# RADIOTHERAPY DATA ANALYSIS AND REPORTING (RADAR) TOOLKIT: AN END-TO-END ARTIFICIAL INTELLIGENCE DEVELOPMENT SOLUTION FOR PRECISION MEDICINE

ZHANG JIANG

PhD

The Hong Kong Polytechnic University

2023

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

# Radiotherapy Data Analysis and Reporting (RADAR) toolkit: an end-to-end artificial intelligence development solution for precision medicine

ZHANG Jiang

A thesis submitted in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

March 2023

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 acknowledgement has been made in the text.

_____(Signed)

ZHANG Jiang (Name of Student)

## Abstract

Radiotherapy is one of the mainstream treatment modalities for cancer. A large amount of structured data, including image, dose, and structure delineations, is produced during treatment planning. Technological advancement, especially artificial intelligence, facilitates the development of more sophisticated quantitative biomarkers from radiotherapy data for improved performance in precision medicine. Nevertheless, challenges in low data processing efficiency, incomplete data usage, and lack of reliability assessment hinder the development and bench-to-bedside translation. This thesis aims to develop an end-to-end and integrated RAdiotherapy Data Analysis and Reporting (RADAR) toolkit for efficient, comprehensive, and reliable quantitative biomarker developments and to evaluate its utility and performance in multiple clinical applications.

RADAR is composed of GUI-equipped semi-independent modules for data curation, feature extraction, and model development. During the development of RADAR, we embedded a new multi-model feature set with innovative designs of anatomical features based on structure delineations and implemented perturbation-based repeatability assessment algorithm. By using the RADAR platform, we investigated radiomic feature (RF) repeatability and its agreements across imaging modalities and head-and-neck cancer subtypes via image perturbations, attempting to provide a direct perceptivity in RF pre-selection for robust model construction. We retrospectively collected contrast-enhanced computed tomography (CECT), contrast-enhanced T1-

weight (CET1-w), T2-weight (T2-w) magnetic resonance (MR) images of 231 nasopharyngeal carcinoma (NPC) patients from Queen Elizabeth Hospital (QEH), and CECT images of 399 oropharyngeal carcinoma (OPC) patients from online database. Randomized translation and rotation were implemented to the images for mimicking scanning position stochasticity. The intra-class correlation coefficient (ICC) was calculated for each RF to assess its repeatability and quantitatively compared to evaluate the repeatability agreement. We also investigated the impact of RF repeatability on generalizable model development on Nasopharyngeal Carcinoma (NPC) cases using CET1-w MR images of 286 NPC patients from QEH for training and 183 from Queen Mary Hospital for external validation. Two separate survival models were developed using high-repeatable and low-repeatable RFs exclusively and compared on their prognostic performance in the validation set. In addition to the two technical studies, we developed two quantitative biomarkers based on anatomical and radiomic features for prognosis and treatment efficacy predictions of NPC patients. Based on the same NPC cohorts, we identified independent prognostic factors from anatomical features of lymph node tumor and constructed a prognostic index with N stage. In the last study, we identified single predictive radiomic feature extracted from primary gross tumor volume for patients receiving concurrent chemoradiotherapy with/without addition of adjuvant chemotherapy (ACT). We further investigated the predictive value of its voxel-wise feature mapping for feature explanation.

We have successfully developed RADAR for efficient, comprehensive, and reliable radiotherapy data analysis for clinical biomarker development, With the help of RADAR, we discovered that more than half of the wavelet-filtered RFs, especially

texture features, were highly susceptible to scanning position variations, irrespective of image modalities or HNC subtypes. It was more prominent when a smaller discretization bin number was used. Using high-repeatable RFs for model development yielded a significantly higher concordance-index (0.63) in the validation cohort than when only low-repeatable RFs were used (0.57, *p*-value= 0.024), suggesting higher model generalizability. For the two developed biomarkers, the anatomy-based prognostic index demonstrated superior cross-institutional performance in disease-free survival (DFS) than the clinical baseline N stage. The predictive radiomic feature, gldm_DependenceVariance in 3mm-sigma LoG filtered image, was discovered, and the high-risk patients who received additional adjuvant chemotherapy achieved a 3-year DFS rate of 90% versus 57% for low-risk patients. The predictive value can also be generalized to the highlighted subvolume of the feature map.

In conclusion, RADAR has been demonstrated as a highly useful tool for efficient analysis of radiotherapy data and effective development of biomarkers for precision medicine. We urge caution when handling wavelet-filtered RFs and advise taking initiatives to exclude low-repeatable RFs during feature pre-selection for generalizable model construction. By using the newly designed anatomical features, the spatial characterization of lymph node tumor anatomy improved the existing N-stage in NPC prognosis. The radiomic signature with its voxel-wise mapping could be a reliable and explainable ACT decision-making tool in clinical practice.

# Dedication

I dedicate this thesis to the people who have played a significant role in my life as a student and researcher. Their unwavering support has made my journey towards attaining a Ph.D. degree a reality.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. CAI Jing who have not only shared their expertise but also inspired and motivated me to reach my goals. His encouragement, advice, and constructive feedback have been instrumental in shaping me into the researcher I am today.

I would also like to extend a heartfelt appreciation to my family members and friends who have stood by me during this journey. Their continuous love and support have been my anchor in times of uncertainty and stress. Their belief in my abilities has reinforced my determination to complete this thesis and pursue my future careers. I recognize the sacrifices they have made along the way, taking up extra responsibilities to allow me to focus on my research work.

I am grateful to the funding agencies that have supported my research. Their financial support has enabled me to conduct my research and given me the opportunity to pursue my academic and professional goals.

I would like to thank my classmates for their help in my research and my life along the way. Their contribution has been invaluable and has significantly impacted the quality of my work. Their willingness to share their knowledge, expertise, and ideas have enriched my research and made it a more collaborative and rewarding experience.

Finally, I am thankful to the colleges in hospitals who have helped us acquire the radiotherapy data. Their contributions have been instrumental in the success of my research.

Publications

1. **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*, 109578.

2. **Zhang, J**., Teng, X., Lam, S., Sun, J., Cheung, A. L. Y., Ng, S. C. Y., ... & Cai, J. (2023). Quantitative Spatial Characterization of Lymph Node Tumor for N Stage Improvement of Nasopharyngeal Carcinoma Patients. *Cancers*, 15(1), 230.

3. **Zhang, J.**, Lam, S., Teng, X., Zhang, Y., Ma, Z., Lee, F., ... & Cai, J. (2022, September). Repeatability of Radiomic Features Against Simulated Scanning Position Stochasticity Across Imaging Modalities and Cancer Subtypes: A Retrospective Multi-institutional Study on Head-and-Neck Cases. *In Computational Mathematics Modeling in Cancer Analysis: First International Workshop, CMMCA 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings* (pp. 21-34). Cham: Springer Nature Switzerland.

4. Ho, L. M., Lam, S. K., **Zhang, J**., Chiang, C. L., Chan, A. C. Y., & Cai, J. (2023). Association of Multi-Phasic MR-Based Radiomic and Dosimetric Features with Treatment Response in Unresectable Hepatocellular Carcinoma Patients following Novel Sequential TACE-SBRT-Immunotherapy. *Cancers*, 15(4), 1105.

5. Zhang, Y., Yang, D., Lam, S., Li, B., Teng, X., **Zhang, J**., ... & 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.

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.

7. Li, B., Zheng, X., **Zhang, J.**, Lam, S., Guo, W., Wang, Y., ... & Cai, J. (2022). Lung Subregion Partitioning by Incremental Dose Intervals Improves Omics-Based Prediction for Acute Radiation Pneumonitis in Non-Small-Cell Lung Cancer Patients. *Cancers*, 14(19), 4889.

8. Li, B., Ren, G., Guo, W., **Zhang, J**., Lam, S. K., Zheng, X., ... & Ge, H. (2022). Function-Wise Dual-Omics analysis for radiation pneumonitis prediction in lung cancer patients. *Computational Intelligence in Personalized Medicine*, 110.

9. Li, W., Lam, S., Li, T., Cheung, A.L.Y., Xiao, H., Liu, C., **Zhang, J.**, Teng, X., Zhi, S., Ren, G. and Lee, F.K.H., 2022, September. Multi-institutional investigation of model generalizability for virtual contrast-enhanced mri synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII* (pp. 765-773). Cham: Springer Nature Switzerland.

10. Teng, X., **Zhang, J.**, Zwanenburg, A., Sun, J., Huang, Y., Lam, S., Zhang, Y., Li, B., Zhou, T., Xiao, H. and Liu, C., 2022. Building reliable radiomic models using image perturbation. *Scientific Reports*, 12(1), pp.1-10.

11. Lam, S.K., **Zhang, J.**, Zhang, Y.P., Li, B., Ni, R.Y., Zhou, T., Peng, T., Cheung, A.L.Y., Chau, T.C., Lee, F.K.H. and Yip, C.W.Y., 2022. A multi-center study of CT-based neck nodal radiomics for predicting an adaptive radiotherapy trigger of ill-fitted thermoplastic masks in patients with nasopharyngeal carcinoma. *Life*, 12(2), p.241.

12. Lam, S.K., Zhang, Y., **Zhang, J.**, Li, B., Sun, J.C., Liu, C.Y.T., Chou, P.H., Teng, X., Ma, Z.R., Ni, R.Y. and Zhou, T., 2022. Multi-organ omics-based prediction for adaptive radiation therapy eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. *Frontiers in oncology*, 11, p.5406.

13. Liu, C., Li, M., Xiao, H., Li, T., Li, W., **Zhang, J.**, Teng, X. and Cai, J., 2022. Advances in MRI-guided precision radiotherapy. *Precision Radiation Oncology*, 6(1), pp.75-84.

14. Li, W., Xiao, H., Li, T., Ren, G., Lam, S., Teng, X., Liu, C., **Zhang, J.**, Lee, F.K.H., Au, K.H. and Lee, V.H.F., 2022. Virtual contrast-enhanced magnetic resonance images synthesis for patients with nasopharyngeal carcinoma using multimodality-guided synergistic neural network. *International Journal of Radiation Oncology* Biology* Physics*, 112(4), pp.1033-1044.

15. Zhang, Y., Lam, S., Yu, T., Teng, X., **Zhang, J.**, Lee, F.K.H., Au, K.H., Yip, C.W.Y., Wang, S. and Cai, J., 2022. Integration of an imbalance framework with

novel high-generalizable classifiers for radiomics-based distant metastases prediction of advanced nasopharyngeal carcinoma. *Knowledge-based systems*, 235, p.107649.

16. Ren, G., Lam, S.K., **Zhang, J.**, Xiao, H., Cheung, A.L.Y., Ho, W.Y., Qin, J. and Cai, J., 2021. Investigation of a novel deep learning-based computed tomography perfusion mapping framework for functional lung avoidance radiotherapy. *Frontiers in Oncology*, 11, p.644703.

17. Ren, G., **Zhang, J.**, Li, T., Xiao, H., Cheung, L.Y., Ho, W.Y., Qin, J. and Cai, J., 2021. Deep learning-based computed tomography perfusion mapping (DL-CTPM) for pulmonary CT-to-perfusion translation. *International Journal of Radiation Oncology* Biology* Physics*, 110(5), pp.1508-1518.

Table of Contents

List of Tables

List of Figures

List of Acronyms

| | |
|---|---|
| **2D** | Two-dimensional |
| **3D** | Three-dimensional |
| **4D** | Four-dimensional |
| **95CI** | 95% confidence interval |
| **ACT** | Adjuvant chemotherapy |
| **AI** | Artificial intelligence |
| **AJCC** | American Joint Committee on Cancer |
| **ANOVA** | Analysis of variance |
| **AUC** | Area under the curve |
| **CCRT** | Concurrent chemoradiotherapy |
| **C-index** | Concordance index |
| **CT** | Computed Tomography |
| **DFS** | Disease-free survival |
| **DICOM** | Digital Imaging and Communications in Medicine |
| **DMFS** | Distant metastasis-free survival |
| **DSC** | Dice similarity coefficient |
| **DVH** | Dose volume histogram |
| **GLCM** | Gray-Level Co-occurrence Matrix |
| **GLDM** | Gray Level Dependence Matrix |
| **gldm_DV** | gldm_DependenceVariance |

| | |
|---|---|
| **GLRLM** | Gray Level Run Length Matrix |
| **GLSZM** | Gray Level Size Zone Matrix |
| **GTVn** | Lymph node gross tumor volume |
| **GTVp** | Primary gross-tumor-volume |
| **GUI** | Graphical User Interface |
| **HD** | Hausdorff distance |
| **HNC** | Head and neck cancer |
| **HR** | Hazard ratio |
| **IBSI** | Image Biomarker Standardisation Initiative |
| **ICC** | Intraclass correlation coefficient |
| **IMRT** | Intensity modulated radiation therapy |
| **KM** | Kaplan-Meier |
| **LA** | Locoregionally advanced |
| **LN** | Lymph node |
| **LoG** | Laplacian-of-Gaussian |
| **LRFS** | Local-regional relapse-free survival |
| **MAD** | Mean absolute difference |
| **MRI** | Magnetic Resonance Imaging |
| **NCCN** | National Comprehensive Cancer Network |
| **NGTDM** | Neighbouring Gray Tone Difference Matrix |
| **NPC** | Nasopharyngeal Carcinoma |
| **OAR** | Organ-at-risk |
| **OARs** | organs-at-risk |

| | |
|---|---|
| **OPC** | Oropharyngeal carcinoma |
| **OS** | Overall survival |
| **OVH** | Overlap volume histogram |
| **PACs** | Picture Archiving and Communication System |
| **Parotid_L** | Left parotid |
| **PC** | Principal components |
| **PCA** | Principal component analysis |
| **PET** | Positron Emission Tomography |
| **PLN** | Parotid lymph node |
| **POV** | Projection overlap volume |
| **QEH** | Queen Elizabeth Hospital |
| **QMH** | Queen Mary Hospital |
| **RADAR** | RAdiotherapy Data Analysis and Reporting |
| **RF** | Radiomic feature |
| **RFS** | Relapse-free survival |
| **RLN** | Retropharyngeal lymph node |
| **ROC** | Receiver operating characteristic |
| **ROI** | Region-of-interest |
| **RT** | Radiotherapy |
| **SEER** | Surveillance, Epidemiology, and End Results |
| **TNM** | Tumor–node–metastasis |
| **UICC** | Union for International Cancer Control |
| **WHO** | World Health Organization |

# Chapter 1.

# Introduction

## 1.1. Background

Cancer is a devastating disease threatening billions of lives worldwide annually[1]. Radiotherapy (RT), as a major cancer treatment approach, has gone through numerous technological advances in the past decade, consolidating its indispensable role in contemporary cancer intervention. Concurrently, the precision medicine initiative has been launched with the aim to enhance healthcare delivery by providing personalized treatment to patients for maximized effectiveness and minimized side effects[2,3]. In this revolutionary phase, translational research in RT, which aims at moving laboratory discoveries to clinical trials or event clinical practice[3], has gained increasing attention due to the rich information of RT data. Thanks to the rapidly evolving big data and quantitative analysis techniques, more sophisticated knowledge can be inferenced from RT data with unprecedented complexity and volume[4].

## 1.2. Radiotherapy Data

RT data has a high variety and highly digitalized due to the increasing complexity and precision of RT treatments[5,6]. Patient clinical records are continuously produced in diagnosis (demographics, TNM stage, histopathology information), treatment delivery (surgery, chemotherapy, RT records) and follow-ups (survival and toxicity information)[7], mostly in numerical, categorical, or free text formats. They can be either undigitized as paper charts[8] or digitalized and standardized by means of, for instance, electronic medical

record[9]. They can be obtained from the home institution or an external or public database such as the Surveillance, Epidemiology, and End Results (SEER) database from the National Cancer Institute. Multiple imaging modalities are applied throughout the treatment course due to the technological advancements and the increasing popularity of image-guided RT[10]. Computed Tomography (CT) is an effective and widely used imaging modality in both diagnosis and treatment planning due to the quantitative and high-quality representation of patient anatomy. Magnetic Resonance Imaging (MRI) shows outstanding contrast to soft tissue with the absence of radiation dose. It has wide applications in diagnosis[11] , treatment planning[12], treatment guidance[13], and response assessment[14] . Positron Emission Tomography (PET) is a nuclear medicine technique that highlights local metabolism by radiotracers. Such functional information makes it highly sensible for viable and early cancer tissues, giving it unique advantages in diagnosis, treatment planning, and recurrence assessment[15]. Despite all the different modalities, the imaging data are deep down multidimensional matrix, mostly stored and transmitted according to the Digital Imaging and Communications in Medicine (DICOM) standard[16]. RT treatment planning data, such as planning CT, structure delineations, calculated dose map, and beam configurations, are generated to guide the radiation dose delivery during the treatment. Planning CT is one specific CT dedicated to patient localization and dose calculation. Structure delineations define patient anatomy by three-dimensional (3D) structures of treatment targets and organs-at-risk (OARs). The dose map, represented by a 3D matrix, simulates the final dose delivered to the patient. Combined with structure delineations, the dose map is used for optimization and evaluation of treatment plans with the aim of a uniform prescription dose in the target region and a minimum dose in the

surrounding normal tissue. The beam setups describe how the beam should be placed and how the radiation should be delivered. The entire set of planning data is standardized by the extended DICOM protocol named DICOM-RT[17] , facilitating intermachine and interinstitutional data sharing.

## 1.3. Radiotherapy Data Analysis

### 1.3.1. Data Curation

As the first step of most RT data analysis pipelines, data curation harmonizes the less structured raw RT data into organized datasets with enhanced value and veracity. For example, standardizations of the structure names in treatment plans are required when studying a multi-center patient cohort with inconsistent naming conventions[18]. Curating a larger RT dataset can be time-consuming and tedious due to the large volume and variety of RT data. Moreover, the highly complex data structure makes the raw RT data less interpretable for non-professionals, which further increases the cost of data curation in both time and manpower. Software tools have been developed to facilitate data curation by breaking the barrier of technical complexity. For example, 3D slicer is one of the widely used open-source software platforms by researchers and clinicians. It provides interactive tools for RT data visualization and manipulation[19]. Similar software tools are Medical Imaging Interactive Toolkit (MITK)[20], Computational Environment for Radiological Research (CERR)[21], Cancer Imaging Phenomics Toolkit (CaPTk)[22], and LIFEx[23]. All of them are equipped with Graphical User Interface (GUI) to be accessible to general public.

### 1.3.2. Biomarker Development

Biomarkers, defined as "objective indications of medical state observed from outside the patient"[24], serve as diagnostic and prognostic factors for guiding treatment managements. Traditional biomarkers are mostly based on histological or genomic measurements with theoretical background in biological or cellular process. Until recently, quantitative biomarkers based on RT data have been proposed and clinically validated. They are also called models which associate the input RT data with the target clinical endpoint through mathematical modeling. They have the unique advantages of non-invasiveness and high accessibility, making precision medicine available to a wider population. More complex quantitative biomarkers, especially ones developed by advanced machine learning techniques, has attracted increasing attention in the field of translation research due to their high performance and less dependance on prior knowledge. Despite the variety of methods in developing biomarkers, both effectiveness and reproducibility[25] need to be guaranteed for clinical utility. One popular approach to construct quantitative biomarkers is Radiomics[26]. It leverages the high-throughput features extracted from imaging data by extracting a large set of quantitative features. They are later filtered and combined based on the correlation with clinical endpoint using machine learning or deep learning algorithms.

Current developments in quantitative biomarker suffered from poor reproducibility[27]. As commented by Steyerberg et al.[28(p3)], among the large number of prognostic models proposed by literature, very few of them have been implemented and externally validated in clinic. Radiomic feature (RF) reproducibility has recently received increasing attention from the research community due to the poor agreements of selected

features between studies[29]. Image Biomarker Standardisation Initiative (IBSI)[27] is an ongoing project aiming to achieve consensus on RF definitions, image processing workflow, quality assurance pipeline, and reporting guidelines through international collaborations. Various software tools with GUI and programming packages have been developed to facilitate standardized RF extraction[23,30,31]. In addition to standardization in data processing, more literatures have focused on the assessment of RF repeatability using experimental[32–35] and perturbation methods. RFs that are less repeatable under the same or similar data acquisition conditions and processing methods are recommended to be excluded from radiomics analysis[36].

Challenges remain in quantitative RT data analysis that inhibit the further development of quantitative biomarkers for precision medicine. First, the current GUI-equipped software tools are mostly designed for single-patient data processing, which can be time consuming for a large patient cohort due to manual operations. Such disadvantage often leads to a small sample size and reduced clinical significance of the drawn conclusion. Second, limited efforts were made in a comprehensive usage of RT data, especially structure delineations. Previous research has analyzed the imaging and planning dose data using Radiomcis and Dosimetrics for various clinical endpoints. Patient anatomical information is mostly assessed qualitatively for cancer diagnosis and prognosis according to established clinical protocols. Quantitative characterization of patient anatomy based on structure delineations was seldom investigated but may contain unique information of disease condition and predictive to patient prognosis. Third, repeatability assessment is not commonly performed in current biomarker developments. The assessment by experiments involves repeated data acquisition, which can be costly

and hazardous to patients. The perturbation-based method is advantageous[37] but has not been widely adopted, possibly due to the insufficient consensus on the benefit on model reliability and the lack of ready-to-use software tool for rapid implementation.

## 1.4. Aim and Objectives

In an attempt to mitigate the remaining challenges introduced before, this thesis aims to (1) develop an end-to-end and integrated RAdiotherapy Data Analysis and Reporting (RADAR) toolkit for efficient, comprehensive, and reliable quantitative biomarker development and (2) to evaluate its utility and performance conduct in multiple clinical applications. Three objectives are achieved in the thesis:

1. To develop RADAR with end-to-end cohort-level RT data analysis capability and technical innovations in feature design and feature repeatability analysis. RADAR is designed as GUI-equipped semi-independent modules for data curation, feature extraction, and model development. It is also optimized for maximum computation speed and minimum human intervention. New feature sets, including anatomical features extracted from structure delineations, are proposed to comprehensively describe the multifarious RT data. Perturbation based feature repeatability assessment is implemented and integrated in RADAR.
2. To study radiomic feature repeatability and its impact on model development using RADAR. Two technical studies are conducted to explore the patterns of feature repeatability in different image modalities and patient cohorts and investigate whether high-repeatable features can benefit model generalizability.

3. To apply RADAR in reliable biomarker developments for Nasopharyngeal Carcinoma (NPC) patients. Two clinical studies were performed for survival prognosis predictions using anatomical features and treatment efficacy predictions using radiomics, respectively. Feature repeatability assessments were incorporated into those two studies to ensure the reliability of the established models.

## 1.5. Thesis overview

This thesis will first review previous literatures on RT data analysis for clinical applications and the ongoing investigation of the reproducibility and repeatability issue. In the next chapter, which focuses on the first objective, the design and technical innovations of RADAR software will be explained in detail. The third chapter contains two published technical studies[36,38] on the analysis of the RF repeatability against patient positioning variations and its impact on model generalizability (the first study is reproduced with permission from Springer Nature). The fourth chapter introduces two clinical studies on the survival prognosis using the newly designed anatomical features and treatment efficacy prediction using Radiomics, where the first study has been published in a peer-reviewed journal[39]. Lastly, we summarize the entire thesis by revisiting the main results and discussing the limitations and future developments.

# Chapter 2.

# Literature Review

## 2.1. Quantitative Radiotherapy Data Analysis for Clinical Application

The rich information contained in the multifarious RT data promoted a broad range of technical and clinical advancements in translational research. Tumor controls and normal tissue toxicity are two most studied clinical outcomes, as they are two major considerations during clinical decision making. Early works on quantitative RT data analysis focused primarily on toxicity prediction and survival prognostication from patient clinicopathological records due to the high data availability and simple data structure. As a result, large patient cohorts are mostly recruited to enhance the model robustness and clinical impact. One sample is the highly influential study published in 2008 by Wang et al. on survival modeling of adjuvant RT for gallbladder cancer using 4180 patients collected from the SEER database[40]. A clinical-ready nomogram and web-based tool were further published to gain recognition and receive external validation. The same research group analyzed further on the survival benefit of adjuvant chemoradiotherapy from 1137 patients by constructing and comparing two survival regression models with and without the target treatment, and the nomograms along with the web-based tool were published[41]. For toxicity prediction, Langius et al. modeled the critical body weight loss after chemoradiotherapy of 910 internal head and neck cancer (HNC) patients by classification and regression tree using age and treatment protocol features[42]. The prediction tree was also presented for full transparency and potential clinical usage. A comprehensive study by Deist et al. attempted to predict multiple

outcomes, including both tumor control and toxicity, of 12 datasets using six different classifiers[43]. Technical recommendations were given instead of clinical solutions.

In addition to clinicopathological information, quantitative biomarkers are being constructed from both imaging and planning data, attempting to give more accurate and cost-effective diagnosis and predict multiple clinical endpoints after treatment. Unlike clinicopathological data, imaging and planning data are higher dimensional and require more complex processing for biomarker development. Some biomarkers from images are being used in routine clinical practice, such as TNM stage inferred from multiple imaging modalities and bone scan index calculated from SPECT[25]. Some have passed the regulatory approval or undergoing clinical trials, but most of them are still in the development stage. The routinely acquired imaging biomarkers, such as TNM staging, have been included into clinicopathological data and will be excluded from the discussions below. One of the most popular quantitative imaging biomarker development techniques is Radiomics. Radiomics leverages a comprehensive set of features extracted from standard imaging and correlates them with the underlying pathophysiology[26]. It has been recognized as a potentially effective tool for diagnosis, survival prognosis, and toxicity prediction. Zhu et al. constructed a CT-based radiomic signature from five features for diagnosing non-small cell lung cancer subtypes with high sensitivity (0.828) and specificity (0.900)[44]. Aerts' work has made a significant contribution to the field by associating prognostic RF with gene patterns and providing evidence of general prognostic imaging biomarkers across disease sites[45]. Several other highly impactful studies have demonstrated the independent predictive power of RF on survival. An early study by Huang et al. analyzed 132 texture RFs from CT images and constructed a

radiomics signature by combining selected features linearly[46]. The radiomics signature was proven to be predictive of disease-free survival of non-small cell lung cancer in addition to the widely applied clinicopathological factors. The study by Peng et al. has shed more light on value of Radiomics in clinical decision making by demonstrating the benefit of the constructed deep-learning-based PET/CT Radiomics prognosis signature in patient stratification for induction chemoradiotherapy[47]. Less studies have applied quantitative information from images to toxicity predictions with the exception of xerostomia prognostication[48]. The baseline or change of RFs from multiple imaging modalities have shown better performance than traditional analytical toxicity models in acute and late xerostomia predictions by several studies[49]. For example, Van Djik et al. found that texture features from pretreatment MRI can be predictive of late xerostomia, supporting the hypothesis of the association between predisposed fat and parotid toxicity[50].

Unlike imaging data, features from dose maps are mostly used in toxicity predictions due to the direct connection between delivered dose to radiation induced damage to normal tissues. Various dose features have been designed based on the anatomy of normal tissues and their radiobiology[51]. Dose volume histogram (DVH) was originally proposed to estimate the normal tissue complication probability and guide plan optimization and quality assurance in clinic[52]. Parameters from the DVH have been extensively studied to model toxicity outcomes[51]. Such histogram-based dose features lacks spatial information encoded in a 3D dose map, especially for tubular normal tissue anatomy such as rectum, esophagus, or pharyngeal mucosa. The DVH alternative — dose-length and dose-circumference — were applied to quantify the dose distribution

along and around the tube direction of the pharyngeal mucosa by Dean et al., and they were shown to be highly predictive to severe acute dysphagia[53]. More spatial dose features were designed to account for the potential spatial heterogeneity of normal tissue's radiosensitivity. Buettner et al. first proposed 3D dose moment invariants, which quantifies the center of mass, spread, and skewness of the dose distribution in the left-right, anterior-posterior and superior-inferior directions[54]. Their superior performance was also validated against traditional models in xerostomia prediction for head-and-cancer. Bourbonne et al. calculated RFs from the dose map and discovered better predictive power for grate ≥ 2 acute and late pulmonary toxicities than clinical and DVH models[55]. Similarly, only dose shape descriptors remained from a combination of DVH parameters, 3D dose invariants, dose gradients, and dose radiomics in a xerostomia prediction study by Gabrys et al[56]. Recent studies have combined dose descriptors with image biomarkers to enhance the toxicity prediction performance. Chopra et al. found better performance of machine learning models with the combination of image and dose RFs from lung subregions than DVH models in predicting radiation pneumonitis[57]. To further consider the complementarity and independency of the imaging and dose features, technical advancements such as multi-view analysis has been suggested to achieve better performance in a lung cancer acute body weight loss study[58]. In addition to toxicity predictions, dose features were demonstrated to be predictive of disease progression after the treatment. In a recent study by Wu et al., the PET/CT image radiomics model was only prognostic of locoregional recurrence for neck and cancer patient after the integration of dose RFs[59].

Compared to dose descriptors, much fewer investigations were performed on the predictions from patient anatomy based on the quantitative descriptions of structure delineations. In the study of early prediction of parotid shrinkage and toxicity by Pota et al., the high-dose target volume, parotid glands structures, lymph node chain, and their overlap volumes were combined with CT/PET for model construction[60]. Some anatomical information was more visible in certain imaging modalities, and the quantification is highly dependent on the imaging data. For example, the tumor location, mesorectal fascia status, and the extramural vascular invasion were manually quantified from MRI landmarks in the recent study by Chen et al[61]. The final constructed model showed an outstanding performance with validation AUC of 0.771 for 3-year disease-free survival (DFS) on locally advanced rectal cancer.

## 2.2. Repeatability and Reproducibility of Radiomic Features

During the fast advancement of RT data analysis techniques, increasing attention has been paid to the precision of quantitative biomarkers retrieved from RT data, especially RF extracted from imaging data. Evidence has suggested a wide range of RFs demonstrated unsatisfactory repeatability and reproducibility, which are precision measured in the same and different conditions[62], despite the effort of standardization in RF definitions by IBSI. Test-retest and four-dimensional (4D) imaging are the two mainstream techniques to assess RF repeatability and reproducibility.

### 2.2.1. Test-Retest Imaging

Test-retest imaging is one popular approach that attempts to reproduce the clinical variability through repeated scans. Less than half of the RF showed low repeatability in the same-day test-retest lung cancer CT imaging experiment conducted by Balagurunathan et al [63]. They were also found to have poor reproducibility under varying imaging conditions such as reconstruction algorithm, voxel size, and image acquisition parameters on CT. et al. assessed the repeatability of MR texture features from test-retest images of 14 patients with rectal cancer; and emphasized that even when the same imaging machine, protocol, and operator were employed, certain high-order texture features exhibited extremely deficient repeatability[33]. Similarly, another recent study conducted by Lu et al. evaluated RFs repeatability on 13 prostate cancer patients via test-retest MR images acquired within two weeks; and concluded that over 90% of the extracted RFs in all the studied MR sequences were not robust[64]. A well-recognized study by Berenguer et al. concluded that many RF were redundant and nonreproducible by test-retest analysis on phantoms[65]. Leijenaar discovered that around 70% of the radiomic features extracted from non-small cell lung cancer PET images were repeatable under test-retest. However, short-interval test-retest imaging it is not routinely practiced in clinic and may introduce extra dose to patients, limiting its application in prospective studies and specific image modalities such as MR and PET imaging. Repeated scans with an even prolonged time interval, in the case of two-week apart, might lead to dramatic disparity in RFs due to enlarged tumor morphological and intra-tumoral microbiologic changes, which may reduce the generalizability of RF repeatability findings[32].

### 2.2.2. Four-Dimensional Imaging

RF repeatability can also be assessed from multi-phase images where each phase corresponds to a snapshot in the patient's breathing cycle. Repeatable RFs across different phases in 4D image are insensitive to respiratory motion. Oliver et al. discovered that half of the included image features extracted from respiratory-gated PET images were susceptible to respiratory motion[66]. Another exploratory study by Larue concluded that low-repeatable features in test-retest could also be identified by 4D imaging[67]. They also pointed out the low generalizability of RF repeatability to cancer site and independency to patient survival. Lafata et al. reported around 30% of RF being sensitive to motion blurring, some of which were predictive to tumor histology when extracted from the end-of-exhalation phase[68]. Therefore, site-specific RF repeatability assessment is recommended for every radiomic study targeting clinical application, and 4D imaging could be an effective method for assessing the RF repeatability assessments when test-retest imaging is not available. On the other hand, 4D imaging is only applied to cancer sites that are largely affected by respiratory motion, such as lung and liver cancer. A more generalized approach is desired for wide-spread implementation of RF repeatability assessment in radiomic studies.

### 2.2.3. Image Perturbation

A new perturbation-based RF repeatability assessment method was proposed by Zwanenburg et al.[37] to overcome the limitations of test-retest imaging and 4D imaging by simulating variabilities in scanning position, image noise, and region-of-interest contouring. They examined RFs repeatability through a set of pre-defined image perturbations using test-retest CT images of 19 HNC and 31 lung cancer patients; and

reported a high concordance with test-retest based RF repeatability[37]. More recently, studies performed by Teng et al.[69,70] demonstrated enhanced radiomic model robustness and internal generalizability, where models were developed by using high-repeatable RFs exclusively. Thus, image perturbation is a promising new technique to assess radiomic feature repeatability from any retrospective cohort. However, there is currently no software tool that implements the image perturbation algorithm, and there is a lack of study investigating the pattern of radiomic feature repeatability and its benefit in improving external generalizability of radiomic model.

# Chapter 3.

# RADAR Development

## 3.1. Introduction

RADAR contains three semi-independent modules including data curation, feature extraction, and model construction. It provides a comprehensive analysis of RT data by new feature designs, embedded perturbation-based feature repeatability assessment algorithms, and standardized modeling workflows. They are also highly efficient with fully streamlined data processing optimized for large patient cohorts. RADAR is also equipped with interactive and intuitive GUI, making it accessible to users with limited skills in programming and computer systems. Herein, we will introduce the first two modules extensively including algorithms, technical specifications, and GUI design.

## 3.2. Start-Up Window

The start-up window is the starting point of RADAR where every module can be directly accessed. **Figure 3-1** is the screen shot of the start-up window. The field named “RT database” specifies the folder path of the raw RT database. It is mandatory for all the three modules. The “Included patient IDs” field is optional for any module, and it is only useful for data curation. User can choose a “csv” file with the first column listing the included patient ids for data curation. “Previous cleaning status” specifies the folder directory generated by the previous data curation that the user wants to recover. It is an optional input parameter for data curation but mandatory for feature extraction. Finally, “Extracted features” tells the location of the previously extracted features. It is optional

for feature extraction but mandatory for model construction. The independent module initialization empowers flexible data analysis workflows and job managements. For example, it is possible to try different feature extraction parameters without repeatedly entering the data curation module. Data curation can be performed in a regular personal computer due to the less requirement in computational power and constant involvement of human adjustments. On the other hand, feature extraction is more suitable for a high-performance computer, and it can be easily transferred given the same database and cleaning record.

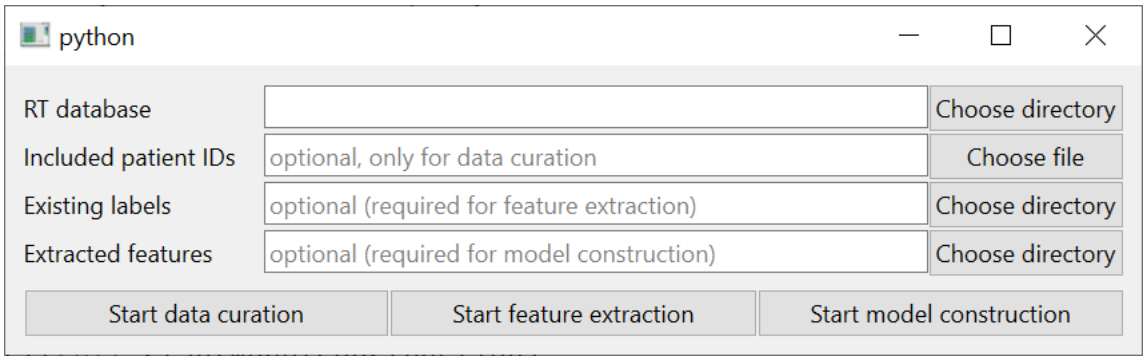


Figure 3-1. Screen-capture of the start-up window.

*It is the starting point of RADAR and can directly access all the data analysis modules including data curation, feature extraction, and model construction.*

## 3.3. Data Curation

### 3.3.1. General Design

The data curation module is a semi-automatic tool for RT data management and labeling. It currently supports image, dose, and structure data in DICOM/DICOM-RT format. The supporting DICOM modalities are CT, MR, and PT for image, DOSE for dose,

RTSTRUCT and SEG for structure delineations, and SEG for registration, as listed in. More modalities and data types will be supported in future developments. The workflow of the data curation module is summarized in **Figure 3-2**. This module starts with extracting or importing DICOM metadata of the entire raw database, depending on whether the local cache has been generated by previous metadata extraction. Registrations between images, dose, and structures are automatically inferred during the metadata extraction and recorded in the cache for first-time analysis. The user sets up the label names and metadata query criteria for the images and/or dose maps with the accompanying masks that are needed for a specific study. The matched data can be automatically screened out by multiple query criteria on DICOM metadata combined by either "and" or "or" logic. Next, the user can manually confirm and change the selected data for each label based on the metadata (listed by "Display metadata" in **Table** 3-1) and image displayed on the user interface. Such labeling strategy emphasizes efficiency through automation while ensuring the accuracy with the flexibility of manual adjustments. The labeling record can be exported anytime during the data curation process. It contains the label query criteria, relative file paths of the selected DICOM data, and other necessary information such as structure index and registration matrix. The labeling record is a snapshot of the data curation status and can be imported to quickly recover previous data curation job. The labeling record is also the final output of the data curation module and the input of the feature extraction module.

