TOWARDS RELIABLE RADIOMICS MODELING: A MULTI-INSTITUTIONAL MULTI-MODALITY FEATURE REPEATABILITY STUDY ON HEAD AND NECK CANCER PATIENTS

MA ZONGRUI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

Towards Reliable Radiomics Modeling: A Multi-institutional Multi-modality Feature Repeatability Study on Head and Neck Cancer Patients

MA Zongrui

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_________________________________(Signed)

____________MA Zongrui__________(Name of Student)

Abstract

Background: Radiomics has shown promise in cancer diagnosis and treatment decision making. However, the reliability of radiomic features and models remains a critical challenge. While feature repeatability has been extensively studied, the relationship between feature stability and model reliability is not well understood. Furthermore, understanding feature repeatability across imaging and data modalities and its relationship with image characteristics is essential for developing robust clinical prediction tools. Current research lacks comprehensive evaluation of feature repeatability across different imaging scenarios, its systematic correlation with image properties, and how feature reliability translates to model stability.

Purpose: This thesis systematically investigates feature repeatability and its impact on radiomics modeling through multi-institutional multi-modality analysis in head and neck cancer. The research encompasses three primary objectives: 1) to systematically quantify and compare feature repeatability across CT and MRI modalities in nasopharyngeal carcinoma, 2) to compare the feature repeatability in CT radiomics and dosiomics features and elucidate the relationships with image characteristics, 3) to validate the beneficial impact of feature repeatability on model performance and reliability through multi-institutional analysis.

Methods and Materials: A multi-institutional investigation of radiomics feature repeatability was conducted across three retrospective cohorts totaling 2,053 patients and nine institutions. Three imaging modalities were utilized including computed

tomography (CT), magnetic resonance imaging (MRI), and radiation dose maps. The study collected pre-treatment CT images of head-and-neck cancer patients from seven institutions via The Cancer Imaging Archive (TCIA), CT and MRI data of nasopharyngeal carcinoma patients from Queen Elizabeth Hospital (2012-2016), and planning CT and dose distributions of cervical cancer patients (2012-2022) from Peking University Third Hospital. A comprehensive perturbation framework was implemented to evaluate feature robustness, incorporating geometric transformations (rotation: ±20°, translation: 0-0.8 pixels) and contour randomization through deformation vector fields. Radiomics features were extracted using PyRadiomics, encompassing first-order statistics, morphological metrics, and texture characteristics derived from original, Laplacian-of-Gaussian, and wavelet-decomposed images. Feature stability was quantified using intraclass correlation coefficients (ICC), with stratified thresholds (0-0.9) for repeatability assessment. The impact of feature reliability on model performance was evaluated through internal cross-validation and external institutional validation using Cox proportional hazards regression. Model discrimination was assessed via concordance indices (C-index), while risk stratification significance was determined through Kaplan-Meier survival analysis. For the cervical cancer cohort, comparative analyses between radiomics and dosiomics feature stability were conducted across multiple regions of interest, providing insights into data modality-specific feature robustness. All feature extraction algorithms were implemented using standardized computational frameworks adherent to the Image Biomarker Standardization Initiative.

Results: Quantitative assessment of feature stability across imaging modalities demonstrated superior repeatability in shape-based features (mean ICC: 0.92, 95% CI: 0.89-0.94), with MRI-derived radiomic features exhibiting significantly higher stability compared to CT-derived features (86.8% vs 42.3% features achieving $\text{ICC}>0.9$, $P<0.001$). In the comparative analysis of CT-radiomics and dosiomics features in cervical cancer specimens, CT radiomic features demonstrated superior repeatability metrics (mean ICC: 0.81, 95% CI: 0.78-0.84) compared to dosiomics features (mean ICC: 0.67, 95% CI: 0.63-0.71), particularly in features extracted from rectum and femoral ROIs (mean ICC: 0.85, 95% CI: 0.82-0.88). Feature repeatability demonstrated strong correlations with image characteristics, specifically entropy ($r=0.76$, $P<0.001$), uniformity ($r=-0.72$, $P<0.001$), and variance ($r=0.74$, $P<0.001$) across all modalities. The integration of highly repeatable features ($\text{ICC} \geqslant 0.9$) consistently enhanced prognostic model performance across different head and neck cancer datasets, demonstrating improved validation metrics (ΔC-index: +0.02 to +0.05, $P<0.01$) and enhanced model generalizability. Notably, stringent repeatability criteria effectively mitigated performance degradation in heterogeneous datasets, suggesting that feature stability is crucial for robust model development.

Conclusions: Our work presents a comprehensive investigation of radiomic feature reliability across multiple imaging and data modalities and institutions in head and neck cancer patients. Through our rigorous multi-institutional analyses and systematic evaluation of feature stability patterns, we characterized the intrinsic relationships between image characteristics and feature repeatability and developed a robust

framework for reliable radiomics modeling. Our findings demonstrate that pre-screening and incorporation of high-reliable features significantly enhances model performance and generalizability, advancing the theoretical and practical foundations of radiomics. Our work establishes a methodological framework for developing more reliable radiomics models and facilitates their translation into clinical practice.

Research Output

1. **Ma, Z.**, Zhang, J., Liu, X., Teng, X., Huang, Y. H., Zhang, X., ... & Cai, J. (2024). Comparative Analysis of Repeatability in CT Radiomics and Dosiomics Features under Image Perturbation: A Study in Cervical Cancer Patients. Cancers, 16(16), 2872. (**Published**)

2. **Ma, Z.**, Zhang, J., Teng, X., Lam, S., Zhang, Y., Huang, Y. H., ... & Cai, J. (2024, October). Assessment of Radiomics Feature Repeatability and Reproducibility and Their Generalizability Across Image Modalities by Perturbation in Nasopharyngeal Carcinoma Patients. In International Workshop on Computational Mathematics Modeling in Cancer Analysis (pp. 110-119). Cham: Springer Nature Switzerland. (**Published**)

3. Zhang, J**.**, Lam, S. K., Teng, X., **Ma, Z.**, Han, X., Zhang, Y., ... & Cai, J. (2023). Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients. Radiotherapy and Oncology, 183, 109578.

4. Zhang, J., Teng, X., Lam, S., **Ma, Z.**, Han, X., Zhang, X., ... & Cai, J. (2024). 2158: A Survival-Driven Radiotherapy Plan Quality Index for Nasopharyngeal Cancer: A Multicenter Study. Radiotherapy and Oncology, 194, S5103-S5105.

5. Lam, S. K., Zhang, Y., Zhang, J., Li, B., Sun, J. C., Liu, C. Y. T., **Ma, Z.**, ... & Cai, J. (2022). Multi-organ omics-based prediction for adaptive radiation therapy

eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. Frontiers in oncology, 11, 792024.

6. Teng, X., Zhang, J., **Ma, Z.**, Zhang, Y., Lam, S., Li, W., ... & Cai, J. (2022). Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. Frontiers in Oncology, 12, 974467.

7. Teng, X., Zhang, J., Han, X., Sun, J., Lam, S. K., Ai, Q. Y. H., **Ma, Z.**, ... & Cai, J. (2023). Explainable machine learning via intra-tumoral radiomics feature mapping for patient stratification in adjuvant chemotherapy for locoregionally advanced nasopharyngeal carcinoma. La radiologia medica, 128(7), 828-838.

8. Teng, X., Zhang, J., Zwanenburg, A., Sun, J., Huang, Y., Lam, S., **Ma, Z.**, ... & Cai, J. (2022). Building reliable radiomic models using image perturbation. Scientific Reports, 12(1), 10035.

9. Zhang, Y., Yang, D., Lam, S., Li, B., Teng, X., Zhang, J., **Ma, Z.**, ... & Cai, J. (2022). Radiomics-based detection of COVID-19 from chest X-ray using interpretable soft label-driven TSK fuzzy classifier. Diagnostics, 12(11), 2613.

10. Li, B., Ren, G., Guo, W., Zhang, J., Lam, S. K., Zheng, X., **Ma, Z.**, ... & Ge, H. (2022). Function-Wise Dual-Omics analysis for radiation pneumonitis prediction in lung cancer patients. Frontiers in Pharmacology, 13, 971849.

11. Zheng, X., Guo, W., Wang, Y., Zhang, J., Zhang, Y., Cheng, C., **Ma, Z.**, ... & Li, B. (2023). Multi-omics to predict acute radiation esophagitis in patients with lung

cancer treated with intensity-modulated radiation therapy. European Journal of Medical Research, 28(1), 126.

12. Huang, Y. H., Teng, X., Zhang, J., Chen, Z., **Ma, Z.**, Ren, G., & Cai, J. (2022). Extracting lung function-correlated information from CT-encoded static textures. arXiv preprint arXiv:2210.16514.

13. Zhang, Y. P., Zhang, X. Y., Cheng, Y. T., Li, B., Teng, X. Z., Zhang, J., **Ma, Z.**, ... & Cai, J. (2023). Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. Military Medical Research, 10(1), 22.

14. Huang, Y. H., Li, Z., Xiong, T., Chen, Z., Li, B., Lou, Z., Dong, Y., Teng, X., **Ma, Z.**, Ge, H., Ren, G., & Cai, J. (2024). Constructing Surrogate Lung Ventilation Maps From 4-Dimensional Computed Tomography-Derived Subregional Respiratory Dynamics. International journal of radiation oncology, biology, physics, S0360-3016(24)03648-4. Advance online publication. https://doi.org/10.1016/j.ijrobp.2024.11.074

Conference abstract

1. **Ma Z.**, Zhang J., Liu X., Teng X., Zhang X., Li J., Pan Y., Sun J., Dong Y., Yip W., Lee F., Cai J. (2024). Comparative Analysis of Repeatability in CT Radiomics and Dosiomics Features under Image Perturbation: A Study in Cervical Cancer Patients. 24TH ASIA-OCEANIA CONGRESS OF

MEDICAL PHYSICS (AOCMP) (**Oral Presentation**)

2. **MA Z.**, Zhang J., Teng X., Lam S., Han X., Xiao H., Liu Chen., Li W., Huang Y., Lee F., Yip W., Cai J. (2022). Radiomics-Based Multiple Prognosis Prediction Using Mutual Information Between Treatment Outcomes as Prior Knowledge in Nasopharyngeal Carcinoma. The 64th Annual Meeting & Exhibition of the American Association of Physicists in Medicine (AAPM) (**Poster Presentation**)

Acknowledgement

I would like to take this opportunity to express my gratitude to everyone who supported me throughout my Ph.D. journey.

Among all, I would like to express my deepest gratitude to my chief supervisor, Prof. Cai Jing. His guidance has been invaluable, not only in my academic pursuits but also in my personal and professional development. Prof. Cai has generously shared his insights and experiences, providing me with constructive advice that has greatly enriched my research. His unwavering support and encouragement have inspired me to push my boundaries and strive for excellence. The years I spent learning under his mentorship have been the most transformative of my life. I sincerely wish Prof. Cai continued success, happiness, and good health in all his future endeavors.

Next, I also extend my heartfelt thanks to my group members, especially Zhang Jiang, Teng Xinzhi, Xiao Haonan, Li Tian, Ren Ge, Tingting, Saikit, Li Wen, Chen Yang, and Yu Hua. Their selfless support and camaraderie have been instrumental in overcoming various challenges throughout my project. Even those who have moved away from Hong Kong will always hold a special place in my heart, and I hope our bonds will remain strong.

I would like to acknowledge my friends from the sports team—Danning, Jingcheng, Guoping, Yunsong, Dingjian, and Hu Yang. We have shared some of the most memorable and challenging times together over the past three years. I wish you all the best in your future paths.

The biggest thanks must go to my family members, especially my mom and grandmom. The guidance and encouragement have shaped who I am today. Without your unwavering support, I would not be where I am. I hope for your continued health and happiness. A special thank you goes to my girlfriend, who has provided me with love and support during both difficult and joyful times. Your companionship has meant the world to me.

I would like to thank my alma mater, PolyU, for nurturing and educating me during this significant chapter of my life.

Lastly, I extend my thanks to the thesis committee and external examiners for their valuable time in reviewing and evaluating my work. I hope my research will make a meaningful contribution to the Radiomic community.

Table of Contents

Abbreviations

CT	Computed Tomography
HU	Hounsfield unit
GLCM	gray-level co-occurrence matrix
PET	Positron Emission Tomography
MRI	Magnetic Resonance Imaging
ROI	Region of Interest
GTV	Gross Tumor Volume
IBSI	Image Biomarker Standardization Initiative
3D/4D	Three-/four-dimensional
CAD	Computer Aided Detection
C-index	Concordance Index
ICC	Intra-class Coefficient of Correlation

LR	Local-/Regional- Recurrence
HNC	Head-and-Neck Carcinoma
OPC	Oropharyngeal Carcinoma
TCIA	The Cancer Imaging Archive
DECT	Dual Energy Computed Tomography
HNSCC	Head and Neck Squamous Cell Carcinoma
DSC	Dice Similarity Coefficients
ADC	Apparent Diffusion Coefficients
IP	Image Perturbation
HD	Hausdorff Distance

Chapter 1. Literature Review

1.1. Introduction

The advent of radiomics represents a significant advancement in precision medicine, leveraging machine learning algorithms to elucidate the complex relationships between cancer imaging phenotypes and their corresponding genotypes or clinical outcomes [1, 2]. This burgeoning field, which has witnessed exponential growth in scholarly publications [3], employs sophisticated computational analyses of medical imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), to extract and analyze hand-crafted features that characterize tumor heterogeneity [4]. While radiomics has demonstrated remarkable potential in tumor characterization, diagnostic assessment [5], and treatment outcome prediction [6], thereby enhancing clinical decision-making processes [7], the translation of radiomic models from research laboratories to clinical practice faces substantial challenges regarding reliability and generalizability. This study addresses these critical limitations through the implementation of simulated perturbations, aiming to enhance the robustness and broader applicability of radiomic models in clinical settings.

1.2. Clinical Value of Radiomics

Medical imaging has traditionally relied on qualitative visual assessment, leaving vast amounts of potentially valuable data unexplored. Radiomics has emerged as an innovative analytical framework that combines the prefix 'radio-' (medical imaging) with '-omics' (comprehensive characterization), representing a paradigm shift in how we analyze medical images. This approach integrates sophisticated feature extraction methodologies with advanced machine learning algorithms to unlock previously inaccessible insights from imaging data [8].

The radiomics workflow consists of two fundamental components: high-dimensional feature extraction and computational analysis [9]. During feature extraction, medical images are transformed into large-scale quantitative datasets, capturing subtle patterns and characteristics that may escape visual detection. These extracted features are then analyzed using machine learning algorithms [10], which can identify complex patterns and relationships within the data to generate predictive models.

Consider a common clinical scenario: two lung cancer patients with matching TNM staging, histological profiles, and similar ages may experience significantly different outcomes (**Table 1**). This observation highlights the limitations of conventional prognostic factors in capturing the full complexity of disease progression. The untapped potential within medical imaging data may hold the key to explaining such divergent outcomes.

A landmark study demonstrated that tumors exhibit distinct phenotypic characteristics that can be captured non-invasively through medical imaging [11]. These phenotypic differences manifest in various ways, including tumor morphology, density distributions (measured in Hounsfield units), and textural patterns, as illustrated in **Figure 1**. By quantifying these characteristics through radiomics analysis, researchers can potentially identify novel patient subgroups and optimize treatment strategies accordingly.

The synergy between radiomics and machine learning algorithms, such as decision trees, enables the transformation of complex imaging data into clinically actionable insights. This approach aligns with the growing emphasis on precision medicine, offering a data-driven pathway to personalized treatment strategies based on comprehensive imaging analysis. Radiomics thus represents a promising avenue for advancing precision oncology by extracting and analyzing the wealth of information embedded within medical images.

This quantitative imaging analysis framework has gained significant traction in radiology and radiation oncology [1], offering a systematic approach to mining previously underutilized imaging data. By enabling the high-throughput extraction of quantitative features from routine medical imaging, radiomics has the potential to enhance clinical decision support systems and improve patient care through more precise characterization of disease states.

Table 1. An example of inadequate biomarkers for predicting survival based on clinical characteristics in non-small cell lung cancer (NSCLC) patients is illustrated by two individuals who, despite being of similar age, having the same TNM staging, histology, and gender, experienced different survival outcomes. Traditional clinical characteristics failed to provide consistent risk stratification between these two patients.

ID	**Age**	**T**	**N**	**M**	**Overall Stage**	**Gender**	Survival (days)	Survival Status
1	71	4	3	0	IIIb	female	2119	alive
2	62	4	2	0	IIIb	female	261	dead

1.3. Challenges in Radiomics Workflow

Although radiomics has demonstrated remarkable potential in oncology with an exponential increase in publications, the reliability of radiomic models remains a fundamental challenge [12]. The radiomics workflow begins with the quantification of features from medical images within regions of interest (ROIs) using predefined mathematical formulations [13, 14]. This quantification process is inherently susceptible to variations in both image acquisition and ROI delineation, with multiple sources of variability introduced throughout the radiomics workflow, from image acquisition to processing.

The importance of reliability in radiomics has been recognized since the field's inception. Early pioneering studies in radiomics intentionally utilized test-retest datasets, such as RIDER Lung CT [15], to evaluate feature repeatability and reproducibility. However, the practical limitations of medical imaging, including radiation exposure concerns in CT and PET, and the time-intensive nature of MRI acquisitions, have restricted the widespread collection of test-retest data. Moreover, the precision of radiomic feature values is influenced by numerous factors, including imaging protocols and acquisition parameters.

Recent years have seen intensive investigation into the repeatability and reproducibility of radiomics features. Factors affecting feature reproducibility, such as imaging protocols and acquisition parameters, have been categorized as controllable factors [16], suggesting they could theoretically be minimized through standardized protocols. However, achieving such standardization across different institutions remains challenging, as these variations typically do not significantly impact routine clinical imaging interpretation.

The radiomics workflow encompasses several distinct steps (**Figure 1**), each presenting unique challenges. These steps include: 1. image acquisition and reconstruction, 2. ROI segmentation, 3. feature extraction, and 4. radiomic modeling. Understanding and addressing the variations introduced at each step is crucial for developing reliable and generalizable radiomic models.

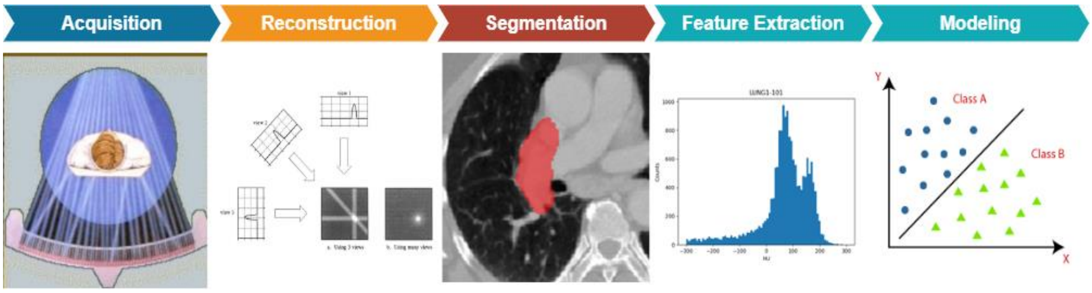

Figure 1. The comprehensive radiomics workflow encompasses multiple sequential steps: image data acquisition, three-dimensional volume reconstruction, region of interest (ROI) segmentation, feature extraction, and machine learning-based modeling. Each step introduces potential sources of variation that may significantly impact the reproducibility and repeatability of extracted features and ultimately affect the consistency of radiomic model outputs. These variations represent critical challenges that must be addressed to ensure reliable and generalizable radiomic models in clinical applications.

1.3.1. Challenges in Image Acquisition and Reconstruction

Contemporary radiomics research primarily employs CT, MR, and PET modalities for image acquisition and reconstruction. However, the standardization of imaging protocols demonstrates significant heterogeneity across institutions. While these protocol variations may not substantially impact conventional diagnostic interpretation, they significantly affect radiomic analyses due to the voxel-level quantification inherent in radiomics [17]. Such variations modulate image noise and texture characteristics, potentially resulting in inconsistent predictions when validating radiomic models across different institutional datasets. Moreover, reconstruction algorithms can substantially affect image quantification and subsequent model performance [18].

The development of institution-independent features represents a crucial foundation for building generalizable models across institutions. A comprehensive investigation employed a phantom study across 17 CT scanners with varying manufacturers and thoracic imaging protocols, demonstrating significant variations in radiomic features under different acquisition parameters [19]. However, most studies remain limited to single tumor sites and imaging modalities, leaving uncertainty regarding feature reliability generalizability across different datasets of interest. Notable studies include a multi-center and multi-vendor investigation examining feature reliability using apparent diffusion coefficient (ADC) maps [20], and research evaluating radiomic feature reproducibility and repeatability in T2-weighted MRI of cervical cancer patients across three distinct settings [21]. These investigations have provided thorough analyses of acquisition parameters' impact on radiomic features and documented comprehensive feature reliability metrics.

