Radiotherapy and Oncology 183 (2023) 109578

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients

Jiang Zhang [a], Sai-Kit Lam [b,c,1], Xinzhi Teng [a,1], Zongrui Ma [a,2], Xinyang Han [a,2], Yuanpeng Zhang [d,1], Andy Lai-Yin Cheung [g,2], Tin-Ching Chau [e,2], Sherry Chor-Yi Ng [e], Francis Kar-Ho Lee [f,2], Kwok-Hung Au [f,2], Celia Wai-Yi Yip [f,2], Victor Ho-Fun Lee [g], Ying Han [h], Jing Cai [a,c,*]

[a] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China; [b] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong, China; [c] Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, China; [d] Department of Medical Informatics, Nantong University, Nantong, Jiangsu, China; [e] Department of Clinical Oncology, Queen Mary Hospital, Hong Kong, China; [f] Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, China; [g] Department of Clinical Oncology, The University of Hong Kong, Hong Kong, China; [h] Department of Clinical Oncology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China

ARTICLE INFO

Article history:
Received 16 July 2022
Received in revised form 9 February 2023
Accepted 17 February 2023
Available online 21 February 2023

Keywords:
Radiomics
Repeatability
Nasopharyngeal Carcinoma
Disease-Free Survival

ABSTRACT

Background and purpose: To investigate the radiomic feature (RF) repeatability via perturbation and its impact on cross-institutional prognostic model generalizability in Nasopharyngeal Carcinoma (NPC) patients.

Materials and methods: 286 and 183 NPC patients from two institutions were included for model training and validation. Perturbations with random translations and rotations were applied to contrast-enhanced T1-weighted (CET1-w) MR images. RFs were extracted from primary tumor volume under a wide range of image filtering and discretization settings. RF repeatability was assessed by intraclass correlation coefficient (ICC), which was used to equally separate the RFs into low- and high-repeatable groups by the median value. After feature selection, multivariate Cox regression and Kaplan-Meier analysis were independently employed to develop and analyze prognostic models. Concordance index (C-index) and P-value from log-rank test were used to assess model performance.

Results: Most textural RFs from high-pass wavelet-filtered images were susceptible to image perturbations. It was more prominent when a smaller discretization bin number was used (e.g., 8, mean ICC = 0.69). Using high-repeatable RFs for model development yielded a significantly higher C-index (0.63) in the validation cohort than when only low-repeatable RFs were used (0.57, $P = 0.024$), suggesting higher model generalizability. Besides, significant risk stratification in the validation cohort was observed only when high-repeatable RFs were used ($P < 0.001$).

Conclusion: Repeatability of RFs from high-pass wavelet-filtered CET1-w MR images of primary NPC tumor was poor, particularly when a smaller bin number was used. Exclusive use of high-repeatable RFs is suggested to safeguard model generalizability for wide-spreading clinical utilization.

Radiomics is an emerging technique that leverages high-throughput feature extraction from medical images for discovering hidden information that is prognostic or predictive of various clinical endpoints. Accumulating evidence has suggested the promising application of Radiomics in the prognosis[1], clinical management[2–4], and treatment response predictions[5,6] of Nasopharyngeal Carcinoma (NPC) from several imaging modalities, including CT[7], MR[8–10], and PET/CT[11,12]. MR was favored in recent publications[13,14] due to its superior soft-tissue contrast. However, the majority of the previous MRI radiomics analysis on NPC were deemed less reliable due to the lack of stability analysis and external validation[15], which impedes the clinical applicability of the research findings[16]. Radiomic features (RFs) repeatability, which indicates the RF stability under the same imaging condition, should be the fundamental requirement of reliable modeling.

Effective assessment of RF repeatability has attracted growing attention in the past decades[17]. Test-retest imaging, which

* Corresponding author at: Y920, Lee Shau Kee Building, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China.
E-mail address: jing.cai@polyu.edu.hk (J. Cai).
[1] This author contributed to manuscript writing and statistical analysis.
[2] This author contributed to data collection and data curation.

https://doi.org/10.1016/j.radonc.2023.109578
0167-8140/

requires two repeated scans on the same patient, is one of the most popular approaches to assess RF repeatability[18–21]. However, short-interval test–retest scans are less applicable in routine clinical practice due to the additional cost of medical resources and potential extra dose to patients. Consequently, most existing test–retest-based RF repeatability studies focused on MR[18–20] and PET imaging[21]. 4DCT scans were also widely studied, where RF repeatability was assessed from multi-phase imaging[22,23]. Repeated scans with an even prolonged time interval, in the case of two-week apart, might lead to dramatic disparity in RFs due to enlarged tumor morphological and intra-tumoral microbiologic changes, which may reduce the generalizability of RF repeatability findings[24]. A new perturbation-based RF repeatability assessment method was proposed by Zwanenburg et al.[25] to overcome such limitations by simulating variabilities in scanning position, image noise, and region-of-interest contouring.

Despite the cumulative evidence of RF repeatability, limited effort has been made to demonstrate the benefit of repeatable features in improving downstream modeling. Several studies have investigated the predictive value of repeatable RFs of lung cancer based on the multi-phase scans of 4DCT[17]. Larue et al. found that only two RFs were associated with survival outcomes when applying multiple test corrections on all the RFs but 108 remained after repeatable RF filtering[22], suggesting the potential benefit of repeatability selection in false discovery control. A follow-up study by Lafata et al. reported high variability of RFs between 3D and 4D lung imaging and its effect on histology classification performance [23]. More recently, studies performed by Teng et al.[26,27] demonstrated enhanced radiomic model robustness and internal generalizability, where models were developed by using high-repeatable RFs exclusively. However, no further investigation has been provided in the body of literature regarding the impact of RF repeatability on cross-institutional model generalizability. More direct evidence on this topic is needed to provide the community with an enhanced understanding on the benefit and usage of RF repeatability in radiomics studies.