Table 3-1. Supported Digital Imaging and Communications in Medicine (DICOM) modalities and the metadata displayed in the user interface.

| **Data Type** | **DICOM Modality** | **Display Metadata** |
|---|---|---|
| Image | CT, MR, PT | Modality, StudyDate, StudyDescription, SeriesInstanceUID, SeriesDescription, SeriesDate, ContrastBolusAgent, NumberOfSlices, DoseSummationType, FrameOfReferenceUID |
| Dose | DOSE | |
| Structure delineation | RTSTRUCT | ROIName, ReferencedSeriesInstanceUID, StructureSetDate, StructureSetLabel, SeriesInstanceUID, FrameOfReferenceUID |

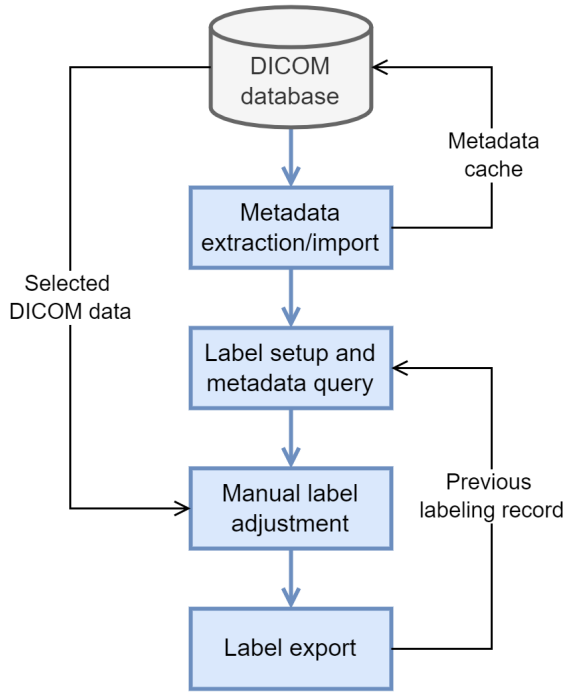


Figure 3-2. General workflow of the data curation module.

Compared to performing data curation patient-by-patient on a general-purpose medical data visualization and analysis tool such as 3D slicer, the data curation module offers its unique advantages in the cohort-based data management. It is designed with a DICOM data import strategy that is fully optimized for large patient cohorts. Only the DICOM metadata is loaded in the initial data import, and the actual data (e.g. image volume) is loaded when requested for visualization during manual label adjustment. Such strategy can save both time and random-access memory space. Compared to converting the labeled data as a clean database, recording the selected DICOM file paths saves the disk space significantly while preventing unwanted data change during conversion. It also allows convenient database relocation as only the relative paths are recorded.

### 3.3.2. Graphical User Interface

The GUI of the data curation module is divided into three sections: query widget, cohort label widget, patient label widget, as shown in screen shot (**Figure 3-3**). The query widget receives commands from the user for adding new image modalities, doses, or structures and modifying the query criteria. It also contains the buttons for operations including loading database, importing previous cleaning status, saving cleaning status, and starting feature extraction. The cohort selection widget summarizes the labeling status of the entire patient cohort by displaying check marks on the labeled patients for each label. It is also where individual patients are selected for patient-level label adjustment and data visualization in patient label widget. The patient label widget contains three areas. The top-left area is a table that displays the metadata of all the image and dose objects. It is also a place where the labeling status, which is indicated by an exclusive check mark, can be manually adjusted by switching it to another data object.

The right area is the place to inspect the metadata and change the labels for structures. The button-left area displays the labeled images, dose maps, and structures. Each labeled image or dose is displayed on the axial, sagittal, and coronal view with all the selected structure masks overlaid. The displaying image or dose can be switched by choosing the desired label on the top bar. Different structures show different colors with an adjustable transparency, and the colors can be reflected in the structure table. User can control when the visualization is updated by pressing the “Update preview” button, as data loading for previous can be time-consuming.

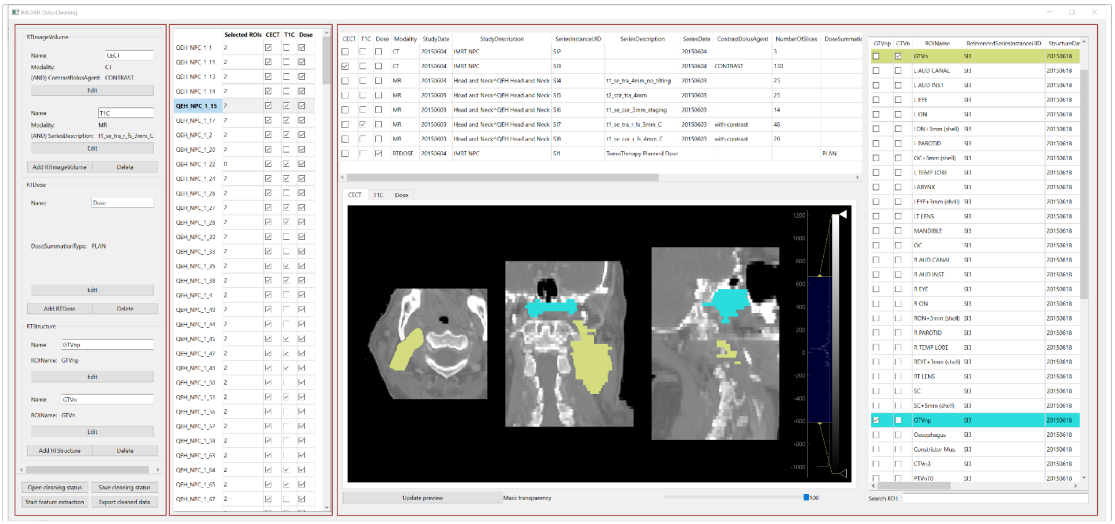


Figure 3-3. The annotated screen caption of the graphical user interface of the data curation module.

*The graphical user interface is divided into three components: query widget, cohort label widget, and patient label widget.*

### 3.3.3. DICOM Volume Interpretation

All the DICOM image, dose, and structure data are interpreted as 3D volumes so that they can share the same processing pipeline. To facilitate further data manipulations such as resampling, they are formatted by SimpleITK (version 2.1.1) Image class[71]. A SimpleITK image is a set of uniform 3D grid points occupying a physical space. We used Python package pydicom (version 2.3) to read individual DICOM files but developed our own algorithms to interpret DICOM data for full transparency.

Image. A complete image series contains multiple DICOM files as different image slices, identified by the DICOM metadata SeriesInstanceUID. By stacking the pixel values of each image slice, a 3D image array can be generated and further processed into a SimpleITK image. Additional metadata including origin, direction, and spacing, which describe the physical space the image grid occupies, are required for the construction of SimpleITK Image. They are converted from the DICOM metadata of ImagePositionPatient, PixelSpacing, and ImageOrientationPatient of the DICOM slices defined on the right-handed left-posterior-head coordinate system (LPH). The conversion of spatial description can be visualized by **Figure 3-4**. ImagePositionPatient records the coordinates (in mm) of the slice origin, PixelSpacing specifies the width and height of each pixel in the image plane, and ImageOrientationPatient contains six values with the first three locating the column unit vector and the last three for the row unit vector. Mathematically, the ImagePositionPatient of slice k is denoted as $\overrightarrow{IPP(k)}$ with the three components along the LPH directions being $IPP_x(k)$, $IPP_y(k)$, and $IPP_z(k)$.

PixelSpacing is denoted as $\overrightarrow{PS} = (PS_i, PS_j)$, assuming uniform planner resolution of each slice. The two vectors from ImageOrientationPatient are denoted as $\overrightarrow{IOP_i} =$

$(IOP_{ix}, IOP_{iy}, IOP_{iz})$ and $\overrightarrow{IOP_J} = (IOP_{jx}, IOP_{jy}, IOP_{jz})$, which are supposed to be constant across slices. The schematic representation of the DICOM image series is drawn in Figure. The slice sequence is determined by ascending slice location $SL(k)$, which is the projection of slice origin along the through-slice direction and can be calculated as $\overrightarrow{IPP(k)} \cdot (\overrightarrow{IOP_\iota} \times \overrightarrow{IOP_J})$. The patient position (x, y, z) of pixel (i, j) at slice k can be calculated as $\overrightarrow{IPP(k)} + PS_i \cdot i \cdot \overrightarrow{IOP_\iota} + PS_j \cdot j \cdot \overrightarrow{IOP_J}$, which can be decomposed into

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} PS_i IOP_{ix} & PS_j IOP_{jx} & 0 & IPP_x(k) \\ PS_i IOP_{iy} & PS_J IOP_{jy} & 0 & IPP_y(k) \\ PS_i IOP_{iz} & PS_j IOP_{jz} & 0 & IPP_z(k) \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \\ 1 \end{bmatrix}. \quad (1)$$

A SimpleITK Image is defined on a grid with homogeneous spacing along image axes. The conversion from image coordinate $(i, j, k)$ to patient coordinate $(x, y, z)$ of a SimpleITK Image can be achieved as the matrix multiplication below

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} S_i D_{ix} & S_j D_{jx} & S_k D_{kx} & O_x \\ S_i D_{iy} & S_j D_{jy} & S_k D_{ky} & O_y \\ S_i D_{iz} & S_j D_{jz} & S_k D_{kz} & O_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \\ 1 \end{bmatrix}, \quad (2)$$

where $\overrightarrow{D_\iota}(D_{ix}, D_{iy}, D_{iz})$, $\overrightarrow{D_J}(D_{jx}, D_{jy}, D_{jz})$, and $\overrightarrow{D_k}(D_{kx}, D_{ky}, D_{kz})$ are the three orthogonal unit vectors defining image direction, $\vec{O}(O_x, O_y, O_z)$ is the image origin, and $S_i$, $S_j$, $S_k$ are the image spacing $\vec{S}$ all in physical space. The following criteria need to be met for a conversion with consistent image geometry:

$$\begin{cases} S_k \overrightarrow{D_k} k + \vec{O} = \overrightarrow{IPP(k)} \\ S_i = PS_i \\ S_j = PS_j \\ \overrightarrow{D_\iota} = \overrightarrow{SOP_\iota} \\ \overrightarrow{D_J} = \overrightarrow{IOP_J} \end{cases}. \quad (3)$$

The first criterion requires a constant ImagePositionPatient difference between adjacent slices

$$\overrightarrow{IPP(k)} - \overrightarrow{IPP(k-1)} = S_k \overrightarrow{D_k}. \tag{4}$$

It can be transformed into constant slice spacing $S_k$ with the assumption of $\overrightarrow{D_k} = \overrightarrow{IOP_i} \times \overrightarrow{IOP_j}$

$$\overrightarrow{IPP(k)} \cdot \overrightarrow{D_k} - \overrightarrow{IPP(k-1)} \cdot \overrightarrow{D_k} = SL(k) - SL(k-1) = S_k, \tag{5}$$

where $SL(k)$ defines the slice location.

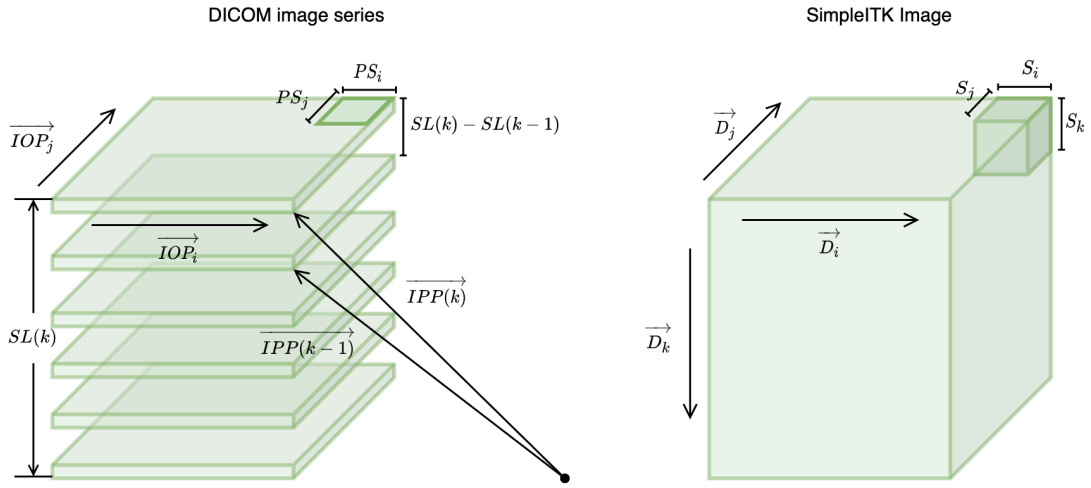


Figure 3-4. The schematic representation of the slice-based Digital Imaging and Communications in Medicine (DICOM) image data structure and the volume-based SimpleITK image data structure.

*The metadata describing the physical space is also marked on both representations to explain the conversion from DICOM image series to SimpleITK image.*

Both slice duplication and uniformity are checked across slices before constructing the 3D image array, as demonstrated in **Figure 3-5**. The image slice is considered as duplicate if the slice spacing $S_k$ is below the tolerance of 1% of the median slice spacing. If the maximum difference of all the slice spacings is within the tolerance of 1% of the median slice spacing, we consider the current image series as uniform, and the final SimpleITK Image can be constructed directly from the original pixel values and the origin, spacing, and direction are calculated based on the above equations. According to the convention, the direction of a SimpleITK image is the inverse of the stacked direction vectors. The final slice spacing $S_k$ is calculated as the average slice spacings to avoid error accumulation, and the image origin $\vec{O}$ is the ImagePositionPatient of the first slice $\overrightarrow{IPP(0)}$. For non-uniform slice spacings, inter-slice interpolation is performed to construct slices with constant spacings. $S_k$ is set as the smallest inter-slice spacing, and the number of slices is the ceiling integer of physical size divided by the final resolution. The original image matrix is interpolated to the non-uniform slice indexes that are linearly mapped from the uniform grid points in physical space. We use the function "interpn" from Python package scipy.interpolation to perform the grid interpolation, and the nearest neighbor interpolation method is chosen to avoid creating any artificial information.

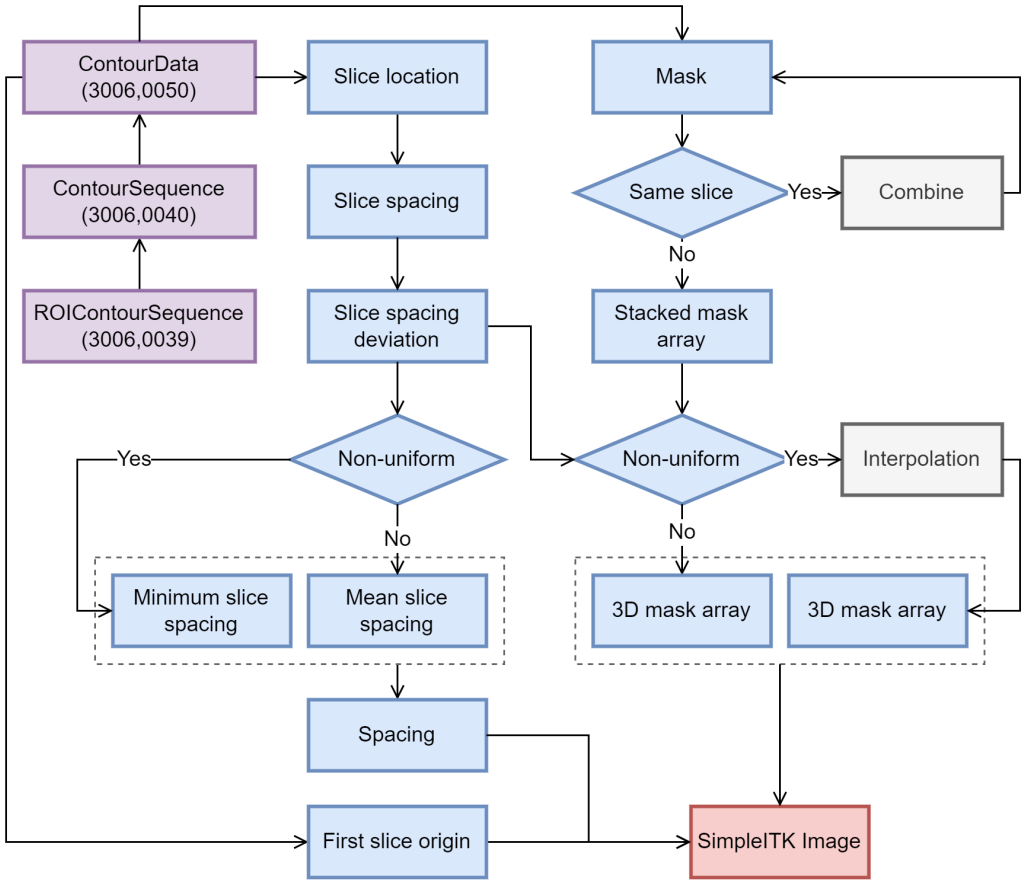


Figure 3-5. Workflow for interpreting DICOM image series and the conversion to a SimpleITK image.

Dose. Unlike DICOM image data, each DICOM DOSE file contains the complete 3D dose array that can be directly used to construct the SimpleITK Image after being rescaled by the "DoseGridScaling". The in-plane resolution is retrieved from "PixelSpacing", and the slice thickness is calculated as the difference of the first two elements of the "GridFrameOffsetVector". Similar to DICOM image data, the "ImagePositionPatient" determines the origin of the dose map SimpleITK Image, and the

direction is calculated as the inverse of the direction vectors acquired from "ImageOrientationPatient". The above conversion is demonstrated by **Figure 3-6**.

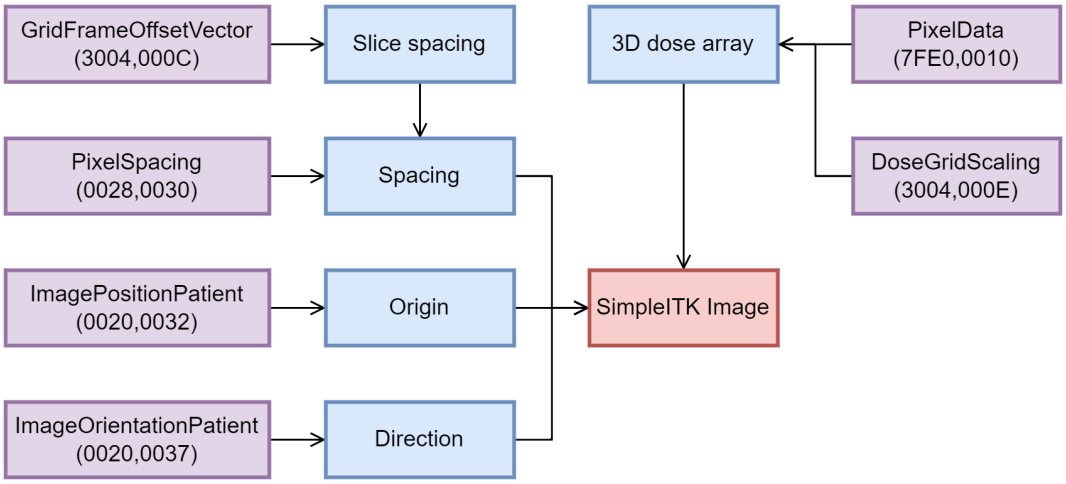


Figure 3-6. Workflow for interpreting DICOM dose and the conversion to a SimpleITK image.

Structure. Multiple DICOM modalities can describe the structure delineation data. DICOM RTSTRUCT is the most commonly used modalities in RT treatment planning. A DICOM RTSTRUCT object contains a collection of structures, each described by the physical space coordinates of the contour points grouped by slice. The contour point coordinates can be retrieved from the "ContourSequence" of each "ROIContourSequence" object. We convert the single-slice contours into a two-dimensional (2D) mask by the function "polygon2mask" provided by the python package scikit-image.draw (version 0.19.2). The resolution of the mask slice is determined by the smallest resolution of the reference image, if provided. Otherwise, the default slice resolution of 0.5mmx0.5mm is used. The mask slices that have the same slice location

are combined into one by the union operation. The mask slices are then stacked into a 3D mask volume according to the rankings of the slice location. The same interpolation strategy is applied as the image data if the slice spacings are not uniform. Finally, the SimpleITK Image objective of the structure mask is constructed from the 3D mask volume using the origin and spacing of the first mask slice and the final slice spacing after slice uniformity check. The direction of the mask is set as the identify matrix. The entire conversion workflow is summarized in **Figure 3-7**.

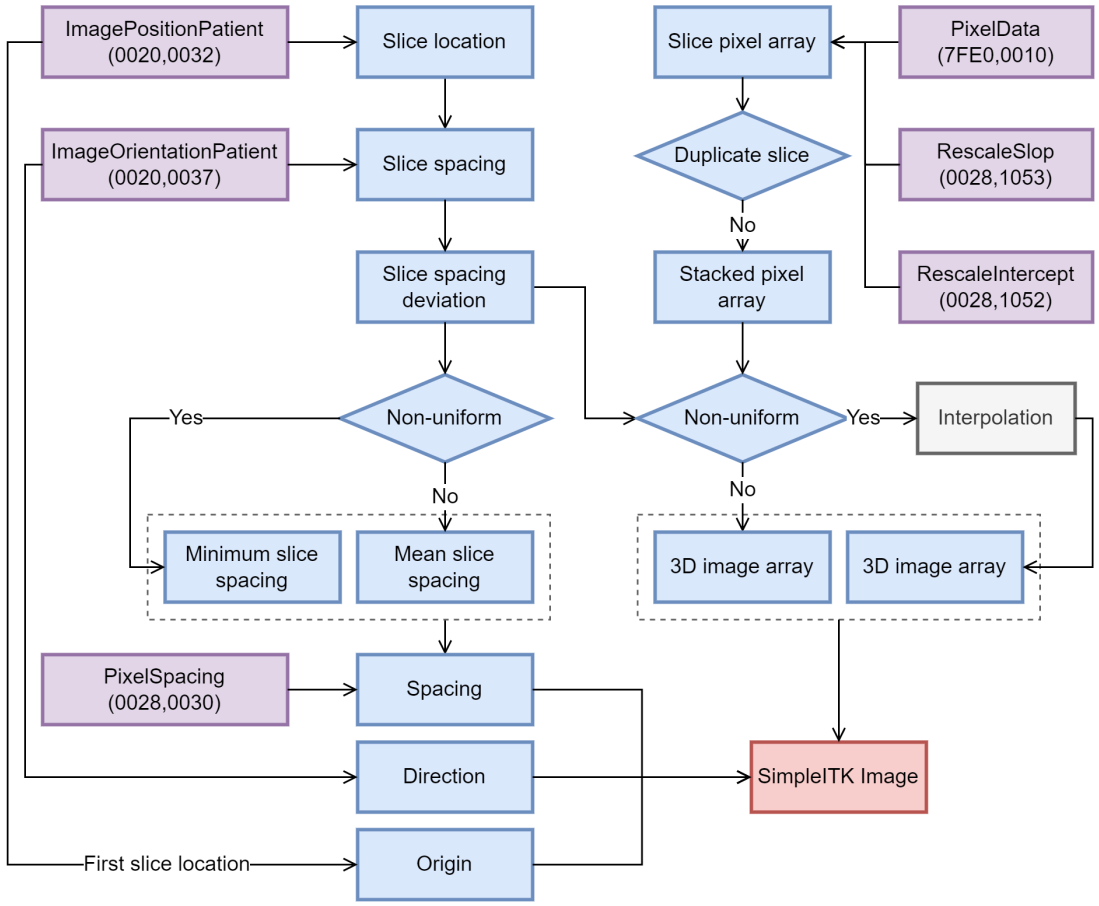


Figure 3-7. Workflow for interpreting DICOM structure and the conversion to a SimpleITK image.

### 3.3.4. Metadata Query and Display

In addition to the volumetric data, other descriptive information is also retrieved from the DICOM metadata for query and display purposes. Most of the display metadata can be directly retrieved from DICOM metadata in the first or deeper level. The StudyDescription, StudyDate, Modality, SeriesDescription, SeriesDate, and SeriesInstanceUID are shared by all the three types of DICOM data. They can also be retrieved from the first level except FrameOfReferenceUID of structure. The ContrastBolusAgent is unique to image and can be retrieved from the first-level DICOM metadata. The NumberOfSlices is inferred from the number of DICOM image files that share the same SeriesInstanceUID and only applicable to image data. For dose data, the only unique display metadata is DoseSummationType, which can be retrieved from the first-level metadata. The ROIName, StructureSetDate, StructureSetLabel, and ReferencedSeriesInstanceUID are unique to structure data, and three of them are retrieved from the DICOM metadata in a deeper level. The ROIName is the second-level metadata from each StructureSetROISequence element. FrameOfReferenceUID is also the second-level metadata from the first element of ReferencedFrameOfReferenceSequence. The ReferencedSeriesInstanceUID is the SeriesInstanceUID acquired following the chain of ReferencedFrameOfReferenceSequence, RTReferencedStudySequence, and RTReferencedSeriesSequence. The acquisition of the display metadata from DICOM metadata can also be explained by. All the metadata with the keyword of “UID” are generally long texts composed by dots and numbers. They are simplified within each patient by re-assigning new IDs with much shorter lengths for the convenience of visual

comparisons. All of the display metadata can be queried based on the data type except for the “UID”s. Numerical metadata, including StudyDate, SeriesDate, NumberOfSlices, StructureSetDate, can be queried based on their maximum or minimum value of each patient. Free text or categorical metadata, including StudyDescription, SeriesDescription, Modality, ContrastBolusAgent, DoseSummationType, ROIName, and StructureSetLabel, are queried by exact match. We assume that most of the patients share the consistent naming convention within one institution, although some of the metadata are based on free text.

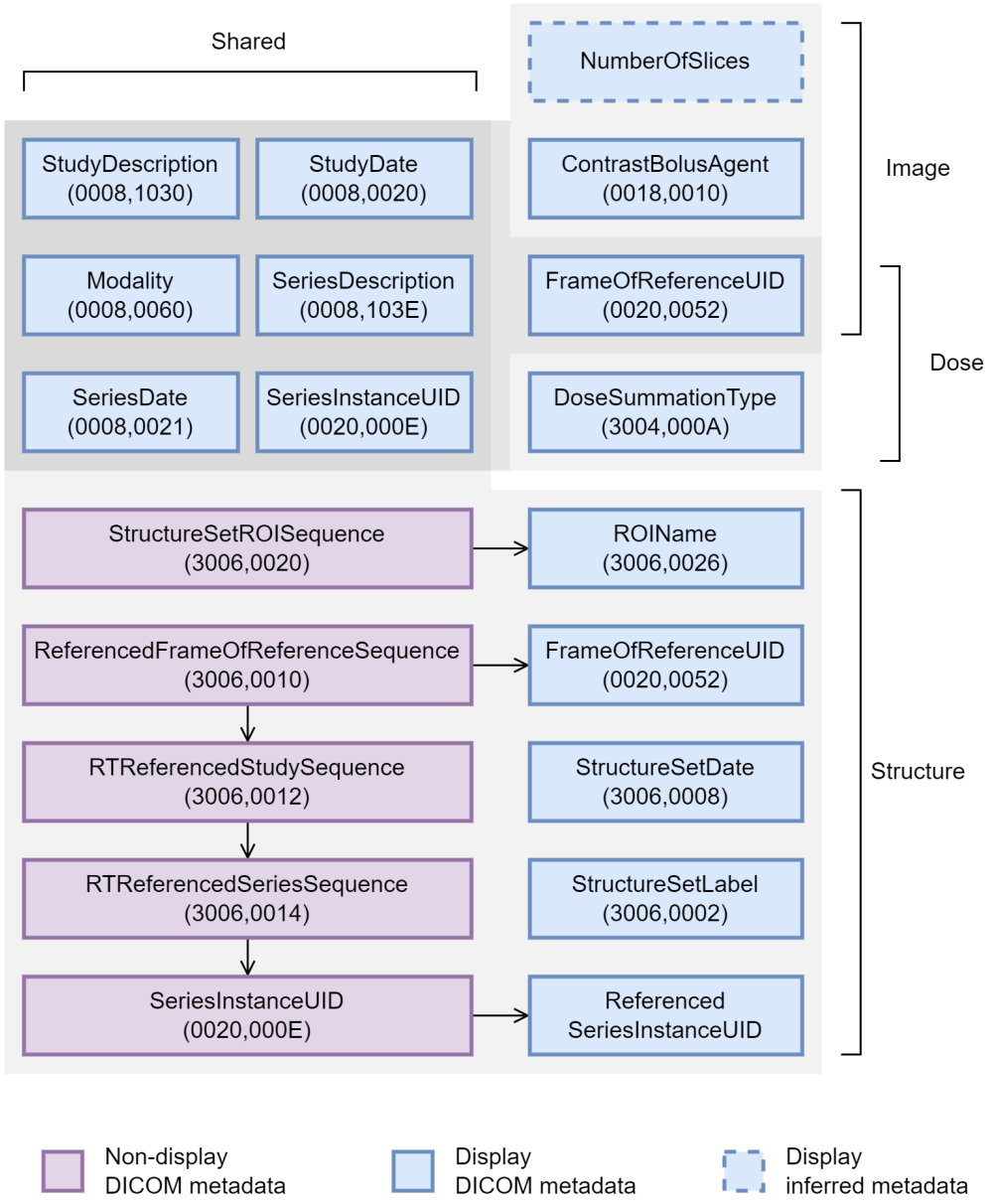


Figure 3-8. The acquisition of the display metadata from DICOM metadata.

*The solid purple rectangles are the intermediate DICOM metadata fields where the display DICOM metadata is acquired. The solid blue rectangles are the DICOM metadata fields that are displayed in the GUI, and the dashed blue rectangle is the inferred information to display in the GUI.*

### 3.3.5. Registration

Registrations are necessary to link different images, dose maps, and structures into one frame of reference (FoR) so that they can share the same anatomy. This module can recognize clusters of FoRs that are directly or indirectly registered and automatically apply the registrations during data loading. FoR is identified by the FrameOfReferenceUID metadata.

First, a bidirectional registration dictionary is constructed by analyzing the direct and indirect connections between FoRs from all the DICOM REG files belonging to one patient. For each DICOM REG file, the FrameOfReferenceUID and MatrixRegistrationSequence are retrieved from each element of RegistrationSequence. All the MatrixSequence in each MatrixRegistrationSequence are combined by the dot product as the registration matrix from the current FoR to the common FoR of the REG object. The final registration matrix connecting every moving and fixed FoR pair within the REG objective is calculated by the dot product of the corresponding registration matrixes. The registration dictionary is thus constructed and expanded as more REG files are found within one patient folder. The dictionary also contains reversed registrations with inverted registration matrixes. Secondary registrations are also evaluated by finding common FoRs in the dictionary.

Next, registrations are clustered with one fixed FoRs surrounded by the moving FoRs. Every unique FoR in the registration dictionary is ranked based the referred DICOM modality and its frequency. “CT” is the first-ranked DICOM modality, followed by “RTSTRUCT”, “MR”, and others. Iterative search of the moving FoRs is performed with the top-ranked FoR as fixed in the registration dictionary until all the FoRs are

exhausted. The FoR is removed from the search pool if it is selected as either the fixed or moving FoR previously.

During metadata analysis, all the moving FoRs are converted to the corresponding fixed FoRs. Only the fixed FoRs are displayed for the FrameOfReferenceUID, and the final registration matrix is recorded and applied during data loading by changing the image origin and direction. This pre-analysis strategy in registration releases the burden of manual registration file selection during data curation, which is another effort in improving efficiency. However, having the same fixed FoR does not guarantee a perfect registration. It is the user's responsibility to ensure the quality of registration by visually inspecting the image/structure in the user interface.

## 3.4. Feature Extraction

### 3.4.1. General Design

The feature extraction module offers high-performance high-throughput feature extraction from the labeled RT data. It can be initiated from the data curation module or directly from the start-up window. It accepts the labeling record generated from the data curation module and outputs the extracted feature values as csv files. The output feature value files can be directly used by the model construction model and other data analysis software, depending on user's preferred workflow.

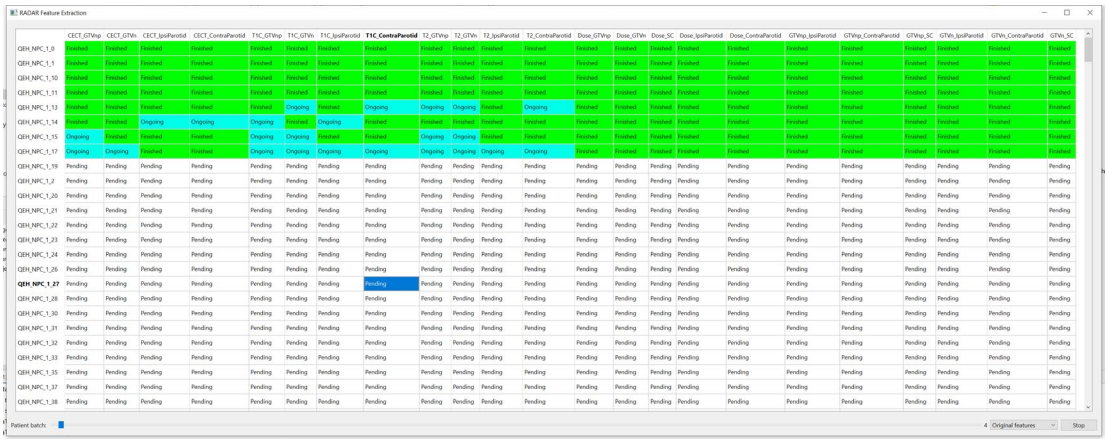

Figure 3-9. Screen capture of the feature extraction module.

Table 3-2. Text and color of the feature extraction status.

| **Status Text** | **Color Name** | **Color Code** |
|---|---|---|
| Pending | White | #fffff |
| Missing | Yellow | #ffff00 |
| Ongoing | Light blue | #00ffea |
| Finished | Green | #00ff00 |
| Failed | Red | #ff0000 |

### 3.4.2. Graphical User Interface

The GUI screen capture of this module during one feature extraction job is shown in **Figure 3-9**. Users can switch between “Original feature” and “Perturbation feature” in

the drop-down combobox to choose whether the features from the original data or ones from the perturbed data are to be extracted. They can also adjust the patient batch number in the slider bar to control how many patients are to be paralleled for each feature extraction iteration. The feature extraction job is initiated by clicking the "start" button. To facilitate monitoring of the feature extraction progress, the main component of the GUI is a table displaying the status of the feature extraction job for each patient and feature category. The text and color of each feature extraction status are listed in **Table 3-2**.

### 3.4.3. Feature Definitions

Volumetric features. A systematic and comprehensive set of features are defined for the multiple types of RT data. Image and dose data share the same feature definitions, but different features were designed for structures due to the uniqueness of the structure data. Both image and dose data present as cubical 3D volumes with continuously varying voxel values. We adopted the conventional RF definitions following the standardization proposed by the IBSI for image and dose data. In addition, spatial-invariant 3D moment and gradients features, which describe the spatial deposition of the volumetric data, are also included. Spatial-invariant 3D moments were first adopted in dosimetric predictions by Buettner et al.[54]. They validated the superior performance of dose moments on in xerostomia prediction than the standard mean-dose model for head-and-cancer cases. Dose gradients were included in the dosimetric feature set proposed by Gabrys et al. in their study on xerostomia predictions[56]. Besides, histogram features are included by defining the point values on the original or cumulative histogram curve.

1. Scale-invariant 3D moments quantify the ordered center-of-mass of the 3D volume within the ROI on the anterior-posterior, medial-lateral, and craniocaudal directions. The translation-invariant 3D moments are defined as

$$\mu_{pqr} = \sum_x \sum_y \sum_z (x - \bar{x})^p (y - \bar{y})^q (z - \bar{z})^r D(x, y, z) I(x, y, z), \quad (6)$$

where $D(x, y, z)$ is the voxel value at location $(x, y, z)$, and $I(x, y, z)$ is the binary indicator of whether the current voxel is in the ROI. $p$, $q$, $r$ are the user defined orders for each dimension. The scale invariance is achieved by normalization:

$$\eta_{pqr} = \frac{\mu_{pqr}}{\mu_{000}^{\frac{p+q+r}{3}+1}}. \quad (7)$$

Four example dose distributions showing contrasting Scale-invariant 3D dose moments are presented in **Figure 3-10**. Two examples are presented by three axial slices in **Figure 3-10(a)**, and the rest two cases are represented by three coronal slices in **Figure 3-10(b)**. A high accumulation of dose on the upper side of the x-y direction can be seen for the high $\eta_{110}$ case, but the distribution is more even for the low $\eta_{110}$ case, although a tendency towards the (-x)-y direction is observed. Similarly, the high $\eta_{003}$ case shows a higher dose accumulation towards the positive direction of the z axis compared to the low $\eta_{003}$ case.

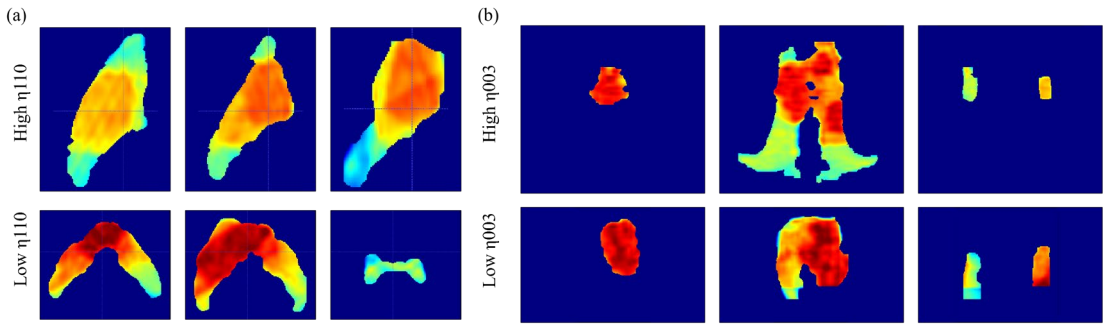


Figure 3-10. Primary planning target volume dose distributions with four representative scale-invariant three-dimensional (3D) dose moments.