A comprehensive review of literature pertaining to image acquisition and reconstruction is presented in **Table 2**. Analysis of 21 publications revealed systematic investigations into the effects of inter-scanner variability, test-retest reliability, image acquisition parameters, and reconstruction algorithms on radiomics feature reproducibility. A consistent finding across these studies indicated that texture features demonstrate greater susceptibility to variations in image acquisition and reconstruction compared to intensity (first-order) features.

However, significant limitations exist in translating these findings to clinical-oriented investigations. The primary challenge lies in quantifying feature reliability and incorporating it effectively into feature selection processes for clinical applications. While seminal work established a methodology utilizing feature reliability and outcome relevance indices for feature ranking, two major constraints persist. First, the majority of studies failed to provide comprehensive tabulation of reproducibility indices for individual features. Second, the requirement for scanner access presents a substantial barrier for many research groups, limiting their ability to conduct independent feature reliability assessments. These limitations underscore the need for more accessible and standardized approaches to feature reliability assessment in clinical radiomics research.

Table 2. The literature examining the image acquisition and reconstruction in CT and MRI. (FO – First Order Feature, TA – Texture Analysis)

Author	**Disease**	**Modality Investigated**	**Sources of variation**	**Feature Categories**
Cabini,2022 [22]	Lung	CT	Image acquisition parameters	Shape, FO, TA
Carbonell,2022 [23]	Liver	MR	1.Test-retest repeatability 2. Inter-scanner 3. Inter-observer segmentation	Shape, FO, TA
Chen,2021 [24]	Hematoma	CT	1. Test-retest repeatability 2. Image acquisition parameters	FO, TA
Chen,2022 [25]	Phantom	CT	1. Test-retest repeatability 2. Scanning modes 3. Inter-scanners	FO, TA
Crombe,2021 [26]	Abdomen	MR	T2-w acquisition methods	FO, TA
Emaminejad,2021 [27]	CA Lung	CT	1. Dose level variation 2. Reconstruction kernel 3. Slice thickness variation	FO, TA
Euler,2021 [28]	Phantom	CT	1. Image acquisition parameters 2. Radiation dose 3. DECT approach	Shape, FO, TA
Fiset,2019 [21]	Cervix	MR	1. Test-retest repeatability 2. Acquisition protocols 3. Inter-observer segmentation	Shape, FO, TA
Gao,2022 [29]	Pulmonary nodules Lung	CT	1. Radiation dose 2. Reconstruction kernels	Shape, FO, TA
Granzier,2022 [30]	Breast	MR	Test-retest repeatability	FO, TA
Ibrahim,2021 [31]	HCC Liver	CT	Imaging phases	Shape, FO, TA

Ibrahim,2021 [32]	Phantom	CT	1. Inter-scanners 2. Scanning parameters	Shape, FO, TA
Lee,2021 [33]	Phantom	MR	1. MRI scanning protocols parameters 2. Scanner types	FO, TA
Lennartz,2022 [34]	Phantom	DECT	Inter-scanners	FO, TA
Mahon,2019 [35]	NSCLC Lung	4DCT, MR	Test-retest repeatability	FO, TA
McHugh,2021 [36]	Colorectal Cancer Liver Metastases	MR	1. MR sequences 2. Pre- and post-contrast 3. Image normalization	Shape, FO, TA
Meyer,2019 [37]	Metastatic liver lesions	CT	1. Radiation dose 2. Reconstruction settings	Shape, FO, TA
Mitchell-Hay,2022 [38]	Brain	MR	1. Inter-scanner 2. Test-retest repeatability for weeks	FO, TA
Reiazi,2021 [39]	Oropharyngeal Cancer Oropharynx	CT	Inter-scanners	Shape, FO, TA
Rinaldi,2022 [40]	NSCLC Lung	CT	1. Tube voltage, scanner model 2. Reconstruction algorithm	Shape, FO, TA
Alis,2020 [41]	Heart	MR	1. Inter-observer reproducibility of radiomics features 2. Cardiac cycle	FO, TA

1.3.2. Challenges in Segmentation

ROI delineation constitutes a fundamental component in radiomics methodology, particularly in determining the precise boundaries for subsequent feature extraction. Contemporary radiation oncology studies predominantly focus on delineating the visible or gross tumor volume (GTV) [42]. Although expert manual contours by radiation oncologists serve as the reference standard, this methodology presents inherent challenges, including substantial time requirements and susceptibility to delineation variability. As demonstrated in **Figure 2**, significant variations exist in GTV delineation among oncologists analyzing prostate cancer CT images, leading to potential inconsistencies in subsequent radiomic feature calculations[43].

Alternative approaches utilizing computational methods for tumor volume delineation have demonstrated promising results regarding consistency and operational efficiency. Notable research has documented superior feature stability using automated approaches compared to manual delineation in non-small cell lung cancer cases [44]. However, automated segmentation methodologies face significant constraints. Their efficacy is primarily demonstrated in anatomically distinct regions with high contrast differentiation, such as pulmonary or prostatic malignancies. More challenging anatomical sites, particularly head-and-neck carcinomas (HNC), where structural complexity is heightened, continue to demonstrate inferior performance compared to expert manual delineation, thus restricting their widespread clinical adoption.

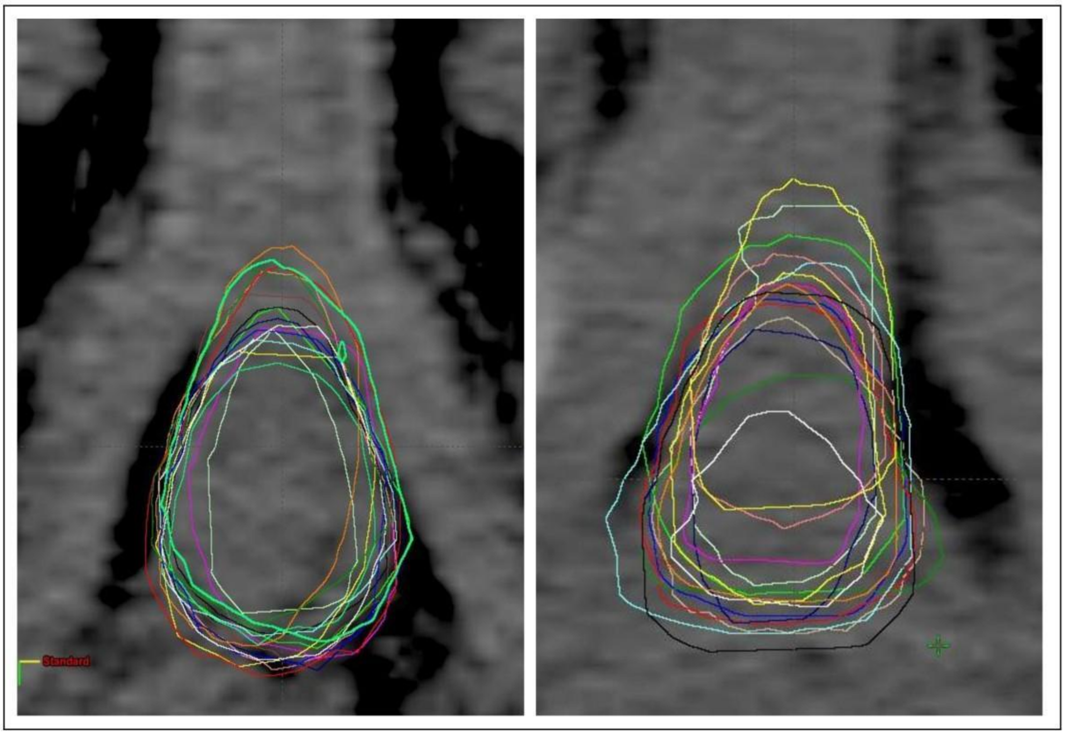

Figure 2. Inter-observer variability in penile bulb delineation is illustrated through CT image examples from two patients. The central axial slices demonstrate how manual segmentation of the penile bulb varies significantly when performed by different oncologists [43]. This comparison highlights the inherent challenges and inconsistencies in organ delineation even among experienced clinicians. The images are reproduced from previous publications with appropriate permissions and no copyright conflicts.

Assessment of radiomic feature stability under segmentation variability has been approached through multiple methodologies. Inter-observer and intra-observer reproducibility studies have evaluated feature robustness using multiple delineations, either from different oncologists or repeated segmentations by the same observer. To address the resource-intensive nature of traditional reproducibility assessments, Zwanenburg et al. introduced an innovative super-voxel methodology for estimating potential segmentation variations [45]. However, these approaches present notable limitations: they fail to account for inherent image variations, and in the case of manual segmentation studies, the substantial resource requirements pose significant challenges for widespread implementation in research settings. Furthermore, the labor-intensive nature of multiple-observer studies may impede their practical application in larger-scale investigations.

Table 3 summarizes 12 studies that investigated feature reproducibility in relation to segmentation variability. While most research primarily examined the impact of inter-observer variability on radiomic features, some also explored intra-observer variability. Only a small number of studies combined the analysis of inter-observer variability with test-retest imaging. These studies share similar limitations as discussed in section 1.3.1: many failed to provide explicit quantitative results, and their findings, often tied to specific clinical contexts, are not directly transferable to other radiomic studies. Additionally, conducting inter-observer variability analysis for radiomic features requires substantial manual effort, making it impractical for routine implementation in every radiomic study.

Table 3. Literatures investigating the impact of segmentation variability on radiomic feature reproducibility and repeatability.

Author	**Site**	**Modalities**	**Sources of variation**	**Feature category**
Bianconi, 2021 [45]	Lung	CT	1. Inter-observer variability 2. Image quantization method	FO, TA
Carbonell, 2022 [23]	Liver	T1-w MR, T2-w MR, ADC	1. Inter-scanner 2. Inter-observer segmentation	Shape, FO, TA
Chen, 2021 [46]	CA Cervix	DWI	1. Inter-observer segmentation 2. Intra-observer segmentation	Shape, FO, TA
Duan, 2022 [47]	HCC Liver	CT, T1-w MR, T2-w MR	Inter-observer segmentation	FO, TA
Granzier, 2020 [48]	Bresat	DCE T1-w MR	Inter-observer variability	Shape, FO, TA
Haniff, 2021 [49]	HCC Liver	T1-w MR	Segmentation method	Shape, FO, TA
Jensen, 2021 [50]	Phantom	CT, T1-w MR, T2-w MR	ROI Size	FO, TA
Kocak, 2019 [51]	Kidney	CT	1. Inter-observer segmentation 2. Intra-observer segmentation	FO, TA
Müller-Franzes, 2022 [52]	Lung, Liver, Kidney, Brain	CT, FLAIR MR	1. Inter-observer segmentation 2. Intra-observer segmentation	Shape, FO, TA
Urraro, 2021 [53]	CA Prostate	T2-w MR, ADC	Inter-observer segmentation	Shape, FO, TA
Wang,2020 [54]	Stomach	CT	1. Intra-observer segmentation 2. Inter-observer segmentation	Shape, FO, TA

1.3.3. Challenges in Image Preprocessing

Following image acquisition and ROI delineation, radiomic feature extraction is implemented through several established platforms, including PyRadiomics, LIFEx, CERR, and IBEX. Research has demonstrated that feature reliability exhibits significant platform dependency, with computational parameters substantially influencing feature calculations [55]. Despite international efforts toward standardization of feature computation [56], challenges persist regarding parametric variables in feature extraction, including resampling methodologies, image interpolation algorithms, and magnetic resonance bias correction protocols.

The Image Biomarker Standardization Initiative (IBSI), a collaborative international endeavor, has established standardized protocols for qualitative image feature definition and implementation, incorporating reference datasets for consensus calculations [56]. This comprehensive initiative, engaging 25 research teams utilizing diverse software platforms, achieved exceptional reproducibility in over 97% of investigated features. While this standardization significantly reduced inter-platform feature variability, notable limitations persist. Specifically, the IBSI framework does not encompass standardization protocols for filtered image features, including Log-Gaussian and wavelet transformations, despite their demonstrated utility in radiomic investigations. These filtered approaches have shown significant predictive value in various radiomic applications, highlighting a critical area for future standardization efforts.

Table 4 presents a review of 10 studies examining the influence of image preprocessing on radiomic features. The majority of these studies investigated the effects of image discretization methods, while others explored normalization techniques, and a smaller subset analyzed the impact of image resampling approaches. Although these studies face similar limitations as discussed in Sections 1.3.1 and 1.3.2, replicating preprocessing variations is relatively more straightforward compared to studying image acquisition or region of interest (ROI) variability.

Table 4. Literatures investigating the impact of image preprocessing on radiomic feature reproducibility and repeatability.

Author	**Site**	**Modalities**	**Sources of variation**	**Feature category**
Duron,2019 [57]	Breast	MR	Image discretization methods	TA
Fornacon,2020 [55]	H&N Lung	CT	Different feature extraction platforms	Shape, FO, TA
Gao,2022 [29]	Lung	CT	Image preprocessing	Shape, FO, TA
Hoebel,2021 [58]	Glioblastoma Brain	MR	Pre-processing Methods	Shape, FO, TA
Li,2020 [59]	Phantom	CT	Image preprocessing parameters	Shape, FO, TA
McHugh,2021 [36]	Liver	MR	1. MR sequence 2. Normalization	Shape, FO, TA
Moradmand,2020 [60]	Glioblastoma Brain	MR	Intensity inhomogeneity correction	Shape, FO, TA
Scalco,2020 [61]	CA Prostate	MR	Image normalization techniques	FO, TA
Schwier,2019 [62]	CA Prostate	MR	1. Image normalization techniques 2. Image discretization	Shape, FO, TA
Simpsons,2020 [63]	Human+ Phantom	MR	Image discretization methods	TA

1.3.4. Challenges in Modeling

Radiomic modeling encompasses two fundamental components: feature selection and model construction. Feature selection aims to reduce dimensionality through dual criteria: maximizing outcome correlation while minimizing inter-feature redundancy. Model construction leverages advanced machine learning algorithms to optimize outcome prediction accuracy.

Contemporary machine learning methodologies facilitate the extraction of clinically relevant information from radiomic features to support medical decision-making. However, model performance demonstrates significant sensitivity to both feature selection strategies and modeling methodologies. This phenomenon was notably documented by Parmar et al., who demonstrated variable model performance across unseen testing cohorts using different feature selection approaches and classification algorithms[64]. Despite this finding, comparative analyses of feature selection and modeling methodologies remain underrepresented in radiomic literature. The absence of universally optimal modeling approaches suggests dataset-specific optimization requirements [65].

It is crucial to distinguish between methodological variations in feature selection and model construction versus those encountered in image acquisition, reconstruction (section 1.3.1), segmentation (section 1.3.2), and preprocessing (section 1.3.3). While earlier workflow stages introduce variations in feature values, modeling methodology variations manifest as differences in model performance metrics. This distinction is fundamental, as diverse feature selection approaches may emphasize different data characteristics, representing complementary rather than problematic methodological diversity.

1.3.5. Summary of Current Challenge

The clinical implementation of radiomic models necessitates consideration of output reliability within acceptable error margins for routine application. The radiomics workflow inherently incorporates multiple sources of variability at each methodological stage, representing a fundamental challenge in the field. This intrinsic vulnerability was recognized during the early development of radiomics, as illustrated in **Figure 3**, which demonstrates the cascade of variations throughout the workflow.

Despite the exponential growth in radiomic literature, systematic evaluation and reporting of feature reliability have often been inadequately addressed by investigators. The absence of comprehensive reliability assessments for individual radiomic features presents a significant impediment to clinical translation. These persistent concerns regarding reliability and reproducibility have substantially decelerated methodological innovation in radiomics and constrained its potential for clinical implementation.

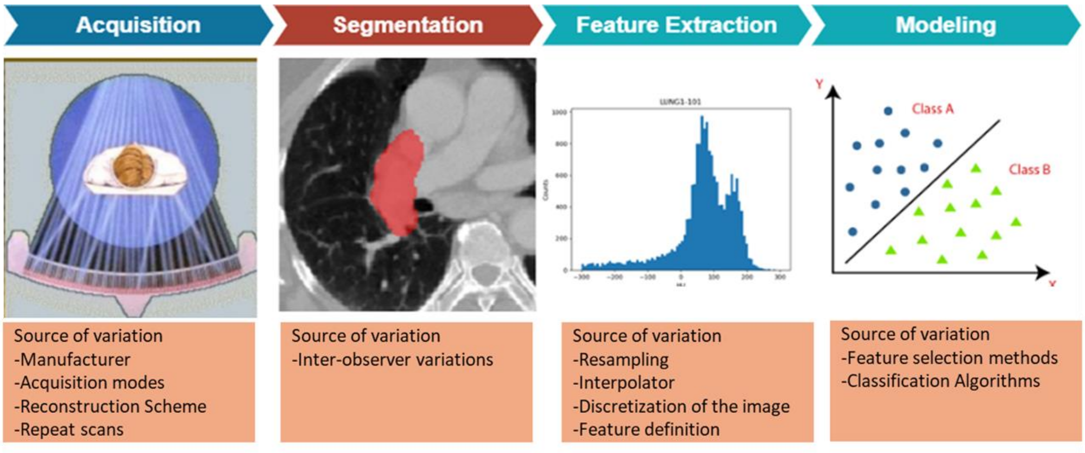

Figure 3. Sources of Variation Throughout the Radiomics Workflow

Chapter 2. Research Objectives

2.1. Research Gap

Despite the proliferation of radiomics research, significant challenges persist in developing robust and generalizable radiomic models for clinical applications. The absence of a standardized, resource-efficient methodology for assessing feature stability and model reliability across diverse imaging modalities and patient cohorts has impeded the translation of radiomics into clinical practice.

Conventional approaches for evaluating radiomic feature reliability, such as test-retest studies, are often impractical due to resource constraints and ethical considerations regarding additional patient radiation exposure. While alternative methods have been proposed, including the utilization of publicly available datasets or multi-observer segmentations, these approaches are limited in their generalizability or scope, and may not adequately capture the full spectrum of variability inherent in radiomics studies.

The potential of simulation-based perturbation methods to evaluate feature reliability without additional image acquisition has been recognized in recent literature. However, the application of such methods in assessing comprehensive model reliability and enhancing model generalizability, particularly in the context of heterogeneous head and neck cancers across multiple imaging modalities, remains insufficiently explored.

Furthermore, there is a dearth of comprehensive studies comparing the stability of radiomics and dosiomics features in head and neck cancers. The potential synergy of these features for enhanced prognostic modeling warrants rigorous investigation, especially considering the complex interplay between tumor characteristics and radiation dose distribution in this anatomical region.

While the significance of feature selection in radiomic modeling is well-established, there is a critical need for a systematic, data-driven approach to identify and eliminate low-reliability features prior to model development. The establishment of a robust, multi-institutional feature robustness databank to guide feature selection across different cancer types and imaging modalities represents a significant lacuna in current research.

Additionally, the impact of feature reliability on model generalizability across diverse patient cohorts and institutional settings remains inadequately characterized. This knowledge gap hinders the development of radiomic models that can perform consistently across heterogeneous clinical scenarios, a prerequisite for widespread clinical adoption.

The interplay between feature stability, model reliability, and clinical outcomes in head and neck cancers has not been thoroughly elucidated. Understanding these relationships is crucial for developing radiomics-based prognostic and predictive models that can meaningfully impact clinical decision-making.

Addressing these research gaps is imperative for advancing the field of radiomics and facilitating its integration into evidence-based clinical workflows. There is a clear exigency for a comprehensive, multi-faceted framework that can rigorously assess feature stability, guide judicious feature selection, and substantially improve model reliability and generalizability across different head and neck cancer subtypes and multimodal imaging paradigms.

2.2. Research Aim

The primary aim of this research was to establish and validate a comprehensive methodological framework for quantifying and enhancing radiomic feature reliability through multi-institutional, multi-modality analysis in head and neck cancer. Utilizing three retrospectively collected cohorts comprising 2,053 patients, we implemented a systematic perturbation-based approach to investigate feature repeatability patterns across diverse imaging protocols and institutional settings. This investigation uniquely synthesizes multi-modal imaging data to elucidate the intrinsic relationships between feature stability, image characteristics, and model performance. Through rigorous statistical analysis and validation, we sought to develop robust strategies for optimizing model reliability and generalizability. To our knowledge, this represents the first comprehensive investigation integrating multi-institutional and multi-modality analyses to systematically evaluate the impact of feature repeatability on radiomics model performance, thereby establishing a theoretical and practical framework for advancing reliable radiomics toward clinical implementation.

2.3. Research Objectives

2.3.1. Objective 1: To systematically quantify and compare feature repeatability across CT and MRI modalities in nasopharyngeal carcinoma.