This study aims to investigate the RF repeatability via perturbation and its impact on the cross-institutional generalizability of the prognostic model for NPC disease-free survival prediction. We attempted to assess cohort-specific RF repeatability by our in-house developed image perturbation framework, taking reference from the previous work carried out by Zwanenburg et al.[25] to mimic a vast amount of scanning position stochasticity on contrast-enhanced T1-weighted (CET1-w) MR in a retrospective NPC cohort. The main objectives of this study were (i) to ascertain the repeatability of a comprehensive set of RFs against scanning position stochasticity via translation and rotation perturbations and (ii) to examine the benefit of repeatable RF in improving cross-institutional generalizability of prognosis modeling by externally validating prognostic models built separately from high- and low-repeatable RFs. Results from this study would provide a direct and conservative perceptiveness of RF repeatability pattern under a wide range of image filtering and discretization settings, offer evidence of its impact on inter-institutional generalizability, and encourage the radiomics community to exclusively adopt high-repeatable RFs for modeling to safeguard model generalizability.

Materials and methods

Patient cohort

We retrospectively recruited two biopsy-proven NPC patient cohorts from Queen Elizabeth Hospital (QEH) between 2012 and 2015 and Queen Mary Hospital (QMH) between 2013 and 2019. Due to the retrospective nature of this study, informed consents from patients were waived during the recruitment. Patients with (1) co-existing cancer or distant metastasis before treatment, (2) radiation therapy only without concurrent chemoradiotherapy, and (3) incomplete clinical record and missing segmentations were excluded from this study. In total, 286 patients from QEH and 183 patients from QMH were included in this study.

Contrast-enhanced T1-weighted (CET1-w) MR images and the planning primary gross tumor volume (GTVp) contours were retrieved from the treatment planning systems. MR scanning and GTVp contouring protocols are listed in Table A1**.** Disease-free survival (DFS) information was collected from patient folders. The time of DFS is defined from the date of treatment to the earliest occurrence of death from any cause, local or regional tumor recurrence, or distant metastasis.

Preprocessing and feature extraction

All the calculations in image preprocessing and feature extraction followed the guidelines proposed by the Image Biomarker Standardization Initiative (IBSI)[28]. They were performed by our in-house developed Python-based (3.7.3) pipeline using the SimpleITK (1.2.4) and PyRadiomics (2.2.0) packages. The workflow is explained by Fig. 1**.** Image preprocessing and feature extraction parameters are listed in Table A2. We extracted all the first-order features and texture features from Gray-Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), and Neighbouring Gray Tone Difference Matrix (NGTDM) from the original, three-dimensional Laplacian-of-Gaussian (LoG) filtered (sigma values of 1, 2, 3, 4, and 5 mm) and all the Coiflet-1 wavelet-filtered images. Each image was discretized by a fixed bin number of 8, 16, 32, 64, and 128 before feature extraction. In total, 6510 RFs were computed per patient.

Perturbation and RF repeatability assessment

Patient position variations were simulated by applying translation and rotation perturbations to each image and GTVp mask simultaneously during image preprocessing. They were implemented following the procedures proposed by Zwanenburg et al. [25], and the parameters are listed in Table A2. In this study, 40 translation and rotation combinations were randomly generated without replacement. The same preprocessing and feature extraction procedures were applied in calculating the RFs under perturbations. Feature repeatability was quantified from the perturbation RFs using the intraclass correlation coefficient (ICC). The one-way, random, absolute-agreement ICC was employed to assess RF repeatability due to the independent assignment of perturbation parameters to patients.

Feature selection

RFs from the unperturbed images were selected based on volume dependency first and then equally separated into high- and low-repeatable groups by the median ICC value before the feature redundancy and outcome relevancy test. The feature selection procedure is also explained in Fig. 1. Since the primary tumor volume has been recognized as a reliable prognostic factor[29], RFs that were highly correlated with GTVp mesh volume were first removed to minimize potential bias in the subsequent analyses [30]. We used the square of the Pearson correlation coefficient (r^2) to quantify the volume correlation, and the threshold of 0.6 was used to filter the volume-independent features. The final features were selected from each repeatability group by the feature redundancy and outcome relevancy test. During the feature redundancy test, r^2 was used to evaluate the correlation between features. For each highly correlated feature pair that has r^2

