across the training and EV groups based on the concordance (C) index gap.

feature stability in radiomic analyses (46). Through systematic perturbation analysis of CT-derived imaging features in esophageal squamous cell carcinoma, their findings substantiate that incorporating high-stability features can enhance model performance in the external validation set. Thomas Louis et al. reported that robust features demonstrated superior predictive potential compared to non-robust features in predicting the outcomes of an external validation dataset (47). These studies collectively highlight the critical importance of feature stability for model generalizability, demonstrating findings consistent with our own results. This convergence of evidence further substantiates the fundamental value of establishing comprehensive RF-RobustDB to support subsequent radiomics research.

The reproducibility of RFs in clinical practice is subject to multiple influencing factors, such as patient positioning variability, segmentation quality, the noise level of medical imaging devices, and

the variations between the performance of devices depending on their model and vendor, leading to differences in CT number values (13). Additionally, the reproducibility of RFs can be affected by scanning parameters and reconstruction algorithms (16, 25). This multifactorial variability explains the observed reduction reliability and generalizability when applying RFs in multi-cohort studies versus single-cohort studies. However, unlike variations in CT scanners or scanning modalities, variations in patient positioning, segmentation, and random noises are similar in various clinical circumstances, which allows perturbation methods to be a universal tool for assessing RF repeatability. Therefore, employing image perturbation across multiple institutions is a feasible approach for evaluating the robustness of RFs, as demonstrated by the enhanced generalizability of our PFS models. Considering these factors, the methodology employed in this study is both feasible and applicable for establishing RF-RobustDBs for other tumor sites and imaging

modalities. The methodology is a promising approach for assessing the repeatability and enhancing the generalizability of radiomic models, thereby facilitating the development of more reliable and robust radiomic models with enhanced clinical translatability.

4.3 Existing limitations analysis

This study has several limitations that should be acknowledged. First, although our perturbation method simulated key variability sources including translations, rotations, random noise, and contour variations, they could not fully replicate all potential sources of variability encountered in clinical practice. For example, transient signal fluctuations that may occur during repeated scans under identical acquisition parameters cannot be effectively modeled. This inherent limitation underscores that image perturbation methods cannot entirely replace traditional test-retest validation approaches. Second, the optimal ICC thresholds for establishing reliable radiomic models remain controversial in the field. Additional investigations are required to establish evidence-based cutoff values for robust feature selection in clinical applications. Third, while our multi-institutional study design strengthened the generalizability of findings, the persistent effects of inter-scanner variability and acquisition parameter differences on feature reproducibility warrant further investigation. Furthermore, the dual use of our dataset for both RF-RobustDB construction and PFS model development may introduce circularity. Future validation should incorporate independent multi-institutional datasets to more rigorously assess the RF-RobustDB's clinical utility. Addressing these limitations through continued research will be essential for optimizing the RF-RobustDB's performance and expanding its applicability across diverse clinical implementations.

5 Conclusion

We have established a RF-RobustDB using an image perturbation approach for CT-derived RFs in OPC patients. The ICCs were calculated to quantify the reliability and repeatability of RFs. Through multi-cohort PFS experiments, we demonstrated the reference value of the RF-RobustDB, demonstrating that preselected highly repeatable RFs improved PFS model generalizability. To quantitatively assess model performance, we introduced a generalizability metric (G-index). The methodology we employed is cost-effective and easily applicable across different institutions, suggesting its potential extension to other lesion areas and imaging modalities. The comprehensive RF-RobustDB can facilitate robust RF selection when only small training datasets or single-institutional data are available, thereby enhancing the reliability, reproducibility, and generalizability of radiomic predictive models.

Data availability statement

The raw image data is available in the cancer image archive (<https://www.cancerimagingarchive.net/collection/head-neck-radiomics-hn1/>, <https://www.cancerimagingarchive.net/collection/head-neck-pet-ct/>, <https://www.cancerimagingarchive.net/collection/hnsccl/>, <https://www.cancerimagingarchive.net/collection/opc-radiomics/>).