We aimed to develop and evaluate a comprehensive framework using image perturbation to assess radiomics feature stability across different imaging modalities. Despite the radiomics community's recognition of feature stability importance, standardized methods for direct measurement across modalities have been lacking. Our study addresses this gap by providing a data-specific, practical method for evaluating feature reliability in nasopharyngeal carcinoma across CT and MRI modalities. The framework simulates variations in imaging and segmentation, quantifying feature stability through statistical measures. This approach offers a more practical alternative to resource-intensive test-retest methods, potentially advancing the field's ability to develop reliable and generalizable radiomic models. While our method shows promise, further validation against conventional approaches may be necessary to establish its efficacy in assessing radiomics feature reliability.

2.3.2. Objective 2: To evaluate and compare the feature repeatability in CT radiomics and dosiomics features and elucidate the relationships with image characteristics through comprehensive perturbation analysis.

This investigation aims to systematically assess and compare the stability of features extracted from planning CT images and radiation dose maps in cervical cancer patients. Through the implementation of image perturbation and contour randomization methodologies, we seek to quantify feature repeatability across different organs at risk (OARs) and analyze their associations with underlying image and dose characteristics. By establishing the first comprehensive reference for dosiomics feature repeatability while simultaneously elucidating confounding factors affecting reproducibility, this study aims to develop robust guidelines for reliable feature selection in both radiomics and dosiomics analyses, ultimately advancing our understanding of their complementary roles in predictive modeling and facilitating the development of more reliable clinical prediction models.

2.3.3. Objective 3: To validate the beneficial impact of feature repeatability on model performance and reliability through multi-institutional analysis.

This investigation aims to systematically quantify the impact of incorporating highly repeatable radiomics features on model performance and reliability through rigorous internal and external validation frameworks. Through the implementation of progressive ICC thresholds in feature selection and comprehensive analysis of their effects on model discrimination and generalizability across diverse head and neck cancer cohorts, we seek to establish the fundamental relationship between feature stability and model robustness. The study endeavors to demonstrate that the strategic prioritization of repeatable features in the modeling process enhances prognostic performance and strengthens the reliability of clinical predictions, with particular emphasis on scenarios characterized by limited sample sizes, thereby advancing the development of more robust and clinically applicable radiomics models.

Chapter 3. Assessment of Radiomics Feature Repeatability and Reproducibility and Their Generalizability Across Image Modalities by Perturbation in Nasopharyngeal Carcinoma Patients

3.1. Introduction

Medical imaging is widely used and has an important role in clinical oncology practice. Biomarkers based on medical imaging can be used for screening, staging, intervention planning, and treatment outcome prediction [11, 66-68]. In the current practice of manual evaluation of medical images, radiologists only semantically annotate a small number of clinically significant radiological features. Tumor phenotypes embedded in medical images may contain more information that cannot be easily processed by the naked eyes [68-71]. Radiomics is a computer-based technology for extracting and analyzing quantitative features from medical images. It surpasses the level of details available to the naked eyes and aims to automatically mark clinically significant tumor phenotypes [72].

There are potential pitfalls in radiomics analysis that could jeopardize the generalizability and robustness of established biomarkers. Several approaches have been proposed to reduce the risk of false discovery [73-75]. In particular, repeatability and reproducibility are the first and foremost criteria towards clinical utility. "Repeatability" refers to features that remain the same when imaged multiple times in the same subject. "Reproducibility" refers to features that remain the same when imaged using different equipment, different software, different image acquisition settings, or different processing settings. They should be incorporated into feature pre-selection strategy and downstream predictive model construction in any radiomic studies. On top of that, identifying the stability of radiomics features (RFs) across different image modalities will provide the radiomics community with direct perceptivity for selecting reliable radiomic features and building robust predictive models for implementing precision medicine.

Efforts attempting to bridge this important gap in knowledge have been mainly focused on test-retest experiments [45, 76], which have considerable shortcomings. First, the impact of tumor segmentation variation is often missed in test-retest studies. However, tumor segmentation variability can propagate into significant variability in radiomics feature stability [77, 78]. Two published studies [49, 79] have shown that MR RFs displayed better stability than CT under segmentation variability. Additionally, the limited sample size owing to the need for recruiting consented patients renders their conclusions less statistically convincible. Last but not least, multi-modality and multi-center based RFs stability study is ignored by the limited dataset.

To address these limitations, we attempted to deploy our in-house developed perturbation(image perturbation and contour randomization) framework, taking reference from previous work by Zwanenburg et al [45], to mimic a vast amount of scanning position and tumor segmentation stochasticity via large patient cohorts of nasopharyngeal carcinoma (NPC) patients. Furthermore, we also compared the RF stability under perturbation across three imaging modalities, which is yet to be explored.

Accordingly, the objectives of this study are: (i) to ascertain the repeatability and reproducibility of radiomics features via perturbation; and (ii) to examine their generalizability across imaging modalities for NPC patients.

3.2. Materials and Methods

3.2.1. Overview

Fig. 4 illustrates the overall study workflow. An internal NPC cohort of 397 patients which consists of contrast-enhanced computed tomography (CECT), contrast-enhanced T1 weighted (CET1-w) MR, and T2 weighted (T2-w) MR were enrolled in this study. Each image modality dataset was processed through preprocessing, image perturbations (rotation and translation), contour randomization and RF extraction before stability evaluation. By comparing the RF stability between CT and MRI, we assess differences in RF stability performance across these imaging modalities.

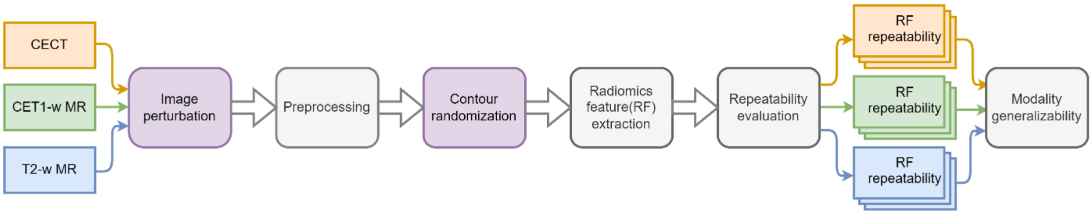

Figure 4. Overall Study Workflow.

3.2.2. Patient Cohort

A total of 397 biopsy-proven (Stage I-IVB) NPC patients who received cancer treatment at the Department of Clinical Oncology of Queen Elizabeth Hospital (QEH) between 2012 and 2016 were retrospectively screened, and 331 patients who had same-institution MR images and eligible target contours were enrolled in the study.

3.2.3. Image Acquisition & Image Preprocessing

All imaging data were acquired in a Digital Imaging and Communications in Medicine (DICOM) format archived using Picture Archiving and Communication System (PACs). All the calculations were performed by our in-house developed Python-based (3.7.3) pipeline using the SimpleITK (1.2.4) [80] and PyRadiomics (2.2.0) package [9]. The detailed workflow is illustrated in **Figure 5(a)**. For MR images, the signal intensity was normalized using the brainstem as a reference structure, and N4B bias correction from SimpleITK was employed for MRI inhomogeneity correction.

3.2.4. Perturbation

We designed the contour randomization as randomized deformation of the original contour. A randomized deformation vector field (DVF) is first generated, followed by normalization and gaussian smoothing controlled by the adjustable kernel size. The deformation field projection on the z-axis is kept constant for the same slice to mimic the slice-by-slice contouring. The final DVF is then scaled by a user-defined factor to control the intensity of the randomization. Finally, the original contour is deformed by the randomized DVF to acquire the randomized contour. The contour randomization will be repeated multiple times, and radiomics features will be extracted from the image masked by the perturbed contours.

Image perturbations and contour randomizations were both applied to each pair of the preprocessed original-resolution image and mask during isotropic (1 mm x 1 mm x 1 mm) resampling after Gaussian anti-aliasing filtering.

Two image perturbation modes, rotation (θ ∈ [-20˚, 20˚], step size = 5, around central z-axis) and translation (μ ∈ [0.00, 0.80], step size = 0.2, along all three dimensions) were implemented following the procedures proposed by Alex et al. [45] to mimic variations in scanning setup positions during image acquisition.

The contour randomization intensity was tuned based on the resulting randomized contour Dice similarities and Hausdorff Distance. For one complete perturbation, the three image perturbation modes are applied to the original image simultaneously using one set of parameters chosen from the total of 125x5x4=2500 combinations. The contour was randomized independently in addition to the image perturbation. Choices of parameters for different patients were independent to generate the broadest range of perturbations with the minimum computational cost.

3.2.5. Feature Extraction

Feature computation was performed on the perturbed images using PyRadiomics. Before feature extraction, the perturbed images were preprocessed by isotropic resampling to 1 mm x 1 mm x 1 mm, and the pixel values were shifted by the same off-set value of 2000 and further discretized into a fixed bin width of 5. In addition to feature extraction on the original image, Laplacian-of-Gaussian (LoG) filters (Sigma values of 1, 2, 3, 4 and 6 mm) and coilf1 wavelet filters (HHH, HLL, LHL, LLH, LHH, HLH, HHL, LLL) were applied to yield advanced features. The entire set of radiomics features, except shape features, were extracted using the widely used Python package PyRadiomics.

A total of 1288 features were computed for each image. The main groupings of texture analysis features were (1) First-order statistics based on pixel gray-level histograms, 18 features; (2) Shape metrics, 14 features; (3) Statistical features derived from texture matrices including gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), neighboring gray tone difference matrix (NGTDM), 73 features; (4) Statistical features derived from texture matrices in Laplacian-of-Gaussian (LoG) filtered domain, 455 features; and (5) Statistical features derived from texture matrices in wavelet filtered domains, 728 features.

3.2.6. Statistical Analysis

Feature stability was quantified using the intraclass correlation coefficient (ICC). Since the perturbation parameters were independently applied to images and masks of different patients, the lower 95% confidence interval of one-way, random, absolute ICC was employed to assess RF repeatability. The calculation was performed by our in-house developed algorithm following the equations presented by McGraw et al[81]. In our study, median ICC values (mICC) under patient subsampling were adopted as the final metric for assessing RF stability to minimize the potential impacts of outlier patients. In detail, all patients were partitioned into subgroups of 40 patients, which was repeated 20 times with shuffling, resulting in around 200 patient subgroups. Here, an ICC of ≥ 0.75–0.89 was considered good reproducibility and an ICC ≥ 0.90 was considered excellent reproducibility as recommended by Koo et al. [82].

To compare RFs repeatability performance in different image modalities under contour randomization, we adopted the pairwise Wilcoxon signed-rank test on the ICC value of RFs in each modality. The p-values of the statistical test for all the modalities were tabulated. The tests were one-sided, p-value <.05 was considered as significant.

The Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) were used as evaluation metrics for the spatial overlap accuracy of the randomized contours for the perturbation images. A Dice Similarity Coefficient of 0 indicated no overlap and a value of 1 corresponded to exact overlap [83]. The Hausdorff Distance which is smaller than 6mm is considered as good overlap accuracy.

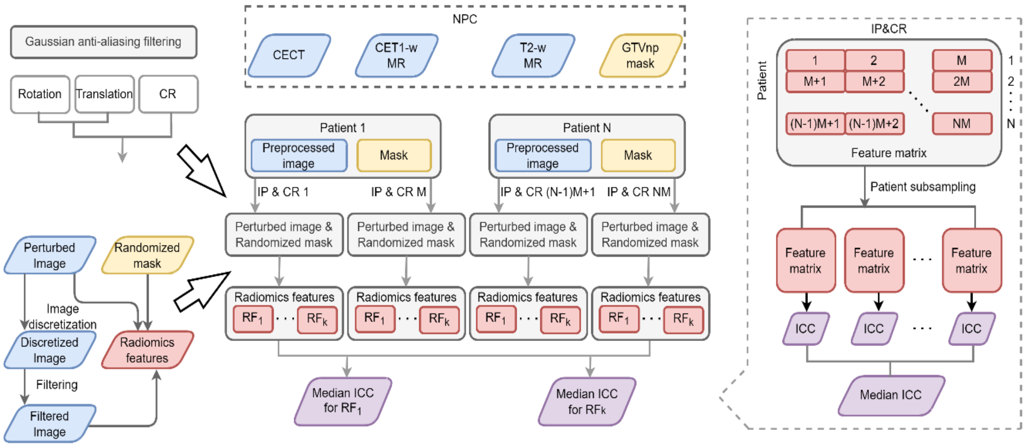

Figure 5. Detailed workflows of (a) radiomics feature extraction under perturbation

and (b) evaluation of its stability

3.3. Results

The Dice Similarity Coefficients and Hausdroff Distance were calculated respectively for the perturbed images on each patient. The Dice Similarity coefficients and Hausdroff Distance were 0.82 IQR [0.79, 0.85] and 3.2 mm IQR [3, 4.2].

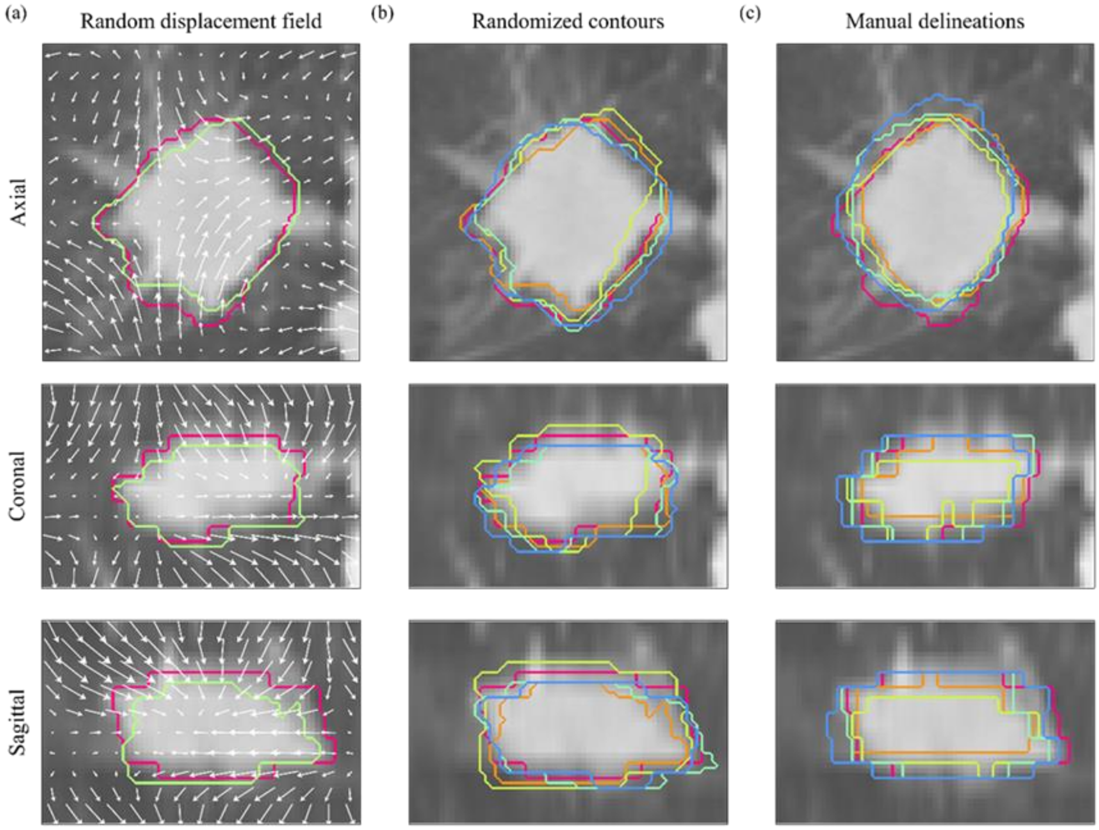

Figure 6. One example of random displacement field of one slice on the three directions were shown in (a), where the original and the corresponding randomized contour were shown by the red and light green lines respectively. A total of 5 randomized contours in changing colors were superimposed in (b). Similar variations

in manuals contours were observed, as shown in (c).

The number of RFs that fell within either the "good" (0.9 > mICC ⩾ 0.75) or "excellent" (mICC ⩾ 0.9) category for each modality is presented in **Table 5**. All the shape metrics features fell into the "excellent" category in both CT and MRI. Overall, the CT-based RFs showed the fewest percentage with "excellent" category, 41.7% of all features. This contrasts with the MRI-based RFs from which 77.6% and 80.2% of features had "excellent" stability in CET1-w and T2-w respectively. Across all three imaging modalities, 1069 common features out of the total 1288 features (including all image domains) had a "good" mICC value ⩾ 0.75, and 497 features had an "excellent" mICC value ⩾ 0.9.

Table 5. Number of features(n) and percentage of their groups (%) which fall into the "excellent" category (mICC≥0.9) and "good" category (mICC≥0.75) for all features and distinct feature types (first-order, shape, texture, LoG filtered and Wavelet filtered)

	CECT		*CET1-w*		*T2-w*	
	n	%	n	%	n	%
All features (1288)						
mICC ⩾0.9	537	41.7	1000	77.6	1033	80.2
mICC ⩾ 0.75	1086	84.3	1235	95.9	1229	95.4
First-order (18)						
mICC ⩾ 0.9	10	55.5	17	94.4	18	100
mICC ⩾ 0.75	16	88.9	18	100	18	100
Shape Metrix (14)						
mICC ⩾ 0.9	14	100	14	100	14	100
mICC ⩾ 0.75	14	100	14	100	14	100
Texture (73)						
mICC ⩾ 0.9	23	31.5	61	83.5	72	98.6
mICC ⩾ 0.75	69	94.5	71	97.3	72	98.6
LoG (455)						
mICC ⩾ 0.9	178	39.1	365	80.2	391	85.9
mICC ⩾ 0.75	400	87.9	444	97.5	440	96.7
Wavelet (728)						
mICC ⩾ 0.9	312	42.9	543	74.6	541	74.3

mICC ⩾ 0.75	587	80.6	688	94.5	684	94

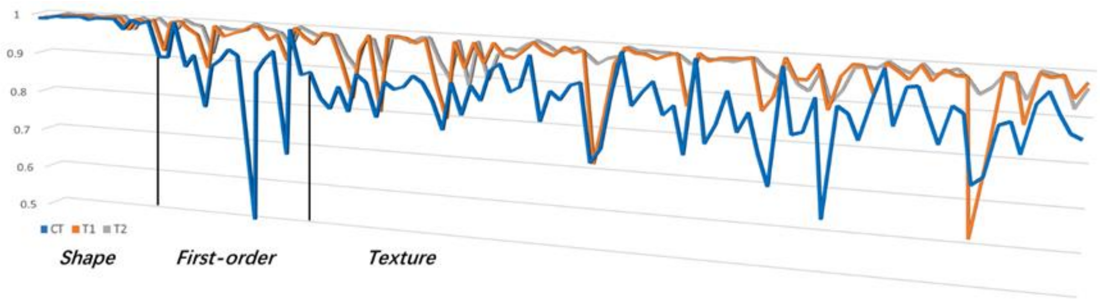

Figure 7. 3D line chart illustrating the for the shape (n = 14), first-order (n = 18) and texture features (n = 73) derived from original image for CT and MR

First-order and texture features were calculated in 13 image domains: the original image, 5 images with LoG filter with kernel sizes(1,2,3,4,6mm), and 8 images from the wavelet decompositions. To explore any variation in feature stability (mICC) by image domains, mICCs for the 18 first-order and 73 texture features were combined by image domain in **Fig. 7**. The mICCs from the features for the original image are included in the graph for comparison.