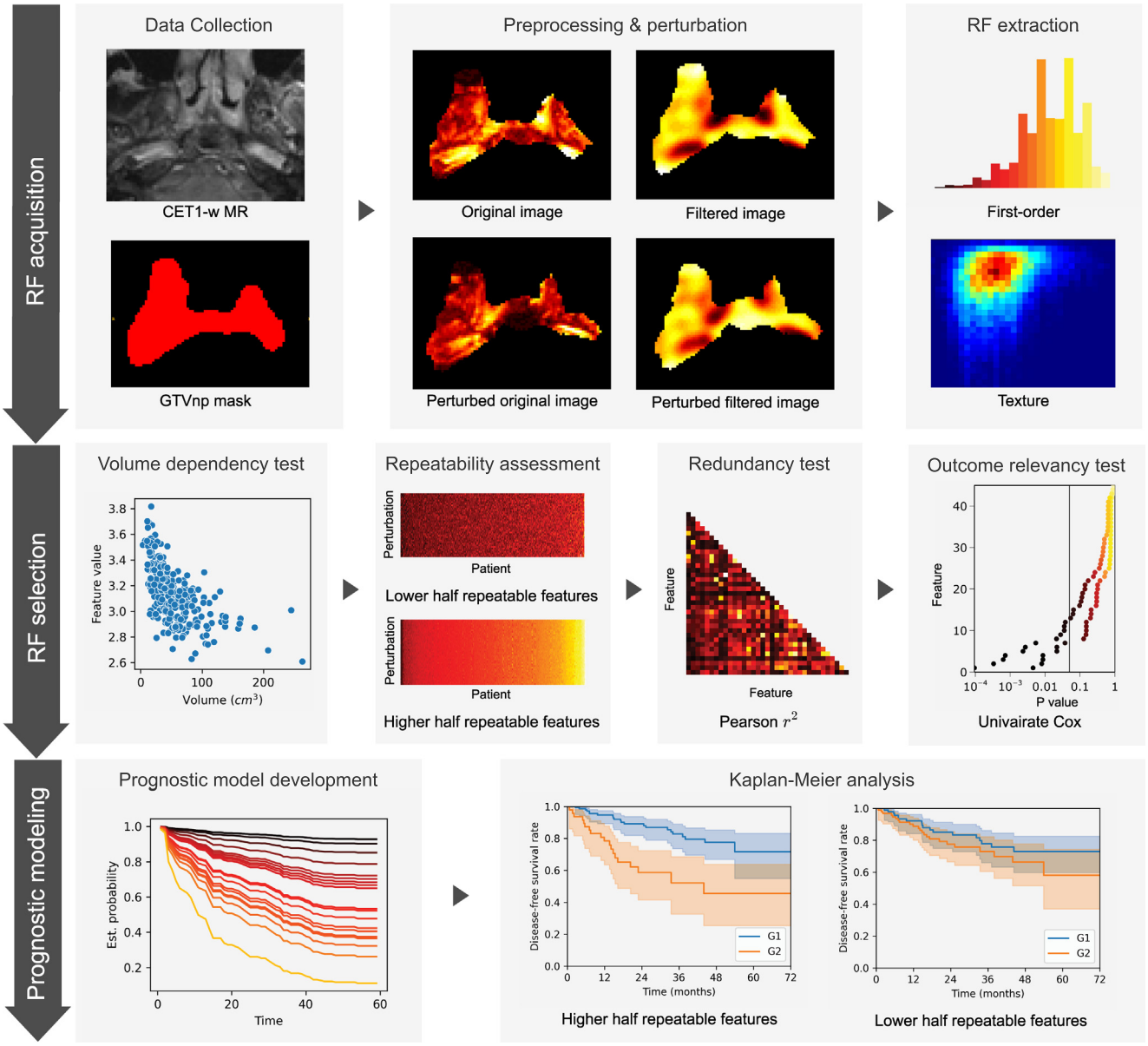

Fig. 1. Study Workflow.

exceeding 0.6, the one that has a larger mean r^2 with the rest of the features were removed. The outcome relevancy was evaluated by univariate Cox regression, and the 10 best features were finally selected, which were defined as the ones having the smallest hazard ratio (HR) P values.

Prognostic model development and evaluation

Two separate prognostic models were developed and evaluated on the selected high- and low-repeatable RFs. DFS survival risks were modeled by multivariate Cox regression on the training cohort, and the concordance index (C-index) was used to evaluate the discriminability on both training and validation. Classification performances at different time points were also assessed by the receiver operating characteristic (ROC) and the area under the curve (AUC) using the function "cumulative_dynamic_auc" provided by the Python package scikit-survival (version 0.18.0). In addition, we conducted 3-fold cross-validation with 10-time random repetitions on the training cohort, in order to assess the internal performance reliability. Independent redundancy test, outcome relevancy test, and Cox regression were performed on each cross-validation iteration. 95% confidence intervals were obtained from 1000-iteration bootstrapping, and P values were assessed by permutation tests where labels of high- and low-repeatable features were randomly shuffled by 1000 times for model performance comparisons. In addition, we evaluated the efficacy of the constructed prognostic model in risk stratification by Kaplan-Meier analysis. Patients were stratified into low- (G1) and high-risk (G2) groups based on the median training prediction, and the log-rank test P value was used to quantify the performance of the risk stratification.

Results

The distributions of the baseline patient characteristics for the two cohorts are listed in Table 1. Consistent distributions of age and sex were found between the training and validation cohort. The overall stage, chemotherapy strategy, and World Health Organization (WHO) histology were significantly different ($P < 0.05$)

Table 1
Baseline patient characteristics of the QEH (training) and QMH (validation) cohort.

		QEH (training)	QMH (external validation)	*P*
Age				
	Median	53	55	0.055
Sex				
	Female	70	40	0.590
	Male	216	143	
Overall stage				
	2	1	29	< 0.001
	3	187	90	
	4	98	64	
Chemotherapy				
	CCRT	178	37	< 0.001
	CCRT + ACT	62	65	
	CCRT + ICT	44	81	
WHO histology				
	Type 2	73	28	0.012
	Type 3	213	155	

Note: Staging was performed according to the 7th edition of the AJCC protocol for the training cohort and switched to the 8th edition after 2017 for the validation cohort. P values were obtained by student t-test for age and chi-square test for the rest of the clinical parameters.
Abbreviations: CCRT, concurrent chemoradiotherapy; ACT, adjuvant chemotherapy; ICT, induction chemotherapy; WHO, World Health Organization.

between the two institutions. The three-year DFS rate was 74.3 % in training and 72.1 % in validation.

RFs with lower repeatability were mostly texture features extracted from high-pass wavelet-filtered images discretized by smaller bin numbers, as visualized by the lighter green colors in the average ICC heatmap (Fig. 2**(a)**). The average ICC of high-pass wavelet-filtered 8-bin discretized texture RFs (Fig. 2**(a)**, red rectangles) was 0.69 but up to 0.99 for the remaining RFs. For image filters (Fig. 2**(b)**), RFs from unfiltered, all the LoG and LLL wavelet-filtered images yielded an average ICCs higher than 0.95, while the rest showed lower average RF repeatability (ICC: 0.73–0.87). Moreover, a decreasing trend of repeatability was found with high-pass wavelet filtering on more image dimensions. The first-order and NGTDM RFs showed the highest average ICC of 0.96 and 0.94, while the rest of the texture classes had mean ICCs below 0.90 (Fig. 2**(c)**). Notably, the GLSZM class had the lowest repeatability with an average ICC of 0.85. An increasing trend of repeatability was observed for larger image gray level discretization bin numbers. Specifically, bin number 8 had the lowest average ICC of 0.88, and the highest repeatability (mean ICC = 0.92) was achieved at 128 bin number (Fig. 2**(d)**). The complete ICC record can be found in **Appendix B**.