*Primary planning target volume dose distributions with four representative scale-invariant 3D dose moments, each presented by three slices. (a) Qualitative comparison of dose distributions between high and low dose moments with the order of 1, 1, 0 on the x, y, and z dimension. Each slice is taken on the x-y plane with the horizontal line as the x-axis. (b) Qualitative comparison of dose distributions between high and low dose moments with the order of 0, 0, 3 on the x, y, and z dimension. Each slice is taken on the x-z plane with the vertical line as the z-axis.*

2. Gradient features are the gradient values of the 3D volume along the anterior-posterior, medial-lateral, and craniocaudal directions, which correspond to the x, y, and z directions on a SimpleITK image. The mathematical definition of the gradient feature for x axis is

$$Gradient_x = \frac{\sum_{x,y,z} D(x+1,y,z)I(x+1,y,z) - D(x-1,y,z)I(x-1,y,z)}{2\sum_{x,y,z} I(x,y,z)}. \quad (8)$$

3. Histogram features record the voxel accumulation or distribution depending on whether the histogram curves are cumulative or not. The histogram features can be defined on both axes of the curves.

Geometric features. RADAR contains a new set of geometric features based on the geometric relationships between structure contours, which describes patient anatomy in a

quantitative manner. Distance and angle features are currently proposed for locating a target relative to the surrounding structures. The geometric relationships between the target volume and the OARs are primarily used in knowledge base treatment planning to predict dose-volume histogram or even 3D dose distributions. Overlap volume histogram (OVH) was first proposed by Kazhdan et al. for quantifying patient geometries[72] and successfully applied by Wu et al. to predict the optimal dose-volume histogram for knowledge-based treatment planning[73]. It summarizes the distances between organ-at-risk (OAR) and the target volume (TV) by recording the fractional OAR volume as a function of the maximum distance from the target surface:

$$OVH(d) = \frac{count_i\left(r\left(v_{OAR}^i, S_{TV}\right)\right)}{V_{OAR}}, \tag{9}$$

where $r\left(v_{OAR}^i, S_{TV}\right)$ is the surface distance defined as the minimum distance from OAR voxel $v_{OAR}^i$ to all the target surface points $v_{TV}^k$:

$$r\left(v_{OAR}^i, S_{TV}\right) = \min_k\left\{\left\|v_{OAR}^i - v_{TV}^k\right\| \middle| v_{TV}^k \in S_{TV}\right\}. \tag{10}$$

The surface distance is positive for an OAR voxel outside TV surface and negative when inside. We used the signed Euclidean distance transform algorithm[74] provided by the Python package SimpleITK (version 2.1.1)[71] to calculate the surface distance map and acquired the OVH as the cumulative histogram within the OAR mask. An example distance map for a lymph node gross tumor volume (GTVn) is visualized by the heat map in **Figure 3-11** where the left parotid (Parotid_L) is drawn as a red contour.

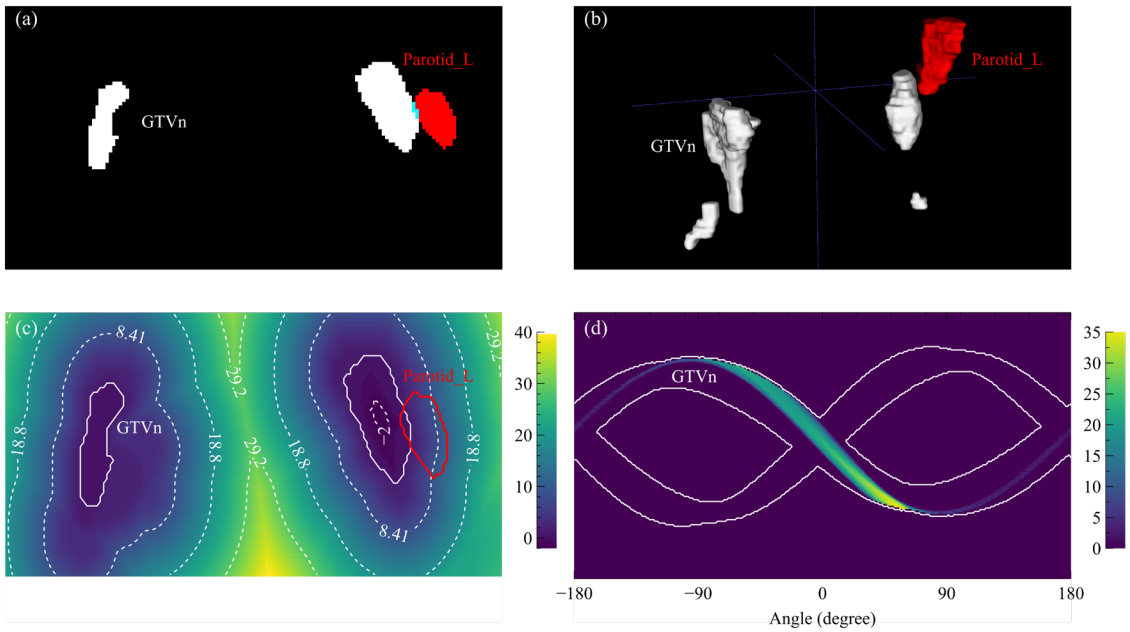


Figure 3-11. Demonstration of distance and angle maps.

*Distance and angle maps based on example lymph node gross tumor volume (GTVn) and left parotid (Parotid_L) structures. (a) One axial slice of the structure masks (white: GTVn, red: Parotid_L) with the overlap region highlighted by blue. (b) The rendered three-dimensional structures. (c) One axial slice of the GTVn distance map with annotated contour lines and the Parotid_L contour. (d) One slice of the Parotid_L angle map masked by the GTVn sinogram edges (white contours).*

Spatial configuration of the TV may not be precisely determined by distance alone due to the potential complex organ structures, especially in the head-and-neck region. We designed the projection overlap volume (POV) histogram to quantify the angular relationships between TV and the surrounding OARs. POV is defined as the relative OAR volume that overlaps with the parallel projection of TV:

$$POV(\alpha) = \frac{\sum_i \chi_{\alpha i}}{V}, \chi_{\alpha i} = f(x) = \begin{cases} 1, & if \ \min_j \theta_{ij} < \alpha < \max_j \theta_{ij} \\ 0, & otherwise \end{cases}, \quad (11)$$

where $V$ is the voxel volume of the OAR and $\theta_{ij}$ is the angle from TV surface point $v_j$ to OAR voxel point $v_i$ on the axial plane. POV histogram is calculated by summing up the masked OAR sinogram along the angle direction. The masked OAR sinogram is the modified radon transform of the OAR mask volume around the axial axis; only the voxels located before TV are counted for each OAR mask volume projection. One Parotid_L masked sinogram is shown in **Figure 3-11(d)**. Customized histogram features can be extracted from OVH and POV curves as geometric features, as well as other dimension reduction methods such as principal component analysis (PCA).

### 3.4.4. Perturbation Feature Extraction

Features can also be extracted under RT data perturbation for feature reproducibility assessment. Currently only translation, rotation, and contour randomization are implemented in the software. More perturbation modes such as noise addition are still under development. Translation displaces the image, dose, and structure mask by a distance within the voxel size, and rotation is performed around the axial axis located at the center of the region-of-interest (ROI) bounding box. The same translation and(or) rotation is performed on the accompanying ROI to ensure the consistent registrations. During the implementation of translation and rotation perturbation, images/dose maps and the ROI masks are resampled with transformation by the resample filter offered by the SimpleITK package. The combination of translation and rotation can simulate the patient position variations during scanning setup. **Figure 3-12** demonstrates three translated and rotated CT images of one patient in three different views.

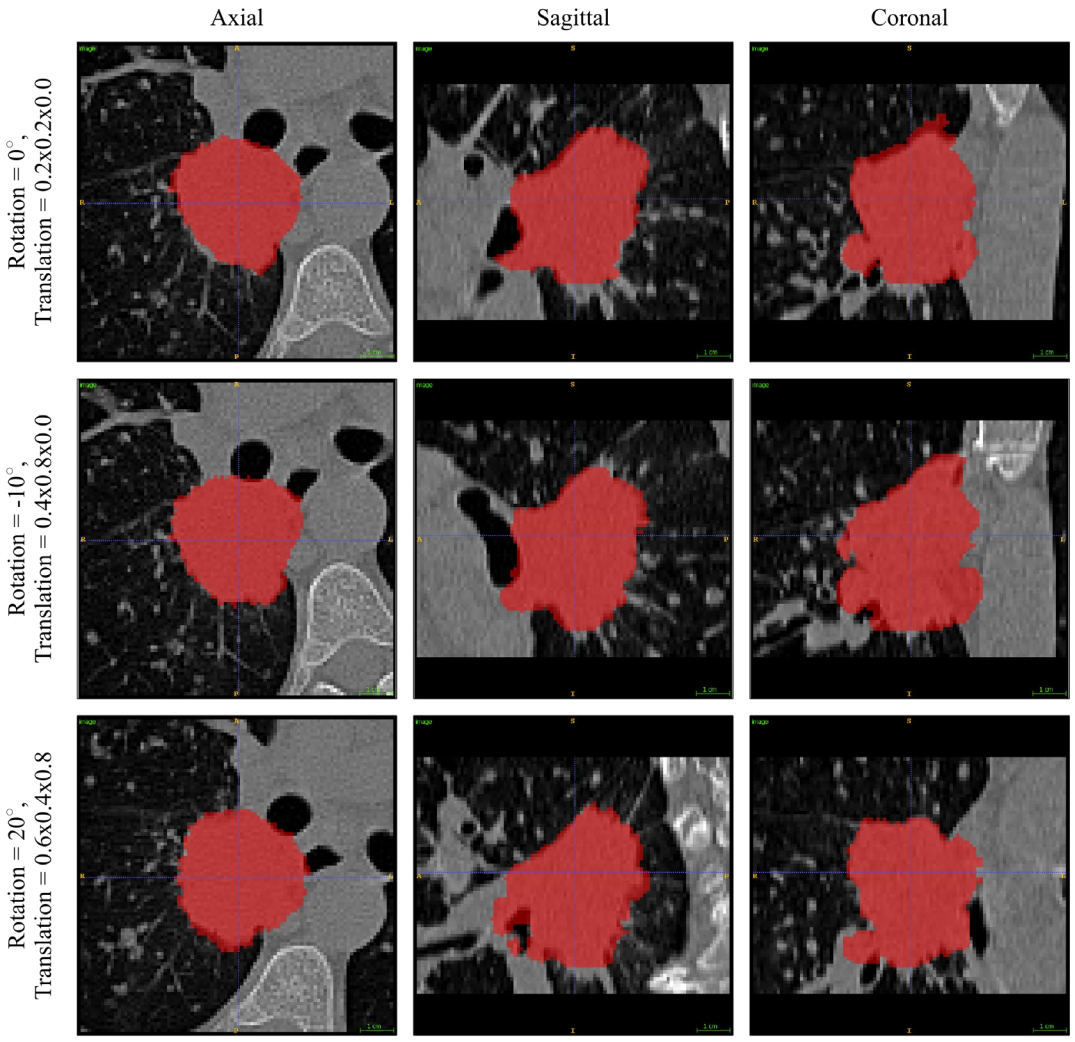


Figure 3-12. Three translated and rotated images and masks of a lung computed tomography (CT).

*Three translated and rotated images with the nodule masked by red of one lung CT image. Every row shows one combination of translation and rotation perturbation, and the three columns are the three views of the 3D image volume.*

Contour randomization simulates multiple delineations of the same structure. A 3D random displacement field deforms the segmented mask and results in a randomized contour. The algorithm of random displacement field generation is adapted from the

methodology proposed by Simard et al.[75]. A random field vector component on each dimension is generated randomly under a uniform distribution between -1 and 1 for each voxel point. All the z-component of the field vectors on the same slice are kept to the same value to mimic the uniform inter-slice contour variations from the slice-by-slice contouring. The field vectors are then normalized on each dimension by the root mean square and scaled by the user-defined intensity value. They are then smoothed by a gaussian filter with user-defined sigma to ensure the continuous change of the random displacement field and avoid sharp changes of the deformed contours. **Figure 3-13(a)** shows one example of random displacement field, 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 are superimposed in **Figure 3-13(b)**, showing similar variations to the actual repeated manual segmentations by five different operators in **Figure 3-13(c)**. The similarity between the original and randomized contour is evaluated by two metrics: Dice similarity coefficient (DSC) and symmetric Hausdorff distance (HD). DSC is defined as the ratio between the union volume and average total volume of the two segmentations $V_\alpha$ and $V_\beta$:

$$DSC = \frac{2|V_\alpha \cap V_\beta|}{|V_\alpha| + |V_\beta|}, \tag{12}$$

where $|V_\alpha|$ denotes the voxel number within the segmentation $V_\alpha$. The calculation of DSC is implemented by LabelOverlapMeasuresImageFilter.GetDiceCoefficient in SimpleITK package. HD is defined as the longest Euclidean distance from every point on one contour ($v_\alpha^i$) to the other contour ($v_\beta$), as formulated in equation below, and the symmetric HD finds the longest distance in both directions:

$$HD = \max_i \left\{ \min_j \left\{ v_\alpha^i - v_\beta^j \right\} \right\}. \quad (13)$$

The calculation of HS is implemented by HausdorffDistanceImageFilter.GetHausdorffDistance in SimpleITK package. Two contour randomization parameters – Intensity and smoothing sigma – can be determined from clinical experience or tuned based on the resulting randomized contour similarities.

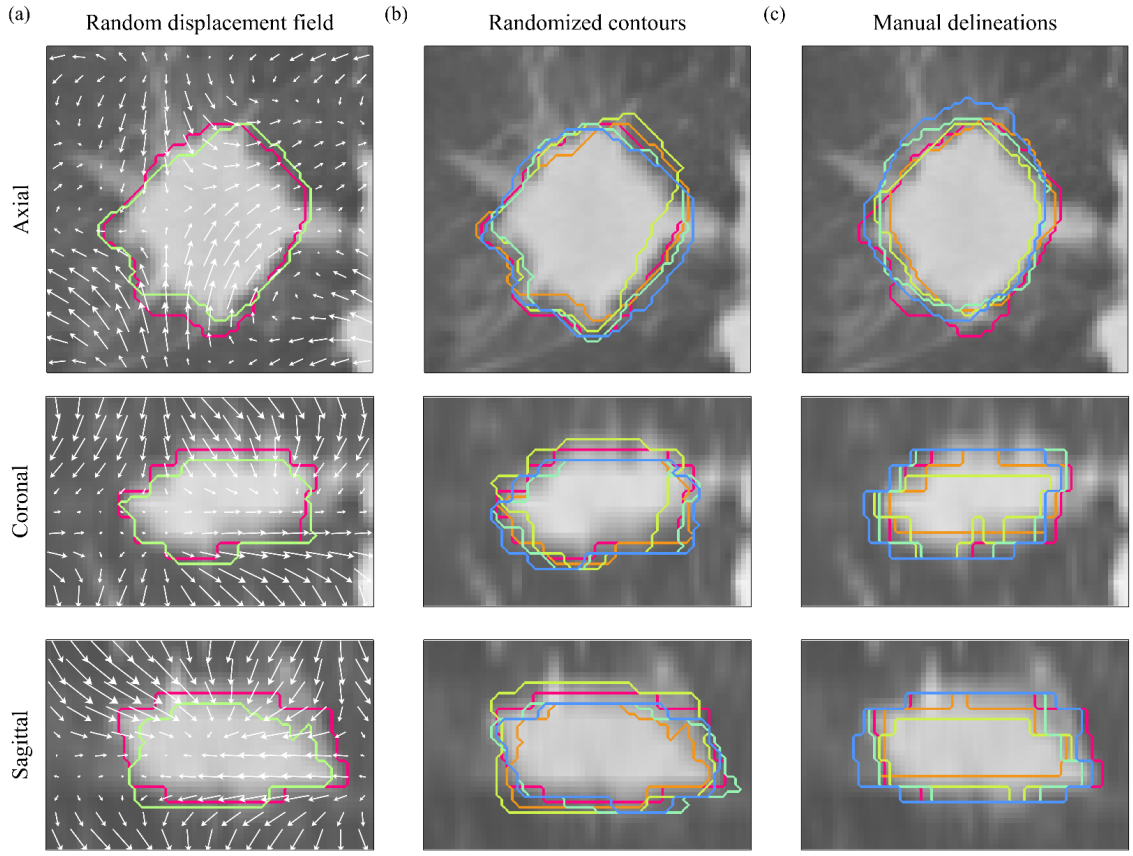


Figure 3-13. Demonstration of the contour randomization method.

*Demonstration of the contour randomization method by one example lung CT image and the nodule contour. (a) The original contour (red) is randomized by the random deformation field (white arrows) into the new contour (green). (b) The original contour (red) and four randomized contours in different colors overlaid. (c) Five manual delineations by different operators.*

## 3.5. Discussion

This chapter presents the design and technical innovations of the RADAR toolkit, with an emphasis on efficiency, flexibility, and usability. The data curation module can reliably parse and reconstruct DICOM files, and facilitate efficient and accurate data selection through data query, image/structure visualization, and manual selection. The feature extraction module is equipped with advanced and innovative feature extraction algorithms tailored for different types of radiotherapy data, as well as image perturbations for feature repeatability analysis. All calculations are accelerated by parallel computation to fully exploit the advantages of multi-core CPUs.

This thesis has not provided detailed information on two modules that are still under development, which are model development and model deployment. The model development module accepts the feature table generated by the feature extraction module and outputs the performance of model candidates, ultimately identifying the best performing model as the final model. Users can build their own model development pipelines by combining different steps of feature selection and classification/regression models. The model deployment module is an independent component for external model sharing and clinical translation. It includes an intuitive user interface that facilitates data selection through data visualization. Most importantly, it can reproduce the entire data preparation, feature extraction, and model predictions specified by the final model.

RADAR offers several key features that greatly benefit model development and clinical translation. One such feature is its support for multi-modal feature extraction. This can be performed simultaneously for different types of radiotherapy data, including images, doses, and structures, within the feature extraction module. The extracted multi-

modal features are then combined into a single feature table and sent to the model development module for building. Users can also choose to build their own models using third-party software tools. The final model contains all the necessary specifications for data preparation, feature extraction, and model prediction. It can be exported for sharing and external deployment, facilitated by the model deployment module. While a typical model development cycle, starting from raw radiotherapy data, can take weeks to complete, the model prediction can be achieved within 5 minutes using the model deployment module. This allows for fast clinical decision-making and minimal waiting time for patients.

# Chapter 4.

# Radiomic Feature Repeatability Under Perturbation

## 4.1. Introduction

Radiomics has been reported to be successful in predicting numerous clinical endpoints through statistical modeling. However, the clinical applicability of these radiomic models has largely been impeded by the lack of studies assessing the RF repeatability in their models[76–80]. As highlighted in several excellent review articles, repeatability and reproducibility of RFs are crucial for reaching reliable and consistent conclusions between studies[35,81,82]. In particular, high repeatability, referring to RFs that remain stable when imaged multiple times if the conditions keep unchanged[33,78], is the first and foremost criterion towards clinical utility. Features with poor repeatability against random changes during the same-condition imaging, where the clinical implication is ensured to be the same, could cause significant uncertainties in the downstream radiomics models, degrading model generalizability in both the same- and multi-institutional settings. As such, RF repeatability should be incorporated into feature pre-selection strategy and downstream predictive model construction in any radiomic studies. For example, excluding low-repeatability RFs by a certain threshold before any other feature selection procedures can ensure only high repeatable RFs are used during model construction.

This chapter includes two technical studies focusing on the pattern of radiomic feature repeatability and its impact on model's external generalizability under simulated patient position stochasticity. Specifically, the first study attempted to identify high or

low repeatable RFs that are generalizable across different cancer subtypes of HNC. Such information will provide the radiomics community with direct perceptivity for selecting reliable radiomic features and building robust predictive models for implementing precision medicine. The second study focused on the investigation of RF repeatability of NPC patients and compared the external generalizability of survival models built using high-repeatable and low-repeatable features. Data curation, feature extraction, and image perturbations in both studies were performed by the RADAR software.

## 4.2. Repeatability of Radiomic Features against Simulated Scanning Position Stochasticity across Imaging Modalities and Cancer Subtypes

### 4.2.1. Introduction

Test-retest imaging is one of the widely applied methods for effective radiomics feature repeatability assessment which underlines the pronounced impacts of scanning position variations on RFs repeatability. Notwithstanding, there are noteworthy shortcomings. First, the impact of segmentation variations on RFs repeatability is often inherent in a test-retest study, where segmentations of region-of-interest are separately delineated on test and retest images, which hinders direct interpretations of the influences on RFs repeatability caused purely by positional discrepancies. Secondly, the prolonged time-interval between test and retest images, in the case of 2-week apart, might disregard the implications of intra-tumoral microbiologic changes during that period of time, which itself might lead to dramatic disparity in RFs between the two scans. Thirdly, the limited

sample size owing to the need for recruiting consented patients renders their conclusions less statistically convincible. Lastly, other test-retest studies looked into the robustness of RFs under a mixture of variables (e.g., image preprocessing steps, scanning protocols, segmentation methods, etc.) [76–80], which, however, pose significant challenges in identifying the culprit of the declined repeatability of particular RFs.

To address these limitations, we attempted to deploy our in-house developed image perturbation framework, taking reference from a previous work by Zwanenburg et al. [37], to mimic a vast amount of scanning position stochasticity via large patient cohorts of NPC and oropharyngeal carcinoma (OPC) . To our best knowledge, the RF repeatability against scanning position variations in HNC is yet to be explored, and there are no relevant publications with multiple imaging modalities. The main objectives of this study were:

1. To ascertain the repeatability of RFs against scanning position stochasticity via image perturbations in both cohorts;
2. To examine their generalizability across CT and MR imaging modalities among NPC patients;
3. To assess their generalizability among HNC subtypes via a publicly available OPC dataset.

### 4.2.2. Methods and Materials

**Figure 4-1** illustrates the overall study workflow. Two HNC cohorts were enrolled in this study: an internal NPC cohort of 250 patients collected from Queen Elizabeth Hospital (QEH) and a publicly available OPC cohort of 492 patients. The NPC cohort consists of three image modalities, which are contrast-enhanced CT (CECT), contrast-enhanced T1

weighted (CET1-w) MR, and T2 weighted (T2-w) MR. Only CECT images were studied for the OPC cohort. Details of data acquisition can be found in **Appendix A**. Each image set was processed through preprocessing, rotation and translation perturbations, and RF extraction before evaluating RF repeatability. By comparing the RF repeatability between each pair of the three imaging modalities in the NPC cohort, we examined repeatability generalizability across NPC imaging modalities. The comparison was also made between the CECT images of NPC and OPC cohorts to evaluate the generalizability across head-and-neck cancer subtypes. Finally, multiple validation experiments were conducted to evaluate bias from feature collinearity, ROI contouring, and ROI volume.

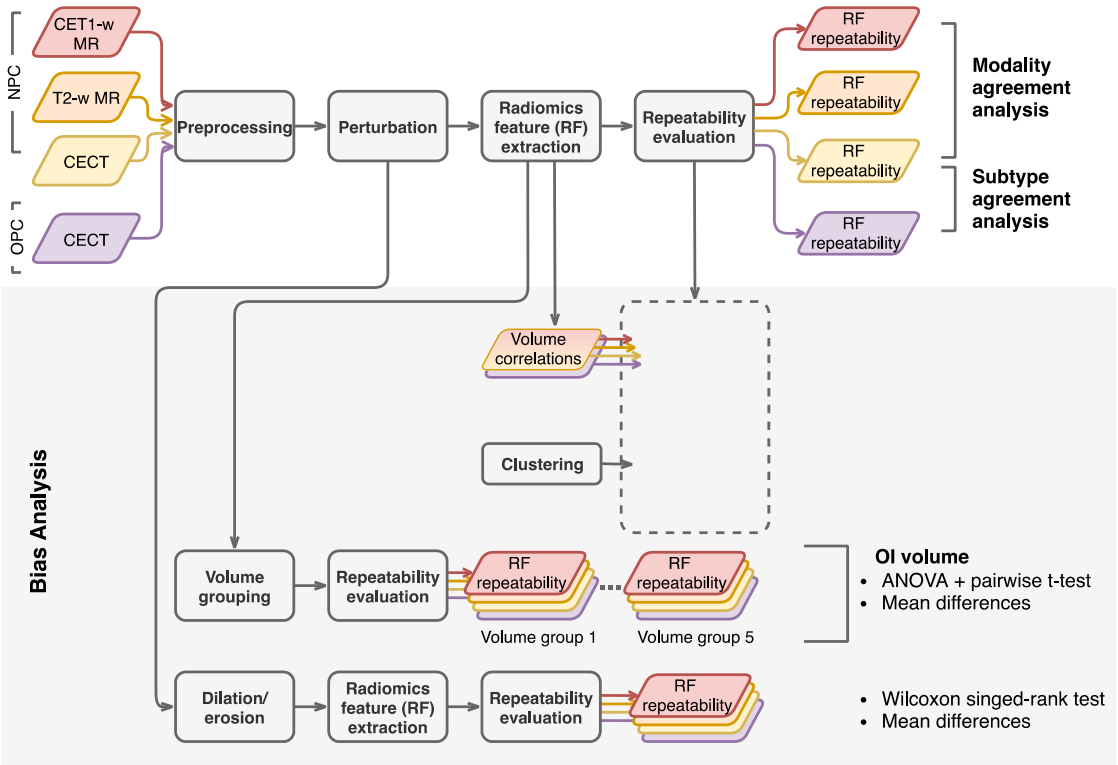


Figure 4-1. Overall study workflow.

Patient cohorts. A total of 250 biopsy-proven (Stage I-IVB) NPC patients who received cancer treatment at the Queen Elizabeth Hospital between 2012 and 2016 were retrospectively screened, and 231 patients that had same-institution MR images and eligible target contours were enrolled in the study.

CECT images of 492 (Stage I-IV) OPC patients treated between 2005 and 2012 were downloaded from The Cancer Imaging Archive[83–85], and 399 patients who have eligible target contours were enrolled in this study. The cancer subsites of origin include base of tongue (n=255), tonsil (n=194), NOS (n=24), glossopharyngeal sulcus (n=11) and soft palate (n=8).

Image acquisition & volume of interest segmentation. All imaging data were acquired in a DICOM format archived using Picture Archiving and Communication System (PACs). In the internal NPC cohort, each primary gross-tumor-volume (GTVp) of NPC was manually delineated on axial CT slices co-registered with MR images by oncologists specialized in head-and-neck cancer with accreditations. In the external OPC cohort, expert radiation oncologists manually segmented GTVp. Details of the image acquisition and contouring protocols can be found in **Appendix A**.

Image preprocessing, perturbation, and feature extraction. All the calculations in image preprocessing, perturbation, and feature extraction were performed by our in-house developed Python-based (3.7.3) pipeline using the SimpleITK (1.2.4) [86] and PyRadiomics (2.2.0) package [30]. All the image processing parameters were listed in **Table 4-1**. Before image perturbation, the signal intensities of MR images were normalized using brainstem as the reference structure with a rescaling factor of 100, and N4B bias correction from SimpleITK was employed for inhomogeneity correction.

Table 4-1. Image processing parameters

| | **CECT** | **CET1-w MR and T2-w MR** |
|---|---|---|
| N4B bias correction maximum iterations | N/A | [50, 40, 30] |
| Normalization reference structure | N/A | Brainstem |
| Normalization rescale factor | N/A | 25 |
| Pixel value offset | 2000 | 2000 |
| Resample pixel size (mm) | [1,1,1] | [1,1,1] |
| Anti-aliasing low-pass filter | Gaussian, $\beta = 0.97$ | Gaussian, $\beta = 0.97$ |
| Image/mask interpolation algorithm | Trilinear | Trilinear |
| CT image intensity rounding | No | N/A |
| Mask partial volume threshold | 0.5 | 0.5 |
| Interpolation grid alignment | Align grid origins | Align grid origins |
| Image thresholding | $\pm 3\sigma$ | $\pm 3\sigma$ |

| | | |
|---|---|---|
| Translation distances (pixel) | [ 0.0, 0.2, 0.4, 0.6, 0.8] | [ 0.0, 0.2, 0.4, 0.6, 0.8] |
| Rotation angles (degree) | [-20,-15,-10,-5,0,5,10,15,20] | [-20,-15,-10,-5,0,5,10,15,20] |
| Image discretization bin size | 10 | 10 |
| Image filters | Unfiltered, Laplacian-of-Gaussian, Wavelet | Unfiltered, Laplacian-of-Gaussian, Wavelet |
| Kernel size of Laplacian-of-Gaussian filter (mm) | [1,2,3,4,6] | [1,2,3,4,6] |
| Wavelet filter starting level | 0 | 0 |
| Wavelet filter total level | 1 | 1 |
| Wavelet filter type | Coilf1 | Coilf1 |
| Wavelet filter decompositions | [LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH] | [LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH] |
| Feature class | shape, firstorder, glcm, glrlm, glszm, gldm, ngtdm | shape, firstorder, glcm, glrlm, glszm, gldm, ngtdm |

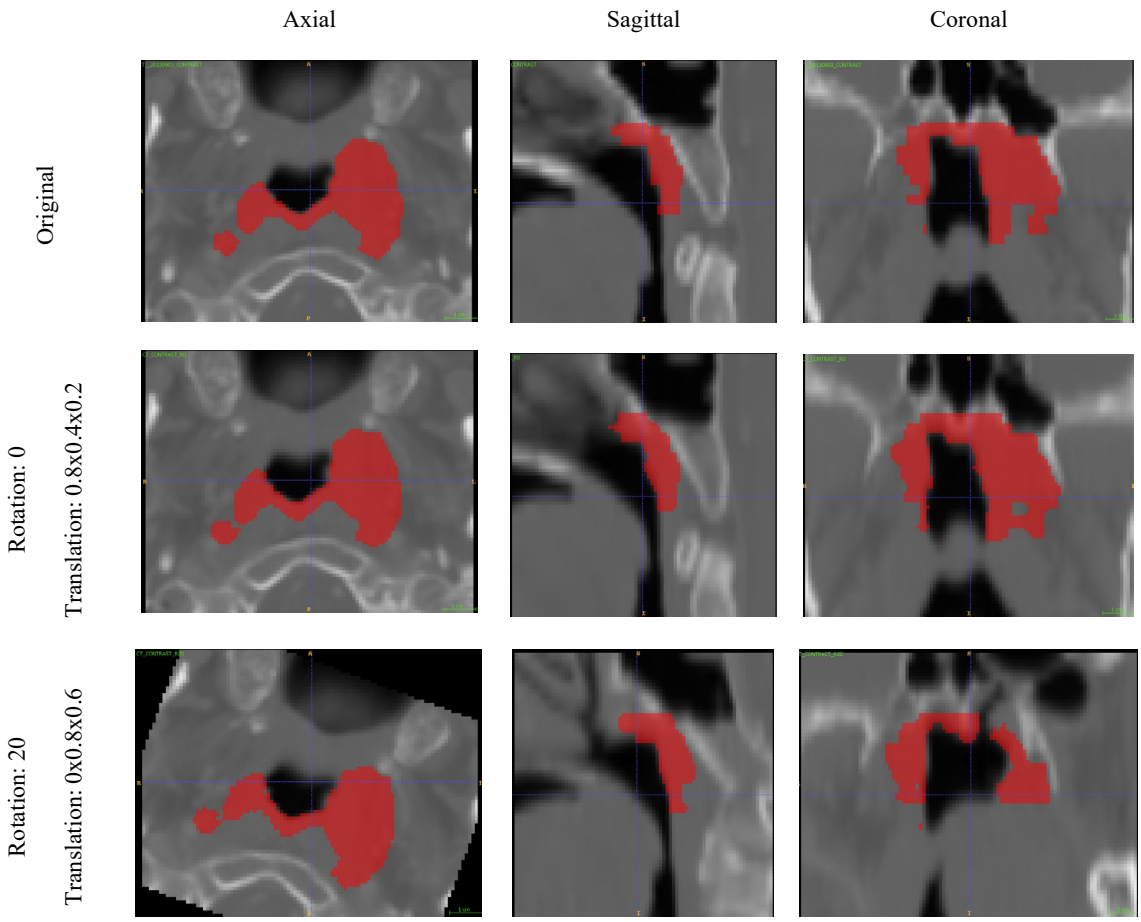


Figure 4-2: Demonstration of translation and rotation perturbation using one sample patient.

*Images are shown in axial, sagittal, and coronal views. The regions-of-interest are drawn by red masks.*

Image perturbations were applied to each pair of the preprocessed original-resolution image and ROI mask during isotropic (1mm x 1mm x 1 mm) resampling after Gaussian anti-aliasing filtering. Two perturbation modes, rotation ($\theta \in [-20^{\circ}, 20^{\circ}]$, step size = 5, around central z-axis) and translation ($\mu \in [0.00, 0.80]$, step size = 0.2, along all three dimensions), were implemented following the procedures proposed by Zwanenburg et al.[37] to mimic variations in scanning setup positions during image acquisition. All the

resampling procedures were performed using SimpleITK. In this study, 40 perturbation parameter sets (θ and μ) were randomly chosen without replacement from the 1125 possible combinations and used to generate 40 sets of perturbed images. Choices of parameters for different patients were independent to generate the broadest range of perturbations with the minimum computational cost. **Figure 4-2** demonstrates the original and two perturbed images and the corresponding GTVp contours in three views.

Table 4-2. Extracted radiomic feature number separated by image filter and feature category.

| **Image Filter** | **Feature Category** | **Feature Number** |
|---|---|---|
| Original | Shape | 14 |
| Original | First-order | 18 |
| Original | Texture | 73 |
| LoG (sigma=1, 2, 3, 4, 6 mm) | First-order, texture | 91x5 = 455 |
| Wavelet (8 compositions) | First-order, texture | 91x8 = 728 |
| Total | | 1288 |

Feature computation was performed on the perturbed images using PyRadiomics, which is compliant with recommendations from IBSI[27]. The perturbed image pixel values were shifted by the same offset value of 2000 and further discretized into a fixed bin width of 25. 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, HHH) were applied to the discretized images for yielding advanced RFs. A total of 1288 RFs (14 shape features, 91 from the unfiltered image, and 91x13 from filtered images) was computed per perturbed image, as reported in **Table 4-2**.

RF repeatability and repeatability agreement. Feature repeatability was quantified using the intraclass correlation coefficient (ICC). Since the perturbation parameters were independently applied to images 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.[87]. The ICC for each RF was binarized to a threshold of 0.9 to classify high and low RF repeatability, as adopted in previous literature [88].

The repeatability agreement between two image sets was assessed using two metrics. The mean absolute difference (MAD) of the ICC was computed between the two compared datasets for each RF category, irrespective of the chosen ICC threshold. We also evaluated the RF repeatability consistency between image sets. It is quantified as the ratio of the commonly high-/low-repeatable RFs binarized by the specified ICC threshold of 0.9.

Bias analysis. We analyzed the applicability of our results by evaluating bias introduced from feature collinearity, ROI contouring, and ROI volume. The relationship between feature collinearity and repeatability was investigated through two sub-analyses. First, we analyzed whether the inter-feature correlation affects the skewness of RF repeatability. For example, significantly more inter-correlated features that are high-repeatable can result in an overestimation of RF repeatability for modeling. We followed the analysis

procedure proposed by Fiset et al. [34] and compared the feature repeatability distributions between all the extracted features and the representative features selected by clustering. KMeans provided by scikit-learn (version 0.23.2)[89] was used to separate the extracted features into clusters. We tested the number of clusters from 50 to 450 with a step size of 50. The smallest number of clusters with more than 75% of in-cluster pairs having the absolute Pearson correlation coefficient exceeds 0.9 in every cluster was chosen. The representative feature in each cluster was selected as the one with the highest median correlation with other cluster members. Quantitatively, we compared the ratios of low-repeatable features (ICC < 0.9) between all the extracted features and the representative features for each image set.

Second, how ROI volume dependency affects repeatability was also investigated. ROI volume is one highly repeatable feature by definition if no contouring variation is introduced. GTVp size is also a common prognostic factor for many disease types[90–92]. Some prognostic RFs have been discovered to be "proxy features" to ROI volume due to their high correlations[29,93]. In this study, we evaluated bias in repeatable feature selection towards volume-correlated features. For each feature category and image set, the high-repeatable portion of volume-independent features was compared with the ratio relative to all the extracted features. The squared value of the Pearson correlation coefficient was used to quantify the volume correlation, and a threshold of 0.6 was chosen to determine whether an RF correlates with volume.

We simulated the systematic contouring protocol deviation by eroding/dilating all the contours twice using a binary 3x3x3-sized structure with one-pixel connectivity. All the RFs under perturbations were recalculated, followed by the same repeatability

evaluations. Wilcoxon signed-rank tests and mean difference calculations on ICCs between the original and the eroded/dilated contours were conducted for each feature category.

Differences in ROI volumes are artificially created by separating patients into five groups by four volume thresholds. The values of the volume thresholds are determined so that each volume group contains the same number of patients. The repeatability score ICC is calculated under bootstrapping (shuffled split) with a 50% test size due to the small patient number in one volume group. The calculations were conducted independently for each volume group. Feature repeatability is compared statistically using repeated-measures one-way analysis of variance (ANOVA) among the five volume groups for each feature category. The pairwise Student *t*-test was adopted as the post-hoc analysis method.