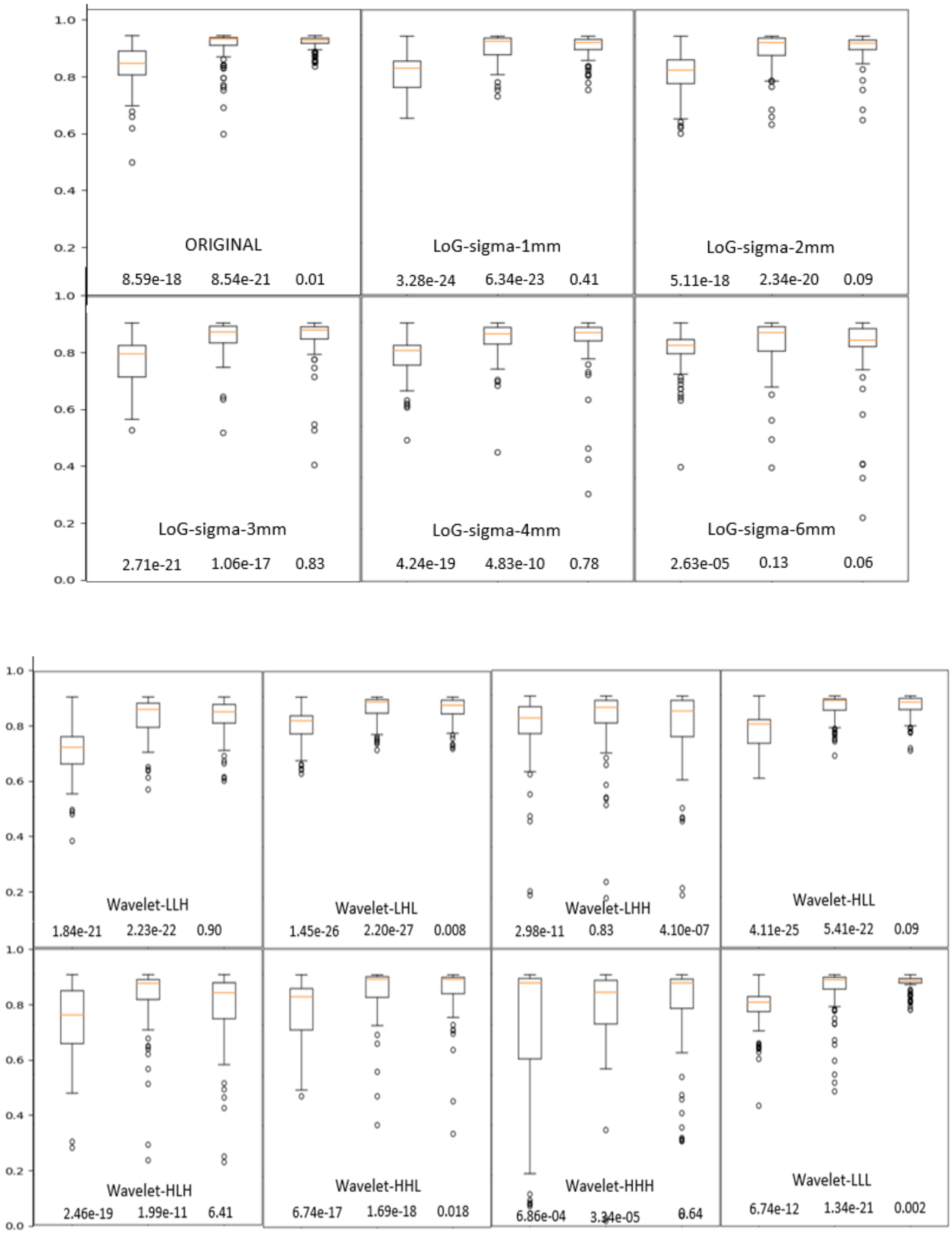

Figure 8. Boxplot of ICC distribution and p-value between each two modalities divided by different image modalities and filter categories. In each small cell, the three boxplots represent CT, T1 and T2 from left to right. The p-values below are calculated by CT-

T1, CT-T2, and T1-T2 in turn.

The CT RFs demonstrated significantly lower mICCs in most LoG-filtered image domains and Wavelet-filtered domains(p-value<0.01). However, in LoG-filtered domains, the percentage of stable MR RFs decreased from 94% to 77% and from 92% to 71% when the kernel size changed from 1mm to 6mm. Furthermore, the percentage of stable MR RFs ranges from 50% to 99% and from 97% to 55% under different wavelet decomposition filters. In particular, the CT and MR RFs exhibited similar stability in LoG-6mm and Wavelet-HHH domains (p-value > 0.1).

3.4. Discussion

Radiomics has emerged as a means of image-based prognostication. Ensuring radiomic feature stability is imperative to the external generalizability of downstream predictive models. It is anticipated that this study could provide actionable insights in the selection of stable radiomic features by providing the information of feature repeatability and reproducibility of radiomic features across different imaging modalities. Specially, the two perturbation modes adopted mimicked the variation during image acquisition and tumor segmentation. The image perturbation aims to simulate unavoidable random positional variations during image acquisition. Meanwhile, contour randomization evaluates the random errors derived from manual tumor delineations in clinical scenario.

Three conclusions can be drawn from this study. Firstly, shape features demonstrated the highest repeatability and reproducibility in all modalities. Shape features are generally reported as highly repeatable and reproducible in the literature and were shown to be less sensitive to CT segmentation variation in a phantom study [84]. Further, MRI-based shape features were found to be stable in test-retest of cervical cancer [85, 86]. A recent systematic review, mostly based on CT studies, concluded that shape features showed higher reproducibility than texture features [87].

Secondly, CT-based RFs are more sensitive to scanning position and tumor segmentation variation than MRI-based RFs. The MRI-based RFs were more stable than CT-based RFs. Specifically, the number of repeatable features derived from CT was fewer than the other two modalities. Of the 537 stable features in CT, 92.5% was also stable in the other two modalities. Furthermore, the MRI-based RFs have overwhelming performance when comparing mICC to CT-based RFs with 86.8% and 83.9% RFs with higher mICC from CET1-w and T2-w respectively. However, it is challenging to compare our findings to previous literatures as multi-modalities features were uncommonly studied, especially considering the variation in both image acquisition and ROI delineation.

Thirdly, there is no substantial difference in feature stability between the original and filtered image domains. Wavelet and LoG-filtered images showed both better and worse reproducibility than the original images in the three modalities in this study. Similarly, Schwier et al. demonstrated no significant improvement in reproducibility with a certain LoG-filter or wavelet decomposition [88]. On the other hand, Timmeren et al. reported that wavelet features were less reproducible than the unfiltered image features in a test-retest scenario [89]. The number of stable RFs (mICC ≥ 0.9) derived from CECT increased from 24 (26.4%) to 55 (60.4%) when the kernel size changed from 1mm to 6mm in LoG-filtered image domains whereas this number decreased for MRI-based RFs, 78 (85.7%) to 67 (73.6%) in CET1-w and 81 (89%) to 72 (79.1%) in T2-w. Additionally, Fave et al. reported coarseness, gray length nonuniformity and run length nonuniformity as reproducible for NSCLC cone-beam CT [90]. Leijenaar et al. reported that GLCM and GLRLM were more reproducible than GLSZM, each of which encompasses at least one feature which appeared in our study as reproducible [91].

We acknowledge limitations in our study. First, our perturbation algorithm may not fully mimic the variation in clinical scenarios owing to technical challenges in fully simulating all the variables. In recent studies, image acquisition, preprocessing, and feature extraction such as image acquisition setting and image reconstruction algorithm were shown to have more significant influence on RFs stability [92]. In addition, this was a single-institutional retrospective study that may not be representative of other institutions or patients. Despite the validation of the PyRadiomics platform, results may differ from other radiomic feature extraction platforms. Additionally, although we used commonly reported cut-offs from the literature for ICC categories (0.75 and 0.9), they may not represent the ideal thresholds for feature inclusion in prognostic models. Finally, considering the similar anatomic environment within head and neck cancer, further investigation of other cancer types (e.g., Oropharyngeal cancer) is warranted.

3.5. Conclusions

Our work is the first study to intentionally scrutinize RF robustness disparity against scanning position and segmentation variations in multi-modality imaging datasets with big sample sizes. In conclusion, CT-based and MRI-based RFs of NPC were evaluated for their repeatability and reproducibility. Shape features emerged as the most stable both in CT and MRI. CT-based RFs displayed higher sensitivity against the scanning position and tumor segmentation stochasticity than MR-based RFs, highlighting the importance of careful feature selection for radiomics generalizability. The feature repeatability results identified by the rather conservative randomizations in this study can be used as the fundamental requirements for building reliable radiomic models in future studies.

Chapter 4. Comparative Analysis of Repeatability in CT Radiomics and Dosiomics Features under Image Perturbation: A Study in Cervical Cancer Patients

4.1. Introduction

Cervical cancer poses a significant healthcare burden, underscoring the necessity for precise and individualized treatment approaches [93]. Texture analysis (TA) technology allows for the evaluation of spatial and statistical voxel intensity distributions within an image, thereby providing valuable information about patterns and voxel correlations [68]. In the context of cervical cancer, TA has demonstrated promising potential in improving diagnostic precision, predicting treatment response, and facilitating personalized treatment planning [93-96].

Radiomics has emerged as a critical component in tailoring personalized treatment strategies and monitoring treatment response by analyzing an extensive array of quantitative features extracted from medical images [9, 72, 97-99]. Chiappa et al. extracted radiomics features from the preoperative ultrasound images and developed machine learning models to accurately predict the malignancies of ovarian masses in the AROMA pilot study [100]. Promising performance of radiomics diagnosis was also reported by the same group in classifying the malignancies of uterine mesenchymal lesions [101]. Radiomics has also demonstrated potential in treatment response prediction, such as the neoadjuvant chemotherapy response for patients with cervical cancer [102]. Dosiomics, a nascent field built upon principles developed in radiomics, focuses on extracting high-dimensional data from three-dimensional radiation dose distributions to aid in clinical decision-making [103]. The integration of dosiomics in radiation therapy has received significant attention due to its potential applications in modeling normal tissue complications, predicting radiation-induced toxicity, and forecasting tumor control outcomes [104, 105]. Dosiomics entails the quantitative evaluation of dosimetric parameters, including maximum dose, mean dose, and dose homogeneity, to comprehensively assess the spatial and dosimetric attributes of the tumor and surrounding tissues during radiation therapy. These features offer insights into the intricate details of dose distribution within the treatment region. Simultaneously, radiomics delves into the complexities of medical images, extracting texture, shape, and intensity features that furnish valuable information about tumor heterogeneity, microenvironment, and underlying molecular characteristics [106, 107].

Notwithstanding the potential advantages of dosiomic and radiomic features in the management of cervical cancer, it is imperative to conduct a comprehensive investigation into their stability. Variations in imaging acquisition protocols, encompassing variances in scanners, imaging parameters, and segmentation methodologies, have the potential to introduce inherent variability into the extracted features [108]. Furthermore, the presence of image noise, stemming from factors such as suboptimal image quality or motion artifacts, can significantly influence the stability and reproducibility of dosiomic and radiomic features [109]. Gaining a comprehensive understanding of the repeatability of dosiomics and radiomics features is crucial for their successful integration into routine clinical practice [110]. This knowledge ensures the reliability and consistency of feature extraction, enabling accurate and robust analysis for tasks such as treatment response prediction, treatment planning optimization, and patient stratification [111, 112]. Ultimately, such insights have profound implications for improving treatment outcomes and optimizing personalized care for cervical cancer patients. While existing research has predominantly focused on the repeatability of radiomics features derived from CT and MR images using test-retest imaging and multiple delineation, limited attention has been given to studying dosiomics features due to the challenges associated with obtaining repeated measurements [113]. Furthermore, inconsistent repeatability results have been observed across different image modalities and cancer sites, thus limiting the generalizability of these findings to new radiomics studies [12, 114].

In this study, we aim to investigate the repeatability of dosiomics and radiomics features extracted from planning CT and dose maps of patients with cervical cancer. We first assessed the repeatability of radiomics and dosiomics features using image perturbation and contour randomizations on different organs at risk (OARs). They were compared both continuously and in binary forms as repeatable and non-repeatable features, followed by analyzing their associations with image/dose appearance. Our findings will provide the first reference of dosiomics feature repeatability for cervical cancer and reveal the confounding factors for radiomics feature reproducibility. This will guide the reproducible dosiomics feature selection for further research endeavors and help to reach a consensus on radiomics feature repeatability under different scenarios.

4.2. Material and Methods

4.2.1. Patient Dataset

We retrospectively recruited cervical cancer patients with age >18 years old and received complete curative radiotherapy courses from 2012 to 2022. Patients with missing treatment planning CT and dose data and metal artifact in the planning CT except from intrauterine contraceptive device were excluded. A total of 304 patients were included in this study, and the planning CT images and dose maps were collected from the treatment planning system in DICOM format. All CT scans were conducted using a Brilliance Big Bore CT scanner (Philips Healthcare, Amsterdam, The Netherlands). The scanning parameters included a tube voltage of 120 kV, an exposure of 300/325 mAs, an image resolution of 512 × 512, a pixel size of 0.98 × 0.98 mm2, and a slice thickness of 5 mm. Five distinct Regions of Interest (ROIs) were delineated manually by radiation oncologists with over 5 years of experience. These ROIs comprised the clinical tumor volume (CTV), Bladder, Rectum, Left Femoral (LFemoral), and Right Femoral (RFemoral). The contouring of the CTV adhered to the updated RTOG protocol released in 2021 [115].

4.2.2. Image Perturbation

Image perturbations were performed by random translations and rotations accompanied by contour randomizations. Each image was translated by 0, 0.4, or 0.8 pixel and rotated by –20, 0, or 20 degrees 40 times. The translation and rotation parameters were chosen randomly for each perturbation. Contour randomizations simulate multiple delineations of the same structure. A 3D random displacement field deforms the segmented mask and results in a randomized contour. The approach for generating random displacement fields is derived from the methodology introduced by Simard et al. [116]. In this adaptation, random vector components for the x and y dimensions were randomly generated following a uniform distribution ranging from −1 to 1 for each voxel point. Notably, no deformations were introduced along the z-axis due to the slice-by-slice contouring procedure typically employed in clinical settings. Subsequently, these field vectors were normalized across all three dimensions using the root mean square method. To ensure a smooth and continuous transition in the random displacement fields and prevent abrupt changes in the deformed contours, a Gaussian filter with a sigma value of 5 was applied. **Figure 9** shows one example of random displacement field for the studied ROIs (except RFemoral since it is highly symmetrical to LFemoral), and the original and the corresponding randomized contour are visualized by the red and green lines, respectively. Four randomized contours in different colors and the original contour in red are superimposed in the second row of **Figure 9**. An in-house developed Python program was used to perform image perturbation.

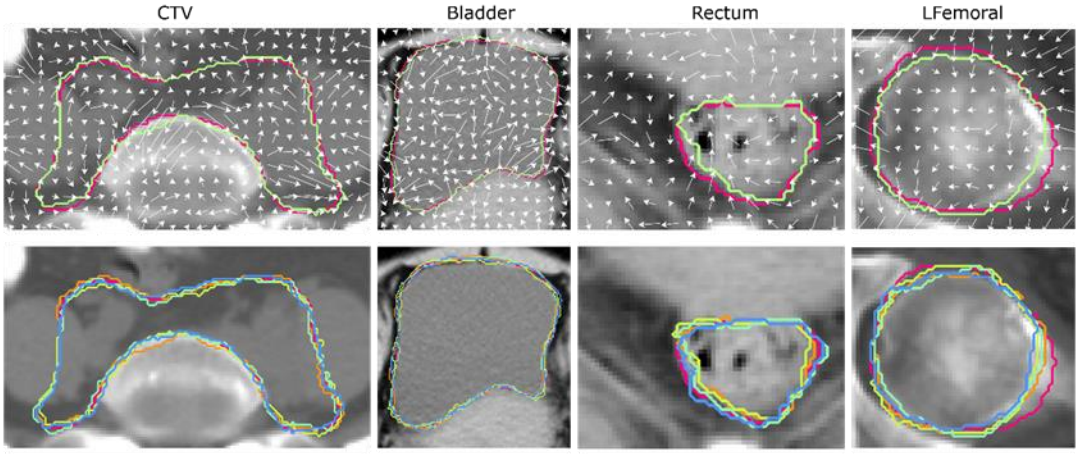

Figure 9. The first row shows rRandom displacement fields (white arrows) in CTV, bladder, rectum, and LFemoral (first row) and two randomized contours for the four ROIs (red: original contour, light green: randomized contour) overlayed with the CT image. The second row shows four different randomized contours with different colored lines in addition to the original contour (red line) (second row) of one example patient.

4.2.3. Radiomics Feature Extraction

The same set of radiomics features were extracted from the CT and dose maps for the five ROIs on each pair of the perturbed images and segmentation. Features extracted from the dose maps were considered as dosiomics in this study. The extracted features encompassed a comprehensive set, including 18 first-order statistics, 24 gray level co-occurrence matrix (GLCM) features, 14 gray level dependency matrix (GLDM) features, 16 gray level run length matrix (GLRLM) features, 16 gray level size zone matrix (GLSZM) features, and 5 neighboring gray-tone difference matrix (NGTDM) features. Each CT image/dose map was filtered using three-dimensional Laplacian-of-

Gaussian (LoG) filters with five different sigma values (1, 2, 3, 4, and 5 mm), as well as the complete set of eight coif1 wavelet filters (different combinations of high- and low-pass on each dimension) [117]. All the original and filtered images were further discretized by a fixed bin number of 32. One example of preprocessed CT and dose images under different filters is shown in **Figure 10**. In total, 1302 features were extracted for each CT/dose map, ROI, and perturbation. The open-source Python package PyRadiomics (version 3.0.0) was used to perform radiomics feature extraction.

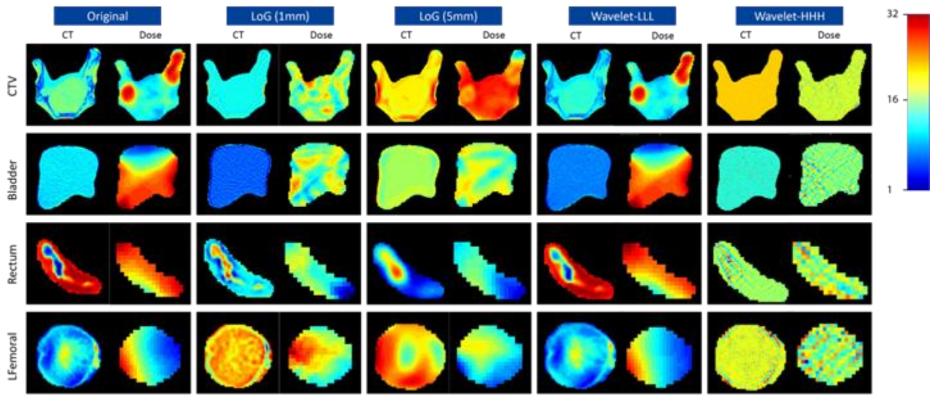

Figure 10. The original, Laplacian-of-Gaussian (LoG) filtered (sigma = 1 and 5 mm), and wavelet (LLL, HHH) filtered images of CT and dose maps within CTV, bladder, rectum, and LFemoral of one example patient. All the images were preprocessed by a 32-bin-number gray level discretization with the final pixel values ranging from 0 to 31. A jet colormap was used to present the voxel values with blue colors for smaller values and red colors for larger values.

4.2.4. Feature Repeatability Analysis

The one-way, random, absolute intra-class correlation coefficient (ICC) was used to assess the repeatability of each radiomics/dosiomics feature against the 40 perturbations. The binarized feature repeatability was measured using the ICC threshold of 0.9, where a feature was considered high-repeatable if ICC ≥ 0.9 and low-repeatable if ICC < 0.9. The threshold was determined based on previous publications [82] on radiomics feature repeatability analysis.

4.2.5. Statistical Analysis

We compared the radiomics and dosiomics repeatability in both continuous and binarized forms. The average ICC value for each image filter and feature class was first calculated and compared. Comparisons on binarized feature repeatability were presented by highlighting the ratios of commonly high-/low-repeatable features and disagreements between CT and dose on different image filters and feature classes. In order to explain the similarities and differences among different image filters and between the two data modalities, we analyzed the correlations between inherent image characteristics and feature repeatability. The mean values of the entropy, uniformity, and variance of the preprocessed image among all the patients were calculated for each image filter and ROI. These three metrics measure the complexity, heterogeneity, and contrast level, respectively, and were directly acquired from the original first-order radiomics features. Their definitions can be found in the PyRadiomics documentation. The Pearson correlation coefficient (r) was then used to quantify the correlations between the image characteristic evaluations and the average ICC values of the radiomics features [82].

4.3. Results

4.3.1. Feature Reliability and Predictability

In general, we have observed higher ICC values of radiomics and dosiomics features extracted from the original, large-sigma LoG filtered, and LLL-/LLH-wavelet filtered images. The rest of the wavelet filters yielded significantly lower feature repeatability with average ICC < 0.75, as shown in **Figure 11**. Fluctuations of mean ICC values were also observed across different feature classes. Specifically, the first-order features exhibited the highest repeatability while the GLSZM features had the lowest. Compared with CT radiomics features, dosiomics feature repeatability were lower, especially after small-sigma LoG and wavelet filtering, and experienced larger deviations across different image filters. One exception on the bladder is that dosiomics features had higher mean ICCs under large-sigma (≥3) LoG filtering. On the contrary, **Figure 11** illustrates that the feature class had a minimum impact on the consistencies between CT radiomics and dosiomics feature repeatability in terms of mean ICC values.