A strong correlation between GTVp volume dependency and repeatability was found on the RFs extracted from QEH T1-w MR images (Figure A3). 673 out of 709 (95 %) volume-dependent features (r^2 > 0.6) had high patient positioning repeatability (ICC > 0.9) whereas 3902 out of 5801 (67.3 %) volume-independent features showed high repeatability. The 709 volume-dependent features were removed from the subsequent analysis.

Distinct distributions of the feature redundancy measured by the mean r^2 to the rest of the features were observed on the high-repeatable and low-repeatable feature groups (Fig. A4 **(a)**), which were split by the median ICC of 0.95. Forty-four percent (1281/2901) of the low-repeatable features appeared to have low redundancy (mean r^2 < 0.1) whereas 15 % (445/2902) had low redundancy for the high-repeatable features. Distributions of the DFS prognosis, which was measured by univariate Cox P value, were similar between the two feature groups, except for the extreme-high prognosis region. Only 286 out of 2901 for the low-repeatable group had high DFS prognosis with P < 0.001 ($-\log_2 P$ > 10.0) while 541 out of 2901 for the high-repeatable group.

Ten RFs were finally selected from both the two feature groups after redundancy and outcome relevancy filtering. Details of the final selected features can be found in Table A3. After redundancy filtering, more low-repeatable features (317) remained with less redundancy but similar outcome relevancy compared to high-repeatable features (116), as shown in Fig. A4**(b)**. Quantitatively, 23 % (72/317) of the low-repeatable features had the mean r^2 larger than 0.05 while up to 84 % (97/116) for the high-repeatable ones, and 28 % (88/317) and 36 % (42/116) had P < 0.05 ($-\log_2 P$ > 4.3) for the two groups.

The discriminability of the multivariate Cox survival regression models developed from both low and high-repeatable features remained stable in the training cohort. As reported in Table 2, the C-index (low-repeatable = 0.65; high-repeatable = 0.67; P = 0.526) and time-dependent AUCs (P > 0.05) were similar in the training cohort. Time-dependent ROCs on the training cohort were also similar between the two feature sets, as shown in Fig. 3. During cross-validation, similar training performances were achieved with a mean C-index of 0.61 (low-repeatable) and 0.63 (high-repeatable). However, the low-repeatable models demonstrated significantly lower C-index values (mean = 0.55) than the high-repeatable ones (mean = 0.60) for internal validation, as shown in Figure A5. Both low and high-repeatable features stratified the training cohort into distinct survival groups (G1 and G2) with similar discriminability (HR = 2.50, 3.19) and statistically significant separations (log-rank P values <= 0.001), as presented in Fig. 4.

The prognostic model based on the high-repeatable features demonstrated significantly higher predictive performance in the validation cohort. Statistically higher C-index (high-repeatable = 0.63; low-repeatable = 0.57; P = 0.024), 1-year AUC (high-repeatable = 0.62; low-repeatable = 0.54; P = 0.031), and 3-year AUC (high-repeatable = 0.70; low-repeatable = 0.58; P = 0.015) were achieved (Table 2), while the 5-year AUC demonstrated weak statistical significance. Fig. 3 demonstrated distinctive differences in ROCs, especially for the 3-year progression event where the deviations were magnified. For survival risk stratifications, the high-repeatable features resulted in a significant separation of survival curves (P < 0.001) whereas a marginal separation (P = 0.054) can be found for the counterpart (Fig. 4).

Discussion

This study directly demonstrated the benefit of the unique information from RF repeatability assessed by translation and rotation perturbations in reducing false discovery and improving cross-