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

YW: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. AZ: Formal Analysis, Writing – review & editing. JZ: Data curation, Methodology, Software, Writing – review & editing. XT: Data curation, Methodology, Writing – review & editing. S-KL: Formal Analysis, Writing – review & editing. JiC: Software, Visualization, Writing – review & editing. ZM: Methodology, Visualization, Writing – review & editing. TZ: Visualization, Writing – review & editing. YZ: Visualization, Writing – review & editing. HG: Supervision, Writing – review & editing. JCa: Writing – review & editing, Conceptualization, Formal Analysis, Funding acquisition, Project administration, Resources, Supervision.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. The Mainland-Hong Kong Joint Funding Scheme (MHKJFS) (MHP/005/20) support evaluating the robustness of radiomic features, Shenzhen Basic Research Program (JCYJ20210324130209023) of Shenzhen Science and Technology Innovation Committee support development of the feature perturbation method, Project of Strategic Importance Fund (P0035421) and Projects of RISA (P0043001) of The Hong Kong Polytechnic University, and Health and Medical Research Fund (HMRF 09200576), the Health Bureau, the Government of the Hong Kong Special Administrative Region supports the study design, data collection, data analysis, and interpretation of the results.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

- Zhang Y, Lam S, Yu T, Teng X, Zhang J, Lee FK, et al. Integration of an imbalance framework with novel high-generalizable classifiers for radiomics-based distant metastases prediction of advanced nasopharyngeal carcinoma. *Knowl Based Syst.* (2022) 235:107649. doi: 10.1016/j.knsys.2021.107649
- Wang P, Wang X, Zhang M, Li G, Zhao N, Qiao Q. Combining the radiomics signature and HPV status for the risk stratification of patients with OPC. *Oral Dis.* (2024) 30:272–80. doi: 10.1111/odi.14386
- Acciavatti RJ, Cohen EA, Maghsoudi OH, Gastounioti A, Pantalone L, Hsieh M-K, et al. Incorporating robustness to imaging physics into radiomic feature selection for breast cancer risk estimation. *Cancers.* (2021) 13:5497. doi: 10.3390/cancers13215497
- Chetan MR, Gleeson FV. Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. *Eur Radiol.* (2021) 31:1049–58. doi: 10.1007/s00330-020-07141-9
- Cheong E-N, Park JE, Park SY, Jung SC, Kim HS. Achieving imaging and computational reproducibility on mul-tiparametric MRI radiomics features in brain tumor diagnosis: phantom and clinical validation. *Eur Radiol.* (2023) 34:2008–23. doi: 10.1007/s00330-023-10164-7
- Zhang J, Lam S-K, Teng X, Ma Z, Han X, Zhang Y, et al. Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on naso-pharyngeal carcinoma patients. *Radiol Oncol.* (2023) 183:109578. doi: 10.1016/j.radonc.2023.109578
- Park JiE, Park SY, Kim HJ, Kim HoS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol.* (2019) 20:1124–37. doi: 10.3348/kjr.2018.0070
- Sollini M, Antunovic L, Chiti A, Kirienko M. Towards clinical application of image mining: A systematic review on artificial intelligence and radiomics. *Eur J Nucl Med Mol Imaging.* (2019) 46:2656–72. doi: 10.1007/s00259-019-04372-x
- Koçak B, Durmaz E.Ş, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: A practical guide for beginners. *Diagn Interv Radiol.* (2019) 25:485–95. doi: 10.5152/dir.2019.19321
- Zhang Y-P, Zhang X-Y, Cheng Y-T, Li B, Teng X-Z, Zhang J, et al. Arti-ficial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. *Military Med Res.* (2023) 10:22. doi: 10.1186/s40779-023-00458-8