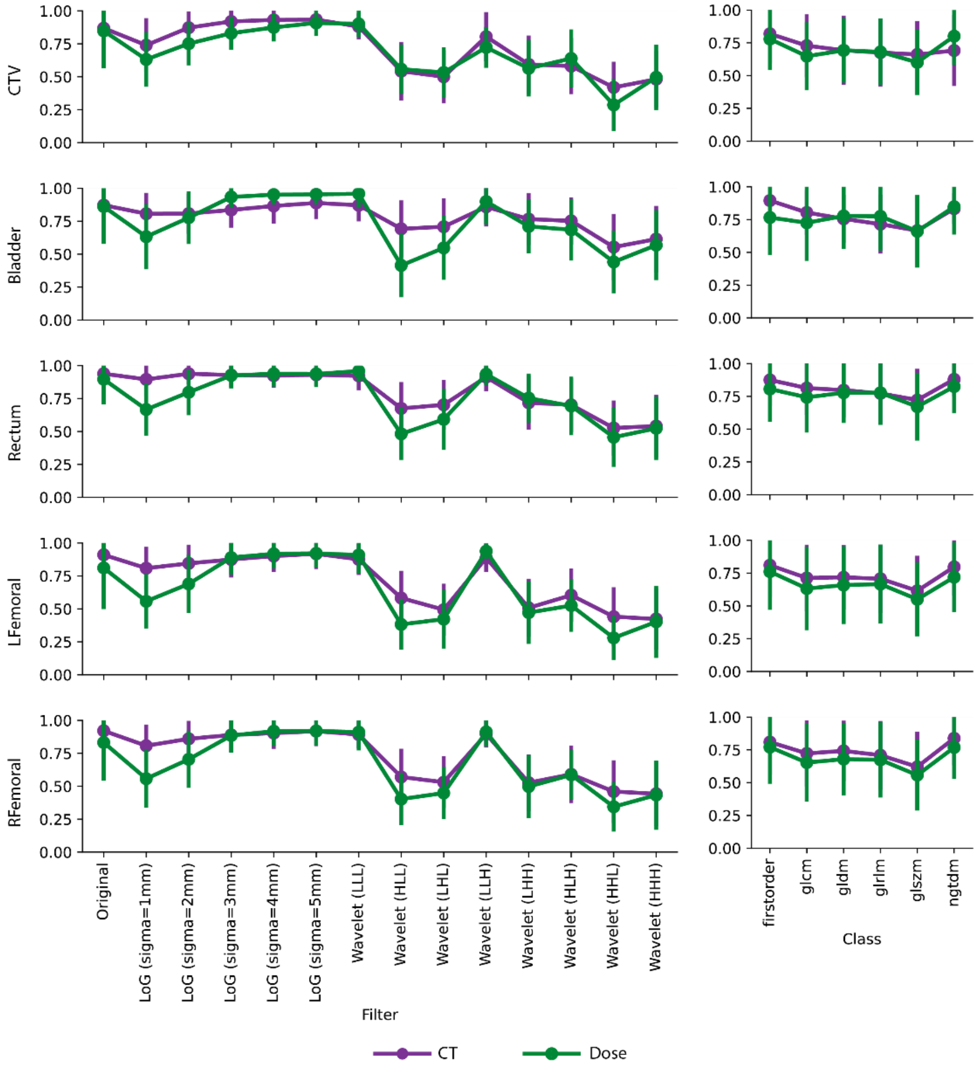

Figure 11. Continuous intraclass correlation coefficient (ICC) comparisons between CT radiomics and dosiomics features on CTV, bladder, rectum, LFemoral, and RFemoral. The mean ICC values averaged on each image filter (left column) and feature class (right column) were plotted as purple (CT) and green (dose) dots with bars indicating the standard deviation. In general, higher mean ICCs were achieved by the CT radiomics compared to dosiomics. Original, large sigma LoG filters, and low-pass wavelet filters resulted in higher mean ICCs compared to other image filters. Rather consistent ICCs were found for different feature classes.

Similar trends can be observed after binarizing the ICC values by the threshold of 0.9, as visualized in **Figure 12**. More repeatable features were found on the original, large-sigma LoG filtered, and LLL-/LLH-wavelet filtered images. Increasing the sigma values of the LoG filter resulted in more repeatable features. On the other hand, minimum repeatable features were found on the rest of the wavelet filtered images. For feature classes, the first-order class had the largest number of repeatable features while the GLSZM features had the smallest. When comparing the binary consistencies of repeatability between CT radiomics and dosiomics features, large deviations (light green/purple bars) can be observed mostly on CTV, bladder, and rectum. Different image filters also affected the consistency patterns. For example, features that are repeatable in dose but non-repeatable in CT (light green) were mostly found in the original image of the three ROIs, which is different from the mean ICC results. Features that are only repeatable in CT (light purple) were more prevalent in CTV under large-sigma LoG filtering and Rectum under small-sigma LoG filtering.

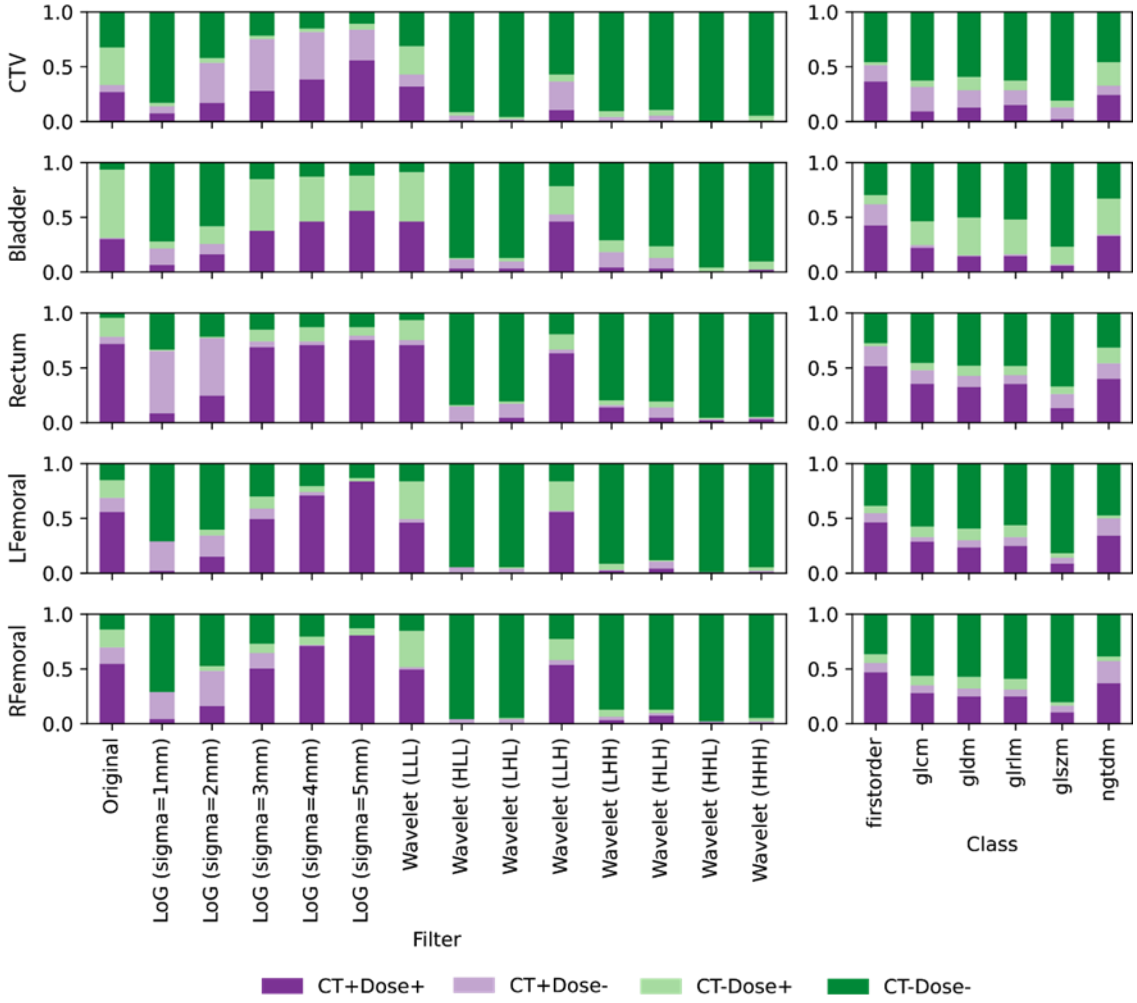

Figure 12. Comparisons of radiomic feature repeatability between CT and planning dose, binarized by the ICC threshold of 0.9. High consistencies can be mostly observed on rectum, LFemoral, and RFemoral for RFs extracted from the original, large sigma (≥3) LoG filtered, and wavelet filtered images/dose maps. Different feature classes demonstrated high consistencies regardless of the ROIs analyzed.

Strong correlations between entropy, uniformity, and variance between the preprocessed images and feature repeatability were discovered, regardless of the data modality (**Figure 13**). The mean entropy, which measures the randomness of the images, had positive correlations with feature repeatability on both CT (r = 0.513) and dose (r = 0.682). A high positive correlation of mean variance (r = 0.617, 0.741) was also observed on CT and dose. On the other hand, uniformity, which measures the image homogeneity, had a negative correlation (r = −0.450, −0.599).

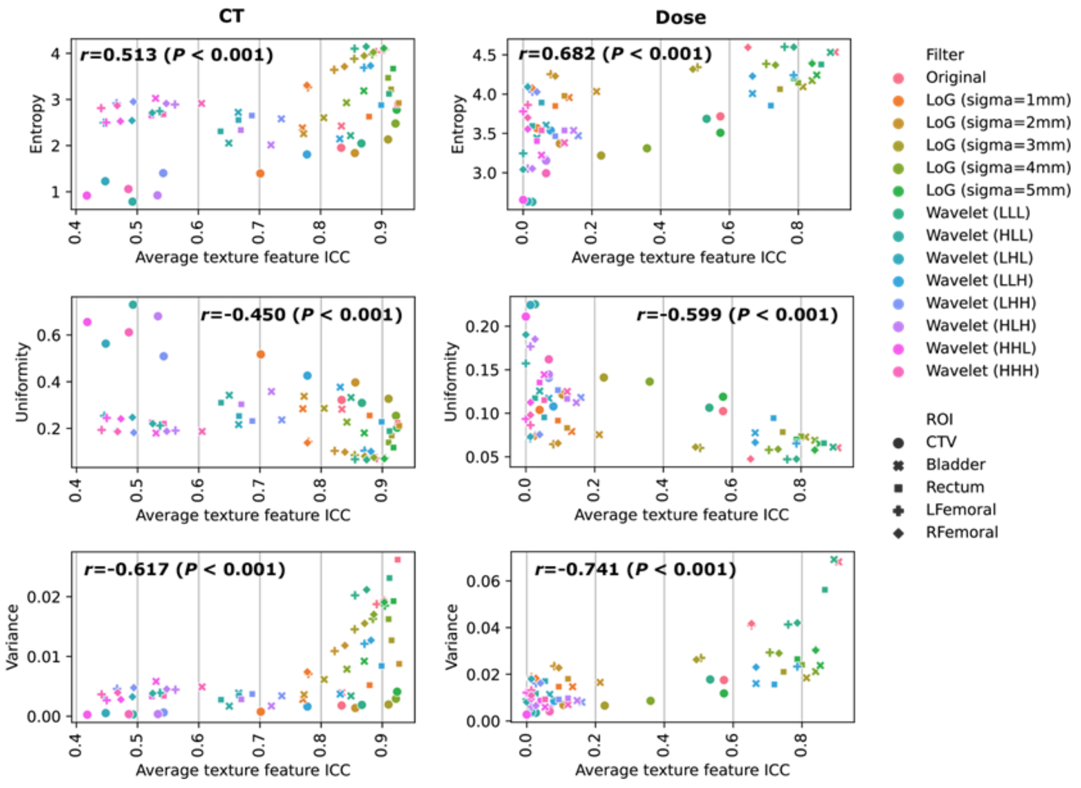

Figure 13. Correlations of mean entropy, uniformity, and variance of preprocessed images with average ICC values of radiomics features at different image filters and ROIs. The Pearson correlation coefficient r and its p-value were given on each plot.

4.4. Discussion

This study, for the first time, assessed and compared the repeatability of radiomics and dosiomics features from the planning CT and dose maps of primary cervical cancer patients using image perturbation. Features extracted from five different ROIs, including CTV, bladder, rectum, LFemoral, and RFemoral, were independently analyzed. A new contour randomization method was introduced to mimic the manual contouring variations by random deformations. In general, features from large-sigma LoG filtered and LLL-/LLH-wavelet filtered images had higher repeatability, both by absolute and binarized ICC. CT radiomics features presented smaller ICC fluctuations across image filters and had higher repeatability compared to dosiomics, especially on small-sigma LoG filtered and wavelet filtered images. Features from different ROIs also had distinctive repeatability patterns. Further analysis discovered that feature repeatability was highly associated with the randomness, heterogeneity, and contrast of the images, regardless of the data modality. Our findings provided a direct reference of repeatability for both CT radiomics and dosiomics for cervical cancer. The distinctive repeatability patterns for features from different filtered images and different ROIs provided valuable guidance for repeatable feature selection, emphasizing the importance of careful consideration when choosing image filters and defining ROIs. The comparison between CT radiomics and dosiomics shed light on the performance and repeatability of features from different image modalities. It further contributed to the understanding of the strengths and limitations of radiomics and dosiomics, which enables researchers to leverage the rich textural information and clinical relevance of

CT radiomics while harnessing the quantitative dose assessment and potential for personalized treatment planning offered by dosiomics, ultimately impacting the development of personalized treatment strategies and improving the reliability and clinical relevance of cancer research.

Our quantitative analysis of the direct impact of image characteristics on feature repeatability may help to explain the different repeatability patterns observed in this study. As suggested in **Figure 13**, features from images with higher entropy, lower uniformity, and higher variance are less susceptible to image perturbations. After gray-level discretization with fixed bin counts, images with a higher pixel complexity, heterogeneity, and contrast levels tend to have reduced noise, which can be directly observed from the example images in **Figure 10**. Therefore, the local pixel connectivity was enhanced, and the resulting texture matrices were more robust against translation and rotation randomizations. Such correlation also raises the importance of image preprocessing where, for example, image resegmentation, image thresholding, and gray-level discretization settings could greatly impact the pixel heterogeneity and noise levels, and careful consideration should be made to balance the repeatability and sensitivity of the extracted features.

For features from LoG filtered images, a higher repeatability was observed as sigma values increased. This is evidenced by the higher mean ICC values and larger repeatable feature numbers with ICC > 0.9. The LoG filter is commonly used for enhancing the visibility of edges and texture in the image. As indicated by **Figures 12 and 15**, larger sigma values resulted in smoother edge enhancements, which increased the pixel complexities, heterogeneities, and contrast levels and eventually improved consistencies in the extracted features. The impact of wavelet filters on feature repeatability is an important consideration in our study. Wavelet filters decompose an image into high/low frequency bands, enabling the identification and study of fine-scale details as well as coarse-scale structures. In our study, we found that specific combinations of wavelet coefficients, such as LLL, and LLH, exhibited the highest repeatability compared to other combinations. This indicates that these coefficients effectively captured the relevant structural and textural information while minimizing noise and artifacts. However, it is noteworthy that the use of high pass filters in the x and y directions resulted in relatively lower feature repeatability; the high pass filters tend to enhance noise and fine-scale features, making the extracted features more susceptible to variability and less consistent across different images. Moreover, it is important to highlight the high repeatability observed even when applying high-pass filtering on the z-direction. This can be attributed to the rotation along the axial direction, which results in minimal pixel variations along the z-direction.

For radiomics features from CT, the higher repeatability observed in the rectum, RFemoral, and LFemoral ROIs compared to bladder and CTV could stem from the inherent anatomical and tissue characteristics of these regions. The rectum and femoral regions had more complex tissue compositions with high electron density differences, as shown in the example images in **Figure 10**. Such tissue characteristics were consistent with the quantitative image descriptions in **Figure 13** where higher entropy, lower uniformity, and higher variance were exhibited, which led to more repeatable radiomics features. On the other hand, the bladder and CTV regions may have greater complexities in shape but rather similar tissue compositions, leading to increased contour variabilities but decreased pixel complexities and subsequently lower repeatability in the extracted features. For radiomics features from dose data, the higher feature repeatability observed in the rectum and bladder ROIs can be attributed to the nature of dosimetry data. Dosiomics involves the analysis of radiation dose distribution within the target area and surrounding organs. The rectum and bladder, being adjacent to the target area, experienced a sharper dose drop-off and higher dose variance within the ROI volume, resulting in the dosiomics features being less susceptible to perturbations.

Several limitations should be acknowledged in this study. Firstly, the relatively small sample size may restrict the generalizability of the findings and limit the statistical power to detect subtle differences. The single-center design introduces the possibility of bias related to patient selection, imaging protocols, and data acquisition techniques, which may affect the external validity of the results. Additionally, the absence of external validation using an independent dataset limits the ability to confirm the repeatability findings in different settings or populations. The focus on specific data modalities, such as CT scans and dose maps, overlooks the potential variability and repeatability of features derived from other imaging modalities, such as MRI or PET. Moreover, the study did not directly assess the clinical impact of feature repeatability on treatment outcomes or patient management decisions. Future research should address these limitations by incorporating larger and more diverse cohorts, conducting multi-center studies with standardized protocols, performing external validation, exploring feature repeatability across various imaging modalities, and investigating the clinical implications of feature repeatability in real-world scenarios.

4.5. Conclusion

In conclusion, this study investigated the repeatability of CT radiomics and dosiomics features under image perturbations. The findings suggest that CT-based radiomics features exhibit higher repeatability compared to features derived from dose maps. The higher repeatability of CT-based radiomics features highlights their potential as reliable and consistent quantitative markers in imaging-based analyses. Our findings contribute to the development of more reliable imaging biomarkers for personalized cancer treatment planning and response assessment. Further research is needed to explore the impact of feature repeatability on predictive performance and clinical utility under different settings and patient populations.

Chapter 5. Systematic Assessment of Feature Repeatability's Impact on Radiomics Model Performance: A Multi-institutional Validation Study

5.1. Introduction

Radiomics has emerged as a transformative computational approach in precision oncology, enabling the high-throughput extraction of quantitative features from medical images to support clinical decision-making [118, 119]. This rapidly evolving field combines advanced image analysis with machine learning techniques to extract potentially thousands of quantitative features from medical images, providing deeper insights into tumor phenotypes and biological characteristics that may not be apparent to the human eye [120, 121]. In head and neck cancer management, where treatment outcomes can vary significantly among patients with similar clinical characteristics, radiomics offers particular promise for improved prognostication and personalized therapy selection [122, 123].

Recent advances in artificial intelligence and computational power have accelerated the development of radiomic approaches, leading to improved feature extraction methodologies and more sophisticated analysis techniques [124, 125]. However, the reliability and reproducibility of radiomic features remain significant challenges in developing robust predictive models [56, 126]. These challenges are particularly pronounced in multi-institutional settings, where variations in imaging protocols, scanner parameters, and reconstruction algorithms can significantly impact feature stability [127, 128].

The stability of radiomic features can be affected by numerous factors throughout the imaging and analysis pipeline, including image acquisition parameters, preprocessing methods, segmentation variability, and institutional protocols [129]. Feature repeatability, typically assessed through metrics such as the Intraclass Correlation Coefficient (ICC), has been identified as a crucial factor in model performance [16, 17][130, 131]. Despite growing recognition of its importance, the relationship between feature repeatability and model generalizability across different cohort sizes and institutional settings remains inadequately understood [132]. Furthermore, the impact of feature selection strategies incorporating repeatability criteria on model performance in large-scale, multi-institutional studies has not been systematically investigated [133].

Previous studies have demonstrated the potential of radiomic features in head and neck cancer prognostication [134, 135], but most have been limited by single-institution datasets, small sample sizes, or lack of external validation [55, 136]. While some multi-institutional studies have shown promising results [137], the variability in feature extraction methods and the absence of standardized repeatability assessment protocols have hampered the clinical translation of these findings [138, 139]. Recent meta-analyses have highlighted the need for robust validation studies that specifically address feature stability across different clinical scenarios and patient populations [140].

The standardization of radiomic feature extraction and analysis has become a critical focus in the field, with initiatives such as the Image Biomarker Standardization

Initiative (IBSI) providing guidelines for reproducible research [1, 11]. However, the optimal approach for incorporating feature repeatability assessments into model development pipelines, particularly in the context of varying dataset sizes and institutional characteristics, remains an open question [141, 142]. Additionally, the relationship between feature repeatability and model performance metrics in both internal and external validation scenarios requires further investigation [143].

This study aims to systematically evaluate the impact of feature repeatability on radiomics model performance in head and neck cancer prognostication across multiple independent datasets. We hypothesize that incorporating highly repeatable features will enhance model generalizability and reduce overfitting, with the magnitude of improvement potentially varying by cohort size. By analyzing four distinct datasets comprising 1,418 patients, we investigate how different ICC thresholds affect model performance in both internal and external validation scenarios. Our comprehensive analysis addresses several key gaps in current knowledge: a) The relationship between feature repeatability and model performance across varying cohort sizes. b) The impact of feature repeatability on model generalizability in multi-institutional settings, c) The role of repeatable feature selection in mitigating overfitting and improving external validation performance.

5.2. Materials and Methods

5.2.1. Patient Cohort

This retrospective study analyzed a dataset of pre-treatment CT images from 1,441 head-and-neck cancer patients obtained from TCIA [144]. Specifically, the dataset comprised data on patients from seven medical institutions: data on 137 patients from the single-institution HEAD-NECK-RADIOMICS-HN1 (HN137) study [11], data on 606 patients from the single-institution Radiomic Biomarkers in Oropharyngeal Carcinoma (OPC606) study [123], data on 298 patients from four institutions in the Head-Neck-PET-CT (PETCT298) study [145], and data on 400 patients from the single-institution Head and Neck Squamous Cell Carcinoma (HNSCC400) study [146, 147].