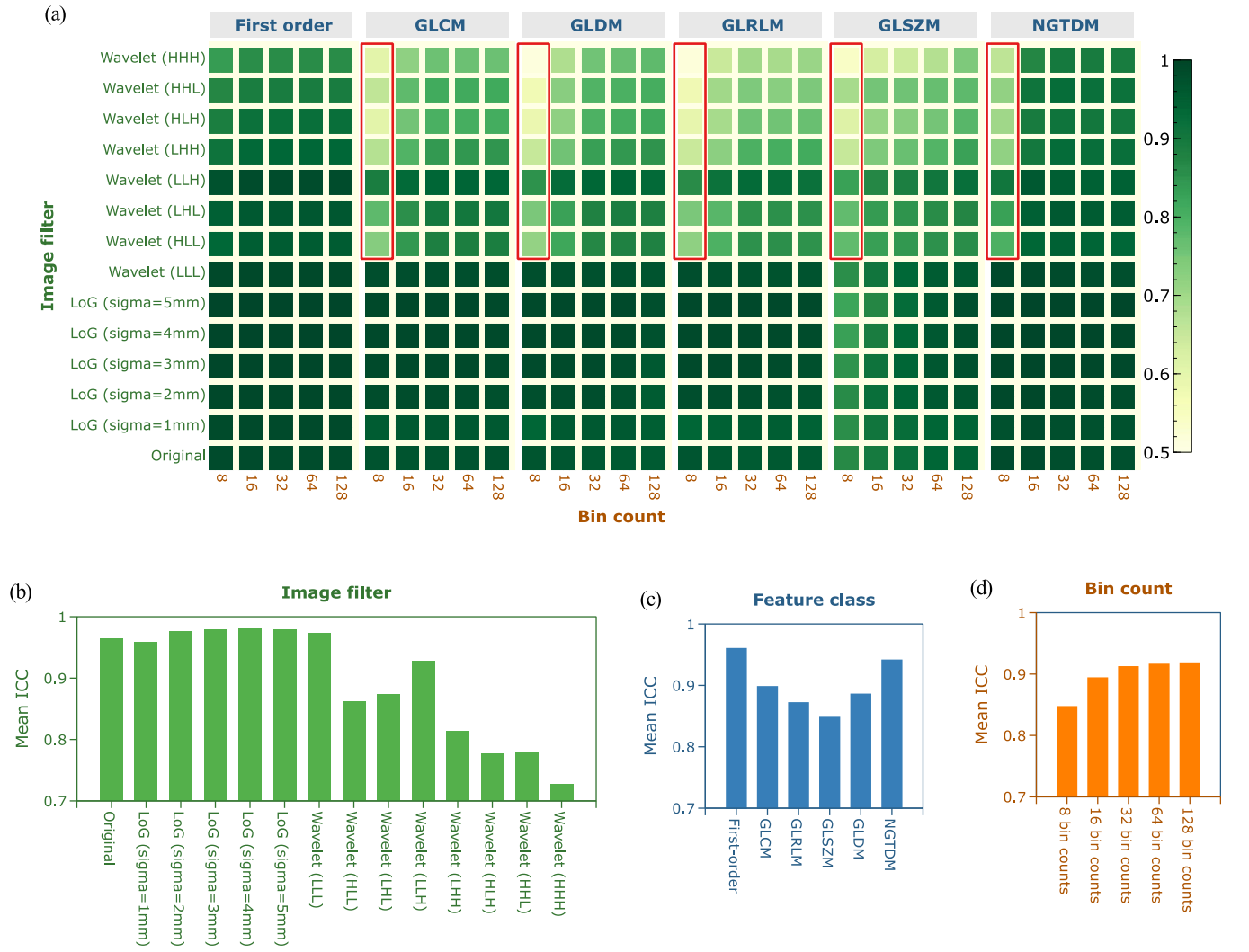

Fig. 2. Mean intraclass correlation coefficient of the extracted Radiomic features subgrouped by image filters, feature classes, and discretization bin numbers. High-pass wavelet-filtered RFs with smaller bin numbers demonstrated significantly lower repeatability with mean ICC = 0.69 for bin number = 8 (red boxes).

Table 2
Training and validation performance of the two constructed Cox survival regression model using high-repeatable and low-repeatable features.

	Training			External validation		
	Low-repeatable	**High -repeatable**	***P***	**Low-repeatable**	**High-repeatable**	***P***
C-index	0.62 (0.57–0.66)	0.67 (0.61–0.72)	0.526	0.57 (0.45–0.67)	0.63 (0.53–0.74)	0.024
1y AUC	0.65(0.60–0.67)	0.64 (0.64–0.72)	0.328	0.54 (0.38–0.72)	0.62 (0.44–0.79)	0.031
3y AUC	0.63 (0.55–0.67)	0.70 (0.62–0.76)	0.216	0.58 (0.46–0.71)	0.70 (0.58–0.81)	0.015
5y AUC	0.53 (0.46–0.58)	0.63 (0.55–0.71)	0.381	0.53 (0.28–0.78)	0.72 (0.46–0.92)	0.427

Abbreviations: C-index, concordance index; 1/3/5y AUC, area under the receiver operating characteristic curve at 1/3/5 year(s).
Note: the range inside each pair of round brackets indicates the 95% confidence interval under 1000 bootstrapping with replacement. P values (two-sided) were calculated by perturbation test where the high and low repeatability labels were randomly shuffled.

institutional generalizability. Results of our study suggested that different image filters, discretization bin numbers, and feature classes displayed heterogeneous patterns of RF repeatability. Notably, texture RFs from high-pass wavelet-filtered images discretized with smaller bin numbers were more susceptible to image perturbations (Fig. 2). After removing the volume-dependent RFs, the low-repeatable features demonstrated less redundancy, but outcome relevancy distributions were similar. The pattern remained unchanged after the redundancy test. Similar prognostic performance was achieved between the high and low-repeatable RFs during model training (Table 2), while the low-repeatable RFs yielded non-significant prognostic stratification on validation (Fig. 4).

Our image preprocessing strategy, especially the homogeneous resampling and gray-level discretization, aimed to minimize the impact of inconsistent image resolutions and intensity levels on feature repeatability and model performance from different scanners and scanning settings within and across institutions. Previous studies have shown the pronounced effect of pixel sizes on radiomic feature variability and suggested resampling to enhance the robustness[31,32]. Gray-level discretization with a fixed bin number could normalize the image intensities and reduce noise