### 4.2.3. Results

Radiomic feature repeatability. As shown in **Table 4-3**, all the shape RFs and most unfiltered RFs (NPC:≥95.6%, OPC:83.5%) and LoG-filtered RFs (NPC:≥93.0%, OPC:93.6%) was highly repeatable against the studied positional variations, which is also visualized as the dominating blue-shaded regions in **Figure 4-3**. However, more than half of the wavelet RFs in all the analyzed image sets had low repeatability (**Table 4-3**). Within the wavelet-filtered categories, we observed that applying high-pass wavelet filters on more dimensions or on the slice direction (from LLL to HLL/LHL to LLH/LHH/HLH/HHL/HHH) caused a significant increase in low-repeatable RFs, as quantified by **Table 4-3** (from 3.3~13.2% to 31.3~41.2% to 69.7~80.0%) and visualized by the increasing fractions of green-shaded regions in **Figure 4-3**.

Table 4-3. Distribution of low repeatability radiomics feature across different imaging modalities and head and neck cancer subtypes.

| **Tumor Subtype** | | | **NPC** | | | **OPC** |
|---|---|---|---|---|---|---|
| **Image Modality** | | | **CET1-w MR** | **T2-w MR** | **CECT** | **CECT** |
| Low repeatable features (ICC > 0.9) | Shape | | 0% | 0% | 0% | 0% |
| | Unfiltered | | 0% | 4.4% | 3.3% | 16.5% |
| | LoG filtered | | 3.5% | 7.0% | 4.0% | 6.4% |
| | Wavelet filtered | LLL | 3.3% | 7.7% | 7.7% | 13.2% |
| | | HLL, LHL | 31.3% | 35.2% | 33.0% | 41.2% |
| | | LLH, LHH, HLH, HHL, HHH | 75.2% | 73.0% | 80.0% | 69.7% |
| | | All wavelet | 55.2% | 55.4% | 59.2% | 55.5% |
| | All | | 32.4% | 34.1% | 35.1% | 34.8% |
| Commonly low-repeatable features (all features) | | | 28.5% | | | N/A |
| | | | 29.7% | | 29.7% | |

*Abbreviations: NPC, nasopharyngeal carcinoma; OPC, oropharyngeal carcinoma; CET1-w MR, contras-enhanced T1-weighted MR; T2-w MR T2-weighted MR; CECT,*

*contrast-enhanced CT; ICC, intra-class correlation coefficient; LoG, Laplacian-of-Gaussian.*

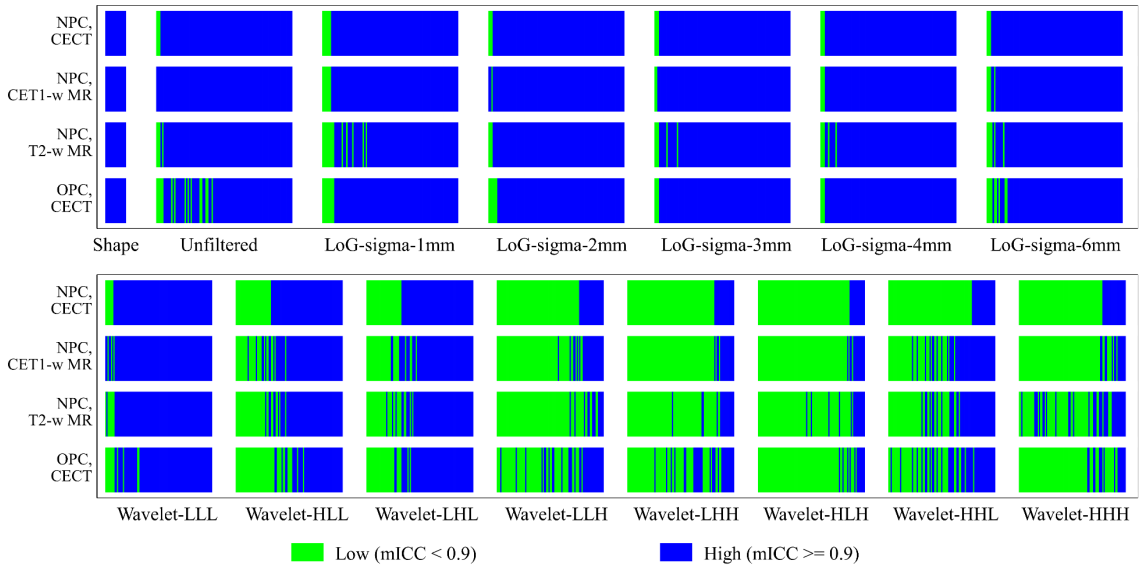


Figure 4-3: Visualization of category-based radiomic feature repeatability, binarized according to a threshold of 0.9 for the intra-class correlation coefficient (ICC).

*The green vertical lines represent low repeatability (ICC < 0.9) and the blue ones represent high repeatability (ICC >= 0.9). Within each category, features are sorted based on the ICCs of NPC CECT images and aligned at the same horizontal positions for all the image datasets.*

Agreement of radiomic feature repeatability across imaging modalities. For all the extracted RFs, high repeatability agreements were observed between any pair of the studied NPC image sets (ICC MAD<0.05, consistency>0.9). As shown in **Figure 4-4(a-c)**, shape, unfiltered, and LoG-filtered RFs expressed the highest repeatability agreements in terms of ICC MAD (<0.02) and consistency (>0.92). Wavelet-LLL/-HLL/-LHL showed the intermediate agreement with small ICC MAD (<0.03) but lower consistency (0.83~0.98). The remaining wavelet-filtered RFs demonstrated the lowest repeatability

agreement in terms of both ICC MAD (0.04-0.14) and consistency (0.70~0.98). The color agreements in **Figure 4-3** visualized such repeatability agreements too. Of note, 28.5% of all the extracted RFs (367/1288) with low repeatability were commonly found across all the imaging modalities within the NPC cohort (**Table 4-3**).

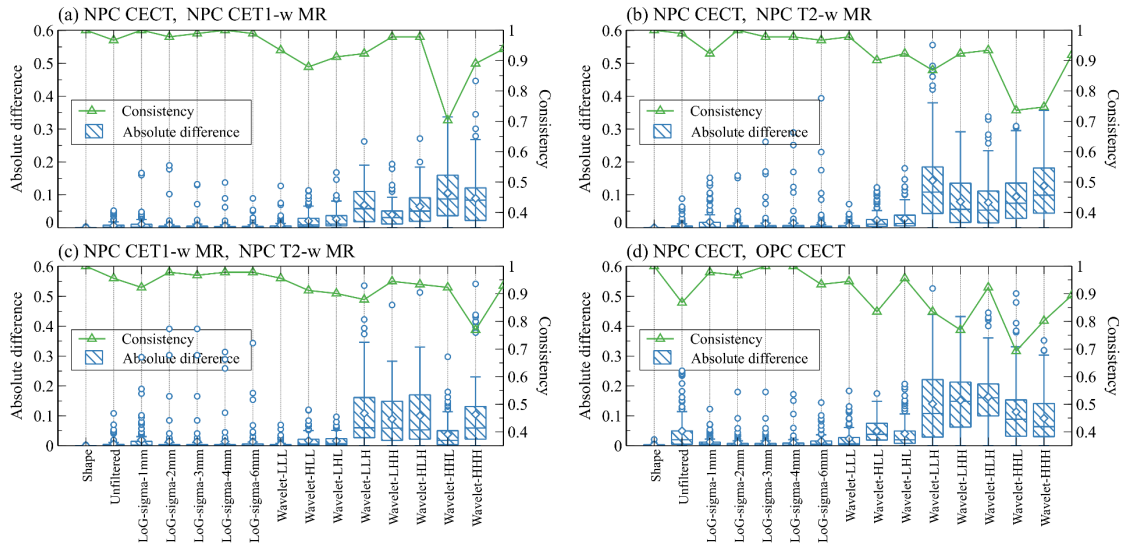


Figure 4-4: Dual y-axis plots demonstrating absolute difference of ICC and the accuracy of binarized repeatability across the studied datasets.

*Distributions of ICC absolute difference across imaging modalities of nasopharyngeal carcinoma (NPC) patients (a-c) and between NPC and oropharyngeal carcinoma contrast-enhanced CT images (d) are represented as blue boxes, and the repeatability accuracies using the threshold of 0.9 are drawn as green curves with triangle points. The median value of absolute ICC differences is represented as a horizontal line within each box, and the means are indicated by the diamonds. The edges of each box represent 25 (lower quartile) and 75 (upper quartile) percentiles of the distributions, and the whisker has a range of 1.5 interquartile.*

Table 4-4. Comparison of high/low repeatable feature counts and ratios between all the extracted features and representative features after clustering

| | **NPC** | | | **OPC** |
|---|---|---|---|---|
| | **CECT** | **CET1-w MR** | **T2-w MR** | **CECT** |
| All high-repeatable | 816 (65%) | 870 (67%) | 849 (66%) | 840 (65%) |
| Representative high-repeatable | 163 (66%) | 172 (66%) | 182 (70%) | 156 (59%) |
| All low-repeatable | 452 (35%) | 418 (32%) | 439 (34%) | 448 (35%) |
| Representative low-repeatable | 84 (34%) | 88 (34%) | 78 (30%) | 107 (41%) |

*Abbreviations: NPC, nasopharyngeal carcinoma; OPC, oropharyngeal carcinoma; CET1-w MR, contras-enhanced T1-weighted MR; T2-w MR T2-weighted MR; CECT, contrast-enhanced CT.*

Agreement of radiomic feature repeatability across head and neck cancer subtypes. RF repeatability was slightly lower between CECTs of the NPC and OPC cohort (ICC MAD = 0.06, consistency=0.89) for all the extracted RFs than the inter-modality repeatability agreements. Similar patterns of repeatability agreements that exist across imaging modalities were also observed across HNC subtypes, as shown in **Figure 4-4(d)**. Shape, unfiltered, and LoG filtered RFs had the highest repeatability agreement (ICC MAD<0.05, consistency≥0.87), followed by wavelet-LLL/-HLL/-LHL (ICC MAD: 0.02-0.05, consistency: 0.83~0.96). RFs from LLH-/LHH-/HLH-/HHL-/HHH-wavelet-filtered

images showed the lowest repeatability agreement in terms of ICC MAD (≥0.1) and consistency (0.69~0.83). Meanwhile, a significant proportion of RFs within the five wavelet-filtered categories had low repeatability (73.0~80% for NPC cohort and 69.7% for OPC cohort, **Table 4-3**). Of note, 30% of all the extracted RFs (383/1288) with low repeatability were commonly found across the CECT images of the two HNC subtypes (**Table 4-3**).

Bias analysis. The optimum number of clusters and the representative features for NPC CECT, NPC CET1-w MR, NPC T2-w MR, and OPC CECT were 250, 350, 450, and 400. The low/high-repeatable feature counts and ratios between the representative and all the extracted features are listed in

**Table** 4-4. Notably, the differences in low-repeatable feature fractions are 0.06 maximum among all the four image sets. During the investigation of bias from volume correlation, the proportions of RFs with high/low correlation and high/low repeatability were compared for each feature category and image set, as demonstrated in **Figure 4-5**. The proportion of RFs with high volume correlation and high repeatability fluctuates between 0.06 and 0.1 for all the feature categories except shape features. In total, less than five low-repeatable features with high volume correlation were observed and only in wavelet categories. Therefore, for every feature category, the portion of low/high repeatable features underwent minimum changes (maximum absolute difference=0.06) after excluding all the volume-correlated features.

Most feature categories (NPC CECT: 13/15, NPC CET1-w MR: 13/15, NPC T2-w MR: 13/15, OPC CECT: 12/15) demonstrated statistically significant (Wilcoxon *p*-value < 0.05) changes of ICC under either erosion or dilation contouring bias simulations.

However, the absolute values of the mean differences for all the feature categories and image sets were below 0.05.

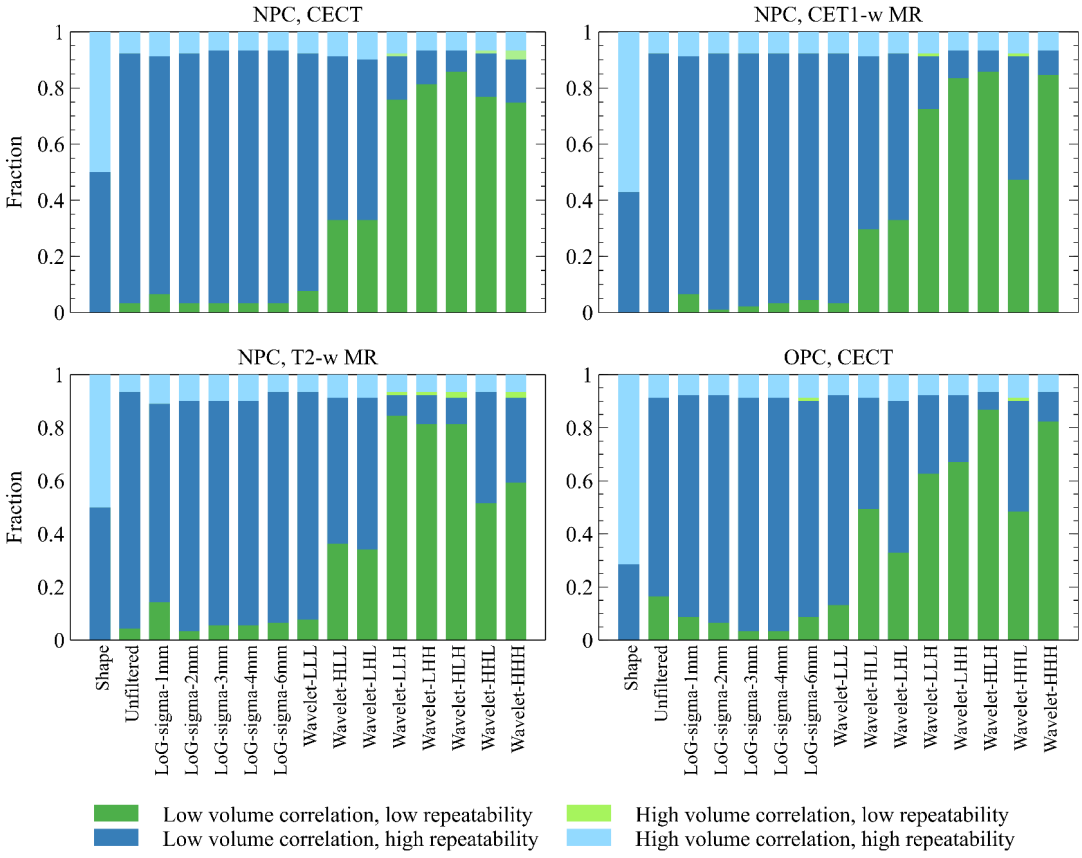


Figure 4-5. Stacked bar plots comparing the fractions of radiomic features with different volume correlation and repeatability levels for the four studies image sets.

*Each bar contains the fraction of low-volume-correlated, low-repeatable (green), low-volume-correlated, high-repeatable (blue), high-volume-correlated, low-repeatable (light green), and high-volume-correlated, high-repeatable (light blue) radiomic features. The portion of high volume-correlated and high repeatable features remained consistent among the image filters. A minimum number of high volume-correlated and low repeatable features was found in each image filter.*

At least 14 out of 15 feature categories have repeated-measures one-way ANOVA *p*-values smaller than 0.05 for all the image sets, suggesting statistically significant

changes of feature repeatability among the different levels of volume. Generally, all the image sets show increasing mean ICCs from smaller to larger volume groups for all the RF categories except shape and wavelet-LLL filtered RFs. Wavelet-filtered RFs demonstrated larger overall ICC increments (up to 0.2) with increasing volumes compared with other RF categories (<0.05). Of all the four analyzed image sets, the OPC CECT image set had the largest ICC increase.

### 4.2.4. Discussion

Results of our study suggested that the majority of the shape, unfiltered, and LoG-filtered RFs (**Table 4-3**) demonstrated high repeatability (ICC≥0.9) in all the studied image modalities and HNC subtypes (**Table 4-3, Figure 4-4**). Notwithstanding, over 50% of the wavelet-filtered RFs exhibited weak repeatability, irrespective of image modalities and HNC subtypes (**Table 4-3**). Notably, we observed numerous interesting fashions within the wavelet-filtered category. One example can be visually perceived in **Figure 4-3**, where images with high-pass filtering on more dimensions demonstrated decreased feature repeatability. Specifically, wavelet-HHH and wavelet-LLL expressed an overwhelming disparity in RF repeatability.

The lower repeatability of RFs from wavelet-filtered images and their distinct patterns could partially be ascribed to the principle of the wavelet filter. Wavelet filter decomposes the original images into eight decompositions in various frequency domains along three possible imaging axes. A high-pass-filter collects noisy and sharp edge signals, while a low-pass-filter smooths the images. Hence, high-pass-filtering could result in a more heterogeneous distribution of pixel values along the dimension where it applies. Our perturbation algorithm translated the high pixel value heterogeneity into the

high RF value variation through interpolation during image/mask resampling. This might elucidate our observation that the more dimensions the high-pass-filter applies to, the fewer repeatable RFs remain, and that HHH-wavelet RFs had the worst performance (**Figure 4-4**). In contrast, LoG-filter combines Laplacian filter for edge detection and Gaussian filter for varying extents of image smoothing; and applies to all image dimensions simultaneously. Compared to wavelet-filter, it, therefore, renders less pixel-value heterogeneity and hence less susceptible to the perturbations. This rationale might shed light on our finding that LLL-wavelet-filtered and LoG-filtered RFs shared similar repeatability performance (**Figure 4-4**).

Apart from the above, our data demonstrated high repeatability agreements (ICC MAD≤0.06, consistency≥0.89) in all the compared image sets (**Figure 4-3, Figure 4-4**) in general. The marginally drop in the agreement between cancer subtypes, compared to the inter-modality agreement, might be attributed to the discrepancies in bin counts during image intensity discretization (**Figure 4-6**) and ROI volumes (**Figure 4-7, Figure 4-8**). We believe that the translation and rotation perturbations change feature values by altering voxel intensities on the ROI edges. We observed fewer bin counts of the unfiltered, LoG filtered, and wavelet-LLL filtered images for OPC CECT than NPC CECT (**Figure 4-6**). The smaller bin counts could introduce more significant changes to image intensity distributions from edge voxel intensity variations, resulting in declined repeatability of the corresponding RFs. The remaining wavelet-filtered OPC CECT images had similar or more bin counts than NPC CECT and would yield RFs with higher repeatability. The OPC cohort has smaller ROI volumes (**Figure 4-7, Figure 4-8**). The higher surface-to-volume ratio caused by smaller volumes led to larger relative variations

of image intensity distributions within ROIs, contributing to the decreased RF repeatability under the applied perturbations. This theory is consistent with our results. The OPC CECT image set showed reduced RF repeatability than NPC CECT without filtering or under LoG and wavelet-LLL filters due to both the fewer bin counts and smaller volumes. The improved RF repeatability under the rest of the wavelet filters could be caused by the more significant improvement from larger bin counts. The theory, again, illuminates our observation that RFs with the high-pass wavelet filter on more dimensions expressed declined repeatability, as the bin counts of the filtered images increased (**Figure 4-6**). Herein, we speculate the impact of the rigid perturbations on RF repeatability might, to a large extent, depend directly on image filters and the inherent image characteristics such as ROI volume and number of gray levels, rather than on the types of image modalities/sequences and cancer subtypes. Nevertheless, there might be other contributing factors and are worthy of further investigation.

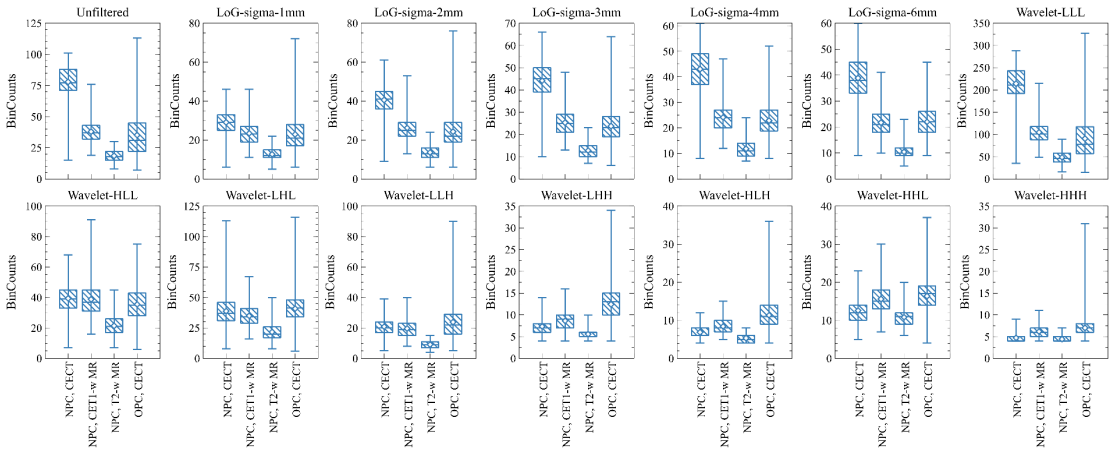


Figure 4-6. Box plots of bin counts of the unfiltered/filtered images for all the image datasets.

*Each image type was drawn as a separate plot, and each plot contains the distributions of the four studied image sets. The whisker edges indicate the maximum and minimum value. Abbreviations: LoG, Laplacian-of-Gaussian.*

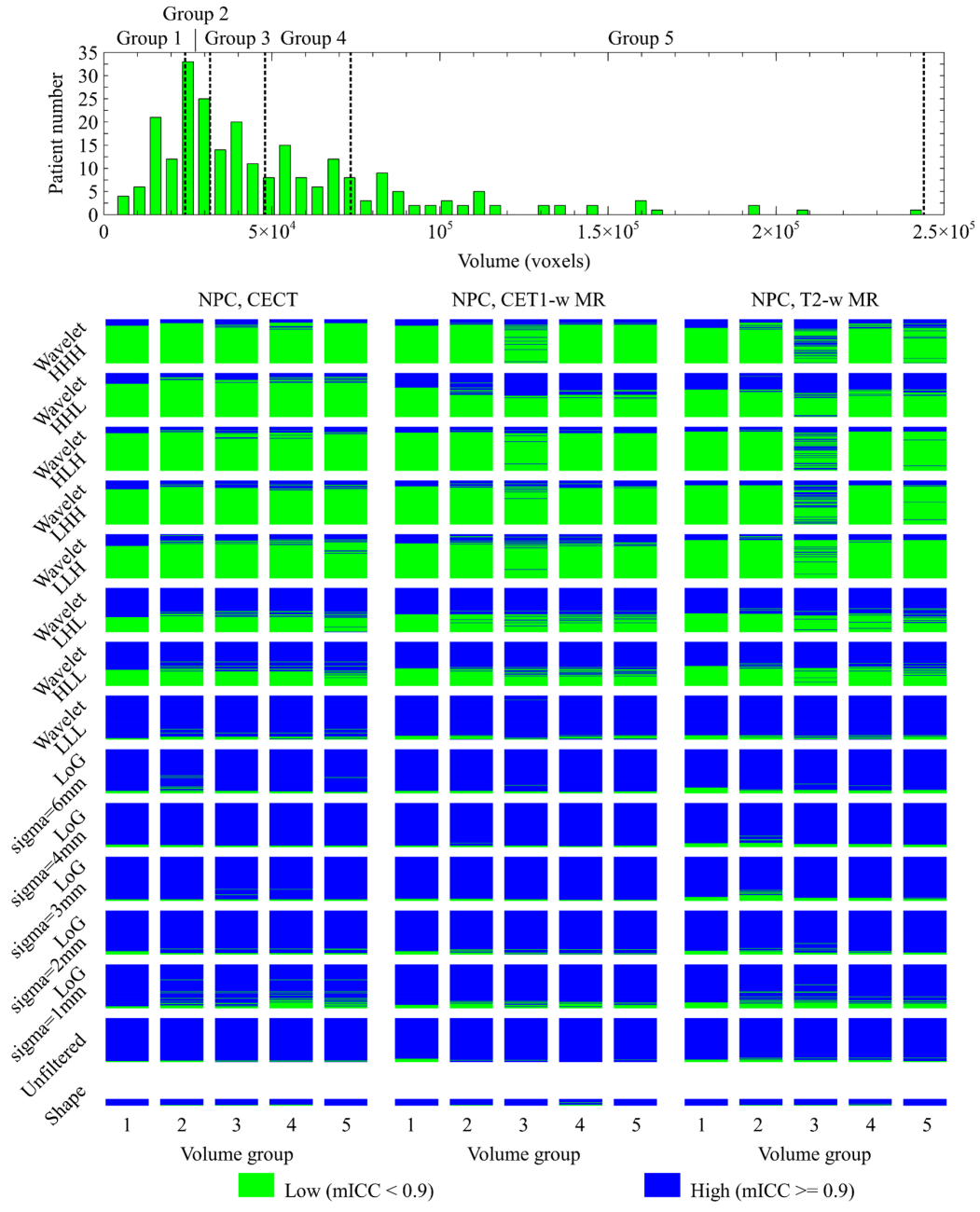


Figure 4-7. Category-based binary radiomics feature repeatability separated by volume groups for contrast-enhanced CT (CECT), contrast-enhanced T1-weighted (CET1-w) MR, and T2-weighted (T2-w) MR of the Nasopharyngeal Carcinoma (NPC) cohort.

*The top figure is the histogram of the primary gross tumor volume for the NPC patient cohort, and the dashed black lines indicate the four threshold values (24213, 31616, 50164, 74242, and 243786) for patient grouping.*

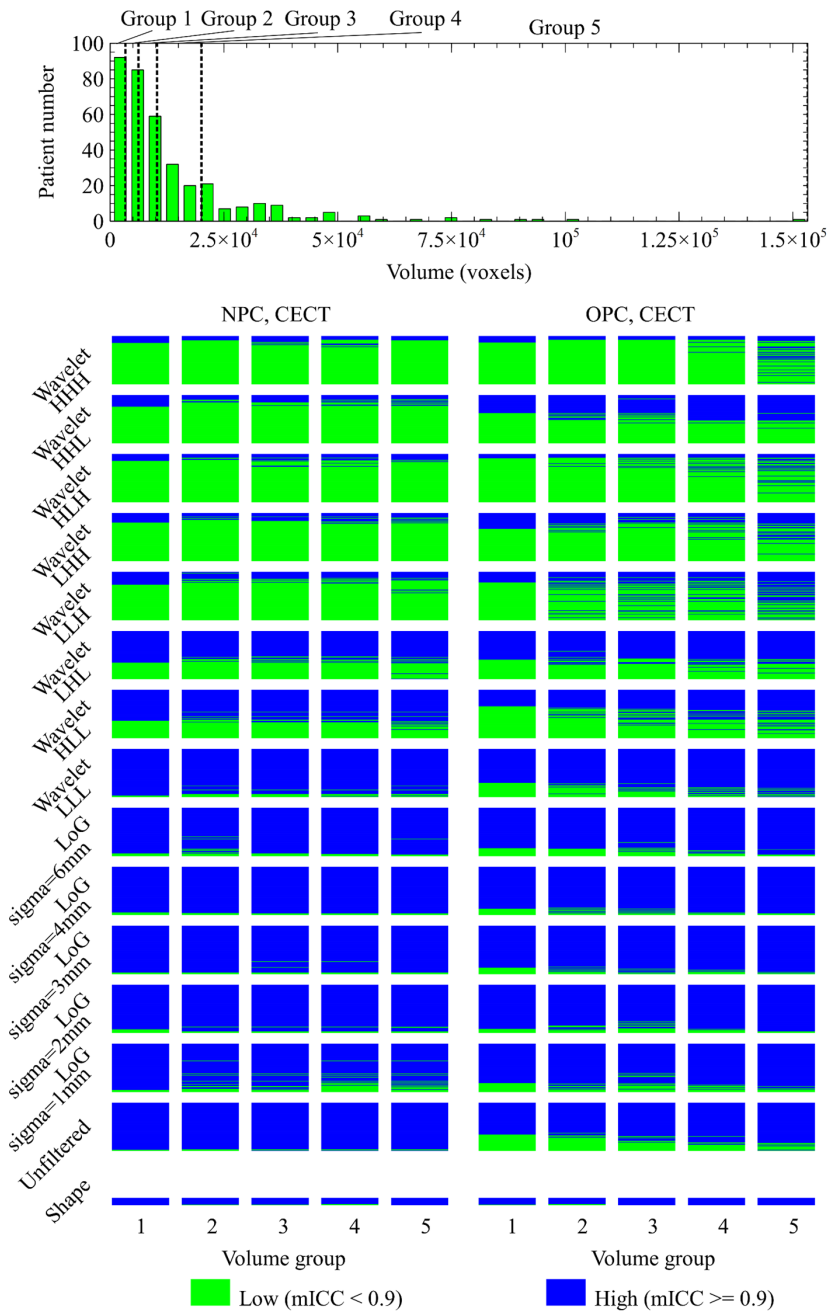


Figure 4-8. Category-based binary radiomics feature repeatability separated by volume groups for CECT of the NPC cohort and CECT of the Oropharyngeal Carcinoma (OPC) cohort.

*The top figure is the histogram of the primary gross tumor volume for the OPC patient cohort, and the dashed black lines indicate the four threshold values (3338, 6256, 10307, 20001, 153301) for patient grouping.*

Multiple validation experiments suggested minimum bias introduced from feature collinearity and ROI contouring, while apparent bias was observed from ROI volume. Notably, the minimum repeatability skewness from feature collinearity is consistent with the conclusion addressed by Fiset et al.[34] Minimum bias from ROI volume correlation was observed, as the proportion of high-repeatable features remains stable after excluding volume-correlated features. Negligible RF repeatability changes in terms of magnitude were observed after the 2-voxel contour dilation and erosion, although they were statistically significant. However, the repeatability variations may be vastly magnified if larger deviations in contouring protocols are introduced. The statistically significant and large mean ICC increment from smaller to larger volume levels showed substantial repeatability bias from ROI volume, especially for RFs from wavelet-filtered images. This positive correlation is another evidence that supports the theory in how ROI volume affects RF repeatability discussed in the previous paragraph. The highest magnitude of repeatability increase for the OPC CECT dataset could be explained by the larger increase rate of the surface-to-volume ratio of ROI under small volumes. The dependency of feature repeatability on ROI volume raises the alarm on the generalizability of radiomics models across treatment sites where the ROI sizes are significantly different.

Many factors could introduce bias to our results besides the studied ones. They were not investigated due to the scope of this study. Image preprocessing procedures, such as bin size/bin counts in image discretization and re-segmentation range, may affect feature repeatability in various ways. As explained before, larger bin size/smaller bin counts would result in declined RF repeatability. Image re-segmentation directly limits the maximum, minimum, and range of pixel intensities within the ROIs. It magnifies the

relative variations of those three features if parts of the patients have truncated image intensities within ROI after re-segmentation. Bias could also arise from other factors such as treatment site and radiomics calculation software.

In light of the progressively increasing adoption of wavelet-filter within the radiomics community in recent years [94–102], our scrutiny of category-based RF repeatability is of paramount importance for RF pre-selection and robust model construction. Of note, various studies reported that over 90% of the key features in their models originated from wavelet-filtered images [97–99]. However, among the selected wavelet RFs reported in the literature for HNC cases in MR images, we only observed high repeatability in 17/36 RFs, while certain extremely underperforming RFs (ICC≤0.5) were noted (**Table 4-5**). Meanwhile, we recognized that a robust model construction is multifactorial, scanning parameters, model of scanners, preprocessing steps, and delineation uncertainties might additively play a role in altering RF repeatability. Although our study intentionally focused on revealing positional variation dependence of RF repeatability, we, herein, argue that an even larger proportion of the underperforming (especially wavelet-filtered) RFs would likely be foreseen when additional factors come into play. Thereby, we stress our pressing concerns on cautious handling of the wavelet-filtered RFs within the radiomics community.

Table 4-5. Repeatability of wavelet feature used as final selected features in previous literature. Low repeatable radiomic features were highlighted as red.

| CET1-w MR | T2-w MR |
|---|---|
| **HLL-wavelet Category** | |
| HLL-first-order-median (0.94) | HLL-glcm-cluster-prominence (0.95) |
| HLL-glrlm-run-percentage (0.98) | HLL-gldm-dependence-entropy (0.98) |
| HLL-glcm-correlation (0.90) | HLL-gldm-small-dependence-low-gray-level-emphasis (0.61) |
| HLL-ngtdm-complexity (0.90) | |
| **LHH-wavelet Category** | |
| LHH-first-order-mean (0.92) | LHH-first-order-mean (0.90) |
| LHH-first-order-median (0.75) | |
| LHH-glszm-gray-level-non-uniformity-normalized (0.30)* | |
| LHH-glszm-small-area-high-gray-level-emphasis (0.50)* | |
| **LLH-wavelet Category** | |
| LLH-first-order-mean (0.98) | LLH-first-order-mean (0.99) |
| LLH-first-order-median (0.93) | |

| | |
|---|---|
| **LLL-wavelet Category** | |
| LLL-glcm-cluster-shade (1.00) | LLL-glcm-cluster-shade (1.00) |
| | LLL-glcm-inverse-variance (1.00) |
| | LLL-glrlm-short-run-low-gray-level-emphasis (0.89) |
| | LLL-glrlm-long-run-high-gray-level-emphasis (0.99) |
| **HHL-wavelet Category** | |
| HHL-glszm-zone-size-non-uniformity-normalized (0.67) | HHL-first-order-mean (0.62) |
| | HHL-glcm-sum-average (0.70) |
| **HLH-wavelet Category** | |
| HLH-first-order-skewness (0.70) | HLH-first-order-rms (0.86) |
| HLH-glcm-informational-measure-of-correlation-1 (0.66) | HLH-glcm-informational-measure-of-correlation-1 (0.62) |
| | HLH-glcm-autocorrelation (0.45)* |
| | HLH-glcm-informational-measure-of-correlation-2 (0.62) |
| **LLH-wavelet Category** | |

| | |
|---|---|
| LLH-glrlm-long-run-high-gray-level-emphasis (0.66) | LLH-glcm-cluster-shade (0.89) |
| LLH-first-order-skewness (0.90) | LLH-glcm-correlation (0.63) |
| LLH-ngtdm-strength (0.75) | |
| LLH-glszm-size-zone-non-uniformity-normalized (0.17)* | |
| LHL-wavelet Category | |
| LHL-glszm-small-area-high-gray-level-emphasis (0.86) | |

*Abbreviations: CET1-w MR, contras-enhanced T1-weighted MR; T2-w MR T2-weighted MR.*

Our study has limitations that need to be addressed in future studies. First, the perturbation algorithm might not fully mimic the positional variations as in real clinical scenarios owing to technical challenges in simulating small deformations of the patient's body between positionings. Second, there exists an inherent inter/intra-observer bias in GTVp delineation, which is commonly encountered in most studies. In view of this, we performed extensive bias analyses regarding the ROI contours used in our study. Third, owing to the limited transparency of previous test-retest studies, detailed results of their RF repeatability/reproducibility are often not provided. This does not allow us to perform comparisons against the results of our study. Besides, a number of works were not accomplished in our research for the sake of maintaining comprehensiveness while minimizing complexity. This includes investigating the implication of RF repeatability to

predictive model building and the agreement of our RF repeatability results in different cancer types or in a phantom study. We encouraged the community to carry out further investigations and will consider an extension of this work in the future.

### 4.2.5. Conclusions

In conclusion, although most RFs from unfiltered and LoG-filtered images demonstrated high repeatability, more than half of the wavelet-filtered RFs had poor repeatability, regardless of imaging modalities and HNC subtypes. Besides, RF repeatability agreements between imaging modalities were outstanding, while slightly worse between cancer subtypes. In particular, we discovered that different sub-categories of the wavelet-filtered RFs exhibited remarkably dissimilar repeatability performance. While LLL-wavelet RFs were the best-performing sub-category, HHH-wavelet RFs expressed diametrically opposite behaviors. Minimum bias was observed from feature collinearity and simulated contouring protocol change. Improvements in RF repeatability were extensive and significant from the ROI volume increase. Herein, we urge caution when handling wavelet-filtered RFs and advise excluding underperforming RFs during feature pre-selection for robust model construction.

## 4.3. Radiomic Feature Repeatability and its Impact on Prognostic Model Generalizability: A Multi-Institutional Study on Nasopharyngeal Carcinoma Patients

### 4.3.1. Introduction

Radiomics is an emerging technique that leverages high-throughput feature extraction from medical images for discovering hidden information that is prognostic or predictive of various clinical endpoints. Accumulating evidence has suggested the promising application of Radiomics in the prognosis[103], clinical management[104–106], and treatment response predictions[107,108] of NPC from several imaging modalities, including CT[109], MR[110–112], and PET/CT[47,113]. MR was favored in recent publications[13,114] due to its superior soft-tissue contrast. However, the majority of the previous MRI radiomics analysis on NPC were deemed less reliable due to the lack of stability analysis and external validation[115], which impedes the clinical applicability of the research findings[116]. RF repeatability, which indicates the RF stability under the same imaging condition, should be the fundamental requirement of reliable modeling.

Effective assessment of RF repeatability has attracted growing attention in the past decades[117]. However, limited effort has been made to demonstrate the benefit of repeatable features in improving downstream modeling, especially on its cross-institutional model generalizability. More direct evidence on this topic is needed to provide the community with an enhanced understanding on the benefit and usage of RF repeatability in radiomics studies.