5.2.2. Image Preprocessing

The preprocessing pipeline consisted of two main steps to ensure feature reproducibility and consistency. Initially, the GTV contours underwent interpolation to generate voxel-based segmentation masks. Subsequently, an isotropic resampling process was applied at 1 mm × 1 mm × 1 mm resolution, utilizing B-spline interpolation for images and nearest-neighbor interpolation for masks. These preprocessing steps were implemented using Python v3.8, incorporating the SimpleITK v2.2.0 and OpenCV packages.

5.2.3. Radiomic Feature Extraction

Feature extraction was performed using the Image Biomarker Standardization Initiative-compliant Pyradiomics v2.2.0 package. The process yielded 5,486 radiomic features from each patient's CT scan GTV, derived from twelve different image types: one unfiltered image, three Laplacian-of-Gaussian filtered images (σ = 1, 3, 6 mm), and eight Coiflet1 wavelet filtered images (LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH). The feature set comprised 14 shape features from GTV segmentation, along with 18 first-order and 73 second-order features extracted from each filtered image. Texture feature extraction incorporated soft-tissue range re-segmentation (-150 to 180 HU) and discretization with fixed bin counts of 100.

5.2.4. Feature Repeatability Assessment

The robustness of each RF was quantified in terms of a one-way, random, absolute-agreement ICC, which was calculated using Equation (1), as follows [82].

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

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. The ICC analysis served as a critical quality control measure, ensuring the selected features demonstrated consistent behavior and reliability across different imaging parameters and acquisition conditions.

5.2.5. Feature Selection

The feature selection methodology incorporated comprehensive repeatability assessment utilizing the Intraclass Correlation Coefficient (ICC). Multiple ICC thresholds (0, 0.2, 0.5, 0.7, and 0.9) were systematically investigated to evaluate feature stability across experimental conditions. Features demonstrating ICC values below the specified threshold were excluded from the feature pool to ensure robust feature representation in subsequent analyses.

To minimize information redundancy within the feature space, inter-feature correlations were quantitatively assessed using Pearson correlation coefficient (r). Feature pairs exhibiting high correlation ($r \geqslant 0.7$) were identified, and within each correlated pair, the feature demonstrating higher mean correlation with the remaining feature set was eliminated. This step was crucial in reducing multicollinearity and improving model stability. Subsequently, the minimum-Redundancy-Maximum-Relevance (mRMR) algorithm was employed to select five optimal features from the remaining feature pool [148]. The mRMR approach optimizes feature selection by simultaneously maximizing the mutual information between selected features and the outcome variable while minimizing redundancy among the selected features, thereby ensuring a complementary and informative feature subset for predictive modeling.

5.2.6. Model Development

The modeling framework consisted of both internal and external validation strategies to comprehensively evaluate the model performance and assess the impact of feature repeatability on model reliability. For both validation approaches, Cox proportional hazards regression was employed as the primary modeling methodology to account for time-to-event outcomes [149].

Internal validation was implemented through a three-fold cross-validation scheme with 30 repetitions to ensure robust performance assessment. This iterative validation approach provided a comprehensive evaluation of the model's predictive capability while minimizing the potential impact of data partitioning bias. The internal validation process was systematically conducted across different ICC thresholds to investigate the relationship between feature repeatability and model overfitting.

External validation was performed through a leave-one-dataset-out approach, where models were iteratively trained on the combined data from all but one dataset and validated on the held-out dataset. This validation strategy was implemented across different ICC thresholds to evaluate the impact of feature repeatability on model generalizability. The external validation framework provided insights into the model's ability to maintain predictive performance across heterogeneous patient populations and varying institutional protocols.

5.2.7. Model Evaluation and Statistical Analysis

Model performance was comprehensively evaluated using the concordance index (C-index), a non-parametric metric that quantifies the model's discriminative ability in survival analysis by assessing the concordance between predicted and observed survival times. In the internal validation framework, the magnitude of model overfitting was systematically assessed through the quantification of discrepancy between training and testing C-indices. The relative performance enhancement achieved through repeatable feature selection was evaluated by comparing models developed with various ICC thresholds (ICC > 0) against the baseline model (ICC = 0) using the Mann-Whitney U test, providing statistical inference on the significance of performance improvements.

For clinical risk stratification, a threshold-based approach was implemented wherein external validation cohorts were dichotomized into high-risk and low-risk groups using the median risk score derived from the training cohort as the stratification threshold. The prognostic significance of this stratification was evaluated through Kaplan-Meier survival analysis, with the statistical significance of survival differences assessed through univariate Cox proportional hazards regression. This analysis yielded hazard ratios (HRs) and corresponding p-values, providing quantitative measures of the magnitude and statistical significance of survival disparities between risk groups, thereby establishing the clinical utility of the developed risk stratification framework.

5.3. Results

5.3.1. Model Performance Across ICC Thresholds

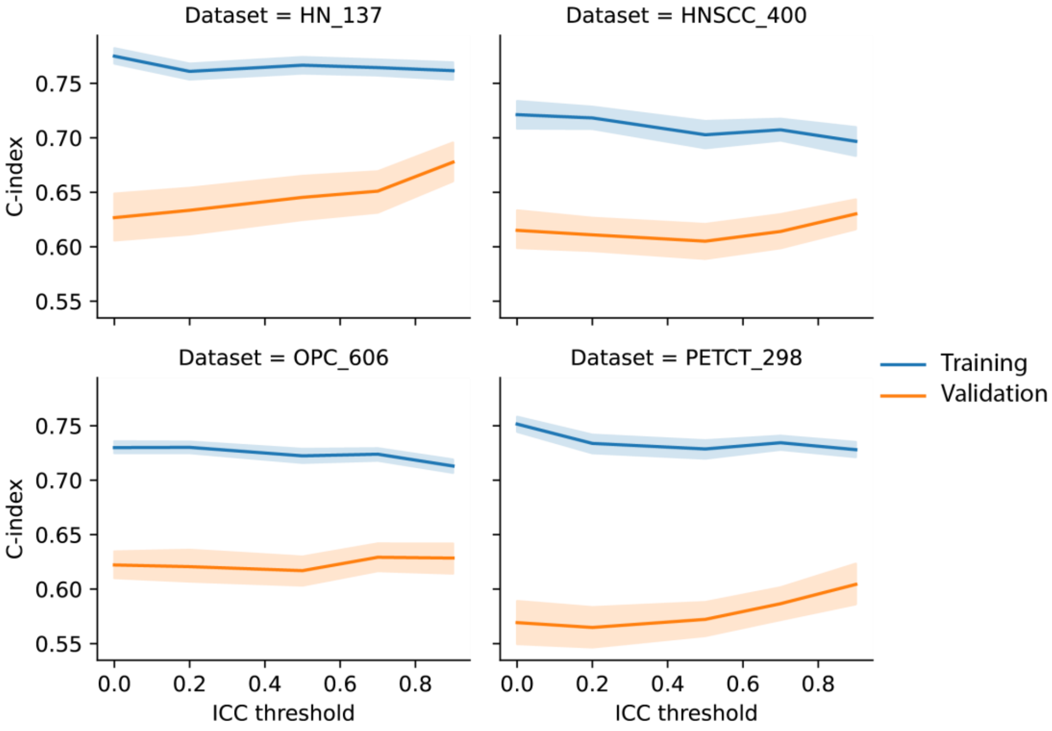

Figure 14. Trends of training and internal validation concordance index (C-index) under different intra-correlation coefficient (ICC) thresholds during repeatable feature selection

Analysis of model performance revealed a consistent pattern characterized by increasing validation C-indices and decreasing training C-indices across escalating ICC thresholds. This trend was particularly pronounced in certain datasets, with HN_137 demonstrating a substantial improvement in validation performance from a mean C-index of 0.627 (SD=0.084) at ICC $\geqslant$ 0 to 0.678 (SD=0.072) at ICC $\geqslant$ 0.9. Similarly, the PETCT_298 dataset exhibited notable improvement, with validation C-index increasing from 0.569 (SD=0.077) to 0.604 (SD=0.078) across the same ICC threshold range. However, the performance enhancement was more modest in the HNSCC_400 and OPC_606 datasets, where the increment in validation C-index from ICC threshold of 0 to 0.9 was less pronounced, with a magnitude of improvement not exceeding 0.02. These findings suggest that the impact of feature repeatability on model performance may vary across different cohorts, potentially influenced by underlying dataset characteristics and heterogeneity.

5.3.2. Model Overfitting Assessment

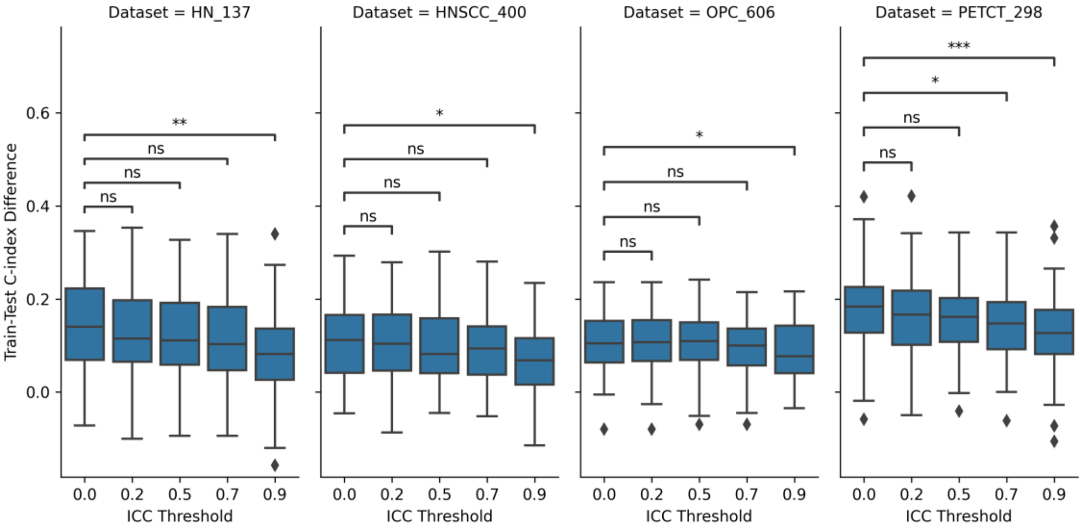

Figure 15. Distributions of C-index difference between training and validation under different ICC thresholds during repeatable feature selection (*: p-value < 0.05, **: p-value < 001, ns: not significant)

Further analysis of model performance metrics revealed a consistent reduction in the disparity between training and validation C-indices as ICC thresholds increased across all four study datasets. The statistical comparison demonstrated significant differences (p-value < 0.05) in these performance gaps between the baseline models (ICC ⩾ 0) and models incorporating highly repeatable features (ICC ⩾ 0.9). This observation suggests that the selection of highly repeatable features effectively mitigates model overfitting. However, intermediate ICC thresholds (0.2, 0.5, and 0.7) did not yield statistically significant differences in performance gaps compared to the baseline, indicating that substantial improvement in model generalizability may require stringent feature repeatability criteria. These findings underscore the importance of implementing rigorous repeatability thresholds in feature selection to optimize model robustness and reliability.

5.3.3. External Validation

Analysis of external validation models demonstrated differential improvements in model performance metrics between training and testing scenarios. While training C-indices showed minimal increases across ICC thresholds, testing performance exhibited substantially larger improvements. The HN-137 dataset demonstrated the most pronounced enhancement in model performance, achieving the highest testing C-index of 0.687 and showing the most substantial improvement from its baseline value of 0.618. The remaining datasets exhibited more moderate improvements in testing performance, with C-indices ranging from 0.595 to 0.656 and absolute improvements ranging from 0.018 to 0.034 from their respective baselines. These findings suggest that the incorporation of repeatable features particularly enhances model generalizability in external validation scenarios, with the magnitude of improvement varying across different cohorts.

Table 6. Training and external testing C-index under different ICC thresholds during repeatable feature selection

	ICC threshold	HN_137	HNSCC_400	OPC_606	PETCT_298
Training	0	0.678 (0.637-0.717)	0.687 (0.637-0.732)	0.720 (0.675-0.762)	0.705 (0.664-0.745)
	0.2	0.678 (0.637-0.717)	0.687 (0.637-0.732)	0.720 (0.675-0.762)	0.698 (0.660-0.738)
	0.5	0.690 (0.652-0.727)	0.674 (0.623-0.720)	0.710 (0.661-0.754)	0.700 (0.660-0.739)
	0.7	0.694 (0.656-0.731)	0.674 (0.623-0.720)	0.710 (0.661-0.754)	0.719 (0.682-0.756)
	0.9	0.700 (0.662-0.738)	0.682 (0.632-0.726)	0.720 (0.672-0.760)	0.707 (0.671-0.746)
External testing	0	0.618 (0.508-0.714)	0.638 (0.552-0.712)	0.561 (0.498-0.637)	0.570 (0.472-0.664)
	0.2	0.618 (0.508-0.714)	0.638 (0.552-0.712)	0.561 (0.498-0.637)	0.570 (0.474-0.659)
	0.5	0.614 (0.512-0.716)	0.659 (0.573-0.727)	0.617 (0.549-0.685)	0.578 (0.485-0.665)
	0.7	0.626 (0.522-0.731)	0.659 (0.573-0.727)	0.617 (0.549-0.685)	0.614 (0.525-0.692)
	0.9	0.687 (0.584-0.791)	0.656 (0.569-0.729)	0.595 (0.532-0.657)	0.602 (0.513-0.688)

5.3.4. Risk Stratification Analysis

Kaplan-Meier (KM) survival analysis provided additional evidence for enhanced model generalizability through repeatable feature selection. Models incorporating highly repeatable features (ICC ⩾ 0.9) demonstrated statistically significant stratification (p-value < 0.05) between risk groups for three datasets: HN_137, HNSCC_400, and PETCT_298. In contrast, baseline models (ICC ⩾ 0) failed to achieve significant risk group separation across all datasets. Notably, the OPC_606 dataset demonstrated an exception to this pattern, showing no significant stratification at either baseline or high ICC threshold (⩾ 0.9). These findings further support the utility of stringent repeatability criteria in feature selection for improving model prognostic performance, while also highlighting potential dataset-specific variations in the effectiveness of this approach for risk stratification.

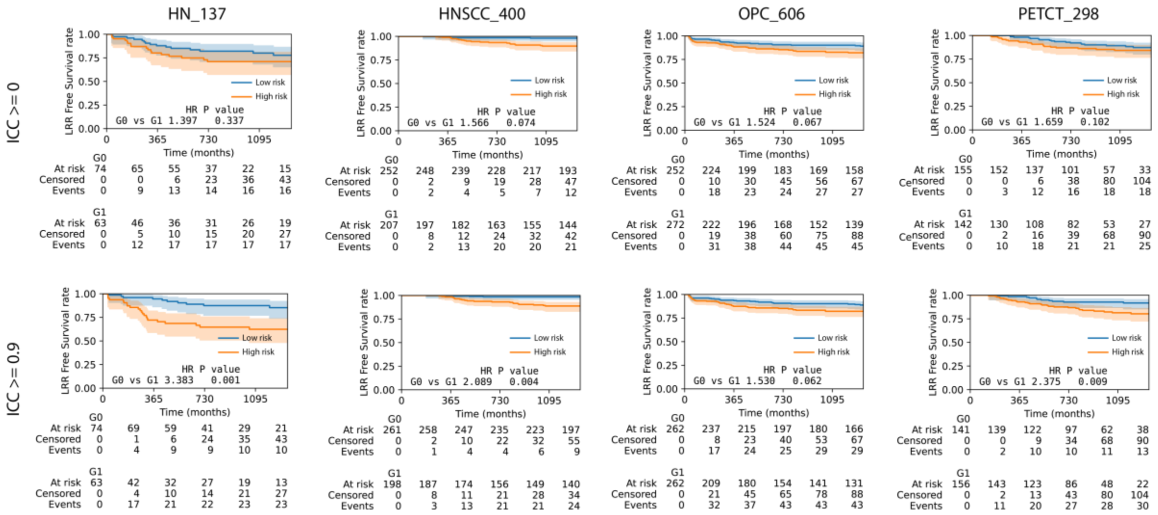

Figure 16. Kaplan-Meier curves of the two risk groups stratified by the external survival models

5.4. Discussion

The differential improvements in validation performance across datasets suggest varying benefits of repeatable feature selection. The modest validation C-index improvements observed in HNSCC_400 and OPC_606 (increment < 0.02) from ICC threshold 0 to 0.9 might be influenced by various factors, including their larger sample sizes. These substantial cohorts possibly provide inherent protection against false discoveries, potentially contributing to more stable feature selection and robust model performance even without strict repeatability criteria. While sample size could be one contributing factor aligning with statistical learning theory, where larger samples tend to yield more reliable feature-outcome relationships, other unmeasured factors may also play important roles in these observations.

The HN_137 dataset's dramatic improvement in both internal and external validation performance underscores the critical role of repeatable feature selection in smaller cohorts. The marked enhancement in validation C-index (from 0.627 to 0.678) and the highest external testing C-index (0.687) demonstrates that stringent repeatability criteria can effectively compensate for the inherent limitations of smaller sample sizes. This finding is particularly relevant for radiomics studies, where large, homogeneous datasets are often challenging to acquire. The implementation of repeatable feature selection appears to serve as a crucial safeguard against overfitting in such scenarios, providing a practical solution for developing robust models despite limited sample sizes.

The performance patterns observed in the OPC_606 dataset highlight the challenges posed by dataset heterogeneity in radiomics modeling. Despite its large sample size, this dataset exhibited the lowest external validation C-index and failed to achieve significant risk stratification at both baseline and high ICC thresholds. This underperformance can be primarily attributed to the mismatch between training and testing cohorts - while the training set comprised various head and neck cancer subtypes, the testing set was restricted to oropharyngeal cancer (OPC) cases. This observation emphasizes that even robust feature selection methods cannot fully overcome fundamental differences in patient populations, suggesting that careful consideration of cohort characteristics is crucial for successful model deployment.

These findings suggest that the optimal approach to radiomics model development should be tailored to specific study characteristics. For smaller datasets, implementing stringent repeatability criteria becomes crucial for model reliability, while larger datasets may allow for more flexible feature selection approaches. The impact of dataset heterogeneity emphasizes the importance of careful cohort selection and validation strategies that account for potential differences in patient characteristics. Future research should focus on developing adaptive feature selection strategies that consider both sample size and dataset heterogeneity to optimize model performance across diverse clinical scenarios.

Several limitations of this study warrant consideration. While we demonstrated the benefits of repeatable feature selection across different sample sizes, our analysis was confined to head and neck cancer datasets. The generalizability of these findings to other cancer types or imaging modalities requires further investigation. The binary classification of features based on ICC thresholds may oversimplify the complex relationship between feature repeatability and model performance. More sophisticated approaches, such as weighted feature selection schemes that incorporate continuous repeatability measures, could potentially yield further improvements. Additionally, our study focused primarily on conventional radiomics features; the integration of deep learning-derived features and their repeatability characteristics presents an interesting avenue for future research.

In summary, this study demonstrates that incorporating highly repeatable radiomics features can effectively enhance model performance in head and neck cancer prognostication. The implementation of strict repeatability criteria led to improved model robustness and generalizability across different datasets, though the magnitude of improvement varied among cohorts. Notably, we observed that sample size might influence the impact of feature repeatability, with smaller datasets showing more pronounced benefits from stringent repeatability criteria. While larger cohorts demonstrated inherent stability in feature selection, the application of repeatability thresholds still contributed to model reliability. Future work should address current limitations and explore more sophisticated approaches to feature selection and dataset harmonization, ultimately advancing the field toward more reliable and clinically applicable radiomics models.

5.5. Conclusion

This study systematically investigated the impact of feature repeatability on radiomics model performance in head and neck cancer prognostication. Our findings demonstrate that incorporating highly repeatable features (ICC ⩾ 0.9) consistently improved model performance across different datasets, with the magnitude of improvement varying by cohort size. The enhancement was particularly prominent in smaller datasets, where strict repeatability criteria effectively improved both internal and external validation performance. While larger cohorts showed more inherent stability in feature selection, the utilization of repeatable features still contributed to model robustness. These results underscore the importance of feature repeatability assessment in radiomics modeling, suggesting that prioritizing stable features in the feature selection process can lead to more reliable and generalizable prognostic models. The findings provide valuable guidance for optimizing radiomics model development strategies and support the integration of repeatability assessment as a crucial step toward achieving robust clinical applications in cancer prognostication.

Chapter 6. Discussion

6.1. Advances in Radiomics

This investigation's findings have substantial implications for both the theoretical foundations of radiomics and its practical clinical applications. The development of our perturbation-based framework represents a fundamental shift in how we conceptualize and evaluate radiomic model reliability.