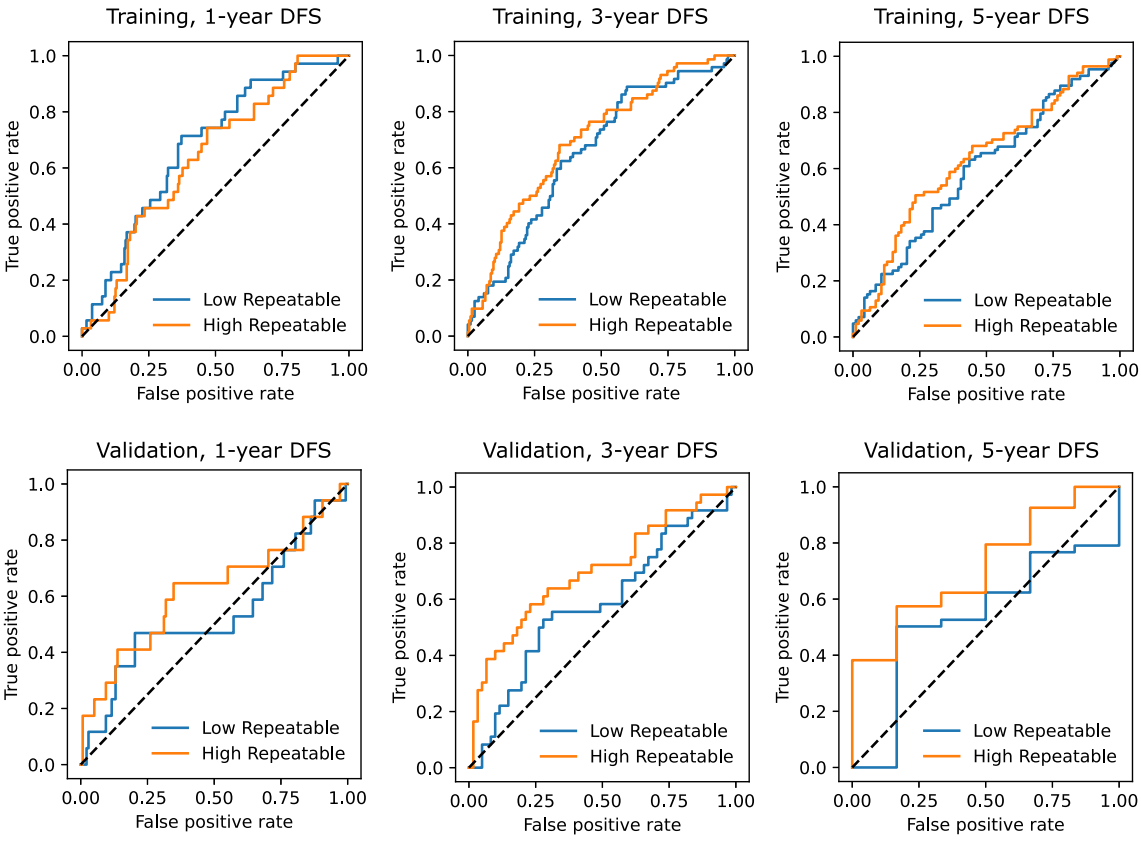

Fig. 3. Time-dependent receiver operating characteristic curves of Cox regression models from low (blue, dashed)) and high repeatable (orange, solid) features on disease-free survival (DFS). Results on one year, three years, and five years were plotted for both training and validation.

simultaneously[33]. It is also recommended by the IBSI for preprocessing image modalities with arbitrary intensities such as MRI [28].

The observed wavelet RF repeatability pattern could be ascribed to multiple factors, including the nature of wavelet filtering, image resampling strategy, and perturbation settings. High-pass wavelet filtering collects high-frequency signals and yields more heterogeneous pixel values. As demonstrated in Figure A2, images with high-pass wavelet filters on more dimensions appeared more heterogeneous, with fewer connected pixels with the same discretized intensity after binning. Notably, the wavelet-LLH image was less heterogeneous than wavelet-HLL and wavelet-LHL, possibly due to the larger slice thickness than in-plane resolution. The uniform 1x1x1mm resampling process up-sampled images along the axial direction, which may create artificial smooth textures. Our perturbation algorithm alters the left–right and anterior-posterior axes by rotation. It may induce drastic changes in pixel distributions under high-pass wavelet filtering on the first two dimensions while much less along the axial direction. It can also be observed in Figure A2 where the texture of wavelet-LLH filtered images was more similar under the two example perturbations than wavelet-HLL and wavelet-LHL. Furthermore, a lower bin number may magnify the discrepancies of texture features due to the smaller size of the gray-level matrix, which is consistent with the results of a previous phantom study[34]. Similar patterns were found in results reported by Larue et al. on a lung 4DCT, RIDER test–retest, and 4D-OES dataset[22] where more statistical high-pass wavelet RFs were highly repeatable than texture features.

The feature selection results suggest that the adopted redundancy and outcome relevancy tests, which are standard approaches in RF reduction, failed to identify the high-repeatable RFs. As expected, a large portion of ROI volume-dependent features were found to be highly repeatable under patient positioning variations, which agrees with previous studies on RF repeatability[35]. The outcome relevancy distributions were consistent between the high and low-repeatable features, but larger differences in redundancy patterns were found. Similar to the previous research[22], a minimum correlation was found between the univariate predictive power and feature repeatability. This suggests that either high or low-repeatable RFs have an equal chance of correlating with the prediction target. The low-repeatable features, which are “noisy” by nature, are more likely to be independent, which may elucidate our finding of the more low-redundant low-repeatable features.

Although the final feature number was strictly controlled, the low-repeatable RFs still suffered severe false discovery. Satisfactory prognostic model performance in training was achieved by the high and low-repeatable RFs with a C-index of 0.67 and 0.65 (Table 2) respectively due to the stringent outcome relevancy test criteria. The significant drop in internal testing performance suggests poor internal generalizability, which is consistent with the previous findings by Teng et al[27]. During external validation, the high-repeatable features yielded slightly lower discriminability of 0.63 in C-index, possibly due to inconsistent patient distributions between the two institutions. However, the low-repeatable RFs showed minimum prognostic power on the unseen data with C-index dropped to 0.57. Consequently, a much less significant survival curve separation of the validation cohort was achieved using only low-repeatable features (P = 0.054), as suggested by Fig. 4.

Our study has limitations that need to be addressed in future studies. First, the perturbation algorithm might not fully mimic the positional variations as in real clinical scenarios owing to technical challenges in simulating small deformations of the patient’s body between positionings. Second, the prognostic model performance, especially during validation, was slightly lower than