This study aims to investigate the RF repeatability via perturbation and its impact on the cross-institutional generalizability of the prognostic model for NPC DFS prediction. We attempted to assess cohort-specific RF repeatability by our in-house developed image perturbation framework, taking reference from the previous work carried out by Zwanenburg et al.[37] to mimic a vast amount of scanning position stochasticity on CET1-w MR in a retrospective NPC cohort. The main objectives of this study were (i) to ascertain the repeatability of a comprehensive set of RFs against scanning position stochasticity via translation and rotation perturbations and (ii) to examine the benefit of repeatable RF in improving cross-institutional generalizability of prognosis modeling by externally validating prognostic models built separately from high- and low-repeatable RFs. Results from this study would provide a direct and conservative perceptiveness of RF repeatability pattern under a wide range of image filtering and discretization settings, offer evidence of its impact on inter-institutional generalizability, and encourage the radiomics community to exclusively adopt high-repeatable RFs for modeling to safeguard model generalizability.

### 4.3.2. Methods and Materials

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

CET1-w MR images and the planning GTVp contours were retrieved from the treatment planning systems. MR scanning and GTVp contouring protocols can be found in **Appendix A.** DFS information was collected from patient folders. The time of DFS is defined from the date of treatment to the earliest occurrence of death from any cause, local or regional tumor recurrence, or distant metastasis.

Preprocessing and feature extraction. All the calculations in image preprocessing and feature extraction followed the guidelines proposed by the IBSI[27]. They were performed by our in-house developed Python-based (3.7.3) pipeline using the SimpleITK (1.2.4) and PyRadiomics (2.2.0) packages. The workflow is explained by **Figure 4-9.** Image preprocessing and feature extraction parameters are listed in **Table 4-6**. We extracted all the first-order features and texture features from Gray-Level Co-occurrence Matrix (GLCM), Gray Level Dependence Matrix (GLDM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), and Neighbouring Gray Tone Difference Matrix (NGTDM) from the original, 3D LoG filtered (sigma values of 1, 2, 3, 4, and 5 mm) and all the Coiflet-1 wavelet-filtered images. Each image was discretized by a fixed bin number of 8, 16, 32, 64, and 128 before feature extraction. In total, 6510 RFs were computed per patient.

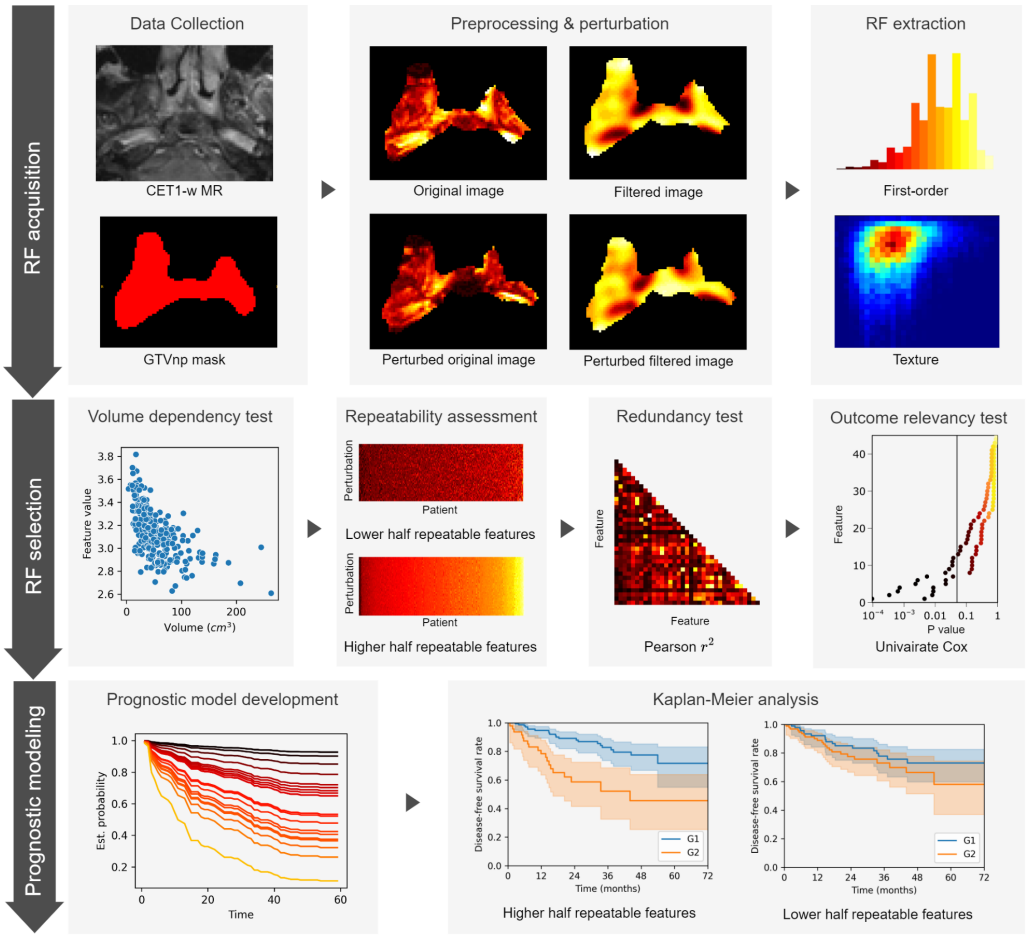


Figure 4-9. Overall study workflow.

Table 4-6. Image preprocessing, perturbation, and feature extraction parameters

| **Parameter** | **Value** |
|---|---|
| N4B bias correction maximum iterations | [50, 50, 50, 50] |
| Normalization scale | 100 |
| Pixel value offset | 0 |

| | |
|---|---|
| Resample pixel size (mm) | [1,1,1] |
| Image/mask interpolation algorithm | BSpline |
| Mask partial volume threshold | 0.5 |
| Interpolation grid alignment | Align grid origins |
| Translation distances (pixel) | [ 0.0, 0.2, 0.4, 0.6, 0.8] |
| Rotation angles (degree) | [-5, 0, 5] |
| Perturbation times | 40 |
| Image discretization bin count | 8, 16, 32, 64, 128 |
| Image filters | Unfiltered, Laplacian-of-Gaussian (3D), Wavelet |
| Kernel size of Laplacian-of-Gaussian filter (mm) | [1,2,3,4,5] |
| Wavelet filter type | Coilf1 |
| Wavelet filter decompositions | [LLL, HLL, LHL, LLH, LHH, HLH, HHL, HHH] |
| Wavelet filter starting level | 0 |
| Wavelet filter total level | 1 |

| | |
|---|---|
| Feature class | First-order, GLCM, GLRLM, GLSZM, GLDM, NGTDM |

Perturbation and radiomic feature repeatability assessment. Patient position variations were simulated by applying translation and rotation perturbations to each image and GTVp mask simultaneously during image preprocessing. They were implemented following the procedures proposed by Zwanenburg et al.[37], and the parameters are listed in **Table 4-6**. In this study, 40 translation and rotation combinations were randomly generated without replacement. The same preprocessing and feature extraction procedures were applied in calculating the RFs under perturbations. Feature repeatability was quantified from the perturbation RFs using the ICC. The one-way, random, absolute-agreement ICC was employed to assess RF repeatability due to the independent assignment of perturbation parameters to patients.

Feature selection. RFs from the unperturbed images were selected based on volume dependency first and then equally separated into high- and low-repeatable groups by the median ICC value before the feature redundancy and outcome relevancy test. The feature selection procedure is also explained in **Figure 4-9**. Since the primary tumor volume has been recognized as a reliable prognostic factor[93], RFs that were highly correlated with GTVp mesh volume were first removed to minimize potential bias in the subsequent analyses[29]. We used the square of the Pearson correlation coefficient ($r^2$) to quantify the volume correlation, and the threshold of 0.6 was used to filter the volume-independent features. The final features were selected from each repeatability group by the feature redundancy and outcome relevancy test. During the feature redundancy test, $r^2$ was used

to evaluate the correlation between features. For each highly correlated feature pair that has $r^2$ exceeding 0.6, the one that has a larger mean $r^2$ with the rest of the features were removed. The outcome relevancy was evaluated by univariate Cox regression, and the 10 best features were finally selected, which were defined as the ones having the smallest hazard ratio (HR) *p*-values.

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

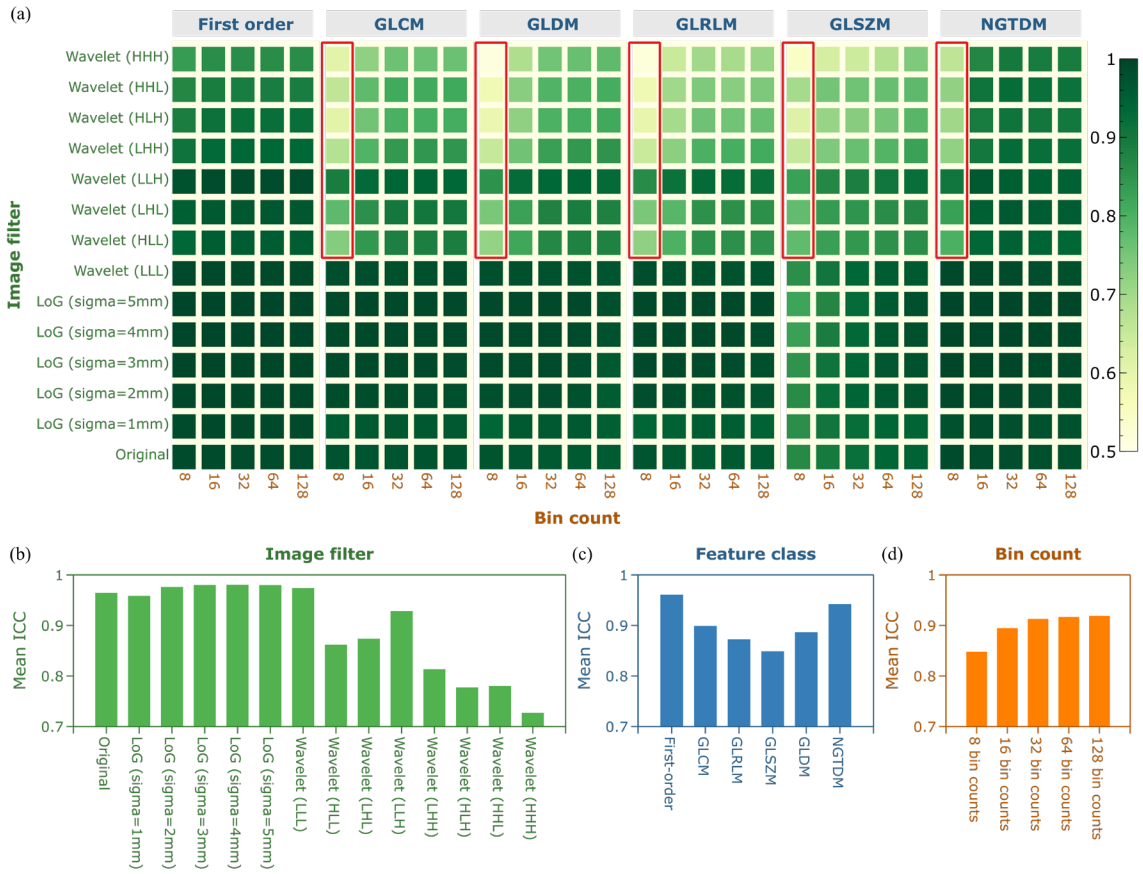


Figure 4-10. Mean ICC of the extracted radiomic features.

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

### 4.3.3. Results

The distributions of the baseline patient characteristics for the two cohorts are listed in **Table 4-7**. Consistent distributions of age and sex were found between the training and validation cohort. The overall stage, chemotherapy strategy, and World Health Organization (WHO) histology were significantly different ($p$-value < 0.05) between the two institutions. The three-year DFS rate was 74.3% in training and 72.1% in validation.

Table 4-7. Baseline patient characteristics of the Queen Mary Hospital (QEH, training) and Queen Mary Hospital (QMH, validation) cohort.

| | **QEH (Training)** | **QMH (External Validation)** | ***p*-value** |
|---|---|---|---|
| Age | | | |
| Median | 53 | 55 | 0.055 |
| Sex | | | |
| Female | 70 | 40 | 0.590 |
| Male | 216 | 143 | |
| Overall stage | | | |
| 2 | 1 | 29 | < 0.001 |
| 3 | 187 | 90 | |
| 4 | 98 | 64 | |
| Chemotherapy | | | |
| CCRT | 178 | 37 | < 0.001 |
| CCRT+ACT | 62 | 65 | |
| CCRT+ICT | 44 | 81 | |
| WHO histology | | | |

| | | | |
|---|---|---|---|
| Type 2 | 73 | 28 | 0.012 |
| Type 3 | 213 | 155 | |

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

RFs with lower repeatability were mostly texture features extracted from high-pass wavelet-filtered images discretized by smaller bin numbers, as visualized by the lighter green colors in the average ICC heatmap (**Figure 4-10(a)**). The average ICC of high-pass wavelet-filtered 8-bin discretized texture RFs (**Figure 4-10(a)**, red rectangles) was 0.69 but up to 0.99 for the remaining RFs. For image filters (**Figure 4-10(b)**), RFs from unfiltered, all the LoG and LLL wavelet-filtered images yielded an average ICCs higher than 0.95, while the rest showed lower average RF repeatability (ICC: 0.73-0.87). Moreover, a decreasing trend of repeatability was found with high-pass wavelet filtering on more image dimensions. The first-order and NGTDM RFs showed the highest average ICC of 0.96 and 0.94, while the rest of the texture classes had mean ICCs below 0.90 (**Figure 4-10(c)**). Notably, the GLSZM class had the lowest repeatability with an average ICC of 0.85. An increasing trend of repeatability was observed for larger image gray level discretization bin numbers. Specifically, bin number 8 had the lowest average ICC of 0.88, and the highest repeatability (mean ICC = 0.92) was achieved at 128 bin number (**Figure 4-10(d)**).

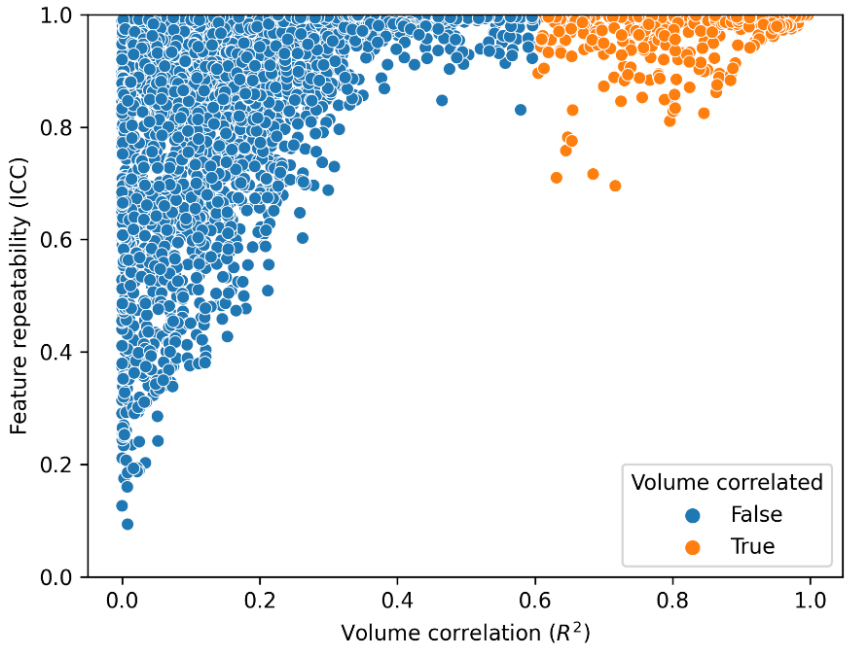


Figure 4-11. Scatter plot of volume correlation versus feature repeatability.

*Volume correlation was quantified by the square Pearson correlation r with the primary gross tumor mesh volume, and feature repeatability was quantified by the intra-class correlation coefficient under translation and rotation perturbations. A cutoff of volume correlation at 0.6 was applied with the volume correlated features indicated by orange and the non-volume-correlated ones by blue.*

A strong correlation between GTVp volume dependency and repeatability was found on the RFs extracted from QEH T1-w MR images (**Figure 4-11**). 673 out of 709 (95%) volume-dependent features ($r^2 > 0.6$) had high patient positioning repeatability (ICC > 0.9) whereas 3902 out of 5801 (67.3%) volume-independent features showed high repeatability. The 709 volume-dependent features were removed from the subsequent analysis.

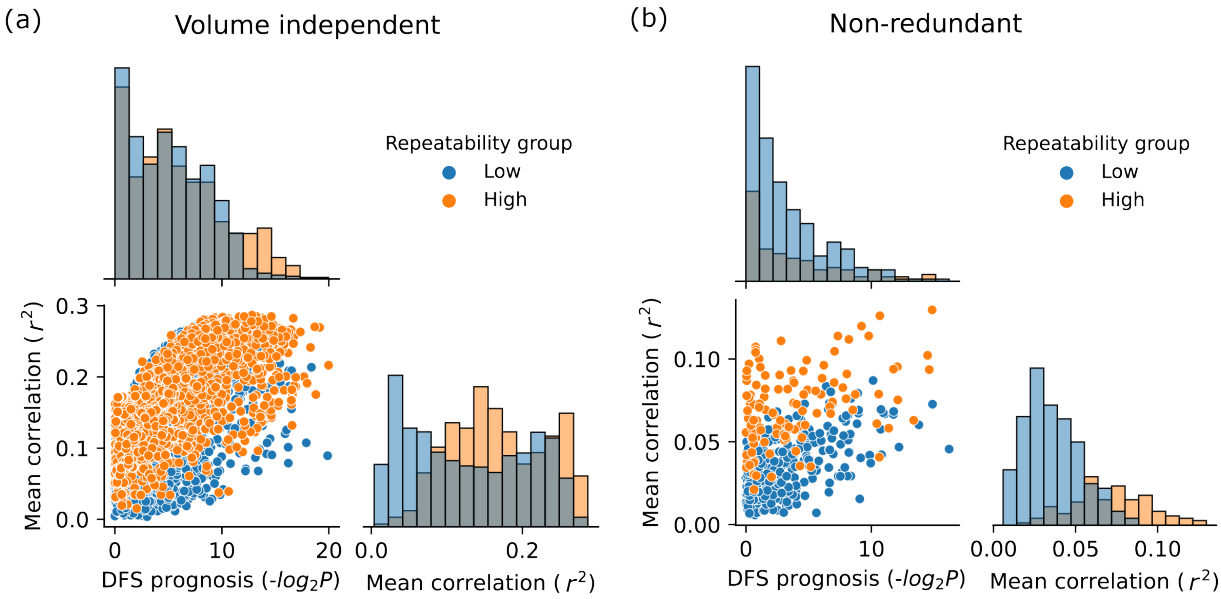


Figure 4-12. Distributions of mean feature correlation and disease-free survival prognosis during feature selection.

*Distributions of mean feature correlation and disease-free survival prognosis after (a) volume dependent feature removal and (b) redundancy test. The high-repeatable feature group is indicated by orange and the low-repeatable group by blue.*

Distinct distributions of the feature redundancy measured by the mean $r^2$ to the rest of the features were observed on the high-repeatable and low-repeatable feature groups (**Figure 4-12(a)**), which were split by the median ICC of 0.95. Forty-four percent (1281/2901) of the low-repeatable features appeared to have low redundancy (mean $r^2 <$ 0.1) whereas 15% (445/2902) had low redundancy for the high-repeatable features. Distributions of the DFS prognosis, which was measured by univariate Cox *p*-value, were similar between the two feature groups, except for the extreme-high prognosis region. Only 286 out of 2901 for the low-repeatable group had high DFS prognosis with *p*-value $< 0.001$ ($-\log_2 P > 10.0$) while 541 out of 2901 for the high-repeatable group.

Table 4-8. Final selected radiomic features and multivariate Cox regression parameters of the low-repeatable and high-repeatable model.

| **Image Filter** | **Feature Class** | **Bin Number** | **Feature Name** | **Hazard Ratio** | ***p*-value** |
|---|---|---|---|---|---|
| **Low-Repeatable** | | | | | |
| Wavelet-HHH | GLRLM | 64 | LongRunHighGrayLevelEmphasis | 1.14 | 0.269 |
| Wavelet-HHH | GLDM | 16 | DependenceVariance | 1.02 | 0.858 |
| Wavelet-LHL | GLDM | 8 | LargeDependenceLowGrayLevelEmphasis | 1.13 | 0.346 |
| Wavelet-LHL | GLRLM | 8 | RunVariance | 0.94 | 0.684 |
| wavelet-HHL | GLSZM | 16 | LargeAreaLowGrayLevelEmphasis | 1.08 | 0.355 |
| wavelet-LLL | GLRLM | 32 | RunVariance | 1.06 | 0.628 |
| LoG (sigma=5mm) | GLCM | 128 | MCC | 1.23 | 0.195 |
| LoG (sigma=2mm) | GLSZM | 64 | ZoneEntropy | 0.97 | 0.848 |

| | | | | | |
|---|---|---|---|---|---|
| LoG (sigma= 4mm) | GLSZM | 32 | ZoneEntropy | 1.19 | 0.300 |
| Original | GLSZM | 8 | ZoneEntropy | 1.15 | 0.373 |
| **High-Repeatable** | | | | | |
| Wavelet -LLH | First-order | 128 | Mean | 1.72 | 0.002 |
| Wavelet -LHL | First-order | 128 | Mean | 0.81 | 0.371 |
| Wavelet -HLL | First-order | 32 | Mean | 1.13 | 0.205 |
| LoG (sigma= 3mm) | GLCM | 32 | InverseVariance | 1.65 | 0.05 |
| LoG (sigma= 3mm) | GLRL M | 8 | LongRunHighGrayL evelEmphasis | 0.91 | 0.590 |
| LoG (sigma= 4mm) | GLCM | 32 | InverseVariance | 1.12 | 0.596 |
| LoG (sigma= 4mm) | GLRL M | 8 | RunEntropy | 0.91 | 0.677 |
| LoG (sigma= 5mm) | GLSZM | 64 | LargeAreaHighGray LevelEmphasis | 1.24 | 0.045 |

| | | | | | |
|---|---|---|---|---|---|
| LoG (sigma= 5mm) | GLRLM | 64 | LongRunHighGrayLevelEmphasis | 1.06 | 0.641 |
| LoG (sigma= 5mm) | GLSZM | 64 | ZoneEntropy | 1.06 | 0.702 |

Ten RFs were finally selected from both the two feature groups after redundancy and outcome relevancy filtering. Details of the final selected features can be found in **Table 4-8**. After redundancy filtering, more low-repeatable features (317) remained with less redundancy but similar outcome relevancy compared to high-repeatable features (116), as shown in **Figure 4-12(b)**. Quantitatively, 23% (72/317) of the low-repeatable features had the mean $r^2$ larger than 0.05 while up to 84% (97/116) for the high-repeatable ones, and 28% (88/317) and 36% (42/116) had *p*-value < 0.05 ($-\log_2 P > 4.3$) for the two groups.

Table 4-9. Training and validation performance of the two constructed Cox survival regression model using high-repeatable and low-repeatable features.

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

*Abbreviations: C-index, concordance index; 1y, 1-year; 3y, 3-year; 5y, 5-year; AUC, area under the curve. Note: the numbers within brackets are 95% confidence intervals.*

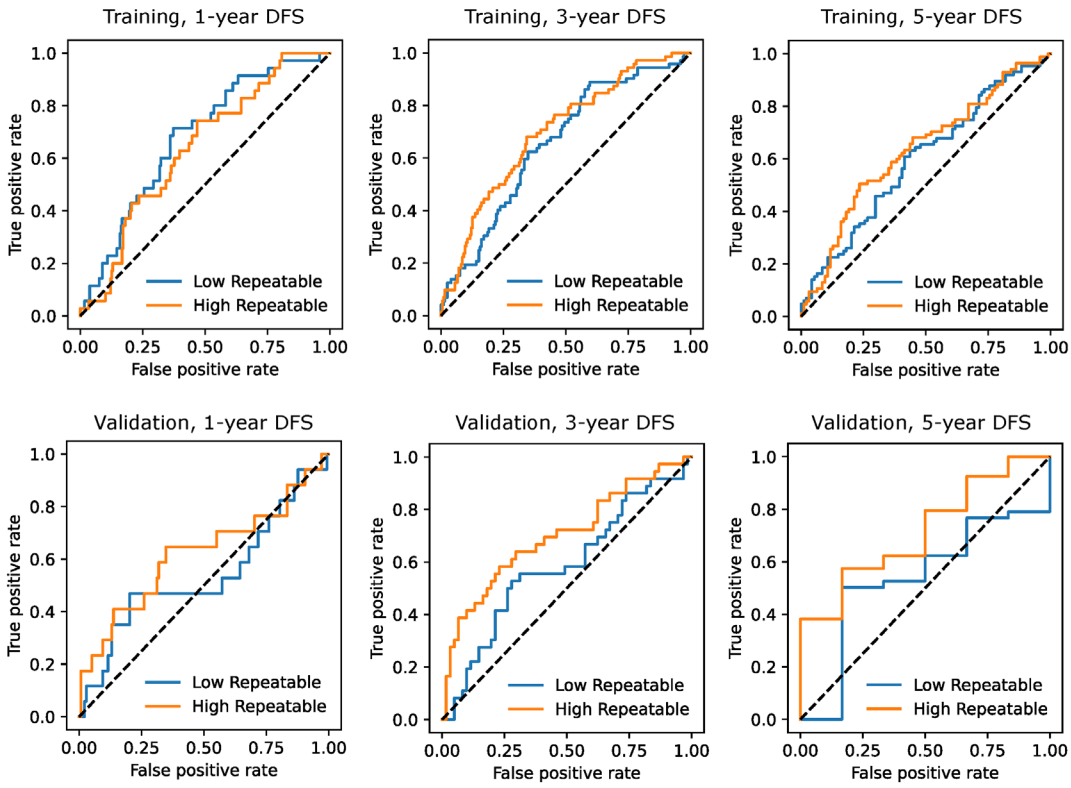


Figure 4-13. Time-dependent receiver operating characteristic curves of low- and high-repeatable Cox regression models.

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

The discriminability of the multivariate Cox survival regression models developed from both low and high-repeatable features remained stable in the training cohort. As reported in **Table 4-9**, the C-index (low-repeatable = 0.65; high-repeatable = 0.67; *p*-value = 0.526) and time-dependent AUCs (*p*-value> 0.05) were similar in the training cohort. Time-dependent ROCs on the training cohort were also similar between the two feature sets, as shown in **Figure 4-13**. During cross-validation, similar training

performances were achieved with a mean C-index of 0.61 (low-repeatable) and 0.63 (high-repeatable). However, the low-repeatable models demonstrated significantly lower C-index values (mean = 0.55) than the high-repeatable ones (mean = 0.60) for internal validation, as shown in **Figure 4-14**. Both low and high-repeatable features stratified the training cohort into distinct survival groups (G1 and G2) with similar discriminability (HR=2.50, 3.19) and statistically significant separations (log-rank *p*-values <= 0.001), as presented in **Figure 4-15**.

The prognostic model based on the high-repeatable features demonstrated significantly higher predictive performance in the validation cohort. Statistically higher C-index (high-repeatable = 0.63; low-repeatable = 0.57; *p*-value = 0.024), 1-year AUC (high-repeatable = 0.62; low-repeatable = 0.54; *p*-value = 0.031), and 3-year AUC (high-repeatable = 0.70; low-repeatable = 0.58; *p*-value = 0.015) were achieved (**Table 4-9**), while the 5-year AUC demonstrated weak statistical significance. **Figure 4-13** demonstrated distinctive differences in ROCs, especially for the 3-year progression event where the deviations were magnified. For survival risk stratifications, the high-repeatable features resulted in a significant separation of survival curves (*p*-value < 0.001) whereas a marginal separation (*p*-value = 0.054) can be found for the counterpart (**Figure 4-15**).

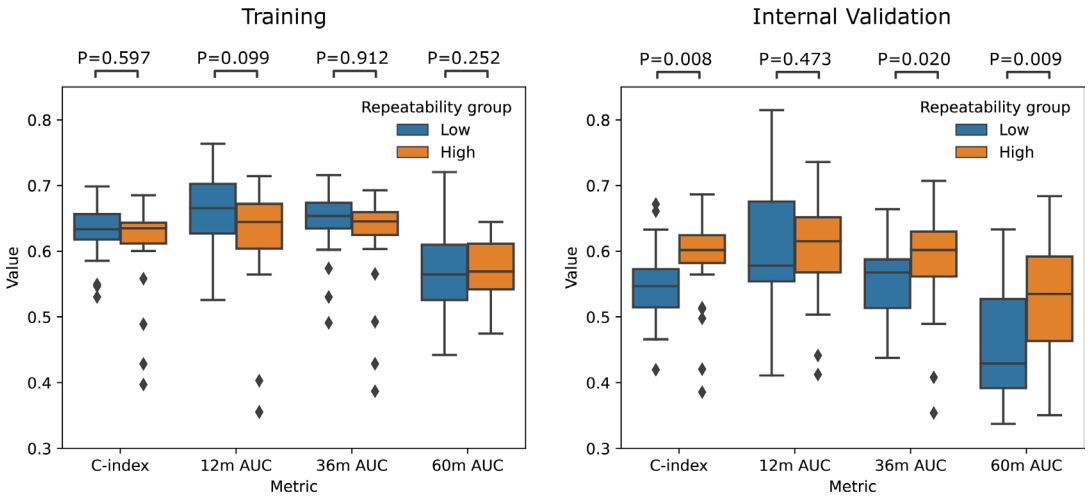


Figure 4-14. Comparison of internal validation performance between low- and high-repeatable features on the training cohort.

*The training performance remained similar while significantly higher internal validation performance for concordance index (C-index) (p-value =0.008), 36-month (36m) area under the curve (AUC) (p-value =0.020), and 60-month (60m) AUC (p-value =0.009) can be observed for the high-repeatable model. P-values were calculated by Mann-Whitney U-test.*

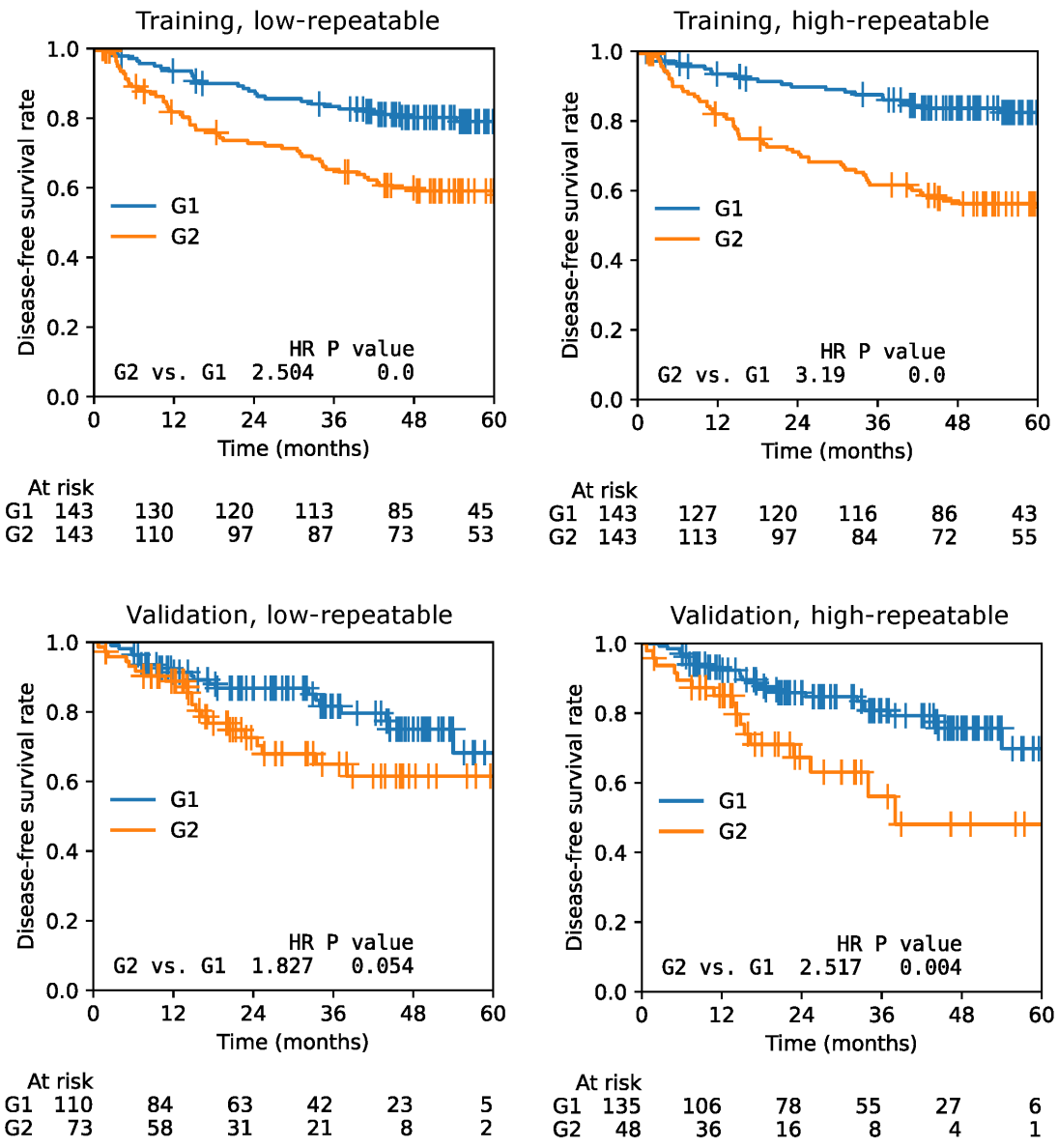


Figure 4-15. Kaplan-Meier analysis of the low (G1) and high (G2) risk groups.

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

### 4.3.4. Discussion

This study directly demonstrated the benefit of the unique information from RF repeatability assessed by translation and rotation perturbations in reducing false discovery and improving cross-institutional generalizability. Results of our study suggested that different image filters, discretization bin numbers, and feature classes displayed heterogeneous patterns of RF repeatability. Notably, texture RFs from high-pass wavelet-filtered images discretized with smaller bin numbers were more susceptible to image perturbations (**Figure 4-10**). After removing the volume-dependent RFs, the low-repeatable features demonstrated less redundancy, but outcome relevancy distributions were similar. The pattern remained unchanged after the redundancy test. Similar prognostic performance was achieved between the high and low-repeatable RFs during model training (**Table 4-9**), while the low-repeatable RFs yielded non-significant prognostic stratification on validation (**Figure 4-15)**.

Our image preprocessing strategy, especially the homogeneous resampling and gray-level discretization, aimed to minimize the impact of inconsistent image resolutions and intensity levels on feature repeatability and model performance from different scanners and scanning settings within and across institutions. Previous studies have shown the pronounced effect of pixel sizes on radiomic feature variability and suggested resampling to enhance the robustness[118,119]. Gray-level discretization with a fixed bin number could normalize the image intensities and reduce noise simultaneously[120]. It is also recommended by the IBSI for preprocessing image modalities with arbitrary intensities such as MRI[27].

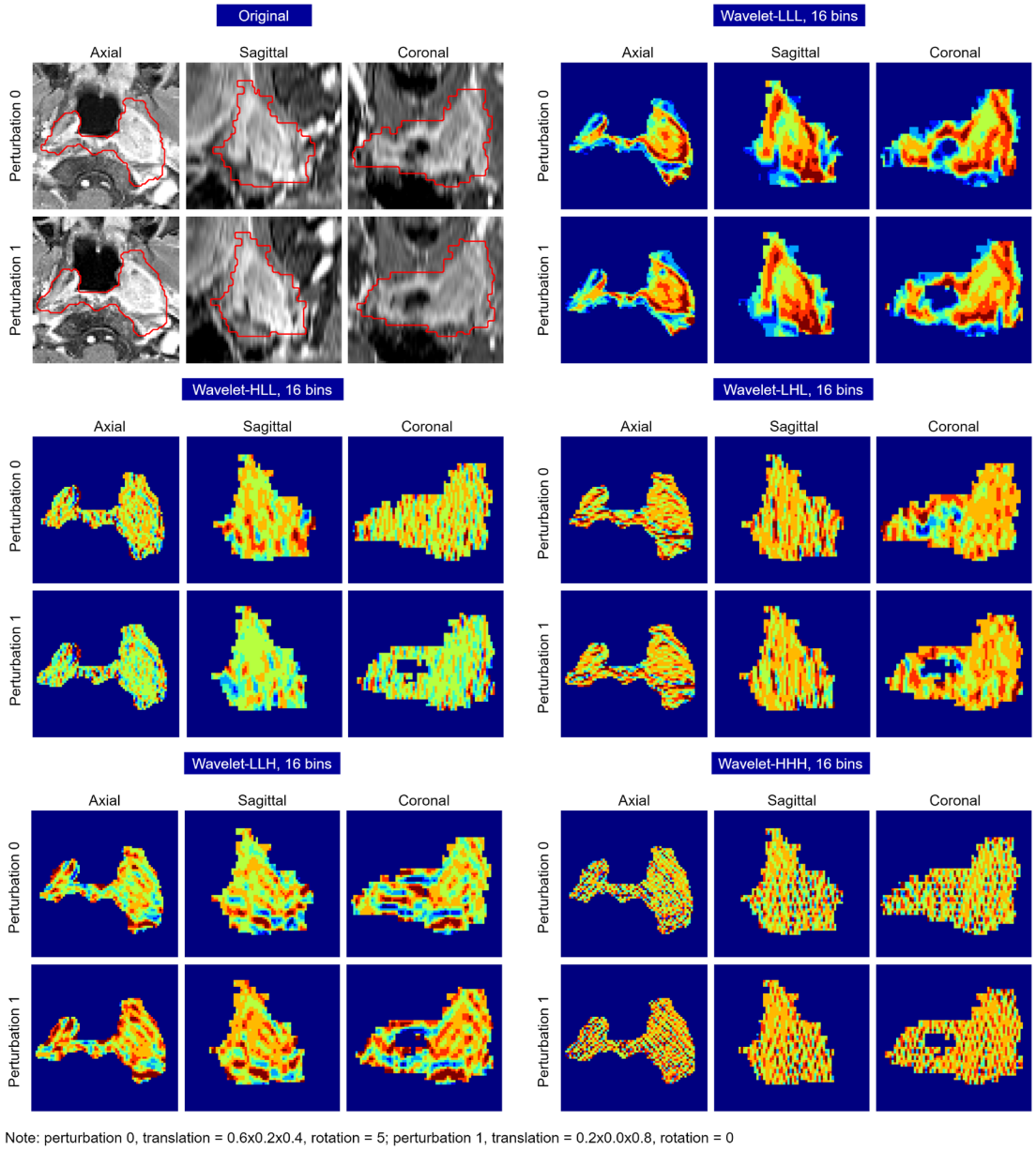


Figure 4-16. Demonstration of wavelet filtering on two example perturbations.