The observed disparity in feature stability between imaging modalities (MRI vs. CT) challenges the current paradigm of modality-agnostic feature extraction. While previous studies have primarily focused on optimizing feature extraction algorithms independently of imaging modalities, our findings suggest that the choice of imaging modality should fundamentally influence feature selection strategies. This observation raises important questions about the standardization of radiomics across different imaging platforms and highlights the need for modality-specific optimization approaches.

The superior stability of certain feature families, particularly in MRI-derived features, provides new insights into the relationship between image acquisition physics and feature reliability. This finding suggests that the underlying physical principles of image formation play a more significant role in feature stability than previously recognized. The implications extend beyond mere feature selection, questioning fundamental assumptions about the transferability of radiomic models across imaging modalities.

Our comparative analysis of CT radiomics and dosiomics features introduces a novel perspective on the integration of different data types in predictive modeling. The observed differences in feature stability between these domains suggest that the traditional approach of treating all quantitative features equally may be suboptimal. Instead, our findings support a more nuanced approach that considers the inherent characteristics and limitations of different data sources.

The demonstrated relationship between feature repeatability and model generalizability has profound implications for the development of clinical radiomics applications. The observation that highly repeatable features contribute to more generalizable models challenges the common practice of feature selection based solely on predictive performance. This finding suggests a need to reconceptualize the feature selection process, incorporating stability metrics as fundamental criteria rather than secondary considerations.

Perhaps most significantly, our work raises important questions about the scalability and reproducibility of radiomic models in clinical settings. The observed dependence of feature stability requirements on sample size has important implications for clinical trial design and the validation of radiomic biomarkers. This relationship suggests that the current approach to model validation may need to be revised, particularly for rare diseases where large cohorts are unavailable.

The temporal stability analysis of radiomic features provides new insights into the potential role of radiomics in longitudinal monitoring. The identification of features that maintain stability over time while remaining sensitive to clinically relevant changes suggests possible applications in treatment response monitoring and disease progression assessment. This finding has particular relevance for personalized medicine approaches, where regular monitoring and treatment adaptation are essential.

6.2. Limitations

The limitations of this investigation can be categorized into two primary areas: implementation challenges and methodological constraints.

The first major limitation concerns the implementation architecture of our perturbation-based framework. Despite its theoretical robustness, the current implementation requires substantial computational expertise and lacks integration with standardized software platforms. Unlike common radiomics tools that offer user-friendly interfaces, our framework demands significant programming knowledge and manual parameter optimization, creating a barrier to widespread clinical adoption. This technical complexity particularly impacts real-time feature stability assessment in clinical workflows, where efficiency and accessibility are crucial. The absence of automated pipelines and standardized software implementation limits the framework's utility for researchers and clinicians who may lack advanced programming expertise.

The second fundamental limitation relates to the framework's scope in simulating radiomics workflow variations. While our approach effectively addresses geometric perturbations, it cannot fully capture the complex interplay of variables in clinical settings. The framework falls short in simulating several critical sources of variation, including vendor-specific scanner characteristics, reconstruction algorithm impacts, inter-institutional protocol differences, and the interaction between image acquisition parameters and tissue characteristics. Moreover, the framework's current implementation may not adequately address challenges posed by emerging imaging modalities and novel feature extraction methodologies, particularly in the context of deep learning-based approaches. While our study demonstrates the importance of feature stability for model generalizability, questions remain about the relationship between feature stability and biological relevance. This raises important considerations about whether the most stable features necessarily represent the most clinically meaningful measurements.

Another notable limitation of this study is the absence of long-term follow-up data, which restricts our ability to assess the biological stability of the identified radiomic features over extended time periods. While our findings demonstrate promising correlations in the short term, the temporal robustness of these imaging biomarkers remains uncertain. Long-term stability is crucial for reliable clinical implementation, as radiomic signatures may evolve with disease progression or treatment response. Additionally, the lack of extended follow-up prevents us from evaluating associations between our radiomic features and important distant clinical outcomes such as overall survival and disease recurrence patterns. Future work should prioritize longitudinal data collection to validate the persistence of these radiomic signatures and their continued predictive value across the disease trajectory.

Chapter 7. Conclusion

This thesis advances the field of radiomics by introducing and validating a comprehensive perturbation-based framework for feature stability assessment, addressing a critical gap in current radiomics methodology. Through rigorous experimental validation across multiple imaging modalities and clinical scenarios, we demonstrate that feature stability significantly influences model generalizability and clinical applicability. Our findings challenge several established paradigms in radiomics, particularly regarding modality-agnostic feature extraction and traditional feature selection approaches. The demonstrated relationship between feature stability and model performance, coupled with our novel insights into temporal stability and modality-specific optimization, provides a foundation for more robust and clinically applicable radiomic models. These contributions not only enhance our understanding of radiomics feature reliability but also establish practical guidelines for future development of quantitative imaging biomarkers, ultimately advancing the field toward more standardized and clinically integrated applications. While limitations in computational implementation and workflow simulation persist, this work represents a significant step toward more reliable and clinically viable radiomics applications, setting a new standard for feature stability assessment in quantitative imaging analysis.

Chapter 8. References

[1] P. Lambin *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," (in English), *Nat Rev Clin Oncol,* vol. 14, no. 12, pp. 749-762, Dec 2017, doi: 10.1038/nrclinonc.2017.141.

[2] Y. Jiang *et al.*, "Noninvasive imaging evaluation of tumor immune microenvironment to predict outcomes in gastric cancer," (in English), *Ann Oncol,* vol. 31, no. 6, pp. 760-768, Jun 2020, doi: 10.1016/j.annonc.2020.03.295.

[3] M. Sollini, L. Antunovic, A. Chiti, and M. Kirienko, "Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics," *Eur J Nucl Med Mol Imaging,* vol. 46, no. 13, pp. 2656-2672, Dec 2019, doi: 10.1007/s00259-019-04372-x.

[4] M. Fan *et al.*, "Radiomic analysis of imaging heterogeneity in tumours and the surrounding parenchyma based on unsupervised decomposition of DCE-MRI for predicting molecular subtypes of breast cancer," *Eur Radiol,* vol. 29, no. 8, pp. 4456-4467, Aug 2019, doi: 10.1007/s00330-018-5891-3.

[5] A. Saha *et al.*, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *Br J Cancer,* vol. 119, no. 4, pp. 508-516, Aug 2018, doi: 10.1038/s41416-018-0185-8.

[6] H. Shen *et al.*, "MRI-based radiomics to compare the survival benefit of induction chemotherapy plus concurrent chemoradiotherapy versus concurrent chemoradiotherapy plus adjuvant chemotherapy in locoregionally advanced nasopharyngeal carcinoma: A multicenter study," *Radiother Oncol,* vol. 171, pp. 107-113, Jun 2022, doi: 10.1016/j.radonc.2022.04.017.

[7] P. Yongfeng *et al.*, "The Usefulness of Pretreatment MR-Based Radiomics on Early Response of Neoadjuvant Chemotherapy in Patients With Locally Advanced Nasopharyngeal Carcinoma," *Oncol Res,* vol. 28, no. 6, pp. 605-613, Mar 16 2021, doi: 10.3727/096504020X16022401878096.

[8] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology,* vol. 278, no. 2, pp. 563-77, Feb 2016, doi: 10.1148/radiol.2015151169.

[9] J. J. M. van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Res,* vol. 77, no. 21, pp. e104-e107, Nov 1 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[10] J. H. Thrall *et al.*, "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," *J Am Coll Radiol,* vol. 15, no. 3 Pt B, pp. 504-508, Mar 2018, doi: 10.1016/j.jacr.2017.12.026.

[11] H. J. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat Commun,* vol. 5, p. 4006, Jun 3 2014, doi: 10.1038/ncomms5006.

[12] A. Ibrahim *et al.*, "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework," *Methods,* vol. 188, pp. 20-29, Apr 2021, doi: 10.1016/j.ymeth.2020.05.022.

[13] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," (in English), *Ieee T Syst Man Cyb,* vol. Smc3, no. 6, pp. 610-621, 1973, doi: Doi 10.1109/Tsmc.1973.4309314.

[14] C. J. Sun and W. G. Wee, "Neighboring Gray Level Dependence Matrix for Texture Classification," (in English), *Comput Vision Graph,* vol. 23, no. 3, pp. 341-352, 1983, doi: Doi 10.1016/0734-189x(83)90032-4.

[15] F. Prior *et al.*, "The public cancer radiology imaging collections of The Cancer Imaging Archive," *Sci Data,* vol. 4, p. 170124, Sep 19 2017, doi: 10.1038/sdata.2017.124.

[16] B. Zhao, "Understanding Sources of Variation to Improve the Reproducibility of Radiomics," *Front Oncol,* vol. 11, p. 633176, 2021, doi: 10.3389/fonc.2021.633176.

[17] B. A. Varghese *et al.*, "Reliability of CT-based texture features: Phantom study," *J Appl Clin Med Phys,* vol. 20, no. 8, pp. 155-163, Aug 2019, doi: 10.1002/acm2.12666.

[18] S. P. Blazis, D. B. M. Dickerscheid, P. V. M. Linsen, and C. O. Martins Jarnalo, "Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system," *Eur J Radiol,* vol. 136, p. 109526, Mar 2021, doi: 10.1016/j.ejrad.2021.109526.

[19] D. Mackin *et al.*, "Measuring Computed Tomography Scanner Variability of Radiomics Features," *Invest Radiol,* vol. 50, no. 11, pp. 757-65, Nov 2015, doi: 10.1097/RLI.0000000000000180.

[20] J. Peerlings *et al.*, "Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial," *Sci Rep,* vol. 9, no. 1, p. 4800, Mar 18 2019, doi: 10.1038/s41598-019-41344-5.

[21] S. Fiset *et al.*, "Repeatability and reproducibility of MRI-based radiomic

features in cervical cancer," *Radiother Oncol,* vol. 135, pp. 107-114, Jun 2019, doi: 10.1016/j.radonc.2019.03.001.

[22] R. F. Cabini *et al.*, "Preliminary report on harmonization of features extraction process using the ComBat tool in the multi-center "Blue Sky Radiomics" study on stage III unresectable NSCLC," *Insights Imaging,* vol. 13, no. 1, p. 38, Mar 7 2022, doi: 10.1186/s13244-022-01171-1.

[23] G. Carbonell *et al.*, "Precision of MRI radiomics features in the liver and hepatocellular carcinoma," *Eur Radiol,* vol. 32, no. 3, pp. 2030-2040, Mar 2022, doi: 10.1007/s00330-021-08282-1.

[24] K. Chen, L. Deng, Q. Li, and L. Luo, "Are computed-tomography-based hematoma radiomics features reproducible and predictive of intracerebral hemorrhage expansion? an in vitro experiment and clinical study," *Br J Radiol,* vol. 94, no. 1121, p. 20200724, May 1 2021, doi: 10.1259/bjr.20200724.

[25] Y. Chen *et al.*, "Robustness of CT radiomics features: consistency within and between single-energy CT and dual-energy CT," *Eur Radiol,* vol. 32, no. 8, pp. 5480-5490, Aug 2022, doi: 10.1007/s00330-022-08628-3.

[26] A. Crombe, X. Buy, F. Han, S. Toupin, and M. Kind, "Assessment of Repeatability, Reproducibility, and Performances of T2 Mapping-Based Radiomics Features: A Comparative Study," *J Magn Reson Imaging,* vol. 54, no. 2, pp. 537-548, Aug 2021, doi: 10.1002/jmri.27558.

[27] N. Emaminejad, M. W. Wahi-Anwar, G. H. J. Kim, W. Hsu, M. Brown, and M. McNitt-Gray, "Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters," *Med Phys,* vol. 48, no. 6, pp. 2906-2919, Jun 2021, doi: 10.1002/mp.14830.

[28] A. Euler *et al.*, "Virtual Monoenergetic Images of Dual-Energy CT-Impact on Repeatability, Reproducibility, and Classification in Radiomics," *Cancers (Basel),* vol. 13, no. 18, Sep 20 2021, doi: 10.3390/cancers13184710.

[29] Y. Gao *et al.*, "Reproducibility of radiomic features of pulmonary nodules between low-dose CT and conventional-dose CT," *Quant Imaging Med Surg,* vol. 12, no. 4, pp. 2368-2377, Apr 2022, doi: 10.21037/qims-21-609.

[30] R. W. Y. Granzier *et al.*, "Test-Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability," *J Magn Reson Imaging,* vol. 56, no. 2, pp. 592-604, Aug 2022, doi: 10.1002/jmri.28027.

[31] A. Ibrahim *et al.*, "Reproducibility of CT-Based Hepatocellular Carcinoma Radiomic Features across Different Contrast Imaging Phases: A Proof of

Concept on SORAMIC Trial Data," *Cancers (Basel),* vol. 13, no. 18, Sep 16 2021, doi: 10.3390/cancers13184638.

[32] A. Ibrahim *et al.*, "The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset," *PLoS One,* vol. 16, no. 5, p. e0251147, 2021, doi: 10.1371/journal.pone.0251147.

[33] J. Lee *et al.*, "Radiomics feature robustness as measured using an MRI phantom," *Sci Rep,* vol. 11, no. 1, p. 3973, Feb 17 2021, doi: 10.1038/s41598-021-83593-3.

[34] S. Lennartz, A. O'Shea, A. Parakh, T. Persigehl, B. Baessler, and A. Kambadakone, "Robustness of dual-energy CT-derived radiomic features across three different scanner types," *Eur Radiol,* vol. 32, no. 3, pp. 1959-1970, Mar 2022, doi: 10.1007/s00330-021-08249-2.

[35] R. N. Mahon, G. D. Hugo, and E. Weiss, "Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome," *Phys Med Biol,* Apr 12 2019, doi: 10.1088/1361-6560/ab18d3.

[36] D. J. McHugh *et al.*, "Image Contrast, Image Pre-Processing, and T(1) Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases," *Cancers (Basel),* vol. 13, no. 2, Jan 11 2021, doi: 10.3390/cancers13020240.

[37] M. Meyer *et al.*, "Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings," *Radiology,* vol. 293, no. 3, pp. 583-591, Dec 2019, doi: 10.1148/radiol.2019190928.

[38] R. N. Mitchell-Hay, T. S. Ahearn, A. D. Murray, and G. D. Waiter, "Investigation of the Inter- and Intrascanner Reproducibility and Repeatability of Radiomics Features in T1-Weighted Brain MRI," *J Magn Reson Imaging,* vol. 56, no. 5, pp. 1559-1568, Nov 2022, doi: 10.1002/jmri.28191.

[39] R. Reiazi *et al.*, "Prediction of Human Papillomavirus (HPV) Association of Oropharyngeal Cancer (OPC) Using Radiomics: The Impact of the Variation of CT Scanner," *Cancers (Basel),* vol. 13, no. 9, May 8 2021, doi: 10.3390/cancers13092269.

[40] L. Rinaldi *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,* vol. 6, no. 1, p. 2, Jan 25 2022, doi: 10.1186/s41747-021-00258-6.

[41] D. Alis, M. Yergin, O. Asmakutlu, C. Topel, and E. Karaarslan, "The influence of cardiac motion on radiomics features: radiomics features of non-enhanced CMR cine images greatly vary through the cardiac cycle," *Eur Radiol,* vol. 31, no. 5, pp. 2706-2715, May 2021, doi: 10.1007/s00330-020-07370-y.

[42] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging-"how-to" guide and critical reflection," *Insights Imaging,* vol. 11, no. 1, p. 91, Aug 12 2020, doi: 10.1186/s13244-020-00887-2.

[43] L. Perna *et al.*, "Inter-observer variability in contouring the penile bulb on CT images for prostate cancer treatment planning," *Radiat Oncol,* vol. 6, p. 123, Sep 24 2011, doi: 10.1186/1748-717X-6-123.

[44] C. Parmar *et al.*, "Robust Radiomics feature quantification using semiautomatic volumetric segmentation," *PLoS One,* vol. 9, no. 7, p. e102107, 2014, doi: 10.1371/journal.pone.0102107.

[45] A. Zwanenburg *et al.*, "Assessing robustness of radiomic features by image perturbation," *Sci Rep,* vol. 9, no. 1, p. 614, Jan 24 2019, doi: 10.1038/s41598-018-36938-4.

[46] H. Chen *et al.*, "Reproducibility of radiomics features derived from intravoxel incoherent motion diffusion-weighted MRI of cervical cancer," *Acta Radiol,* vol. 62, no. 5, pp. 679-686, May 2021, doi: 10.1177/0284185120934471.

[47] J. Duan *et al.*, "Reproducibility for Hepatocellular Carcinoma CT Radiomic Features: Influence of Delineation Variability Based on 3D-CT, 4D-CT and Multiple-Parameter MR Images," *Front Oncol,* vol. 12, p. 881931, 2022, doi: 10.3389/fonc.2022.881931.

[48] R. W. Y. Granzier *et al.*, "MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability," *Sci Rep,* vol. 10, no. 1, p. 14163, Aug 25 2020, doi: 10.1038/s41598-020-70940-z.

[49] N. S. M. Haniff, M. K. Abdul Karim, N. H. Osman, M. I. Saripan, I. N. Che Isa, and M. J. Ibahim, "Stability and Reproducibility of Radiomic Features Based Various Segmentation Technique on MR Images of Hepatocellular Carcinoma (HCC)," *Diagnostics (Basel),* vol. 11, no. 9, Aug 30 2021, doi: 10.3390/diagnostics11091573.

[50] L. J. Jensen, D. Kim, T. Elgeti, I. G. Steffen, B. Hamm, and S. N. Nagel, "Stability of Radiomic Features across Different Region of Interest Sizes-A CT and MR Phantom Study," *Tomography,* vol. 7, no. 2, pp. 238-252, Jun 8 2021, doi: 10.3390/tomography7020022.

[51] B. Kocak, E. S. Durmaz, O. K. Kaya, E. Ates, and O. Kilickesmez, "Reliability of Single-Slice-Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility," *AJR Am J Roentgenol,* vol. 213, no. 2, pp. 377-383, Aug 2019, doi: 10.2214/AJR.19.21212.

[52] G. Muller-Franzes *et al.*, "Reliability as a Precondition for Trust-Segmentation Reliability Analysis of Radiomic Features Improves Survival Prediction," *Diagnostics (Basel),* vol. 12, no. 2, Jan 19 2022, doi: 10.3390/diagnostics12020247.

[53] F. Urraro *et al.*, "MRI Radiomics in Prostate Cancer: A Reliability Study," *Front Oncol,* vol. 11, p. 805137, 2021, doi: 10.3389/fonc.2021.805137.

[54] L. Wang *et al.*, "Assessment of liver metastases radiomic feature reproducibility with deep-learning-based semi-automatic segmentation software," *Acta Radiol,* vol. 62, no. 3, pp. 291-301, Mar 2021, doi: 10.1177/0284185120922822.

[55] I. Fornacon-Wood *et al.*, "Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform," *Eur Radiol,* vol. 30, no. 11, pp. 6241-6250, Nov 2020, doi: 10.1007/s00330-020-06957-9.

[56] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology,* vol. 295, no. 2, pp. 328-338, May 2020, doi: 10.1148/radiol.2020191145.

[57] L. Duron *et al.*, "Gray-level discretization impacts reproducible MRI radiomics texture features," *PLoS One,* vol. 14, no. 3, p. e0213459, 2019, doi: 10.1371/journal.pone.0213459.

[58] K. V. Hoebel *et al.*, "Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma," *Radiol Artif Intell,* vol. 3, no. 1, p. e190199, Jan 2021, doi: 10.1148/ryai.2020190199.

[59] Y. Li, G. Tan, M. Vangel, J. Hall, and W. Cai, "Influence of feature calculating parameters on the reproducibility of CT radiomic features: a thoracic phantom study," *Quant Imaging Med Surg,* vol. 10, no. 9, pp. 1775-1785, Sep 2020, doi: 10.21037/qims-19-921.

[60] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, "Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma," *J Appl Clin Med Phys,* vol. 21, no. 1, pp. 179-190, Jan 2020, doi: 10.1002/acm2.12795.

[61] E. Scalco *et al.*, "T2w-MRI signal normalization affects radiomics features reproducibility," *Med Phys,* vol. 47, no. 4, pp. 1680-1691, Apr 2020, doi: 10.1002/mp.14038.

[62] M. Schwier *et al.*, "Repeatability of Multiparametric Prostate MRI Radiomics Features," *Sci Rep,* vol. 9, no. 1, p. 9441, Jul 1 2019, doi: 10.1038/s41598-019-45766-z.

[63] G. Simpson *et al.*, "Impact of quantization algorithm and number of gray level intensities on variability and repeatability of low field strength magnetic resonance image-based radiomics texture features," *Phys Med,* vol. 80, pp. 209-220, Dec 2020, doi: 10.1016/j.ejmp.2020.10.029.