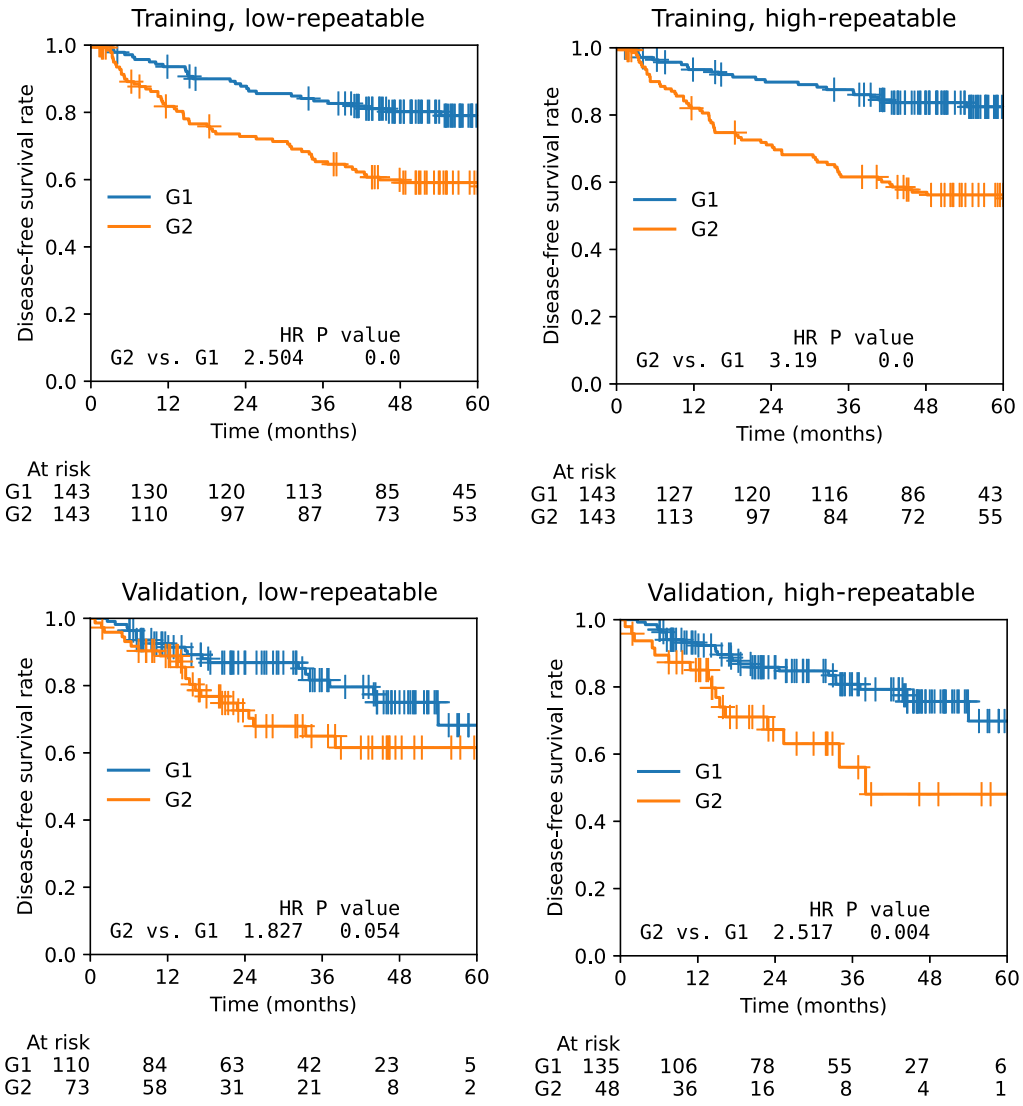

Fig. 4. Kaplan-Meier analysis of the low (G1) and high (G2) risk groups determined by the survival regression model from low-repeatable and high-repeatable features. Both features yielded similar survival curves on training, but a non-significant separation was found on validation with low-repeatable features.

previous radiomics research on NPC prognosis, where a range of C-index between 0.72 to 0.85 for NPC survival prognosis was reported[13]. This could be caused by the omission of clinical factors and lymph node tumor RFs in our final model. Nevertheless, our study did not intend to construct the best-performing model for clinical utility. Finally, several works were not accomplished in our research to maintain comprehensiveness while minimizing complexity. They include investigations under different imaging modalities, cancer types, feature extraction settings, or in a phantom study. We encouraged the community to carry out further investigations and to consider extending this work in the future.

Conclusions

Most textural RFs from high-pass wavelet-filtered CET1-w MR images of primary NPC tumor had poor repeatability under patient position variations, especially under a smaller bin number discretization. The prognostic model developed by low-repeatable RFs had significantly lower performance than high-repeatable RFs in the validation cohort, suggesting poor cross-institutional generalizability. We urge caution when handling high-pass wavelet-filtered RFs and advise exclusive use of high-repeatable RFs for prognostic model development to safeguard generalizability.

Funding information

This research was partly supported by research grants of Shenzhen-Hong Kong-Macau S&T Program (Category C): SGDX20201103095002019, Shenzhen Basic Research Program: JCYJ20210324130209023, Innovation and Technology Fund - Mainland-Hong Kong Joint Funding Scheme (ITF-MHKJFS), MHP/005/20, Project of Strategic Importance Fund of The Hong Kong Polytechnic University: P0035421, and Project of RISA of The Hong Kong Polytechnic University: P0043001.

Data availability

The patients' clinical and DICOM data are not publicly available for patient privacy protection purposes. Requests to access these datasets should be directed to the corresponding author.

IRB statement

The use of data was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB), reference number UW21-412, and the Research Ethics Committee (Kowloon Central/Kowloon East), reference number KC/KE-18-0085/ER-1.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109578.

References

[1] Zhang Y, Lam S, Yu T, 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:. https://doi.org/10.1016/j.knosys.2021.107649107649.

[2] Lam SK, Zhang J, Zhang YP, et al. A multi-center study of CT-based neck nodal radiomics for predicting an adaptive radiotherapy trigger of ill-fitted thermoplastic masks in patients with nasopharyngeal carcinoma. Life 2022;12:241. https://doi.org/10.3390/life12020241.

[3] Lam SK, Zhang Y, Zhang J, et al. Multi-organ omics-based prediction for adaptive radiation therapy eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. Front Oncol 2022;11:. https://doi.org/10.3389/fonc.2021.792024792024.

[4] Yu TT, Lam SK, To LH, et al. Pretreatment prediction of adaptive radiation therapy eligibility using MRI-based radiomics for advanced nasopharyngeal carcinoma patients. Front Oncol 2019;9:1050. https://doi.org/10.3389/fonc.2019.01050.

[5] Hou J, Li H, Zeng B, et al. MRI-based radiomics nomogram for predicting temporal lobe injury after radiotherapy in nasopharyngeal carcinoma. Eur Radiol 2021;32:1106–14. https://doi.org/10.1007/s00330-021-08254-5.