- Chaddad A, Liang X. Stability of radiomic models and strategies to enhance reproducibility. *IEEE Trans Radiat Plasma Med Sci.* (2024) 8:540–55. doi: 10.1109/TRPMS.2024.3365778
- Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* (2020) 295:328–38. doi: 10.1148/radiol.2020191145
- Chen Y, Zhong J, Wang L, Shi X, Lu W, Li J, et al. Robustness of CT radiomics features: consistency within and between single-energy CT and dual-energy CT. *Eur Radiol.* (2022) 32:5480–90. doi: 10.1007/s00330-022-08628-3
- Rinaldi L, De Angelis SP, Raimondi S, Rizzo S, Fanciullo C, Rampinelli C, et al. Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters. *Eur Radiol Exp.* (2022) 6:2. doi: 10.1186/s41747-021-00258-6
- Pandey U, Saini J, Kumar M, Gupta R, Ingahlhalikar M. Normative baseline for radiomics in brain MRI: evaluating the robustness, regional variations, and reproducibility on FLAIR images. *J Magn Reson Imaging.* (2021) 53:394–407. doi: 10.1002/jmri.27349
- Muenzfeld H, Nowak C, Riedlberger S, Hartenstein A, Hamm B, Jahnke P, et al. Intra-scanner repeatability of quantitative imaging features in a 3D printed semi-anthropomorphic CT phantom. *Eur J Radiol.* (2021) 141:109818. doi: 10.1016/j.ejrad.2021.109818
- Haniff NSM, Abdul Karim MK, Osman NH, Saripan MI, Che Isa IN, Ibadim MJ. Stability and reproducibility of radiomic features based various segmentation technique on MR images of hepatocellular carcinoma (HCC). *Diagnostics.* (2021) 11:1573. doi: 10.3390/diagnostics11091573
- Tunali I, Hall LO, Napel S, Cherezov D, Guvenis A, Gillies RJ, et al. Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions. *Med Phys.* (2019) 46:5075–85. doi: 10.1002/mp.13808
- Wang H, Zhou Y, Wang X, Zhang Y, Ma C, Liu Bo, et al. Reproducibility and repeatability of CBCT-derived radiomics features. *Front Oncol.* (2021) 11:773512. doi: 10.3389/fonc.2021.773512
- Jha AK, Mithun S, Jaiswar V, Sherkhane UB, Purandare NC, Prabhaskar K, et al. and reproducibility study of radiomic features on a phantom and human cohort. *Sci Rep.* (2021) 11:2055. doi: 10.1038/s41598-021-81526-8
- Bologna JM, Tenconi C, Corino VDA, Annunziata G, Orlandi E, Calareso G, et al. Repeatability and reproducibility of MRI-radiomic features: A phantom experiment on a 1.5 T scanner. *Med Phys.* (2023) 50:750–62. doi: 10.1002/mp.16054
- Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy Oncol.* (2019) 135:107–14. doi: 10.1016/j.radonc.2019.03.001
- Kocak B, Durmaz ES, Kaya OK, Ates E, Kiliçkesmez O. Reliability of single-slice-based 2D CT texture analysis of renal masses: influence of intra- and interobserver manual segmentation variability on radiomic feature reproducibility. *Am J Roentgenol.* (2019) 213:377–83. doi: 10.2214/AJR.19.21212
- Traverso A, Wee L, Dekker A, Robert G. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys.* (2018) 102:1143–58. doi: 10.1016/j.ijrobp.2018.05.053
- Escudero Sanchez L, Rundo L, Gill AB, Hoare M, Mendes Serrao E, Sala E. Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Sci Rep.* (2021) 11:8262. doi: 10.1038/s41598-021-87598-w
- Li Y, Tan G, Vangel M, Hall J, Cai W. Influence of feature calculating parameters on the reproducibility of CT ra-diomic features: A thoracic phantom study. *Quant Imaging Med Surg.* (2020) 10:1775–85. doi: 10.21037/qims-19-921
- Scalo E, Belfatto A, Mastropietro A, Rancati T, Avuzzi B, Messina A, et al. T2w-MRI signal nor-malization affects radiomics features reproducibility. *Med Phys.* (2020) 47:1680–91. doi: 10.1002/mp.14038
- Euler A, Laqua FC, Cester D, Lohaus N, Sartoretto T, Pinto dos Santos D, et al. Virtual monoenergetic images of dual-energy CT—Impact on repeatability, reproducibility, and classification in radiomics. *Cancers.* (2021) 13:4710. doi: 10.3390/cancers13184710