*The original images and primary gross tumor volume contours were also presented. More heterogeneous pixel distributions were observed with high-pass wavelet filterings on more image dimensions, and larger deviations were found between the two perturbations. Notably, applying high-pass filter on the third axis (superior-inferior) yielded more smooth images than the first two dimensions.*

The observed wavelet RF repeatability pattern could be ascribed to multiple factors, including the nature of wavelet filtering, image resampling strategy, and perturbation settings. High-pass wavelet filtering collects high-frequency signals and yields more heterogeneous pixel values. As demonstrated in **Figure 4-16**, images with high-pass wavelet filters on more dimensions appeared more heterogeneous, with fewer connected pixels with the same discretized intensity after binning. Notably, the wavelet-LLH image was less heterogeneous than wavelet-HLL and wavelet-LHL, possibly due to the larger slice thickness than in-plane resolution. The uniform 1x1x1mm resampling process up-sampled images along the axial direction, which may create artificial smooth textures. Our perturbation algorithm alters the left-right and anterior-posterior axes by rotation. It may induce drastic changes in pixel distributions under high-pass wavelet filtering on the first two dimensions while much less along the axial direction. It can also be observed in **Figure 4-16** where the texture of wavelet-LLH filtered images was more similar under the two example perturbations than wavelet-HLL and wavelet-LHL. Furthermore, a lower bin number may magnify the discrepancies of texture features due to the smaller size of the gray-level matrix, which is consistent with the results of a previous phantom study[121]. Similar patterns were found in results reported by Larue et al. on a lung 4DCT, RIDER test-retest, and 4D-OES dataset[67] where more statistical high-pass wavelet RFs were highly repeatable than texture features.

The feature selection results suggest that the adopted redundancy and outcome relevancy tests, which are standard approaches in RF reduction, failed to identify the high-repeatable RFs. As expected, a large portion of ROI volume-dependent features were found to be highly repeatable under patient positioning variations, which agrees

with previous studies on RF repeatability[122]. The outcome relevancy distributions were consistent between the high and low-repeatable features, but larger differences in redundancy patterns were found. Similar to the previous research[67], a minimum correlation was found between the univariate predictive power and feature repeatability. This suggests that either high or low-repeatable RFs have an equal chance of correlating with the prediction target. The low-repeatable features, which are "noisy" by nature, are more likely to be independent, which may elucidate our finding of the more low-redundant low-repeatable features.

Although the final feature number was strictly controlled, the low-repeatable RFs still suffered severe false discovery. Satisfactory prognostic model performance in training was achieved by the high and low-repeatable RFs with a C-index of 0.67 and 0.65 (**Table 4-9**) respectively due to the stringent outcome relevancy test criteria. The significant drop in internal testing performance suggests poor internal generalizability, which is consistent with the previous findings by Teng et al[70]. During external validation, the high-repeatable features yielded slightly lower discriminability of 0.63 in C-index, possibly due to inconsistent patient distributions between the two institutions. However, the low-repeatable RFs showed minimum prognostic power on the unseen data with C-index dropped to 0.57. Consequently, a much less significant survival curve separation of the validation cohort was achieved using only low-repeatable features ($p$-value= 0.054), as suggested by **Figure 4-15**. Meanwhile, we observed a higher 5-year AUC in the validation cohort compared to training (**Table 4-9**), possibly due to random effect caused by small sample size and bias in staging between the two patient cohorts.

Our study has limitations that need to be addressed in future studies. First, the perturbation algorithm might not fully mimic the positional variations as in real clinical scenarios owing to technical challenges in simulating small deformations of the patient's body between positionings. Second, the prognostic model performance, especially during validation, was slightly lower than previous radiomics research on NPC prognosis, where a range of C-index between 0.72 to 0.85 for NPC survival prognosis was reported[114]. This could be caused by the omission of clinical factors and lymph node tumor RFs in our final model. Nevertheless, our study did not intend to construct the best-performing model for clinical utility. Finally, several works were not accomplished in our research to maintain comprehensiveness while minimizing complexity. They include investigations under different imaging modalities, cancer types, feature extraction settings, or in a phantom study. We encouraged the community to carry out further investigations and to consider extending this work in the future.

### 4.3.5. Conclusions

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

# Chapter 5.

# Biomarker Development for Nasopharyngeal Carcinoma Patients

## 5.1. Introduction

NPC is a radiosensitive epithelial malignancy in Southern China[123,124]. With the development of the intensity modulated radiation therapy (IMRT) technique, better survival patterns can be achieved for patients with early and late stage NPC, especially local and regional tumor control[5,125]. However, locoregional recurrence and distant metastasis remained the primary failure patterns with a high occurrence rate in five years, especially for patients with advanced lymph node (LN) metastasis[126,127]. Of the newly diagnosed 133,354 NPC worldwide in 2020[1], over 70% of diagnosed cases were classified as locoregionally advanced diseases. Concurrent chemoradiotherapy (CCRT) and CCRT following adjuvant chemotherapy (ACT) is recommended for locoregionally advanced NPC by the National Comprehensive Cancer Network (NCCN) guidelines. However, the individualized treatment regimen remains debated in clinical practice. Thus, effective prognosis stratification, especially for the LN tumor, and treatment efficacy predictions by development of new quantitative biomarkers are necessary to guide more accurate clinical decision-making for personalized treatments[128,129]. This chapter presents two clinical studies focusing on the development of geometric and radiomics signatures for improved prognosis and treatment efficacy predictions on NPC patients. Both of them utilized RADAR for data cleaning and feature extraction, and feature repeatability assessments were included for optimal model reliability.

## 5.2. Quantitative Spatial Characterization of Lymph Node Tumor for N Stage Improvement of Nasopharyngeal Carcinoma Patients

### 5.2.1. Introduction

N stage, which belongs to the tumor–node–metastasis (TNM) staging system jointly proposed by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC), is one of the most robust and widely used LN classifications[130]. The current edition (8th) for NPC is based on anatomical characterization, including size, laterality, and location. However, N stage has been suggested to be less comprehensive and precise due to the qualitative definitions[131].

Over the past decades, various new LN anatomical descriptors have been proposed to improve the current N staging system[132]. For instance, parotid lymph node (PLN) involvement was found to be associated with a poor prognosis in distant metastasis, and an upgrade to the N3 classification was recommended[133,134]. Besides, the current N-staging system categorizes retropharyngeal lymph node (RLN) involvement (<6 cm) as N1 disease. However, Huang et al. suggested an upgrade of patients with bilateral retropharyngeal lymph node involvement to N2 due to the distinctive prognostic performance within N1[135]. Other anatomical characteristics of LN, such as extra-nodal extension[136–138] and positive LN numbers[131,139] have been proposed to improve the existing N stage classification system for NPC.

Despite the tremendous efforts made, the development of a more accurate N staging system was still hindered by the rather complex LN anatomical environment. In the era of IMRT, detailed tumor and normal tissue delineations have become the standard

procedure for treatment planning with the increasing availability of advanced imaging techniques such as MRI and PET[140–142]. Quantitative spatial characterization of metastatic LN may provide more accurate descriptions of its anatomy, enabling the holistic discovery of anatomical prognostic factors by a data-driven approach.

Therefore, this study aims to investigate the feasibility of improving the prognosis stratification of N staging system from quantitative spatial characterizations of metastatic LN. We designed two types of geometric histograms based on the distances and angles of LN tumor volume to surrounding normal tissues. Independent prognostic factors were extracted by principal component analysis and combined into one prognostic index. A new risk stratification from the combined index was proposed and evaluated on multiple survival endpoints, including DFS, overall survival (OS), relapse-free survival (RFS) and distant metastasis-free survival (DMFS) both internally and externally. Our methodology may promote accelerated improvement of the LN classification for NPC and can be potentially generalized to other cancer sites.

### 5.2.2. Materials and Methods

Two cohorts of biopsy-proven NPC patients receiving chemoradiotherapy were retrospectively recruited from Hong Kong QMH between 2013 and 2019 and Hong Kong QEH between 2012 and 2015, respectively. Informed consents from patients were waived due to the retrospective nature of this study. The total number of included patients was 194 from QMH and 284 from QEH after excluding patients with (1) co-existing cancer or distance metastasis before treatment, (2) radiation therapy only without concurrent chemoradiotherapy, (3) patients in stage N0 who do not have visible tumor in the lymph node region and (4) incomplete clinical record and missing segmentations. Patients from

the QMH cohort were used for deriving independent prognostic factors and development of prognostic index, while the QEH cohort was used solely for external validation.

Clinical factors, including age, sex, T stage, N stage, M stage, overall stage, chemotherapy strategy, and survival information were collected from patient folders. The time of OS, RFS, DMFS, and DFS is defined from the date of treatment to the earliest occurrence of death from any cause, local or regional tumor recurrence, distant metastasis, and the combination of above all, respectively. The TNM stage was administered according to the 7th edition of the AJCC protocol for the QEH cohort and switched to the 8th edition after 2017 for the QMH cohort. Treatment planning structure sets were retrieved from the PACs in DICOM format. GTVn was contoured from CECT fused with MRI in QEH and an extra imaging modality of PET/CT in QMH by oncologists with at least five years of experience.

Both OVH and POV (see Chapter III-Feature Extraction-Feature Definitions) were adopted to describe the spatial configuration of GTVn relative to the surrounding OARs. OARs that were consistently delineated across the two institutions, including SpinalCord, Parotids (combined Left and Right Parotid), Mandible, Larynx, and Brainstem, were included in this study. Dimensions of the OVH and POV histograms were reduced by PCA, where the components that explained the greatest variance across patients were highlighted. This study included the smallest number of principal components (PCs) of OVH and POV that explained 75% of the cumulative variance for each OAR. The coefficients of the PCs were extracted as the potential prognostic factors.

Independent prognostic factors were identified from the selected PCs by univariate Cox regression on DFS followed by the covariate independency test with N

stage through multivariate Cox regression. The final prognostic index was built by combining the independent prognostic factors with N stage through multivariate Cox regression and evaluated by C-index. The confidence interval and *p*-values for baseline N stage comparison were determined by 1000-iteration bootstrapping. Risk stratification performance was assessed by KM analysis, where patients were equally stratified into high (G1), median (G2), and low (G3) risk groups based on the prognostic index in the discovery cohort. The stratification thresholds were applied to the testing cohort as well for the three-grade stratification. HRs with 95CI and the log-rank *p*-values between risk groups were acquired from univariate Cox regression. All Cox regressions and KM analysis were implemented by the Python package lifelines (version 0.27.0)[143], and the *p*-value of 0.05 was considered significant.

### 5.2.3. Results

Baseline patient characteristics. Distributions of the baseline patient characteristics for the two cohorts were listed in **Table 5-1**. Consistent distributions of age, sex, overall stage, chemotherapy strategy, and WHO histology were found between the discovery and validation cohort. The T stage and N stage were significantly different ($p$-value < 0.05) between the two institutions. The median follow-up time of the discovery cohort is 2.5 years and 4.6 years for the validation cohort. Of the 194 discovery patients within the follow-up period, 22 developed local recurrence, 17 with regional recurrence, 29 with distant metastases, and 25 died. The three-year DFS, OS, RFS, and DMFS rates were 72.1%, 90.0%, 82.4%, and 82.4%, respectively. In the validation cohort, 34, 25, 44, and 40 patients of 284 developed local recurrence, regional recurrence, distant metastasis, and death, and the five-year DFS, OS, RFS, DMFS are 74.3%, 94.0%, 85.0%, and 86.2%.

Table 5-1. Baseline patient characteristics of the discovery and validation cohort.

| | **Discovery Cohort** | **Validation Cohort** | ***p*-value** |
|---|---|---|---|
| Age | | | |
| Mean | 53.39 | 52.16 | 0.249 |
| Sex | | | |
| Female | 41 | 70 | 0.667 |
| Male | 153 | 214 | |
| N stage | | | |
| N1 | 62 | 17 | 0.035 |
| N2 | 93 | 228 | |
| N3 | 39 | 39 | |
| Chemotherapy | | | |
| CCRT | 33 | 178 | 0.330 |
| CCRT+ACT | 78 | 61 | |
| CCRT+ICT | 83 | 43 | |
| WHO histology | | | |

| | | | |
|---|---|---|---|
| Type 2 | 27 | 74 | 0.142 |
| Type 3 | 167 | 210 | |

*Note: Staging was performed according to the 7th edition of the AJCC protocol for the validation cohort and switched to the 8th edition after 2017 for the discovery cohort. Abbreviations: CCRT, concurrent chemoradiotherapy; ACT, adjuvant chemotherapy; ICT, induction chemotherapy; WHO, World Health Organization.*

Prognostic lymph node spatial factors. Thirty-one PCs were extracted from the OVH and POV histograms in total, including four OVH PC and three POV PC of SpinalCord, five OVH PC and three POV PC of Parotids, two OVH PC and two POV PC of Brainstem, three OVH OC and three POV PC of Larynx, and four OVH PC and two POV OC of Mandible. After univariate and multivariate Cox regressions, two spatial factors including the first PC of spinal cord OVH ($OVH_{SC,PC1}$) and the third PC of spinal cord POV ($POV_{SC,PC3}$) were selected as independently prognostic to DFS. Between the two spatial factors, $OVH_{SC,PC1}$demonstrated a higher discriminability to DFS with C-index of 0.66 at discovery and 0.56 at external validation, while 0.57 at discovery and 0.54 at external validation for $POV_{SC,PC3}$.

As listed in **Table 5-2**, $POV_{SC,PC3}$ contributed the highest positive hazard (HR = 3.35, 95CI: 1.41--7.99), followed by the N stage (HR = 2.26, 95CI: 1.46--3.49). On the other hand, $OVH_{SC,PC1}$ had the negative impact of survival hazard (HR = 0.63, 95CI: 0.48--0.83). **Figure 5-1(a)** presents the distributions of the two spatial factors of the 3-year disease and non-disease progressed patients at both discovery and validation. Patients who developed disease progression within three years had significantly lower $OVH_{SC,PC1}$ (mean: -0.80 vs. -0.07, $p$-value = 0.007) and higher $POV_{SC,PC3}$ (mean: 0.082

vs. 0.057, $p$-value = 0.012) at discovery, but smaller differences were found on the validation cohort ($OVH_{SC,PC1}$: 0.46 vs. 0.74, $p$-value = 0.032}; $POV_{SC,PC3}$: -0.18 vs. -0.25, $p$-value = 0.089). Moreover, the spinal cord OVH appeared to be overall larger in the validation but smaller for the POV. After binarizing the two spatial factors by the median values in the discovery cohort, more patients in the validation cohort fell into the low-risk groups, as indicated by **Figure 5-1(b)**. The odds ratios were 0.30 ($p$-value = 0.006}) for $OVH_{SC,PC1}$ and 2.21 ($p$-value = 0.052) for $POV_{SC,PC3}$ in the discovery cohort. They were less significant for $OVH_{SC,PC1}$ (odds ratio = 0.60, $p$-value = 0.275) but more significant for $POV_{SC,PC3}$ (odds ratio = 2.83, $p$-value = 0.004) in the validation cohort.

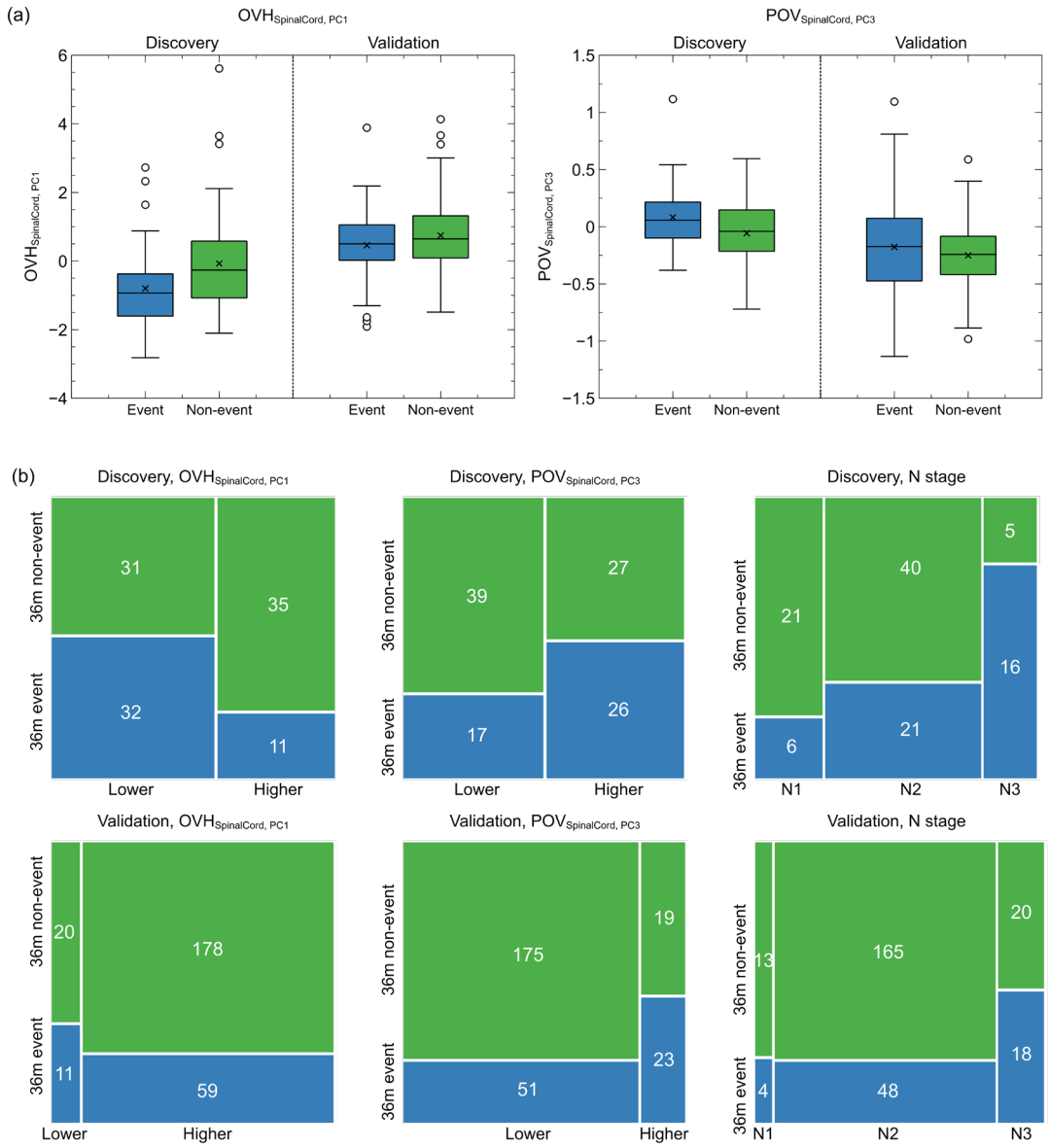


Figure 5-1. Continuous and binarized spatial factor distributions and N stage distributions for 3-year disease progressed and non-disease progressed patients in the discovery and validation cohort.

*(a) Box plots of continuous spatial factor distributions. Patients with disease progression within three years had lower spinal cord overlap volume histogram (OVH) factor values and higher spinal cord projection overlap volume (POV) values at both discovery and validation. (b) Mosaic plots of the binarized spatial factor and N stage distributions of patients with and without 3-year disease progression.*

Table 5-2. Hazard ratios and *p*-values of the selected spatial factors and N stage from multivariate Cox regression on disease-free survival.

| **Covariate** | **HR (95CI)** | ***p*-value** |
|---|---|---|
| $OVH_{SC,PC1}$ | 0.63 (0.48-0.83) | < 0.001 |
| $POV_{SC,PC3}$ | 3.35 (1.40-7.99) | 0.006 |
| N stage | 2.26 (1.46-3.49) | < 0.001 |

*Abbreviations: HR, hazard ratio; 95CI, 95% confidence interval.*

Combined prognostic index. The combined prognostic index had better discriminability than N stage on all the survival endpoints but showed statistical significance mainly in DFS and RFS, as reported in **Table 5-3**. C-index in DFS increased from 0.654 (training) and 0.568 (external validation) to 0.722 (training) and 0.603 (external validation) when combining the two new spatial factors with N stage. Such improvement was significant in training (*p*-value = 0.020) while much less in external validation (0.086). On the other hand, the training and validation improvements were both significant in RFS with C-index reaching 0.723 (*p*-value = 0.020) and 0.603 (*p*-value = 0.019), respectively.

Table 5-3. Hazard ratios and *p*-values of the selected spatial factors and N stage from multivariate Cox regression on disease-free survival.

| | | **DFS** | **OS** | **RFS** | **DMFS** |
|---|---|---|---|---|---|
| Discovery cohort | Anatomical Index (95CI) | 0.72 (0.65-0.79) | 0.75 (0.63-0.84) | 0.72 (0.62-0.82) | 0.72 (0.63-0.81) |
| | N-stage (95CI) | 0.65 (0.57-0.73) | 0.72 (0.64-0.80) | 0.64 (0.54-0.73) | 0.65 (0.54-0.76) |
| | *p*-value | 0.02 | 0.245 | 0.02 | 0.062 |
| Validation cohort | Anatomical Index (95CI) | 0.60 (0.54-0.67) | 0.59 (0.48-0.71) | 0.60 (0.52-0.69) | 0.57 (0.47-0.67) |
| | N-stage (95CI) | 0.56 (0.52-0.62) | 0.58 (0.50-0.67) | 0.53 (0.47-0.60) | 0.57 (0.50-0.65) |
| | *p*-value | 0.086 | 0.395 | 0.019 | 0.536 |

*Abbreviations: DFS, disease-free survival; OS, overall survival; RFS, relapse-free survival; DMFS, distant metastasis-free survival; 95CI: 95% confidence interval.*

Better risk stratifications were achieved by the combined prognostic index in DFS and DMFS than N stage itself, as shown by the KM curves in **Figure 5-2**. **Table 5-4** reports the hazard ratios and the corresponding *p*-values between different risk groups as well as the three-year survival rates in DFS, OS, RFS, and DMFS. On the discovery cohort, the DFS survivals of the three new risk groups were statistically different (*p*-value $< 0.05$) whereas much lower statistical significance was found between the N1 and N2

groups ($p$-value = 0.139). Higher hazard ratios were observed between G2 (4.49) and G3 (9.07) to G1 compared to the N stage (N1 vs. N2: 1.83, N1 vs. N3: 5.19). However, the HR was less between G3 to G2 (1.913) compared to the one between N3 to N2 (2.988). A similar trend was found in DMFS where G2 (4.11) and G3 (10.41) were better separated from G1 but worse between G2 and G3 (2.26). In the validation cohort, the HRs between G2 (DFS: 1.71, $p$-value = 0.021; DMFS: 1.72, $p$-value = 0.101}) and G3 (DFS: 4.02, $p$-value < 0.01; DMFS: 2.93, $p$-value = 0.014) to G1 also increased significantly compared to that between N2 (DFS: 0.772, $p$-value = 0.518; DMFS: 0.552, $p$-value = 0.271) and N3 to N1 (DFS: 1.821, $p$-value = 0.171; DMFS: 1.876, $p$-value = 0.216) in both DFS and DMFS. Similarly, a less HR was found between G2 and G3 (DFS: 2.44, $p$-value = 0.006; DMFS: 1.74, $p$-value = 0.219) than between N2 and N3 (DFS: 2.66, $p$-value < 0.001; DMFS: 3.17, $p$-value = 0.001).

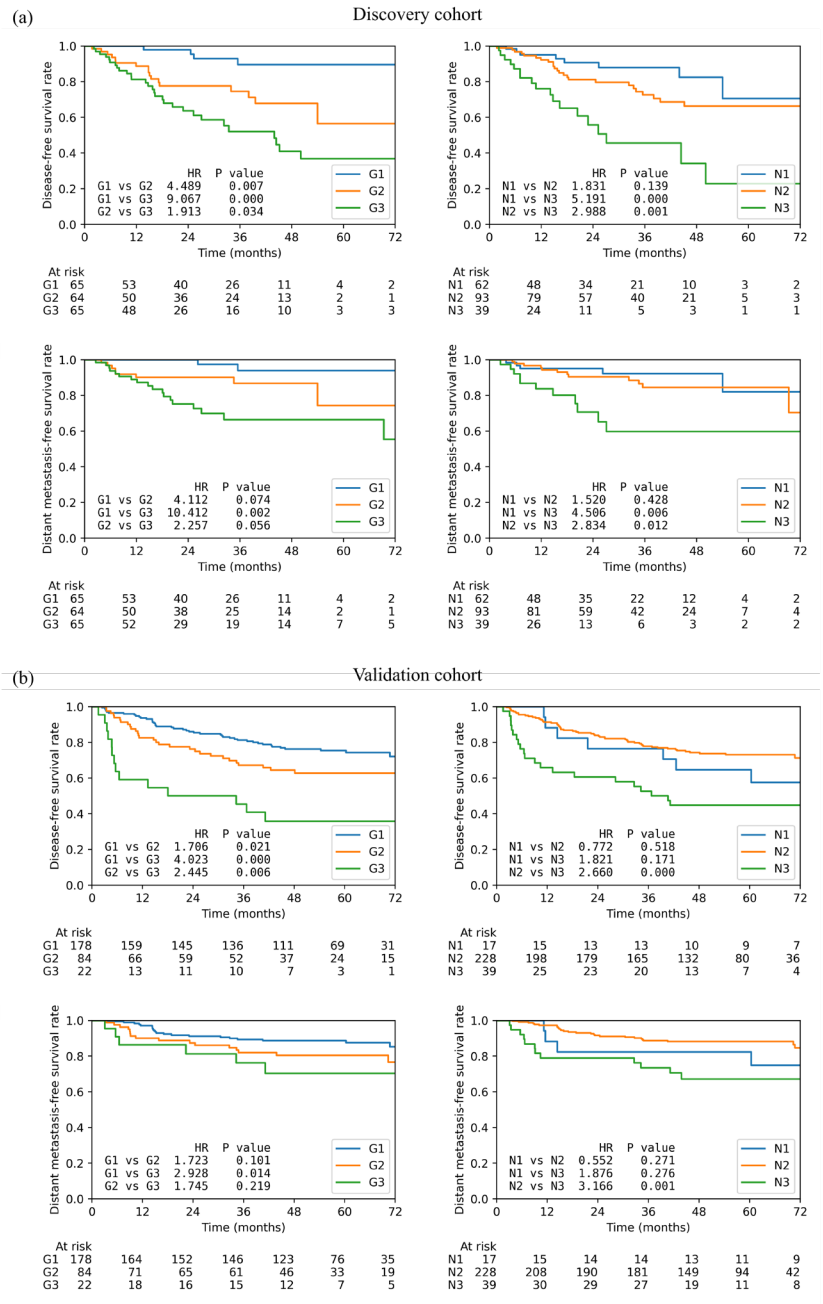


Figure 5-2. Kaplan-Meier curves of the three-risk patient groups based on the new spatial index and N stage.

*Kaplan-Meier curves of the low-(G1), median-(G2), and high-risk (G3) patient groups based on the new spatial index and the three N stages on (a) disease-free survival and (b) distant metastasis-free survival. Each plot also contains the hazard ratio (HR) and the corresponding p-value between each two groups.*

Table 5-4. Risk stratification performance of the new risk groups and N stage in multiple survival endpoints and discovery and validation cohort.

| | | **DFS** | **OS** | **RFS** | **DMFS** |
|---|---|---|---|---|---|
| **Discovery Cohort** | | | | | |
| G1 | HR | – | – | – | – |
| | *p*-value | – | – | – | – |
| | 3-y survival rate | 89.60% | 97.30% | 89.60% | 94.00% |
| G2 | HR | 4.49 | 7.66 | 2.23 | 4.11 |
| | *p*-value | 0.007 | 0.055 | 0.181 | 0.074 |
| | 3-y survival rate | 74.60% | 92.70% | 85.30% | 86.80% |
| G3 | HR | 9.07 | 13.98 | 4.76 | 10.41 |
| | *p*-value | <0.001 | 0.011 | 0.005 | 0.002 |
| | 3-y survival rate | 52.10% | 79.70% | 72.20% | 66.50% |
| N1 | HR | – | – | – | – |
| | *p*-value | – | – | – | – |

| | | | | | |
|---|---|---|---|---|---|
| | 3-y survival rate | 87.90% | 100.00% | 93.00% | 92.20% |
| N2 | HR | 1.83 | 3.33 | 2.64 | 1.52 |
| | *p*-value | 0.139 | 0.115 | 0.079 | 0.428 |
| | 3-y survival rate | 72.60% | 89.40% | 79.50% | 84.50% |
| N3 | HR | 5.19 | 11.62 | 4.59 | 4.51 |
| | *p*-value | <0.001 | 0.002 | 0.014 | 0.006 |
| | 3-y survival rate | 45.60% | 72.60% | 74.90% | 59.80% |
| **Validation cohort** | | | | | |
| G1 | HR | – | – | – | – |
| | *p*-value | – | – | – | – |
| | 3-y survival rate | 81.20% | 95.20% | 88.70% | 89.30% |
| G2 | HR | 1.71 | 1.36 | 1.46 | 1.72 |
| | *p*-value | 0.021 | 0.384 | 0.219 | 0.101 |

| | | | | | |
|---|---|---|---|---|---|
| | 3-y survival rate | 67.20% | 93.50% | 82.90% | 82.00% |
| G3 | HR | 4.02 | 2.28 | 3.69 | 2.93 |
| | *p*-value | <0.001 | 0.076 | 0.001 | 0.014 |
| | 3-y survival rate | 45.50% | 85.90% | 62.70% | 76.20% |
| N1 | HR | – | – | – | – |
| | *p*-value | – | – | – | – |
| | 3-y survival rate | 76.50% | 87.80% | 87.80% | 82.40% |
| N2 | HR | 0.77 | 1.56 | 0.84 | 0.55 |
| | *p*-value | 0.518 | 0.548 | 0.736 | 0.271 |
| | 3-y survival rate | 77.80% | 95.30% | 85.90% | 88.70% |
| N3 | HR | 1.82 | 2.57 | 1.2 | 1.88 |
| | *p*-value | 0.171 | 0.223 | 0.764 | 0.276 |

| | | | | | |
|---|---|---|---|---|---|
| | 3-y survival rate | 52.70% | 89.00% | 78.20% | 73.50% |

*Note: HR and P value were relative to the low-risk group (G1) or N1. Abbreviations: HR, hazard ratio; DFS, disease-free survival; OS, overall survival; RFS, relapse-free survival; DMFS, distant metastasis-free survival; 95CI: 95% confidence interval.*

The remaining survival endpoints showed heterogeneous patterns under the new risk stratification (**Table 5-4**). Significant HR improvements were observed in OS, but marginal in RFS for the discovery cohort. On the other hand, RFS showed significantly higher stratification performance in the validation cohort, but no improvement in OS was observed. Moreover, the validation cohort demonstrated higher 3-year survival rates on G1 and lower on G2 for RFS and DMFS, whereas marginal improvement of 3-year survival rates was found in the discovery cohort.

Representative cases. To further explain the contribution of the two anatomical factors in better identifying the risk of disease progression, we selected two representative cases from the discovery cohort with the same N stage but distinct risks based on the spatial index. The high-risk patient was classified as G1 and the low-risk one as G3, both having the same N stage (N2) and chemotherapy strategy (CCRT + ACT). The high-risk patient developed distant metastases at 32.3 months, while the low-risk patient showed no signs of disease progression for at least 34.3 months. **Figure 5-3(a)** presents the 2D axial masks and the 3D volumes of GTVn and three OARs for the high and low-risk patient. Anatomically, both patients had metastatic retropharyngeal LN, but a significantly larger {extent} of the right cervical LN tumor was observed in the high-risk patient. Meanwhile, distinct patterns of the spinal OVH and POV curves were found, as drawn in **Figure**

**5-3(b)**, where the selected PC vectors were also included. The OVH curve of the high-risk patient was significantly higher than that of the low-risk patient with the largest overlap volume difference emphasized at around the global minimum (~75 mm) of the first PC vector. The POV at the first local maximum (~25 degrees) of the PC vector was much higher in the high-risk patient, exceeding the higher POV of the low-risk patient at the second local maximum (~125 degrees).

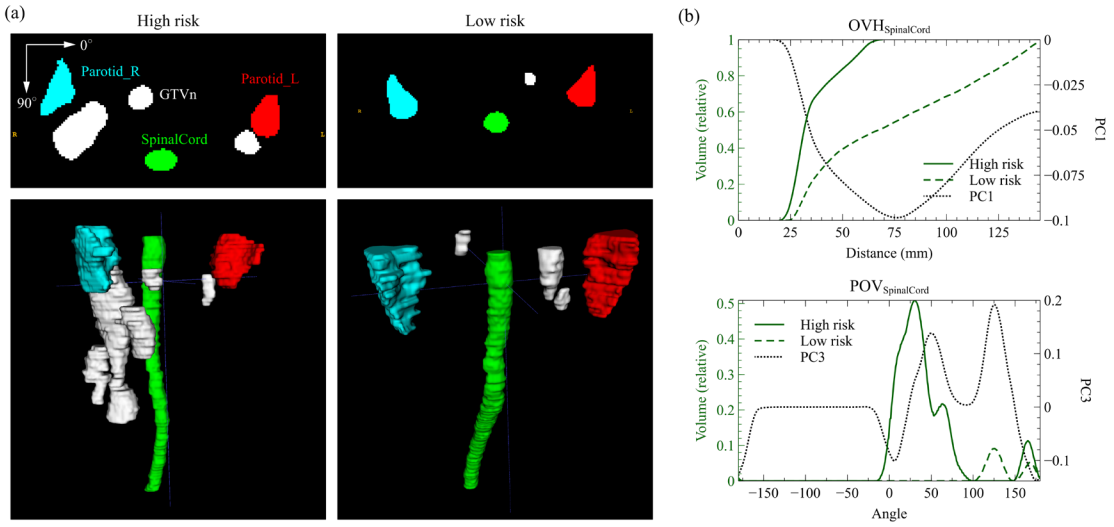


Figure 5-3. Quantitative anatomical characterizations of the high-risk and low-risk patient.

*(a) The axial slice masks and rendered 3D volumes of lymph node gross tumor volume (GTVn), Parotid_L, Parotid_R, and SpinalCord structures. (b) The SpinalCord overlap volume histogram (POV) and projection overlap volume (POV) of the two patients and the corresponding selected principal component (PC) vector. Significant differences in lymph node anatomy were captured by the large variations in the histograms and highlighted by the PCs.*

### 5.2.4. Discussion

This study demonstrated the feasibility of discovering new prognostic factors from quantitative spatial characterization of LN tumor for better LN risk stratification with

high cross-site generalizability. Two histograms precisely characterized the LN tumor anatomy by distances (OVH) and angles (POV). PCA effectively reduced the high-dimensional histograms into several informative and independent anatomical factors, and two final independent prognostic factors were discovered by Cox regressions in DFS. The prognostic index that combines the independent prognostic spatial factors and the N stage achieved better new three-level risk stratifications than the N stage itself in DFS and DMFS at both discovery and external validation.

Only the spinal cord spatial factor $OVH_{SC,PC1}$ and $POV_{SC,PC3}$ were identified as the independent prognostic factors to DFS. $OVH_{SC,PC1}$ highlights the overlap of the lower spinal cord with the expansion of isotropic LN tumor by approximately 75 mm (**Figure 5-3(b)**), indicating a smaller axial expansion of LN. The PC vector of $POV_{SC,PC3}$ has two peaks at around 25 and 125 degrees and reaches local minimums at 0 and 180 degrees (**Figure 5-3(b)**). Higher projection overlaps at the peak angles indicate more volume of LN tumor in the anterior direction of the spinal cord, whereas the valley angles suggest less involvement of the LN tumor on the lateral sides. Additionally, both factors are correlated with the axial extent of the LN tumor due to the thin cylindrical structure of the spinal cord. Such correlation was also demonstrated by the two example patients in **Figure 5-3(a)** where the high-risk patient with lower $OVH_{SC,PC1}$ and higher $POV_{SC,PC3}$ had a significantly larger axial extent of cervical LN.