[64] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," (in English), *Sci Rep-Uk,* vol. 5, Aug 17 2015, doi: ARTN 13087

10.1038/srep13087.

[65] D. A. P. Delzell, S. Magnuson, T. Peter, M. Smith, and B. J. Smith, "Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data," *Front Oncol,* vol. 9, p. 1393, 2019, doi: 10.3389/fonc.2019.01393.

[66] N. Emaminejad *et al.*, "Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients," *IEEE Trans Biomed Eng,* vol. 63, no. 5, pp. 1034-1043, May 2016, doi: 10.1109/TBME.2015.2477688.

[67] V. Popovici, E. Budinska, L. Dusek, M. Kozubek, and F. Bosman, "Image-based surrogate biomarkers for molecular subtypes of colorectal cancer," *Bioinformatics,* vol. 33, no. 13, pp. 2002-2009, Jul 1 2017, doi: 10.1093/bioinformatics/btx027.

[68] E. Scalco and G. Rizzo, "Texture analysis of medical images for radiotherapy applications," *Br J Radiol,* vol. 90, no. 1070, p. 20160642, Feb 2017, doi: 10.1259/bjr.20160642.

[69] S. Alobaidli, S. Mcquaid, C. South, V. Prakash, P. Evans, and A. Nisbet, "The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning," (in English), *Brit J Radiol,* vol. 87, no. 1042, Oct 2014, doi: ARTN 20140369

10.1259/bjr.20140369.

[70] S. Chicklore, V. Goh, M. Siddique, A. Roy, P. K. Marsden, and G. J. Cook, "Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture

analysis," *Eur J Nucl Med Mol Imaging,* vol. 40, no. 1, pp. 133-40, Jan 2013, doi: 10.1007/s00259-012-2247-0.

[71] K. A. Miles, B. Ganeshan, and M. P. Hayball, "CT texture analysis using the filtration-histogram method: what do the measurements mean?," *Cancer Imaging,* vol. 13, no. 3, pp. 400-6, Sep 23 2013, doi: 10.1102/1470-7330.2013.9045.

[72] P. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," (in English), *Eur J Cancer,* vol. 48, no. 4, pp. 441-446, Mar 2012, doi: 10.1016/j.ejca.2011.11.036.

[73] V. Kumar *et al.*, "Radiomics: the process and the challenges," (in English), *Magn Reson Imaging,* vol. 30, no. 9, pp. 1234-1248, Nov 2012, doi: 10.1016/j.mri.2012.06.010.

[74] K. Baumann, "Cross-validation as the objective function for variable-selection techniques," (in English), *Trac-Trend Anal Chem,* vol. 22, no. 6, pp. 395-406, Jun 2003, doi: 10.1016/S0165-9936(03)00607-1.

[75] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement (vol 162, pg 55, 2015)," (in English), *Ann Intern Med,* vol. 162, no. 8, pp. 600-600, Apr 21 2015, doi: 10.7326/L15-0078-4.

[76] R. B. Haynes, K. A. McKibbon, N. L. Wilczynski, S. D. Walter, S. R. Werre, and H. Team, "Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey," (in English), *Bmj-Brit Med J,* vol. 330, no. 7501, pp. 1179-1182a, May 21 2005, doi: DOI 10.1136/bmj.38446.498542.8F.

[77] R. T. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. van Elmpt, "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures," *Br J Radiol,* vol. 90, no. 1070, p. 20160665, Feb 2017, doi: 10.1259/bjr.20160665.

[78] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review," *PLoS One,* vol. 10, no. 5, p. e0124165, 2015, doi: 10.1371/journal.pone.0124165.

[79] C. Haarburger, G. Muller-Franzes, L. Weninger, C. Kuhl, D. Truhn, and D. Merhof, "Author Correction: Radiomics feature reproducibility under inter-rater variability in segmentations of CT images," *Sci Rep,* vol. 11, no. 1, p. 22670, Nov 16 2021, doi: 10.1038/s41598-021-02114-4.

[80] R. Beare, B. Lowekamp, and Z. Yaniv, "Image Segmentation, Registration

and Characterization in R with SimpleITK," *J Stat Softw,* vol. 86, Aug 2018, doi: 10.18637/jss.v086.i08.

[81] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlations coefficients (vol 1, pg 30, 1996)," (in English), *Psychol Methods,* vol. 1, no. 4, pp. 390-390, Dec 1996, doi: Doi 10.1037//1082-989x.1.4.390.

[82] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J Chiropr Med,* vol. 15, no. 2, pp. 155-63, Jun 2016, doi: 10.1016/j.jcm.2016.02.012.

[83] K. H. Zou *et al.*, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad Radiol,* vol. 11, no. 2, pp. 178-89, Feb 2004, doi: 10.1016/s1076-6332(03)00671-8.

[84] B. S. Zhao, Y. Q. Tan, W. Y. Tsai, L. H. Schwartz, and L. Lu, "Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study," (in English), *Transl Oncol,* vol. 7, no. 1, pp. 88-93, Feb 2014, doi: 10.1593/tlo.13865.

[85] P. Hu *et al.*, "Reproducibility with repeat CT in radiomics study for rectal cancer," *Oncotarget,* vol. 7, no. 44, pp. 71440-71446, Nov 1 2016, doi: 10.18632/oncotarget.12199.

[86] M. C. Desseroit *et al.*, "Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort," *J Nucl Med,* vol. 58, no. 3, pp. 406-411, Mar 2017, doi: 10.2967/jnumed.116.180919.

[87] P. Therasse *et al.*, "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J Natl Cancer Inst,* vol. 92, no. 3, pp. 205-16, Feb 2 2000, doi: 10.1093/jnci/92.3.205.

[88] P. Woznicki *et al.*, "Multiparametric MRI for Prostate Cancer Characterization: Combined Use of Radiomics Model with PI-RADS and Clinical Parameters," *Cancers (Basel),* vol. 12, no. 7, Jul 2 2020, doi: 10.3390/cancers12071767.

[89] C. V. Zwirewich, S. Vedal, R. R. Miller, and N. L. Muller, "Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation," *Radiology,* vol. 179, no. 2, pp. 469-76, May 1991, doi: 10.1148/radiology.179.2.2014294.

[90] T. P. Coroller *et al.*, "Radiomic phenotype features predict pathological

response in non-small cell lung cancer," *Radiother Oncol,* vol. 119, no. 3, pp. 480-6, Jun 2016, doi: 10.1016/j.radonc.2016.04.004.

[91] R. T. Leijenaar *et al.*, "Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability," *Acta Oncol,* vol. 52, no. 7, pp. 1391-7, Oct 2013, doi: 10.3109/0284186X.2013.812798.

[92] Y. Balagurunathan *et al.*, "Reproducibility and Prognosis of Quantitative Features Extracted from CT Images," (in English), *Transl Oncol,* vol. 7, no. 1, pp. 72-87, Feb 2014, doi: 10.1593/tlo.13844.

[93] C. A. Burmeister *et al.*, "Cervical cancer therapies: Current challenges and future perspectives," (in English), *Tumour Virus Res,* vol. 13, Jun 2022, doi: ARTN 200238

10.1016/j.tvr.2022.200238.

[94] J. Meng *et al.*, "Texture Analysis as Imaging Biomarker for recurrence in advanced cervical cancer treated with CCRT," (in English), *Sci Rep-Uk,* vol. 8, Jul 30 2018, doi: ARTN 11399

10.1038/s41598-018-29838-0.

[95] S. Pedraza *et al.*, "The value of metabolic parameters and textural analysis in predicting prognosis in locally advanced cervical cancer treated with chemoradiotherapy," (in English), *Strahlenther Onkol,* vol. 198, no. 9, pp. 792-801, Sep 2022, doi: 10.1007/s00066-022-01900-x.

[96] I. Yamada *et al.*, "Texture Analysis of Apparent Diffusion Coefficient Maps in Cervical Carcinoma: Correlation with Histopathologic Findings and Prognosis," (in English), *Radiol-Imag Cancer,* vol. 2, no. 3, May 2020, doi: ARTN e190085

10.1148/rycan.2020190085.

[97] Y. P. Zhang *et al.*, "Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling," (in English), *Military Med Res,* vol. 10, no. 1, May 16 2023, doi: ARTN 22

10.1186/s40779-023-00458-8.

[98] T. S. Jiang *et al.*, "Radiomics signature of osteoarthritis: Current status and perspective," (in English), *J Orthop Transl,* vol. 45, pp. 100-106, Mar 2024, doi: 10.1016/j.jot.2023.10.003.

[99] N. Horvat, N. Papanikolaou, and D. M. Koh, "Radiomics Beyond the Hype: A Critical Evaluation Toward Oncologic Clinical Use," *Radiol Artif Intell,* vol.

6, no. 4, p. e230437, Jul 2024, doi: 10.1148/ryai.230437.

[100] V. Chiappa *et al.*, "The Adoption of Radiomics and machine learning improves the diagnostic processes of women with Ovarian MAsses (the AROMA pilot study)," (in English), *J Ultrasound,* vol. 24, no. 4, pp. 429-437, Dec 2021, doi: 10.1007/s40477-020-00503-5.

[101] V. Chiappa *et al.*, "Using rADioMIcs and machine learning with ultrasonography for the differential diagnosis of myometRiAL tumors (the ADMIRAL pilot study). Radiomics and differential diagnosis of myometrial tumors," (in English), *Gynecol Oncol,* vol. 161, no. 3, pp. 838-844, Jun 2021, doi: 10.1016/j.ygyno.2021.04.004.

[102] V. Chiappa *et al.*, "Using Radiomics and Machine Learning Applied to MRI to Predict Response to Neoadjuvant Chemotherapy in Locally Advanced Cervical Cancer," (in English), *Diagnostics,* vol. 13, no. 19, Oct 2023, doi: ARTN 3139

10.3390/diagnostics13193139.

[103] B. Liang *et al.*, "Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis," (in English), *Frontiers in Oncology,* vol. 9, Apr 12 2019, doi: ARTN 269

10.3389/fonc.2019.00269.

[104] S. H. Lee *et al.*, "Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy," (in English), *Physics in Medicine and Biology,* vol. 65, no. 19, Oct 7 2020, doi: ARTN 195015

10.1088/1361-6560/ab8531.

[105] A. Q. Wu *et al.*, "Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases," (in English), *Oral Oncol,* vol. 104, May 2020, doi: ARTN 104625

10.1016/j.oraloncology.2020.104625.

[106] A. Eloyan, M. S. Yue, and D. Khachatryan, "Tumor heterogeneity estimation for radiomics in cancer," (in English), *Stat Med,* vol. 39, no. 30, pp. 4704-4723, Dec 30 2020, doi: 10.1002/sim.8749.

[107] W. D. Kang *et al.*, "Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis," (in English), *J Transl Med,* vol. 21, no. 1, Sep 6 2023, doi: ARTN 598

10.1186/s12967-023-04437-4.

[108] L. E. Sanchez, L. Rundo, A. B. Gill, M. Hoare, E. M. Serrao, and E. Sala, "Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle," (in English), *Sci Rep-Uk,* vol. 11, no. 1, Apr 15 2021, doi: ARTN 8262

10.1038/s41598-021-87598-w.

[109] Y. Soleymani, A. R. Jahanshahi, A. Pourfarshid, and D. Khezerloo, "Reproducibility assessment of radiomics features in various ultrasound scan settings and different scanner vendors," (in English), *J Med Imaging Radiat,* vol. 53, no. 4, pp. 664-671, Dec 2022, doi: 10.1016/j.jmir.2022.09.018.

[110] T. A. D'Antonoli *et al.*, "Reproducibility of radiomics quality score: an intra- and inter-rater reliability study," (in English), *European Radiology,* vol. 34, no. 4, pp. 2791-2804, Apr 2024, doi: 10.1007/s00330-023-10217-x.

[111] X. Z. Teng *et al.*, "Building reliable radiomic models using image perturbation," (in English), *Sci Rep-Uk,* vol. 12, no. 1, Jun 16 2022, doi: ARTN 10035

10.1038/s41598-022-14178-x.

[112] L. Ubaldi, S. Saponaro, A. Giuliano, C. Talamonti, and A. Retico, "Deriving quantitative information from multiparametric MRI via Radiomics: Evaluation of the robustness and predictive value of radiomic features in the discrimination of low-grade versus high-grade gliomas with machine learning," (in English), *Phys Medica,* vol. 107, Mar 2023, doi: ARTN 102538

10.1016/j.ejmp.2023.102538.

[113] A. Jahanshahi, Y. Soleymani, M. F. Ghaziani, and D. Khezerloo, "Radiomics reproducibility challenge in computed tomography imaging as a nuisance to clinical generalization: a mini-review," (in English), *Egypt J Radiol Nuc M,* vol. 54, no. 1, May 9 2023, doi: ARTN 83

10.1186/s43055-023-01029-6.

[114] J. Zhang *et al.*, "Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients," (in English), *Radiotherapy and Oncology,* vol. 183, Jun 2023, doi: ARTN 109578

10.1016/j.radonc.2023.109578.

[115] W. Small *et al.*, "NRG Oncology/RTOG Consensus Guidelines for

Delineation of Clinical Target Volume for Intensity Modulated Pelvic Radiation Therapy in Postoperative Treatment of Endometrial and Cervical Cancer: An Update," (in English), *Int J Radiat Oncol,* vol. 109, no. 2, pp. 413-424, Feb 1 2021, doi: 10.1016/j.ijrobp.2020.08.061.

[116] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," (in English), *Proc Int Conf Doc,* pp. 958-962, 2003. [Online]. Available: <Go to ISI>://WOS:000185624000176.

[117] P. Whybra *et al.*, "The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights," (in English), *Radiology,* vol. 310, no. 2, Feb 2024, doi: ARTN e231319

10.1148/radiol.231319.

[118] M. Avanzo *et al.*, "Machine and deep learning methods for radiomics," (in English), *Medical Physics,* vol. 47, no. 5, pp. E185-E202, Jun 2020, doi: 10.1002/mp.13678.

[119] R. Obuchowicz, M. Strzelecki, and A. Piórkowski, "Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing-A Review," (in English), *Cancers,* vol. 16, no. 10, May 2024, doi: ARTN 1870

10.3390/cancers16101870.

[120] R. Sun *et al.*, "A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study," (in English), *Lancet Oncol,* vol. 19, no. 9, pp. 1180-1191, Sep 2018, doi: 10.1016/S1470-2045(18)30413-3.

[121] M. E. Mayerhoefer *et al.*, "Introduction to Radiomics," (in English), *Journal of Nuclear Medicine,* vol. 61, no. 4, pp. 488-495, Apr 1 2020, doi: 10.2967/jnumed.118.222893.

[122] R. T. H. Leijenaar *et al.*, "Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study," (in English), *Brit J Radiol,* vol. 91, no. 1086, 2018, doi: ARTN 2017049811075

10.1259/bjr.20170498.

[123] J. Y. Y. Kwan *et al.*, "Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in HPV-related Oropharyngeal Carcinoma," (in English), *Int J Radiat Oncol,* vol. 102, no. 4, pp. 1107-1116, Nov 15 2018, doi: 10.1016/j.ijrobp.2018.01.057.

[124] W. Rogers *et al.*, "Radiomics: from qualitative to quantitative imaging," (in English), *Brit J Radiol,* vol. 93, no. 1108, 2020, doi: ARTN 20190948

10.1259/bjr.20190948.

[125] A. Jethanandani *et al.*, "Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review," (in English), *Frontiers in Oncology,* vol. 8, May 14 2018, doi: ARTN 131

10.3389/fonc.2018.00131.

[126] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," (in English), *Int J Radiat Oncol,* vol. 102, no. 4, pp. 1143-1158, Nov 15 2018, doi: 10.1016/j.ijrobp.2018.05.053.

[127] R. Da-ano *et al.*, "Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies," (in English), *Sci Rep-Uk,* vol. 10, no. 1, Jun 24 2020, doi: ARTN 10248

10.1038/s41598-020-66110-w.

[128] A. Ibrahim *et al.*, "The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization," (in English), *Cancers,* vol. 13, no. 8, Apr 2021, doi: ARTN 1848

10.3390/cancers13081848.

[129] R. N. Mahon, G. D. Hugo, and E. Weiss, "Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome," (in English), *Physics in Medicine and Biology,* vol. 64, no. 14, Jul 2019, doi: ARTN 145007

10.1088/1361-6560/ab18d3.

[130] J. E. Park *et al.*, "Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement," (in English), *European Radiology,* vol. 30, no. 1, pp. 523-536, Jan 2020, doi: 10.1007/s00330-019-06360-z.

[131] E. Pfaehler *et al.*, "Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts," (in English), *Journal of Nuclear Medicine,* vol. 61, no. 3, pp. 469-476, Mar 1 2020, doi: 10.2967/jnumed.119.229724.

[132] M. L. Welch *et al.*, "Vulnerabilities of radiomic signature development: The

need for safeguards," (in English), *Radiotherapy and Oncology,* vol. 130, pp. 2-9, Jan 2019, doi: 10.1016/j.radonc.2018.10.027.

[133] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, "Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomics," (in English), *Radiology,* vol. 291, no. 1, pp. 52-58, Apr 2019, doi: 10.1148/radiol.2019182023.

[134] S. W. Mes *et al.*, "Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures," (in English), *European Radiology,* vol. 30, no. 11, pp. 6311-6321, Nov 2020, doi: 10.1007/s00330-020-06962-y.

[135] R. B. Ger *et al.*, "Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies," (in English), *Sci Rep-Uk,* vol. 8, Aug 29 2018, doi: ARTN 13047

10.1038/s41598-018-31509-z.

[136] M. Pavic *et al.*, "Influence of inter-observer delineation variability on radiomics stability in different tumor sites," (in English), *Acta Oncologica,* vol. 57, no. 8, pp. 1070-1074, 2018, doi: 10.1080/0284186x.2018.1445283.

[137] P. Whybra, C. Parkinson, K. Foley, J. Staffurth, and E. Spezi, "Assessing radiomic feature robustness to interpolation in

F-FDG PET imaging," (in English), *Sci Rep-Uk,* vol. 9, Jul 4 2019, doi: ARTN 9649

10.1038/s41598-019-46030-0.

[138] I. Buvat and F. Orlhac, "The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results," (in English), *Journal of Nuclear Medicine,* vol. 60, no. 11, pp. 1543-1544, Nov 1 2019, doi: 10.2967/jnumed.119.235325.

[139] C. Parmar, J. D. Barry, A. Hosny, J. Quackenbush, and H. J. W. L. Aerts, "Data Analysis Strategies in Medical Imaging," (in English), *Clin Cancer Res,* vol. 24, no. 15, pp. 3492-3499, Aug 1 2018, doi: 10.1158/1078-0432.Ccr-18-0385.

[140] S. Starke *et al.*, "2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma," *Sci Rep,* vol. 10, no. 1, p. 15625, Sep 24 2020, doi: 10.1038/s41598-020-70542-9.

[141] M. Bogowicz, S. Tanadini-Lang, M. Guckenberger, and O. Riesterer, "Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer," *Sci Rep,* vol. 9, no. 1, p. 15198, Oct 23 2019, doi: 10.1038/s41598-019-51599-7.

[142] I. Zhovannik *et al.*, "Learning from scanners: Bias reduction and feature correction in radiomics," *Clin Transl Radiat Oncol,* vol. 19, pp. 33-38, Nov 2019, doi: 10.1016/j.ctro.2019.07.003.

[143] D. Ou *et al.*, "Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status," *Oral Oncol,* vol. 71, pp. 150-155, Aug 2017, doi: 10.1016/j.oraloncology.2017.06.015.

[144] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," (in English), *J Digit Imaging,* vol. 26, no. 6, pp. 1045-1057, Dec 2013, doi: 10.1007/s10278-013-9622-7.

[145] M. Vallieres *et al.*, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Sci Rep,* vol. 7, no. 1, p. 10117, Aug 31 2017, doi: 10.1038/s41598-017-10371-5.

[146] A. J. Grossberg *et al.*, "Author Correction: Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy," *Sci Data,* vol. 5, no. 1, p. 1, Nov 27 2018, doi: 10.1038/s41597-018-0002-5.

[147] M. M. D. A. C. C. Head and G. Neck Quantitative Imaging Working, "Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges," *Sci Data,* vol. 4, p. 170077, Jul 4 2017, doi: 10.1038/sdata.2017.77.

[148] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "mRMRe: an R package for parallelized mRMR ensemble feature selection," (in English), *Bioinformatics,* vol. 29, no. 18, pp. 2365-2368, Sep 15 2013, doi: 10.1093/bioinformatics/btt383.

[149] E. Motakis, A. V. Ivshina, and V. A. Kuznetsov, "Data-driven approach to predict survival of cancer patients: estimation of microarray genes' prediction significance by Cox proportional hazard regression model," *IEEE Eng Med Biol Mag,* vol. 28, no. 4, pp. 58-66, Jul-Aug 2009, doi: 10.1109/MEMB.2009.932937.