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2025.1464884/full#supplementary-material>

29. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. *Sci Rep.* (2019) 9:614. doi: 10.1038/s41598-018-36938-4
30. Teng X, Zhang J, Zwanenburg A, Sun J, Huang Y, Lam S, et al. Building reliable radiomic models using image perturbation. *Sci Rep.* (2022) 12:10035. doi: 10.1038/s41598-022-14178-x
31. Teng X, Zhang J, Ma Z, Zhang Y, Lam S, Li W, et al. Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. *Front Oncol.* (2022) 12:974467. doi: 10.3389/fonc.2022.974467
32. Zhang J, Teng X, Zhang X, Lam S-K, Lin Z, Liang Y, et al. Comparing effectiveness of image perturbation and test retest imaging in improving radiomic model reliability. *Sci Rep.* (2023) 13:18263. doi: 10.1038/s41598-023-45477-6
33. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7
34. Koo TK, Li MYA. Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* (2016) 15:155–63. doi: 10.1016/j.jcm.2016.02.012
35. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* (2014) 5:4006. doi: 10.1038/ncomms5006
36. Wee L, Dekker A. Data from HEAD-NECK-RADIOMICS-HN1 [Data set]. *Cancer Imaging Arch.* (2019). doi: 10.7937/tcia.2019.8kap372n
37. Kwan JYY, Su J, Huang SH, Ghoraie LS, Xu W, Chan B, et al. Radiomic biomarkers to refine risk models for distant metastasis in HPV-related oropharyngeal carcinoma. *Int J Radiat Oncol Biol Phys.* (2018) 102:1107–16. doi: 10.1016/j.ijrobp.2018.01.057
38. Kwan JYY, Su J, Huang SH, Ghoraie LS, Xu W, Chan B, et al. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma. *Cancer Imaging Arch.* (2019). doi: 10.7937/tcia.2019.8dho2gls
39. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep.* (2017) 7:10117. doi: 10.1038/s41598-017-10371-5
40. Vallières M, Kay-Rivest E, Perrin Léo J, Liem X, Furstoss C, Khaouam N, et al. Data from head-neck-PET-CT. *Cancer Imaging Arch.* (2017). doi: 10.7937/K9/TCIA.2017.8oje5q00
41. Grossberg AJ, Mohamed ASR, Elhalawani H, Bennett WC, Smith KE, Nolan TS, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci Data.* (2018) 5:180173. doi: 10.1038/sdata.2018.173
42. Elhalawani H, Mohamed ASR, White AL, Zafereo J, Wong AJ, Berends JE, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci Data.* (2017) 4:170077. doi: 10.1038/sdata.2017.77
43. Grossberg A, Elhalawani H, Mohamed A, Mulder S, Williams B, White AL, et al. HNSCC [Dataset]. *Cancer Imaging Arch.* (2020). doi: 10.7937/k9/tcia.2020.a8sh-7363
44. Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep.* (2015) 5:11075. doi: 10.1038/srep11075
45. van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472.CAN-17-0339
46. Gong J, Wang Q, Li J, Yang Z, Zhang J, Teng X, et al. Using high-repeatable radiomic features improves the cross-institutional generalization of prognostic model in esophageal squamous cell cancer receiving definitive chemoradiotherapy. *Insights into Imaging.* (2024) 15:239. doi: 10.1186/s13244-024-01816-3
47. Louis T, Lucia F, Cousin F, Mievie C, Jansen N, Duysinx B, et al. Identification of CT radiomic features robust to acquisition and segmentation variations for improved prediction of radiotherapy-treated lung cancer patient recurrence. *Sci Rep.* (2024) 14:9028. doi: 10.1038/s41598-024-58551-4