Previous clinical observations on the prognostic power of the anatomy of LN tumors were highly correlated with our quantitative findings. The results of our survival analysis suggest an increased risk of disease progression with lower $OVH_{SC,PC1}$ (adjusted HR = 0.63, 95CI: 0.48--0.83; $p$-value < 0.001) and higher $POV_{SC,PC3}$ (adjusted HR =

3.35, 95CI: 1.41--7.99), regardless of the N stage. Their independent prognostic power could be explained by the two example patients in whom the high-risk one developed early distant metastases despite their identical N stage. As discussed in the previous paragraph, a higher prognostic index value suggests a higher axial expansion and extent of the LN tumor, which supports the ongoing discussion of the high prognostic value of the quantitative LN burden. Previous clinical studies reported the number of metastatic LN regions as an independent predictor of DMFS [131,139]. For $POV_{SC,PC3}$, a higher value may also indicate retropharyngeal LN metastasis with a larger size or bilateral involvement. Retropharyngeal LN has also been suggested to indicate worse in DFS and DMFS [135,144]. Specifically, the size of the metastatic retropharyngeal LN with a cutoff axial diameter of 6mm has been identified as a significant prognostic factor for OS and DMFS [145,146]. It was also suggested that the bilateral involvement of the retropharyngeal lymph nodes should be upgraded to N2 disease due to the worse 5-year OS and DMFS [135]. These anatomical characteristics have been partially included in the definition of the N1 classification of the 7th and 8th N staging system [147], where metastasis is limited above the caudal border of cricoid cartilage and/or retropharyngeal lymph node(s) does not exceed 6mm in greatest dimension. Our quantitative anatomical factors may provide more precise descriptions of various LN anatomy characterizations, thus independent of the existing N stage classifications.

The two final spatial factors were predictive of three-year DFS and DMFS at both discovery and validation. However, the binarization thresholds were less generalizable from discovery to validation due to the overall different magnitudes of the spatial factor values. As a result, much higher low-risk patients were classified in the validation cohort

when using the median values in the discovery as the binarization thresholds. The systematic cross-institutional variations in the spatial factor magnitudes could be attributed to the inconsistent spinal cord volume definitions, especially the starting and ending point. A higher spinal cord extent may lead to a lower relative overlap volume for both OVH and POV at the same absolute distance and angle, and the resulting PC coefficients are expected to be smaller. For clinical utility, consistent organ and tumor segmentations are important to ensure a reliable quantitative spatial characterization. Further adjustments in the spatial factor definitions for enhanced robustness are needed in future studies.

Despite the promising performance of the spatial characterization of lymph node tumors in survival prognosis, the analysis involves standardized tumor and OAR segmentations[148] as well as complex computations of distance and angle histograms for thorough characterization, which often require specific training. The potential long learning curve for clinicians may hinder the clinical application of the proposed predictors. Integration of auto-segmentation [149] and dedicated calculation scripts into the existing treatment planning system could be one solution for fast implementation in daily clinical practice. On the other hand, other types of biomarkers, which are easier to implement in clinics, have been proposed as strong survival predictors for patients with NPC and other HNC diseases. Systematic inflammation indicators, which can be directly measured from blood test results, have been reported to be prognostic in multiple HNC subtypes. For example, pre-treatment neutrophil-to-lymphocyte ratio (NLR) has been investigated, and a strong statistical correlation was observed with positive neck occult metastasis in laryngeal squamous cell carcinoma [150]. Another study by Orabona et al.

confirmed the independent prognostic power of the systemic immune-inflammation index (SII) and the systemic inflammation response index (SIRI) on OS of patients who received malignant salivary gland tumor surgery[151].

The constructed prognostic index results in improved risk stratifications in DFS and DMFS compared to the existing N stage both internally and externally. It is consistent with previous findings on the improved DMFS prognostication of the LN tumor region number [131] and the involvement of the retropharyngeal LN tumor [135]. Better risk stratifications on OS were only observed on the discovery cohort and RFS on the external validation cohort. Several reasons could contribute to the heterogeneous results. First, the thresholds of the prognostic index for the three-class risk classification could be suboptimal and less generalizable. The threshold optimization method for risk stratification requires a more careful design and wide validation for clinical practice. As discussed in the previous paragraph, the overall magnitudes of the spatial factors were inconsistent, which may contribute to the reduced generalizability of the prognostic index and the resulting risk groups. Second, some patient characteristics, such as stages and chemotherapy treatments, are rather different between discovery and external validation. They may affect the generalizability of the risk stratification performance due to the different baseline performances. Third, the sample sizes and follow-up durations are limited, especially in the discovery cohort. Less patients remained as uncensored samples, resulting in less reliable results. Increasing the sample size with more complete follow-up information is needed in future studies to enhance the clinical evidence of our findings.

### 5.2.5. Conclusions

This study used the distance histogram OVH and the newly proposed angle histogram POV to quantitatively characterize the anatomy of the LN tumor in relation to the surrounding spinal cord and parotids. Independent prognostic factors on DFS were discovered from the principal components of the anatomical histograms and combined with the N stage into a spatial index. It surpassed the N stage itself in risk discrimination and stratification. The proposed quantitative approach may facilitate the discovery of new anatomical characteristics in a more holistic and precise way to improve patient staging in other diseases.

## 5.3. Explainable Machine Learning via Intra-Tumoral Radiomics Feature Mapping for Patient Stratification in Adjuvant Chemotherapy for Locoregionally Advanced Nasopharyngeal Carcinoma

### 5.3.1. Introduction

The extraction and analysis of high-dimensional image features, known as radiomics, provides a unique non-invasive tool to quantify intra-tumor heterogeneity that are nearly impossible to be perceived by human eyes[45]. The emerging evidence has confirmed that radiomic features have the potential to predict the treatment response in NPC and contribute to individualized treatment decision-making without extra medical procedures for patients. Peng et al.[47] and Zhong et al.[152] utilized deep learning algorithms with radiomic features from pre-treatment PET/CT and multi-parametric MR images to identify high-risk locoregionally advanced (LA) NPC patients who may potentially

benefit from induction chemotherapy over CCRT-alone. Shen et al.[153] also constructed a joint multi-parametric MR-based radiomic and clinical signature to identify patients who benefit more from induction chemotherapy or adjuvant chemotherapy. Despite the promising performance from previous publications, the lack of explainability in radiomic features and transparency in radiomic signatures prevent the further validation of the signature and hinder the clinical application of the radiomic model. Severn et al.[154] proposed an intra-tumoral heterogeneity measurement by a voxel-wised calculation of radiomic features. An intra-tumoral heterogeneity map visualized the spatial response to radiomic features within the tumor, however, the prognosis and treatment guidance characteristics of the intra-tumoral signatures have not been investigated.

In this study, we aim to identify and validate quantitative intra-tumor heterogeneity signatures from pre-treatment CET1-w MR images and investigate the application of their voxel-wise mapping for personalized adjuvant chemotherapy decision-making in LA NPC patients. Heterogeneity signatures were selected from repeatable texture radiomics features with high predictive value. They were further mapped locally to account for the spatial variations of tumor tissue heterogeneity. We further evaluated the predictive value of tumor subvolumes highlighted on the voxel-wise mappings, which could potentially serve as direct signature visualizations and explainable treatment decision making tool. To maximize the transparency of the signature, we provided an end-to-end signature calculator from the image to treatment decision recommendations.

### 5.3.2. Materials and Methods

Patients. This retrospective study was approved by Institutional Review Board (ethics approval number: KC/KE-18-0085/ER-1). Signed informed consent form was waived due to the retrospective nature of this study. A total of 398 patients with biopsy-confirmed NPC from the Queen Elizabeth Hospital (hospital 1) were screened. We recruited patients with initial diagnosis of NPC confirmed by pathologists in hospital 1 from 2014–2016. The inclusion criteria were: (1) patients who received definitive IMRT with the following treatment modes: CCRT or CCRT+ACT , (2) patients diagnosed with stage III–IVA NPC re-staged according to the 8th Edition of AJCC Cancer Staging Manual, (3) patients in good performance status (KPS $\geqslant$ 70), without serious medical or surgical diseases, and no other malignant tumors, (4) patients with available complete initial medical history, chemotherapy and radiation therapy data and treatment planning MRI. The patients from hospital 1 were divided into discovery and validation with diagnosed time. The flowchart of patients' inclusion is found in **Figure 5-4**.

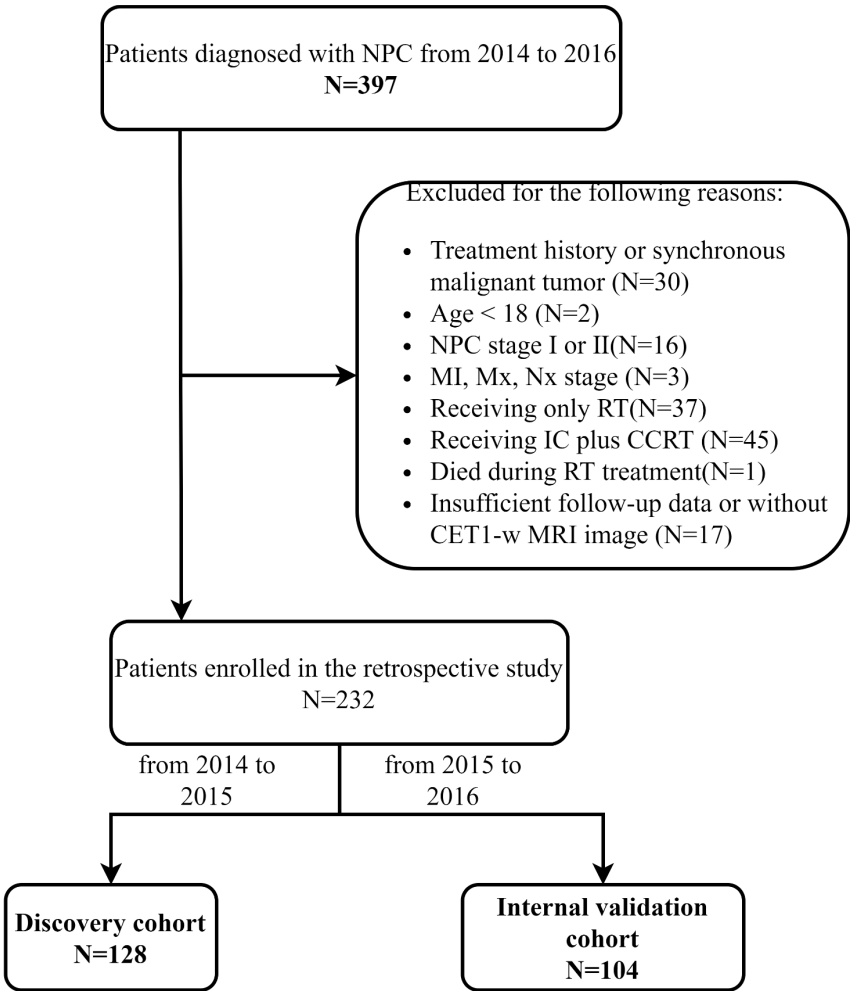


Figure 5-4. Flowchart of patients' inclusion of the study.

All patients received CCRT with or without ACT. For radiation therapy, the prescribed dose was 68 to 76 Gy for the GTVnp, 60 to 71 Gy for any involved cervical lymph nodes, 60 to 66 Gy for the high-risk region, and 54 to 60 Gy for the low-risk region in 30 to 33 fractions over 6 to 7 weeks. All patients were treated with 2 to 3 cycles of chemotherapy concomitant with RT, with or without 2–3 cycles of ACT. Patients have all received 100 mg/m2 of cisplatin (3-weekly in D1, D22, and D43), or 40 mg/m2

cisplatin every week, while CCRT + ACT group received another 2-3 cycles of adjuvant cisplatin plus fluorouracil (cisplatin 80 mg/m2 on the first day and 5-fluorouracil (5FU) 1000 mg/m2 daily on days 1-4 (or continuous intravenous infusion for 96 hours) every 4 weeks. For those with a contraindication to cisplatin, carboplatin was offered alternatively.

Radiation therapy data collection. All the enrolled patients underwent pre-treatment planning CECT, CET1-w MR, and RT plan. The images and contoured structures were retrospectively retrieved in DICOM format from the PACs.

Image acquisition. All patients in hospital 1 were scanned with 3T MRI (Achieve, Philips Healthcare). The contrast-enhanced T1-weighted sequence of pre-treatment nasopharyngeal and neck MR images were collected for each patient. Scanning parameters for CET1-w MR images acquisition were as follows: contrast agent (gadolinium-based Dotarem, Gd-DOTA); axial CET1-w spin-echo MR sequence (repetition time [TR]: 4.8-9.4ms, echo time [TE]: 2.4-8.0ms, field-of-view [FOV] = 24 x 24 cm, number of acquisition = 1, slice thickness = 0.7-4 mm x 48 slices, spacing: 0-3mm, matrix: 280-640.

Segmentation. The gross nasopharyngeal tumor, GTVnp for radiation therapy, was collected as the volume of interest for the radiomic feature extraction. The GTVnp was delineated by experienced (at least 5-year clinical experience) oncologists on the planning CECT images with reference to CET1-w MR or T2-w MR or PET if available. The GTVnp contours were reviewed and approved by a 15-year clinical experienced oncologist for RT treatment planning. In previous publications, the primary

nasopharyngeal tumor volume has shown its prognostic value[99,153] and predictive role in treatment response[47,152,153].

Follow-up and clinical endpoint. After completion of treatment, patients were assessed every three months during the first two years, every six months for third year, and annually thereafter. The information obtained was used to evaluate patient survival, patterns of relapse, and other clinical symptoms.

For the endpoint definition, the primary endpoint was DFS, defined as the time from starting RT until first progression (locoregional failure or distant failure) or death from any cause, whichever occurred first[155,156].OS was defined as the time from RT starting date until death from any cause. The endpoints of local-regional relapse-free survival (LRFS) and DMFS were determined by the patient's first relapse of a local or nodal tumor and the occurrence of distant metastasis.

Feature extraction. A total of 140 texture radiomic features were extracted from GTVnp on CET1-w MR images, with and without LoG filters. They are categorized by 14 GLDM, 24 GLCM, 16 GLRLM, and 16 GLSZM. Both the original and LoG-filtered images were preprocessed by a fixed 64 bin count discretization before feature extraction. Both image preprocessing and feature extraction were performed by PyRadiomics [30], which is compliant with IBSI[27]. The radiomic features was extracted from CET1-w MR images within the volume of GTVnp for radiation therapy. The GTVnp contours were propagated from CECT image to CET1-w MR with rigid registration using SimpleITK (2.1.1.2). The radiomic features was extracted from the original images and LoG filters (1mm, 3mm and 5mm as sigma value).

Predictive signature identification. We identified the repeatable heterogeneity signatures with high predictive value in CCRT+ACT treatment response based on the signature-treatment interaction. **Figure 5-5** lays out the signature discovery workflow. Firstly, the intra-class coefficient of correlation, ICC(1,1), was used to evaluate the feature robustness against random errors and inter-observer variabilities simulated by image perturbations and contour randomization[37,69,157]. Features with ICC(1,1) > 0.9 were considered repeatable and remained for further analysis. Secondly, features with variance less than $10^{-3}$ were removed. Third, we binarized the signature value by median and calculated the interaction between signature and treatment through multiplication. Features with significant association of itself and its treatment interaction to 3-year DFS in multivariate analysis remained as the final signatures.

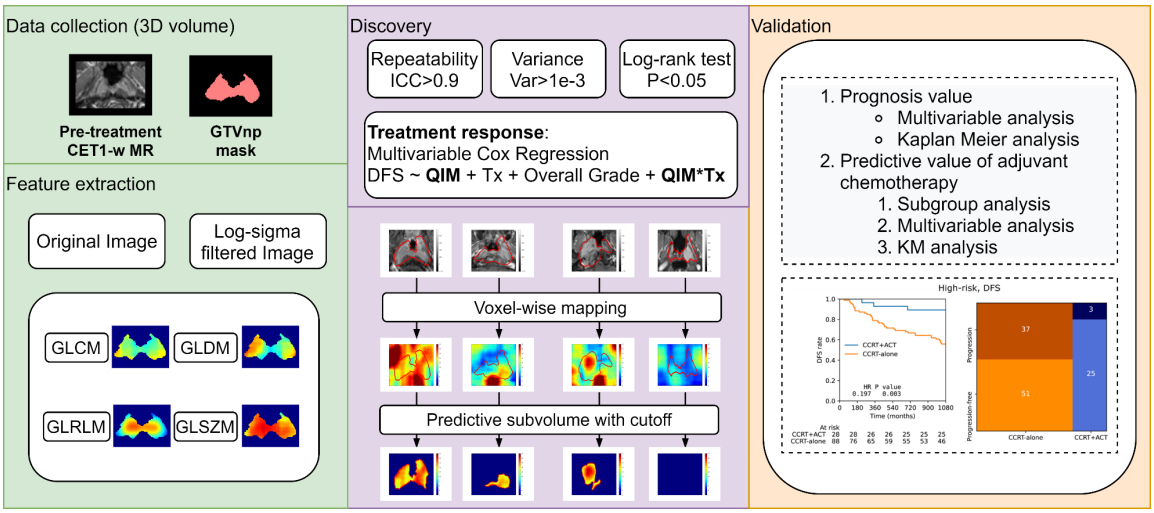


Figure 5-5. Study workflow.

Voxel-wise mapping and predictive subregion acquisition. The voxel-wised mappings of the identified heterogeneity signatures were performed on the entire patient cohort. Signature values were calculated for each voxel on a sliding window with kernel size of 21mm. Predictive subregions were acquired by thresholding the signature maps within GTVnp by the mean map value across the entire cohort. The relative volume of the prediction subregion, which are named predictive subvolume, was calculated and binarized by the median value across the entire cohort. It was proposed as the explainable decision-making signature. Source code of signature, heterogeneity map, and high heterogeneity subregion calculations can be found in our GitHub page for standardization and reproducibility (https://github.com/vivixinzhi/ACT_Decision_Making_With_CET1-wMR_Radiomic_Feature).

Survival analysis. Multiple survival analysis procedures were performed to assess the prognosis and predictive value of both heterogeneity signature and its predictive subregion from voxel-wise mapping. The prognostic value of the signature was evaluated on the training and validation set separately using univariate, multivariate, and KM analysis. To assess the predictive value of both heterogeneity signatures and predictive subvolume, subgroup analysis was performed on the entire cohort due to limited sample size. We evaluated and compared the prognostic value of the two treatments in each patient subgroup through univariate and multivariate analyses, followed by KM analysis were the curves and survival rates were reported.

Statistical analysis. Categorical variables were compared using the Chi-square or Fisher exact test. KM method was used to estimate the cumulative survival rates, and survival

curves were compared using the log-rank test. HRs with 95% CIs were calculated using the Cox proportional hazards model. Univariate and multivariate analyses with Cox proportional hazards models were performed to evaluate the independent significance of the treatments and other potential prognostic factors, including age, sex, and overall tumor stage. All tests were 2-sided, and a $p$-value < 0.05 was considered significant. Bonferroni multi-test correction is performed with the false discovery rate set as 0.05.

### 5.3.3. Results

Patient population. We included 232 patients with locally advanced NPC patients received CCRT (N = 177, 74.06%) or CCRT plus ACT (N = 62, 25.94%) between 2014-2016. Patient baseline characteristics in the discovery and validation cohort were listed in **Table 5-5**. No statistically significant differences were observed in patient demographics and tumor characteristics. There were no statistically significant differences between the two treatments regarding gender, N stage, T stage, and overall grade.

Table 5-5. Baseline characteristics comparison between the discovery and validation patients.

| **Characteristic** | **Discovery Cohort** | **Validation Cohort** | ***p*-value** |
|---|---|---|---|
| Total patient number | 128 | 104 | |
| Age | | | |
| Mean | 53.73 | 52.12 | 0.130 |

| | | | |
|---|---|---|---|
| Range | 30-73 | 19-75 | |
| Gender | | | |
| Male | 102 (79.69%) | 71(68.27%) | 0.056 |
| Female | 26(20.31%) | 33(31.73%) | |
| T stage | | | |
| T1 | 7(5.47%) | 4(3.85%) | 0.422 |
| T2 | 2(1.56%) | 6(5.77%) | |
| T3 | 105(82.03%) | 83(79.81%) | |
| T4 | 14(10.94%) | 11(10.58%) | |
| N stage | | | |
| N1 | 9(7.03%) | 5(4.81%) | 0.815 |
| N2 | 101(78.91%) | 85(81.73%) | |
| N3 | 18(14.06%) | 14(13.46%) | |
| Overall stage | | | |
| III | 96(75.00%) | 80(76.92%) | 0.842 |
| IV | 32(25.00%) | 24(23.08%) | |

| | | | |
|---|---|---|---|
| Treatment | | | |
| CCRT | 97(75.78%) | 77(74.04%) | 0.996 |
| CCRT+ACT | 31(24.22%) | 27(25.96%) | |

*Abbreviation: CCRT, concurrent chemoradiotherapy; ACT, adjuvant chemotherapy; Note: Stage was given according to the 7th edition of AJCC protocol for the validation cohort and switched to the 8th edition after 2017 for the discovery cohort.*

Discovery and validation of heterogeneity signature. Only one radiomic feature, gldm_DependenceVariance (gldm_DV), extracted from GTVnp on log-sigma filtered CET1-w image, with a median binarization cutoff of 7.10, was discovered and validated as robust, prognostic and predictive for 3-year DFS. **Figure 5-6** visualizes the CET1-w image, log-sigma filtered CET1-w image, gldm matrix and voxel-wised mapping for 8 random selected patients with T3N2M0 NPC. We observed that the patients with lower 3-year DFS tend to have larger gldm matrix (**Figure 5-6**, third column) as well as larger highlighted regions within the GTVnp on the voxel-wised gldm_DV map (**Figure 5-6**, fourth column). The highlighted tumor subregions tend to have more homogeneous appearance in the original image.

The signature values (**Figure 5-7**) in progression group and progression-free group were averaged to be 8.06 (range: 3.37 to 13.94) and 6.68 (range: 2.70 to 13.08), $P < 0.001$. The feature robustness index ICC was 0.97 (95CI: 0.96 – 0.98), indicating the excellent reliability of the feature measurement under image resampling and inter-observer GTVnp segmentation variability.

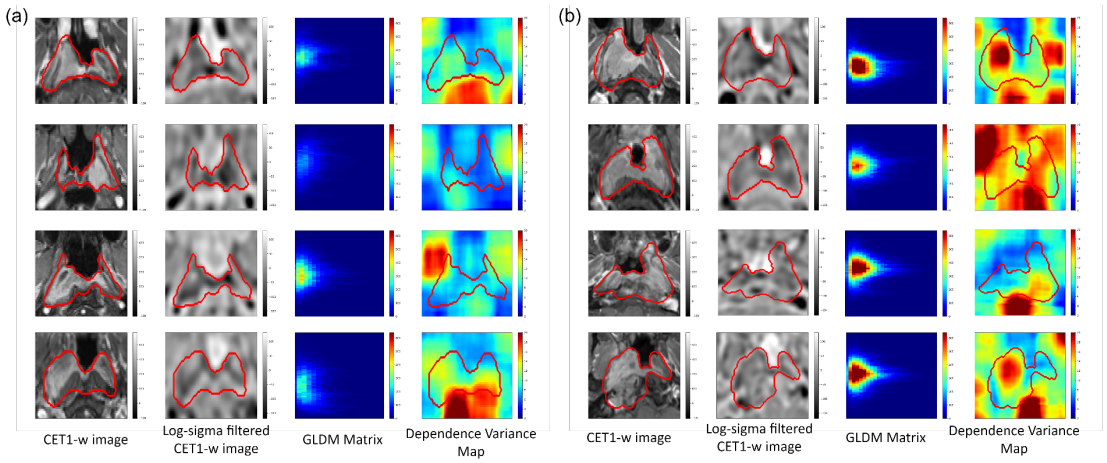


Figure 5-6. Voxel-wised intra-tumoral heterogeneity mapping (4th column) and the intermediate graphs for heterogeneity mapping.

*The first column are the CET1-w images, the second column are the log-sigma filtered CET1-w images, the third column shows the gldm matrix and the fourth column shows the selected quantitative image marker (Dependence variance). The contours in the red represent the GTVnp for radiation therapy. Plot (a) shows the T3N2M0 NPC patients without disease progression in three years and (b) shows the T3N2M0 patients with disease progression in three years. Eight patients with T3N2M0 were chosen randomly. The window and level were fixed across patients.*

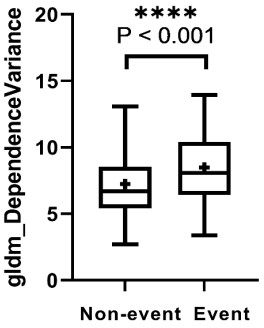


Figure 5-7. Box plot of gldm_DependenceVariance feature value distribution between the event and non-event group.

*Statistically larger feature values were observed for patient with three-year disease progression than the patients without.*

In the discovery cohort, patients with positive gldm_DV status (≥7.10) demonstrated higher progression risk than patients with negative status (<7.10). As shown in **Figure 5-8**, the high-risk patients achieved worse 3-year DFS rate (65% vs. 85%; OR, 3.12; 95CI, 1.33 – 7.35; *p*-value = 0.009) and LRFS rate (80% vs. 94%; OR, 4.00; 95CI, 1.21 – 13.17; *p*-value = 0.023). The univariate and KM analysis results suggest a high DFS HR (2.65; 95CI = 1.18 – 5.07) with a significant survival curve separation (*p*-value = 0.008). Similarly, the signature had a LRFS HR of 3.74 (95CI = 1.21 – 11.73) with log-rank test *p*-value of 0.014. No statistically significant association for 3-year OS and 3-year DMFS, were found in the discovery cohort (*p*-value > 0.05). In a multivariable cox regression on DFS (**Table 5-6**), the signature's HR was 2.20 (95CI: 1.04 – 4.63, P = 0.038) while the treatment (CCRT+ACT vs. CCRT-alone) did not demonstrate a statistically significant prognostic power (*p*-value = 0.221). For the treatment predictive value discovery (**Table 5-7**), the treatment-signature interaction was added to the multivariable cox regression model. The HR of the signature maintained in 3.34 (95CI: 1.38 – 8.09, *p*-value = 0.007), and the HR for treatment-signature interaction was 0.14 (95CI: 0.02 – 0.92, *p*-value = 0.040).

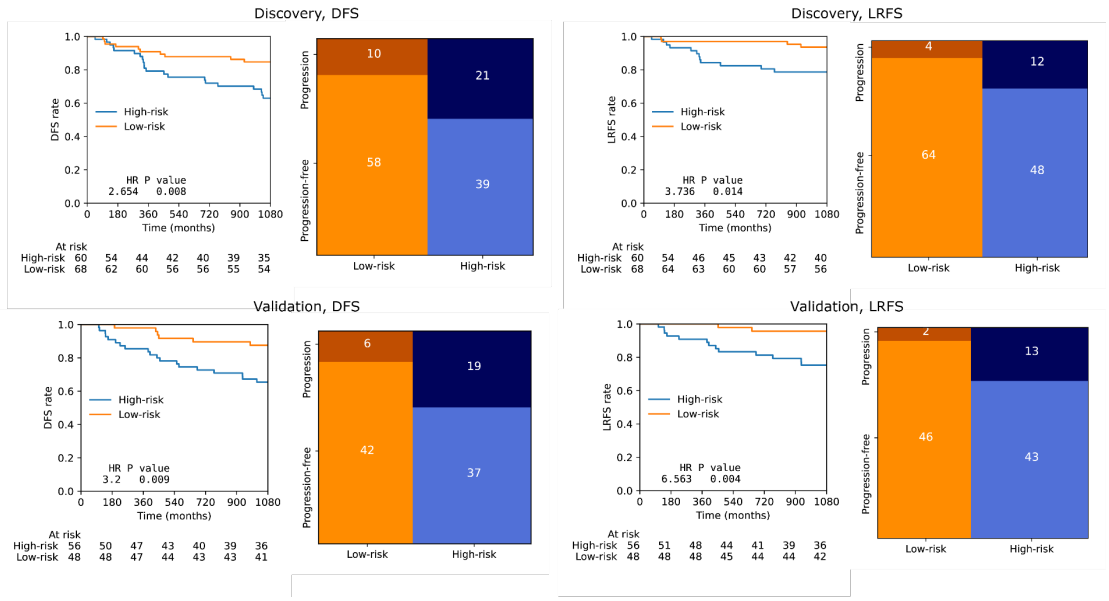


Figure 5-8. Kaplan Meier curves and mosaic plots (3-year event) on disease-free survival (DFS) and local-regional relapse-free survival (LRFS) of discovery and validation groups.

*The high-risk group patients, stratified by the predictive signature gldm_DependenceVariance (>7.10) showed higher progression risk with statistically lower DFS and LRFS rates than the low-risk patients in both discovery and validation cohort.*

Table 5-6. Multivariable stratified cox regression analysis for three-year disease-free survival for prognosis.

| **Variable** | **Discovery** | | **Validation** | |
|---|---|---|---|---|
| | **HR (95CI)** | ***p*-value** | **HR (95CI)** | ***p*-value** |
| Stage (IVA vs. III) | 2.43 (1.19-4.98) | **0.015** | 2.64 (1.18-5.90) | **0.018** |
| ACT (yes vs. no) | 0.57 (0.23-1.40) | 0.221 | 0.56 (0.19-1.64) | 0.293 |

| | | | | |
|---|---|---|---|---|
| gldm_DependenceVariance (>=7.10 vs. < 7.10) | 2.20 (1.04-4.63) | **0.038** | 2.66 (1.05-6.74) | 0.040 |

*Abbreviation: ACT, adjuvent chemotherapy; HR, hazard ratio; 95CI, 95% confidence interval. Note: Stage was given according to the 7th edition of AJCC protocol for the validation cohort and switched to the 8th edition after 2017 for the discovery cohort.*

Table 5-7. Multivariable Cox regression analysis of three-year disease-free survival (DFS) in the discovery and validation cohorts with the interaction term.

| **Variable** | **Discovery** | | **Validation** | |
|---|---|---|---|---|
| | **HR (95CI)** | ***P*-value** | **HR (95CI)** | ***P*-value** |
| Stage (IVA vs. III) | 2.30 (1.38-8.09) | **0.022** | 2.78 (1.23-6.25) | **0.014** |
| ACT (yes vs. nno) | 1.73 (0.51-5.92) | 0.381 | 2.51 (0.50-12.48) | 0.262 |
| gldm_DependenceVariance (>=7.10 vs. < 7.10) | 3.34 (1.38-8.09) | **0.007** | 4.82 (1.41-16.48) | **0.012** |
| Interaction (ACT=yes and DependenceVariance >= 7.10 vs. ACT=no or DependenceVariance < 7.10 | 0.14 (0.02-0.92) | **0.040** | 0.07 (0.01-0.91) | **0.042** |

*Abbreviation: ACT, adjuvant chemotherapy; HR, hazard ratio; 95CI, 95% confidence interval. Note: Stage was given according to the 7th edition of AJCC protocol for the validation cohort and switched to the 8th edition after 2017 for the discovery cohort. Note: p-values less than 0.05 were bolded.*

The findings on prognostic and predictive value of feature gldm_DV were consistent in the validation cohort. Patients with positive signature status achieved worse DFS compared to a patient with negative status (**Figure 5-8**) with 3-year survival rate of 67% vs. 87% (OR, 3.59; 95CI, 1.30 – 9.96; *p*-value = 0.014). The 3-year LRFS rate was 77% for the high-risk patients compared with 96% for the low-risk ones (OR, 6.95; 95CI, 1.48 – 32.62; *p*-value = 0.014). Univariate HRs of the signature were 3.20 (95CI = 1.28 – 8.01) and 6.56 (95CI = 1.48 – 29.08) and log-rank test *p*-values were 0.009 and 0.004 for DFS and LRFS, respectively. During multivariable cox regression (**Table 5-6**), the HR of the signature was 2.66 (95CI: 1.05 – 6.74, *p*-value = 0.040), and the treatment (CCRT+ACT vs. CCRT-alone) did not demonstrate a statistically significant prognostic power (*p*-value = 0.262). For the predictive value validation (**Table 5-7**), the HR of the signature and interaction was 4.82 (95CI: 1.41 – 16.48, *p*-value = 0.012) and 0.07 (95CI: 0.01 – 0.91, *p*-value = 0.042), respectively.

Table 5-8. Multivariable analysis on 3-year DFS for high-/low-risk patient subgroups stratified by the heterogeneity signature and predictive subvolume.

| | **High risk** | | **Low risk** | |
|---|---|---|---|---|
| | **HR (95CI)** | ***p*-value** | **HR (95CI)** | ***p*-value** |
| **Heterogeneity Signature** | | | | |
| Age (< 53 vs. ⩾ 53 years) | 0.77 (0.41 - 1.45) | 0.418 | 0.25 (0.08 - 0.76) | 0.002 |

| | | | | |
|---|---|---|---|---|
| Gender (female vs. male) | 1.55 (0.54 - 4.48) | 0.418 | 1.63 (0.51 - 5.14) | 0.408 |
| T stage (T3-4 vs. T1-2) | 1.15 (0.15 - 8.57) | 0.895 | 0.53 (0.11 - 2.60) | 0.437 |
| N stage (N2-3 vs. N1) | 0.60 (0.13 - 2.67) | 0.499 | 0.53 (0.11 - 2.60) | 0.437 |
| Overall stage (IVA vs. III) | 1.83 (0.95 - 3.53) | 0.072 | 5.19 (1.82 - 14.79) | 0.002 |
| Treatment (CCRT+ACT vs. CCRT alone) | 0.21 (0.06 - 0.68) | 0.009 | 1.28 (0.46 - 3.56) | 0.636 |
| **Predictive Subvolume** | | | | |
| Age (< 53 vs. ⩾ 53 years) | 0.52 (0.26-1.04) | 0.065 | 0.47 (0.18-1.22) | 0.120 |
| Gender (female vs. male) | 3.10 (0.71-13.58) | 0.133 | 1.30 (0.48-3.51) | 0.607 |
| T stage (T3-4 vs. T1-2) | 1.25 (0.37-4.20) | 0.712 | N/A | 0.996 |
| N stage (N2-3 vs. N1) | 0.19 (0.03-0.99) | 0.049 | 1.20 (0.25-5.66) | 0.821 |
| Overall stage (IVA vs. III) | 2.80 (1.40-5.57) | 0.003 | 2.57 (0.96-6.86) | 0.059 |

| | | | | |
|---|---|---|---|---|
| Treatment (CCRT+ACT vs. CCRT alone) | 0.27 (0.09-0.80) | 0.017 | 1.13 (0.35-3.61) | 0.836 |

*Note: Hazard ratio of N stage in the predictive sub-volume low risk group cannot be calculated because of no progression event for N1 patients. Abbreviations: HR, hazard ratio;95CI, 95% confidence interval.*

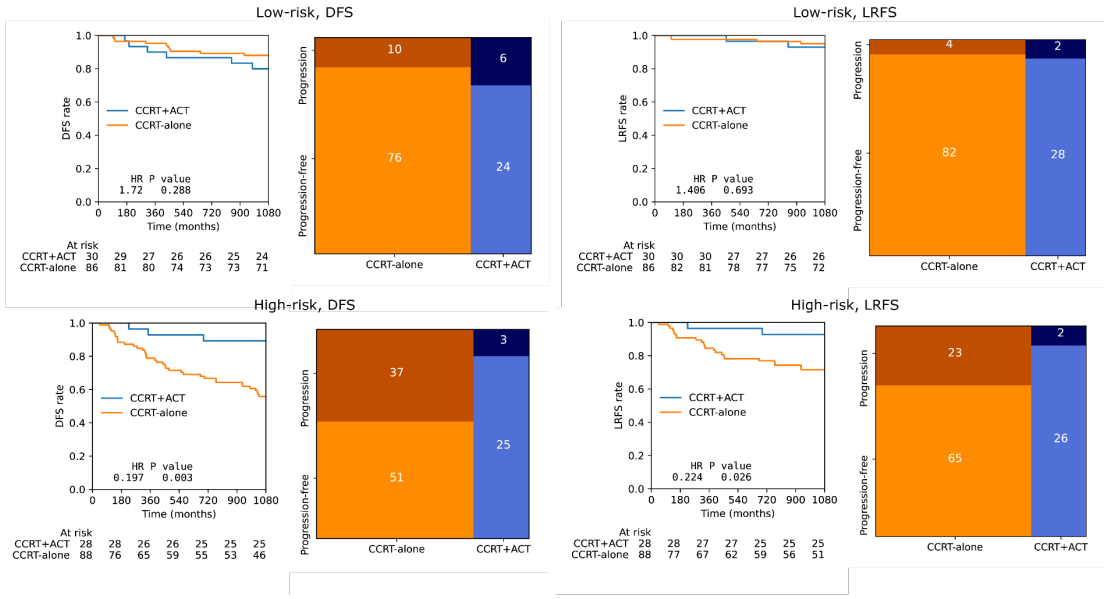


Figure 5-9. Kaplan Meier curves and mosaic plots of low-risk and high-risk patient groups stratified by the heterogeneity signature.

*Results on both 3-year disease-free survival (DFS) and local-regional free survival (LRFS) were reported. Among the high-risk group patients stratified by the heterogeneity signature gldm_DependenceVariance (>7.10), patients who underwent CCRT+ACT had significantly higher DFS and LRFS than patients how underwent CCRT alone However, minimum survival differences were found on the low-risk group for both survival endpoints.*

Subgroup analysis of the predictive value. For the whole cohort, all the clinical outcomes were comparable between CCRT+ACT and CCRT alone groups. Within the high-risk

group identified by the positive status of the signature gldm_DV, the HR of CCRT+ACT vs. CCRT-alone was 0.21 (95CI: 0.06 – 0.68, *p*-value = 0.009) in multivariate analysis on DFS (**Table 5-8**). During KM analysis, patients who received CCRT+ACT achieved a better 3-year DFS (**Figure 5-9(b)**) with the survival rate of 90% versus 57% (HR, 0.20; 95CI, 0.06 – 0.64; *p*-value = 0.007) and 3-year LRFS with 93% versus 72% (HR, 0.22; 95CI, 0.05 – 0.94; *p*-value = 0.042). On the other hand, no statistically significant difference was observed between patients treated with CCRT+ACT and CCRT-alone (*p*-value > 0.05) in the low-risk group (**Figure 5-10**).