[6] Hu C, Zheng D, Cao X, et al. Application value of magnetic resonance radiomics and clinical nomograms in evaluating the sensitivity of neoadjuvant chemotherapy for nasopharyngeal carcinoma. Front Oncol 2021:11. https://doi.org/10.3389/fonc.2021.740776.

[7] Zhu C, Huang H, Liu X, et al. A clinical-radiomics nomogram based on computed tomography for predicting risk of local recurrence after radiotherapy in nasopharyngeal carcinoma. Front Oncol 2021:11. https://doi.org/10.3389/fonc.2021.637687.

[8] Shen H, Wang Y, Liu D, et al. Predicting progression-free survival using MRI-based radiomics for patients with nonmetastatic nasopharyngeal carcinoma. Front Oncol 2020:10. https://doi.org/10.3389/fonc.2020.00618.

[9] Zhao L, Gong J, Xi Y, et al. MRI-based radiomics nomogram may predict the response to induction chemotherapy and survival in locally advanced nasopharyngeal carcinoma. Eur Radiol 2020;30:537–46. https://doi.org/10.1007/s00330-019-06211-x.

[10] Zhuo EH, Zhang WJ, Li HJ, et al. Radiomics on multi-modalities MR sequences can subtype patients with non-metastatic nasopharyngeal carcinoma (NPC) into distinct survival subgroups. Eur Radiol 2019;29:5590–9. https://doi.org/10.1007/s00330-019-06075-1.

[11] Peng L, Hong X, Yuan Q, Lu L, Wang Q, Chen W. Prediction of local recurrence and distant metastasis using radiomics analysis of pretreatment nasopharyngeal [18F]FDG PET/CT images. Ann Nucl Med 2021;35:458–68. https://doi.org/10.1007/s12149-021-01585-9.

[12] Peng H, Dong D, Fang MJ, et al. Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. Clin Cancer Res 2019;25:4271–9. https://doi.org/10.1158/1078-0432.CCR-18-3065.

[13] Li S, Deng YQ, Zhu Z, Hua HL, Tao ZZ. A comprehensive review on radiomics and deep learning for nasopharyngeal carcinoma imaging. Diagnostics 2021:11. https://doi.org/10.3390/diagnostics11091523.

[14] Liu C, Li M, Xiao H, et al. Advances in MRI-guided precision radiotherapy. Precis Radiat Oncol 2022;6:75–84. https://doi.org/10.1002/pro6.1143.

[15] Spadarella G, Calareso G, Garanzini EM, Ugga L, Cuocolo A, Cuocolo R. MRI based radiomics in nasopharyngeal cancer: Systematic review and perspectives using radiomic quality score (RQS) assessment. *European journal of radiology*. 140:109744. doi:10.1016/j.ejrad.2021.109744.

[16] Jia X, Ren L, Cai J. Clinical implementation of AI technologies will require interpretable AI models. Med Phys 2020;47:1–4. https://doi.org/10.1002/mp.13891.

[17] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol*Biol*Phys 2018;102:1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.

[18] Li Z, Duan H, Zhao K, Ding Y. Stability of MRI radiomics features of hippocampus: an integrated analysis of test-retest and inter-observer variability. IEEE Access 2019;7:97106–16. https://doi.org/10.1109/ACCESS.2019.2923755.

[19] Shiri I, Hajianfar G, Sohrabi A, et al. Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: test–retest and image registration analyses. Med Phys 2020;47:4265–80. https://doi.org/10.1002/mp.14368.

[20] Sun M, Baiyasi A, Liu X, et al. Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study. Phys Med 2022;96:130–9. https://doi.org/10.1016/j.ejmp.2022.03.002.

[21] Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52:1391–7. https://doi.org/10.3109/0284186X.2013.812798.

[22] Larue RTHM, Van De Voorde L, van Timmeren JE, et al. 4DCT imaging to assess radiomics feature stability: an investigation for thoracic cancers. Radiother Oncol 2017;125:147–53. https://doi.org/10.1016/j.radonc.2017.07.023.

[23] Lafata K, Cai J, Wang C, Hong J, Kelsey CR, Yin FF. Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology. Phys Med Biol 2018;63:. https://doi.org/10.1088/1361-6560/aae56a225003.

[24] van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography 2016;2:361–5. https://doi.org/10.18383/j.tom.2016.00208.

[25] Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep 2019;9:1–10. https://doi.org/10.1038/s41598-018-36938-4.

[26] Teng X, Zhang J, Zwanenburg A, et al. Building reliable radiomic models using image perturbation. Sci Rep 2022;12:10035. https://doi.org/10.1038/s41598-022-14178-x.

[27] Teng X, Zhang J, Ma Z, et al. Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. Front Oncol 2022;12:. https://doi.org/10.3389/fonc.2022.974467974467.

[28] Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295:328–38. https://doi.org/10.1148/radiol.2020191145.

[29] Fave X, Zhang L, Yang J, et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 2016;5:349–63. https://doi.org/10.21037/8709.

[30] Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019;130:2–9. https://doi.org/10.1016/j.radonc.2018.10.027.

[31] Park SH, Lim H, Bae BK, et al. Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer. Cancer Imaging 2021;21:19. https://doi.org/10.1186/s40644-021-00388-5.

[32] Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31:1460–70. https://doi.org/10.1007/s00330-020-07174-0.

[33] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 2020;11:91. https://doi.org/10.1186/s13244-020-00887-2.

[34] Wichtmann BD, Harder FN, Weiss K, et al. Influence of Image Processing on Radiomic Features From Magnetic Resonance Imaging. *Invest Radiol*. 2022; Publish Ahead of Print. doi:10.1097/RLI.0000000000000921.

[35] Jha AK, Mithun S, Jaiswar V, et al. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. Sci Rep 2021;11:2055. https://doi.org/10.1038/s41598-021-81526-8.