Similar patterns were found for the patient stratification determined by the predictive tumor subvolume derived from the voxel-wise heterogeneity mapping. The cohort mean of the heterogeneity map, which was 12, was used to generate the predictive tumor subvolume within PTV, and the relative volume threshold of 0.25 was used to binarize the predictive subvolume into high- and low-risk groups. As listed in **Table 5-8**, the multivariate DFS HR of CCRT+ACT vs. CCRT-alone was 0.27 (95CI: 0.09 – 0.80, *p*-value = 0.017) on the high-risk group, while minimum prognostic was found on the low-risk group (HR = 0.73, 95% CI = 0.22-2.36, *p*-value = 0.601). KM analysis also suggested statistically significant separations of survival curves between patient receiving CCRT+ACT and CCRT alone in the high-risk group (DFS: HR=0.36, *p*-value=0.025; LRFS: HR=0.25, *p*-value=0.042) while minimum significant was found in the low-risk group (DFS: HR=1.05, *p*-value=0.930; LRFS: HR=0.86, *p*-value=0.845).

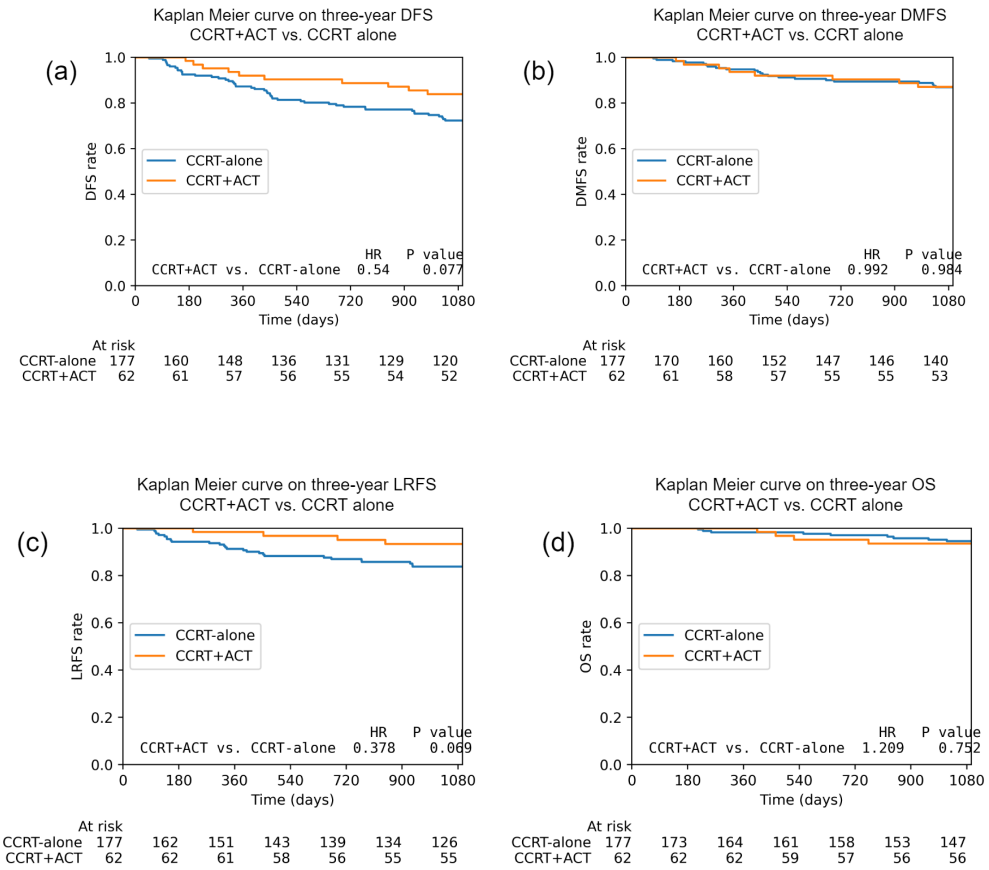


Figure 5-10. Kaplan Meier curves comparing all patients receiving concurrent chemoradiotherapy (CCRT)+adjuvant chemotherapy (ACT) vs. CCRT alone.

*No statistically significant differences were observed in (a) three-year DFS, (b) three-year DMFS, (c) three-year LRFS, and (d) three-year OS.*

### 5.3.4. Discussions

This study successfully discovered and validated an independent prognostic and predictive image-based tumor heterogeneity signature, gldm_DV, for locally advanced NPC patients in a retrospectively collected cohort. The acquisition of the tumor heterogeneity signature only requires the CET1-w images and the GTVnp contour from RT planning, rendering it non-invasive, economic, and fully automatic. The predictive

value of the signature was also confirmed by the tumor subvolume derived from voxel-wised mapping, which is a direct visualization and explanation of the signature. Our results suggest that patients with the positive status of the image marker gldm_DV as well as the relative size of the predictive subvolume could benefit more in 3-year DFS and 3-year LRFS when receiving adjuvant chemotherapy.

The clinical question of the adjuvant chemotherapy decision is still in debating after the landmark intergroup 0099 trial[158] establishing the chemo-radiation therapy as the standard treatment for local advanced NPC. Several network meta-analyses[159–161] reported potential survival benefit in CCRT+ACT compared with CCRT alone, but the differences were not statistically significant. Routine usage of adjuvant chemotherapy for all advanced NPC likely represents an over-treatment. This approach results in high cumulative cisplatin exposure, which could lead to irreversible long-term complications such as peripheral neuropathy, renal impairment, and ototoxicity, hence severely impair the quality of life of survivors. Therefore, novel biomarker guided ACT treatment selection is in demand to identify patients who may benefit from ACT after CCRT. Early attempt in the use of post-treatment plasma EBV-DNA level to risk-stratify NPC patients for ACT has not been proven useful[162]. Further clinical trials, such as the NRG-HN001, are ongoing to explore this approach. In our study, the single radiomic feature, gldm_DV, was externally validated to be independent prognostic factor in addition to overall stage (Spearman r = 0.16, $p$-value < 0.05). Significant DFS and LRFS benefits from the addition of ACT were observed among patients who demonstrated positive level of the features. However, the benefit in 3-year OS and DMFS were not observed. There are two potential reasons. First, our image signature quantifies the status of the primary

nasopharyngeal tumor, but the status of the lymph node involvement was not considered. The primary tumor characteristics, such as orbital or intracranial invasion, were reported as independent prognostic factors for local failure, whereas the N stage-related factors such as retropharyngeal lymph node involvement and cervical lymph node involvement were reported as the independent prognostic factor for distant metastasis[163]. As the image marker was quantified within the primary nasopharyngeal tumor, it is more likely that the quantified image phenotype was associated with the recurrence of the tumor. Secondly, the adjuvant chemotherapy is likely to offer local tumor control benefit than distant control. A network meta-analysis which compared various chemotherapy sequences in localized NPC[159] reported ACT following CCRT ranked the highest in LRFS. Therefore, it is more likely to identify a subgroup of patients who were benefit ACT in terms of local and regional control.

During predictive signature identification, we emphasized repeatability as the first-line selection criteria to minimize the chance of false discovery in the following observations. Previous studies have shown that using repeatable radiomic-based image markers could improve both robustness and the generalizability of the model predictions. In our observations, the prognostic value of the final identified signature gldm_DV shared statistically significant multi-variable HRs in both the discovery and validation cohort. The reproducibility of gldm_DV in image acquisition phase has also been confirmed before with an ICC of 0.933 under different scanners, image reconstruction parameters, and contrast medium[122]. It has also been suggested to be intrinsically reproducible in different image preprocessing settings[164]. Overall, we believe the proposed tumor heterogeneity signature gldm_DV is repeatable and reproducible in each

step of the signature acquisition workflow. However, the robustness of the prognostic and predictive values requires further validation in inter-group and inter-institutional settings.

This study confirmed the applicability of the voxel-wise mapping of the tumor heterogeneity signature gldm_DV in ACT decision making using the highlighted tumor subvolume. Although the global signature gldm_DV has been validated to be predictive, whether the predictive value can still retain locally in different tumor sublocations is another important measure of reliability. In this study, we used the relative volume of the highlighted region of gldm_DV voxel-wise mapping within GTV (predictive subvolume) to locally assess the predictive value of the signature. This assessment also avoids the bias from volume-confounding effect[29] as a fixed kernel size is used for local signature calculations. The example patients demonstrated a strong correlation between global signature value and predictive subvolume, indicating a high consistency of global and local tumor heterogeneity pattern (**Figure 5-6**). The survival analysis results suggests that the predictive subvolume demonstrated high predictive value in subgroup analysis where the high-risk grouped demonstrated better DFS and LRFS under the treatment of CCRT+ACT, suggesting a high reliability in ACT efficacy prediction (**Figure 5-11**). Moreover, the voxel-wise tumor heterogeneity mapping also serves as a direct visualization and explanation of the signature which can be used more confidently in clinical practice (**Figure 5-6**).

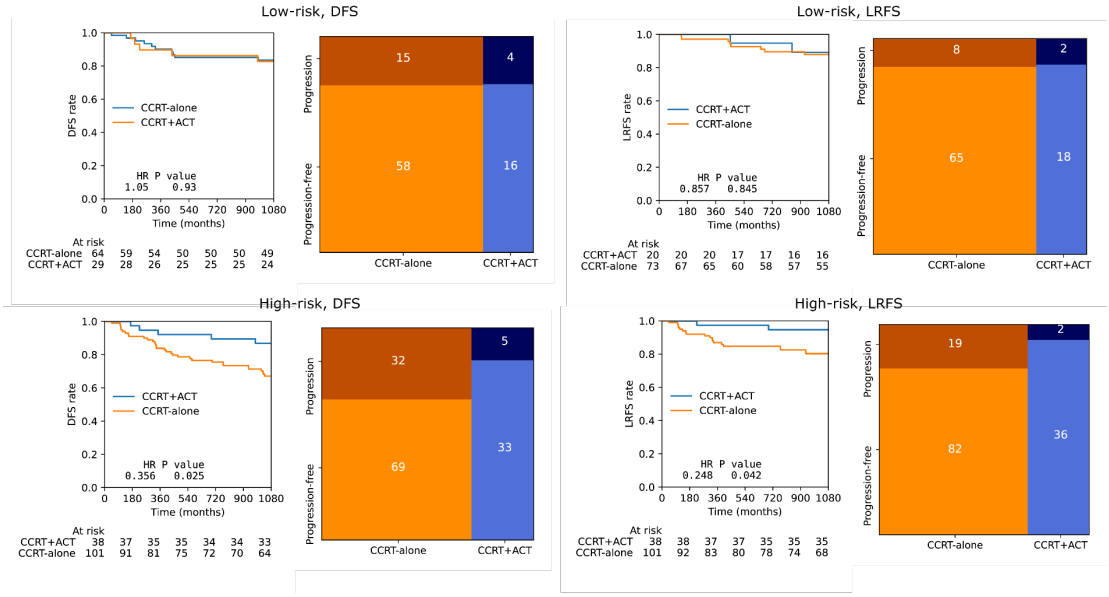


Figure 5-11. Kaplan Meier curves and mosaic plots of low-risk and high-risk patient groups stratified by the predictive tumor subvolume.

*Results on both 3-year disease-free survival (DFS) and local-regional free survival (LRFS) were reported. Patients were stratified by the predictive tumor subvolume based on signature mapping (map threshold = 12, subvolume threshold = 0.25) The high-risk group patients showed statistically significant DFS and LRFS rate differences between CCRT+ACT and CCRT. However, minimum survival differences were found on the low-risk group for both survival endpoints.*

This study has its limitations. First, the treatment efficacy evaluation was conducted on a retrospectively collected cohort, which it is difficult to control the treatment efficacy related confounding factors, such as the adjuvant chemotherapy preference by physicians and patients. Nevertheless, the efficacy analysis on the high-risk and low-risk patients provided extra credit in the potential clinical value of our identified image marker. Secondly, a larger validation cohort is required to validate the prognosis value and predictive value of identified signature and its voxel-wise mapping for better understanding the performance under different clinical settings. Furthermore, validations

from different groups and institutions are still necessary. Thirdly, the EBV-DNA was not included in the analysis due to limited data, therefore, we cannot rule out the correlation between identified image signature and EBV-DNA, and further analysis is demanded.

### 5.3.5. Conclusion

We discovered and validated a robust, prognostic, and predictive quantitative tumor heterogeneity signature, gldm_DependenceVariance, on CET1-w MR images for local-regionally advanced NPC patients. Patients with larger signature values were identified as high-risk group, which was associated with poorer 3-year DFS and LRFS, and benefited from the addition of ACT to CCRT-alone. Meanwhile, the high-risk patients with bigger highlighted regions on the voxel-wised feature mapping were also observed to derive more benefit from ACT. The proposed signature could be a potential candidate for reliable and explainable non-invasive image biomarker for ACT decision making in locoregionally advanced NPC.

# Chapter 6. Summary

This thesis presents the development and clinical application of an end-to-end RADAR toolkit for quantitative biomarker development for prediction medicine. The third chapter introduces the design and algorithms of RADAR, with emphasis on the innovative features that enhance the curation and analysis of radiotherapy data. First, we developed our own radiotherapy data processing algorithm to ensure reliable data transcoding. Using a semi-automatic data curation strategy, we demonstrated that the efficiency and accuracy of data curation can be greatly improved by batch data query at the cohort level, followed by patient-specific data selection guided by embedded data visualization. We also developed batch feature extraction capabilities from the curated radiotherapy data for maximizing efficiency. The RADAR toolkit also includes a user-friendly GUI dashboard, enabling real-time monitoring of each feature extraction task for fast trouble-shooting as needed. Additionally, we developed an extended feature set that provides a comprehensive description of multi-model radiotherapy datasets. Specifically, we integrated new histogram features based on the distance and angular projection distributions, which quantitatively characterizes geometric relationships between structure delineations. Lastly, we implemented an image perturbation-based feature repeatability assessment method and integrated it into RADAR, which is further optimized by parallel computation for optimal speed.

The fourth chapter presents two technical studies that shed light on the patterns of RF repeatability and their impact on the generalizability of prognostic models under stochastic patient positioning. In the first study, we analyzed different image modalities

and cancer subtypes of HNC to identify patterns of RF repeatability. We found that the majority of the shape, unfiltered, and LoG-filtered RFs demonstrated high repeatability (ICC$\geqslant$0.9), whereas more than half of the wavelet-filtered RFs displayed poor repeatability. Further analysis of NPC MR images revealed heterogeneous patterns of RF repeatability under different image filters, discretization bin numbers, and feature classes. Specifically, texture RFs extracted from wavelet-filtered images showed the lowest repeatability to image perturbation, which was consistent with the results of the first study. During the development of the survival model, we found that low-repeatable RFs achieved similar prognostic performance as high-repeatable ones during training, but resulted in non-significant survival stratification on external validation.

Chapter five includes two clinical studies focusing on the development of clinical biomarkers for prognosis and treatment efficacy predictions of NPC patients. In the first study, we utilized newly designed geometric features between GTVn and OARs, discovering two independent prognostic factors based on spinal cord and parotids. This led to improved LN risk stratifications in both the discovery and external validation cohort. In the second study, we discovered an independent prognostic and predictive image-based signature, gldm_DV, from CET1-w MR for locally advanced NPC patients. It describes intra-tumor heterogeneity, and the subvolume identified with heterogeneous CCRT+ACT mapping also demonstrated strong predictive value.

This thesis has some limitations that require further investigation and development. Despite the highly structured DICOM format for data encoding, different naming conventions of the radiotherapy data, such as ROI names, imposes challenges for external sharing and collaborations. Development and integration of automatic naming

standardization following the protocol provided by TG report 263 into the data curation module could be one solution to further enhance the usability and impact of RADAR toolkit. When using RADAR to perform feature extraction, less-experienced users may find it challenging to configure the settings using the parameter file due to its complexity. Therefore, future RADAR development should focus on migrating to GUI-based settings to improve user-friendliness during configuration. Another limitation is that the perturbation-based feature repeatability assessment only considers translation, rotation, and contour randomization. As a result, we will conduct research to simulate image noise using multiple models proposed for different image modalities. The effects of noise addition on feature repeatability and model generalizability will also be studied. Furthermore, we only analyzed RFs in the two repeatability studies, and other feature categories such as dosiomics and geometric features need to be comprehensively assessed for their susceptibility to perturbations and their impact on model performance. Regarding biomarker developments, the sample size and follow-up durations may limit our results' clinical significance. We will expand our data collection efforts to include data from more institutions and externally validate and fine-tune our established image signatures for potential clinical utility.

In conclusion, RADAR can be a valuable tool for efficiency, comprehensive, and reliable radiotherapy data analysis and predictive model development. We urge caution when handling high-pass wavelet-filtered RFs and advise exclusion of low-repeatable RFs for predictive model development to safeguard generalizability. The newly designed anatomical features may facilitate the discovery of new anatomical characteristics in a more holistic and precise way, thereby improving patient prognosis predictions.

Additionally, the image signature gldm_DV, which we discovered, could serve as a valuable biomarker for reliable and explainable ACT decision-making for patients with locoregionally advanced NPC.

# Appendix A: Image Acquisition and Contouring Protocols

## Nasopharyngeal Carcinoma Cohort

For CECT images acquisition, each patient was immobilized in a supine position with a thermoplastic cast. Intravenous (IV) contrast-enhanced CT simulation was performed at 3-mm intervals from the vertex to 5-cm below the sternoclavicular notch with a 16-slice Brilliance Big Bore CT (Philips Medical Systems, Cleveland, OH). For patients to be treated under the Tomotherapy machine, the field-of-view (FOV) was set to be sufficiently large to cover the CT couch with a 2-cm margin (i.e., FOV>570-mm). CECT acquisition parameters were as follows: scan mode: Helical, voltage = 120-kVp, X-ray tube current = 264-mA, exposure = 325-msec, pixel spacing = 1.152x1.152 mm, slice thickness = 3-mm, matrix = 512x512 pixels. Two types of IV contrast agents were available: (i) OMNIPAQUE TM 350 mg I/ml and (ii) VISIPAQUE TM 320 mg I/ml; either one of them was prescribed to each eligible patient and was injected at a rate of 2-ml/sec for 70-ml, followed by scanning after a 30-sec delay.

For MR images, all patients were scanned with 1.5-T MRI (Avanto, Siemens, Germany). Acquisition parameters for T2-w MR images were as follows: axial T2-w using short-tau-inversion-recovery (STIR) MR sequence (repetition time [TR]/ echo time [TE]: 7640/97 ms, field-of-view [FOV] = 24 x 24 cm, number of acquisition = 1, slice

thickness = 4 mm x 25 slices, spacing: 0.75mm x 0.75mm x 4.4mm, matrix: 320). Scanning parameters for CET1-w MR images acquisition were as follows: contrast agent (gadolinium-based Dotarem, Gd-DOTA); axial CET1-w spin-echo MR sequence (repetition time [TR]/ echo time [TE]: 739/17 ms, field-of-view [FOV] = 24 x 24 cm, number of acquisition = 1, slice thickness = 3 mm x 48 slices, spacing: 0.938mm x 0.938mm x 3.3mm, matrix: 256).

All the GTV contours were manually delineated slice-by-slice by oncologists specialized in head-and-neck cancer with accreditations using Eclipse Aria 13 (Varian Medical Systems). The GTVp and GTVn were determined from the imaging, clinical and endoscopic findings.

## Online Oropharyngeal Carcinoma Cohort

Data was downloaded from https://doi.org/10.7937/tcia.2020.2vx6-fy46. CECT acquisitions involved scanners from various manufacturers, a couple of tube voltage settings, and multiple image resolutions. Details are tabulated in **Table 6-1**. The contouring of the primary gross-tumor-volumes follows the protocol specified in ICRU62/83 using VelocityAI 3.0.1, according to the original publication [165].

Table A-1. Image acquisition parameters for the online Oropharyngeal Carcinoma cohort

| **Image acquisition parameter** | **Parameter values** | **Patient number** |
|---|---|---|
| Scanner | Aquilion | 9 |

| | | |
|---|---|---|
| | BrightSpeed | 2 |
| | Brilliance16 | 3 |
| | Brilliance 16P | 1 |
| | CT/e | 2 |
| | Definition | 1 |
| | Discovery CT750 HD | 8 |
| | ECLOS | 1 |
| | Emotion 16 | 1 |
| | Emotion 6 | 1 |
| | Gemini TF 64 | 1 |
| | LightSpeed Plus | 38 |
| | LightSpeed Pro 16 | 1 |
| | LightSpeed QX/i | 15 |
| | LightSpeed Ultra | 1 |
| | LightSpeed VCT | 90 |
| | LightSpeed16 | 217 |

| | | |
|---|---|---|
| | Mx8000 IDT 16 | 2 |
| | Sensation 16 | 2 |
| | Sensation 64 | 2 |
| | Volume Zoom | 1 |
| KVP | 120 | 205 |
| | 130 | 2 |
| | 135 | 1 |
| | 191 | 140 |
| Exposure time (msec) | 400-1921 | 399 |
| Tube current (mA) | 61-599 | 399 |
| Exposure (mAs) | 0-21103 | 399 |
| Slice thickness (mm) | 1, 1.25 | 345 |
| | 2, 2.5 | 18 |
| | 3, 3.75 | 33 |
| | 5 | 3 |
| Pixel spacing (mm) | 0.34-0.63 | 398 |

| | | |
|---|---|---|
| | 0.98 | 1 |
| Matrix | 512x512 | 399 |

# References


1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-249. doi:10.3322/caac.21660

2. Jameson JL, Longo DL. Precision Medicine — Personalized, Problematic, and Promising. *N Engl J Med*. 2015;372(23):2229-2234. doi:10.1056/NEJMsb1503104

3. Quon H, McNutt T, Lee J, et al. Needs and Challenges for Radiation Oncology in the Era of Precision Medicine. *International Journal of Radiation Oncology*Biology*Physics*. 2019;103(4):809-817. doi:10.1016/j.ijrobp.2018.11.017

4. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181

5. Lee AWM, Ng WT, Chan LLK, et al. Evolution of treatment for nasopharyngeal cancer – Success and setback in the intensity-modulated radiotherapy era. *Radiotherapy and Oncology*. 2014;110(3):377-384. doi:10.1016/j.radonc.2014.02.003

6. Chan C, Lang S, Rowbottom C, Guckenberger M, Faivre-Finn C. Intensity-Modulated Radiotherapy for Lung Cancer: Current Status and Future Developments. *Journal of Thoracic Oncology*. 2014;9(11):1598-1608. doi:10.1097/JTO.0000000000000346

7. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiotherapy and Oncology*. 2014;110(2):370-374. doi:10.1016/j.radonc.2013.11.001

8. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol*. 2013;108(1):174-179. doi:10.1016/j.radonc.2012.09.019

9. Ajami S, BagheriTadi T. Barriers for Adopting Electronic Health Records (EHRs) by Physicians. *Acta Inform Med*. 2013;21(2):129. doi:10.5455/aim.2013.21.129-134

10. Jaffray DA. Image-guided radiotherapy: from current concept to future perspectives. *Nat Rev Clin Oncol*. 2012;9(12):688-699. doi:10.1038/nrclinonc.2012.194

11. Al-Sukhni E, Milot L, Fruitman M, et al. Diagnostic Accuracy of MRI for Assessment of T Category, Lymph Node Metastases, and Circumferential Resection

Margin Involvement in Patients with Rectal Cancer: A Systematic Review and Meta-analysis. *Ann Surg Oncol*. 2012;19(7):2212-2223. doi:10.1245/s10434-011-2210-5

12. Brunt JNH. Computed Tomography–Magnetic Resonance Image Registration in Radiotherapy Treatment Planning. *Clinical Oncology*. 2010;22(8):688-697. doi:10.1016/j.clon.2010.06.016

13. Liu C, Li M, Xiao H, et al. Advances in MRI-guided precision radiotherapy. *Precision Radiation Oncology*. 2022;6(1):75-84. doi:10.1002/pro6.1143

14. Padhani AR, Koh DM. Diffusion MR Imaging for Monitoring of Treatment Response. *Magnetic Resonance Imaging Clinics*. 2011;19(1):181-209. doi:10.1016/j.mric.2010.10.004

15. Bomanji J, Costa D, Ell P. Clinical role of positron emission tomography in oncology. *The Lancet Oncology*. 2001;2(3):157-164. doi:10.1016/S1470-2045(00)00257-6

16. Pianykh OS. What Is DICOM? In: Pianykh OS, ed. *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide*. Springer; 2012:3-5. doi:10.1007/978-3-642-10850-1_1

17. Law MYY, Liu B. DICOM-RT and Its Utilization in Radiation Therapy. *RadioGraphics*. 2009;29(3):655-667. doi:10.1148/rg.293075172

18. Matuszak M, Moran J, Xiao Y, et al. SU-E-P-22: AAPM Task Group 263 Tackling Standardization of Nomenclature for Radiation Therapy. *Medical Physics*. 2015;42(6Part4):3231-3231. doi:10.1118/1.4923956

19. Kikinis R, Pieper SD, Vosburgh KG. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In: Jolesz FA, ed. *Intraoperative Imaging and Image-Guided Therapy*. Springer; 2014:277-289. doi:10.1007/978-1-4614-7657-3_19

20. Wolf I, Vetter M, Wegner I, et al. The Medical Imaging Interaction Toolkit. *Medical Image Analysis*. 2005;9(6):594-604. doi:10.1016/j.media.2005.04.005

21. Deasy JO, Blanco AI, Clark VH. CERR: A computational environment for radiotherapy research. *Medical Physics*. 2003;30(5):979-985. doi:10.1118/1.1568978

22. Davatzikos C, Rathore S, Bakas S, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *JMI*. 2018;5(1):011018. doi:10.1117/1.JMI.5.1.011018

23. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Res*. 2018;78(16):4786-4789. doi:10.1158/0008-5472.CAN-18-0125

24. Strimbu K, Tavel JA. What are biomarkers? *Current Opinion in HIV and AIDS*. 2010;5(6):463-466. doi:10.1097/COH.0b013e32833ed177

25. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology*. 2017;14(3):169-186. doi:10.1038/nrclinonc.2016.162

26. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2015;278(2):563-577. doi:10.1148/radiol.2015151169

27. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020;295(2):328-338. doi:10.1148/radiol.2020191145

28. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381

29. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*. 2019;130:2-9. doi:10.1016/j.radonc.2018.10.027

30. Griethuysen JJM van, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

31. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports*. 2017;7(1):10117. doi:10.1038/s41598-017-10371-5

32. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*. 2016;2(4):361-365. doi:10.18383/j.tom.2016.00208

33. Gourtsoyianni S, Doumou G, Prezzi D, et al. Primary Rectal Cancer: Repeatability of Global and Local-Regional MR Imaging Texture Features. *Radiology*. 2017;284(2):552-561. doi:10.1148/radiol.2017161375

34. Fiset S, Welch ML, Weiss J, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and Oncology*. 2019;135:107-114. doi:10.1016/j.radonc.2019.03.001

35. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*. 2018;102(4):1143-1158. doi:10.1016/j.ijrobp.2018.05.053

36. Zhang J, Lam SK, Teng X, et al. Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients. *Radiotherapy and Oncology*. 2023;183:109578. doi:10.1016/j.radonc.2023.109578

37. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. *Scientific Reports*. 2019;9(1):1-10. doi:10.1038/s41598-018-36938-4

38. Zhang J, Lam S, Teng X, et al. Repeatability of Radiomic Features Against Simulated Scanning Position Stochasticity Across Imaging Modalities and Cancer Subtypes: A Retrospective Multi-institutional Study on Head-and-Neck Cases. In: Qin W, Zaki N, Zhang F, Wu J, Yang F, eds. *Computational Mathematics Modeling in Cancer Analysis*. Vol 13574. Lecture Notes in Computer Science. Springer Nature Switzerland; 2022:21-34. doi:10.1007/978-3-031-17266-3_3

39. Zhang J, Teng X, Lam S, et al. Quantitative Spatial Characterization of Lymph Node Tumor for N Stage Improvement of Nasopharyngeal Carcinoma Patients. *Cancers*. 2022;15(1):230. doi:10.3390/cancers15010230

40. Wang SJ, Fuller CD, Kim JS, Sittig DF, Thomas CR, Ravdin PM. Prediction Model for Estimating the Survival Benefit of Adjuvant Radiotherapy for Gallbladder Cancer. *JCO*. 2008;26(13):2112-2117. doi:10.1200/JCO.2007.14.7934

41. Wang SJ, Lemieux A, Kalpathy-Cramer J, et al. Nomogram for Predicting the Benefit of Adjuvant Chemoradiotherapy for Resected Gallbladder Cancer. *JCO*. 2011;29(35):4627-4632. doi:10.1200/JCO.2010.33.8020

42. Langius JAE, Twisk J, Kampman M, et al. Prediction model to predict critical weight loss in patients with head and neck cancer during (chemo)radiotherapy. *Oral Oncology*. 2016;52:91-96. doi:10.1016/j.oraloncology.2015.10.021

43. Deist TM, Dankers FJWM, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics*. 2018;45(7):3449-3459. doi:10.1002/mp.12967

44. Zhu X, Dong D, Chen Z, et al. Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. *Eur Radiol*. 2018;28(7):2772-2778. doi:10.1007/s00330-017-5221-1

45. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*. 2014;5(1):4006. doi:10.1038/ncomms5006

46. Huang Y, Liu Z, He L, et al. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. *Radiology*. 2016;281(3):947-957. doi:10.1148/radiol.2016152234

47. Peng H, Dong D, Fang MJ, et al. Prognostic Value of Deep Learning PET/CT-Based Radiomics: Potential Role for Future Individual Induction Chemotherapy in Advanced Nasopharyngeal Carcinoma. *Clin Cancer Res*. 2019;25(14):4271-4279. doi:10.1158/1078-0432.CCR-18-3065

48. Desideri I, Loi M, Francolini G, Becherini C, Livi L, Bonomo P. Application of Radiomics for the Prediction of Radiation-Induced Toxicity in the IMRT Era: Current State-of-the-Art. *Front Oncol*. 2020;10:1708. doi:10.3389/fonc.2020.01708

49. Carbonara R, Bonomo P, Di Rito A, et al. Investigation of Radiation-Induced Toxicity in Head and Neck Cancer Patients through Radiomics and Machine Learning: A Systematic Review. Ahmad N, ed. *Journal of Oncology*. 2021;2021:1-9. doi:10.1155/2021/5566508

50. van Dijk LV, Thor M, Steenbakkers RJHM, et al. Parotid gland fat related Magnetic Resonance image biomarkers improve prediction of late radiation-induced xerostomia. *Radiotherapy and Oncology*. 2018;128(3):459-466. doi:10.1016/j.radonc.2018.06.012

51. Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. *Frontiers in Oncology*. 2020;10. Accessed February 26, 2022. https://www.frontiersin.org/article/10.3389/fonc.2020.00790

52. Lyman JT. Complication Probability as Assessed from Dose-Volume Histograms. *Radiation Research Supplement*. 1985;8:S13-S19. doi:10.2307/3583506

53. Dean J, Wong K, Gay H, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. *Clinical and Translational Radiation Oncology*. 2018;8:27-39. doi:10.1016/j.ctro.2017.11.009

54. Buettner F, Miah AB, Gulliford SL, et al. Novel approaches to improve the therapeutic index of head and neck radiotherapy: An analysis of data from the PARSPORT randomised phase III trial. *Radiotherapy and Oncology*. 2012;103(1):82-87. doi:10.1016/j.radonc.2012.02.006

55. Bourbonne V, Da-ano R, Jaouen V, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. *Radiotherapy and Oncology*. 2021;155:144-150. doi:10.1016/j.radonc.2020.10.040

56. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. *Front Oncol*. 2018;8. doi:10.3389/fonc.2018.00035

57. Chopra N, Dou T, Sharp G, Sajo E, Mak RH. A Combined Radiomics-Dosiomics Machine Learning Approach Improves Prediction of Radiation Pneumonitis

Compared to DVH Data in Lung Cancer Patients. *International Journal of Radiation Oncology, Biology, Physics*. 2020;108(3):e777. doi:10.1016/j.ijrobp.2020.07.231

58. Lee SH, Han P, Hales RK, et al. Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy. *Phys Med Biol*. 2020;65(19):195015. doi:10.1088/1361-6560/ab8531

59. Wu A, Li Y, Qi M, et al. Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases. *Oral Oncology*. 2020;104:104625. doi:10.1016/j.oraloncology.2020.104625

60. Pota M, Scalco E, Sanguineti G, et al. Early prediction of radiotherapy-induced parotid shrinkage and toxicity based on CT radiomics and fuzzy classification. *Artificial Intelligence in Medicine*. 2017;81:41-53. doi:10.1016/j.artmed.2017.03.004

61. Chen S, Tang Y, Li N, et al. Development and Validation of an MRI-Based Nomogram Model for Predicting Disease-Free Survival in Locally Advanced Rectal Cancer Treated With Neoadjuvant Radiotherapy. *Front Oncol*. 2021;11:784156. doi:10.3389/fonc.2021.784156

62. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. *Radiology*. 2015;277(3):813-825. doi:10.1148/radiol.2015142202

63. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Translational Oncology*. 2014;7(1):72-87. doi:10.1593/tlo.13844

64. Lu H, Parra NA, Qi J, et al. Repeatability of Quantitative Imaging Features in Prostate Magnetic Resonance Imaging. *Front Oncol*. 2020;10:551. doi:10.3389/fonc.2020.00551

65. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361

66. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. *Translational Oncology*. 2015;8(6):524-534. doi:10.1016/j.tranon.2015.11.013

67. Larue RTHM, Van De Voorde L, van Timmeren JE, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. *Radiotherapy and Oncology*. 2017;125(1):147-153. doi:10.1016/j.radonc.2017.07.023

68. Lafata K, Cai J, Wang C, Hong J, Kelsey CR, Yin FF. Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology. *Phys Med Biol*. 2018;63(22):225003. doi:10.1088/1361-6560/aae56a

69. Teng X, Zhang J, Zwanenburg A, et al. Building reliable radiomic models using image perturbation. *Sci Rep*. 2022;12(1):10035. doi:10.1038/s41598-022-14178-x

70. Teng X, Zhang J, Ma Z, et al. Improving radiomic model reliability using robust features from perturbations for head-and-neck carcinoma. *Front Oncol*. 2022;12:974467. doi:10.3389/fonc.2022.974467

71. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The Design of SimpleITK. *Front Neuroinform*. 2013;7. doi:10.3389/fninf.2013.00045

72. Kazhdan M, Simari P, McNutt T, et al. A Shape Relationship Descriptor for Radiation Therapy Planning. In: Yang GZ, Hawkes D, Rueckert D, Noble A, Taylor C, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*. Vol 5762. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2009:100-108. doi:10.1007/978-3-642-04271-3_13

73. Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Medical Physics*. 2009;36(12):5497-5505. doi:10.1118/1.3253464

74. Maurer CR, Rensheng Qi, Raghavan V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans Pattern Anal Machine Intell*. 2003;25(2):265-270. doi:10.1109/TPAMI.2003.1177156

75. Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Vol 1. IEEE Comput. Soc; 2003:958-963. doi:10.1109/ICDAR.2003.1227801

76. Bianchini L, Santinha J, Loução N, et al. A multicenter study on radiomic features from T2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. *Magnetic Resonance in Medicine*. 2020;00:1-14. doi:10.1002/mrm.28521

77. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Scientific Reports*. 2019;9(1):4800. doi:10.1038/s41598-019-41344-5

78. Schwier M, van Griethuysen J, Vangel MG, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Scientific Reports*. 2019;9(1):9441. doi:10.1038/s41598-019-45766-z

79. Liu R, Elhalawani H, Radwan Mohamed AS, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clinical and Translational Radiation Oncology*. 2020;21:11-18. doi:10.1016/j.ctro.2019.11.005

80. Kalendralis P, Traverso A, Shi Z, et al. Multicenter CT phantoms public dataset for radiomics reproducibility tests. *Medical Physics*. 2019;46(3):1512-1518. doi:10.1002/mp.13385

81. Nie K, Al-Hallaq H, Li XA, et al. NCTN Assessment on Current Applications of Radiomics in Oncology. *International Journal of Radiation Oncology*Biology*Physics*. 2019;104(2):302-315. doi:10.1016/j.ijrobp.2019.01.087

82. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean Journal of Radiology*. 2019;20(7):1124-1137. doi:10.3348/kjr.2018.0070

83. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

84. Kwan JYY, Su J, Huang SH, et al. Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma. Published online 2019. doi:10.7937/TCIA.2019.8DHO2GLS

85. Kwan JYY, Su J, Huang SH, et al. Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in HPV-related Oropharyngeal Carcinoma. *International Journal of Radiation Oncology*Biology*Physics*. 2018;102(4):1107-1116. doi:10.1016/j.ijrobp.2018.01.057

86. Beare R, Lowekamp B, Yaniv Z. Image Segmentation, Registration and Characterization in R with SimpleITK. *Journal of Statistical Software*. 2018;86(1):1-35. doi:10.18637/jss.v086.i08

87. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1(1):30-46. doi:10.1037/1082-989X.1.1.30

88. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Molecular Imaging and Biology*. 2016;18(5):788-795. doi:10.1007/s11307-016-0940-2

89. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825-2830.

90. Chao KSC, Ozyigit G, Blanco AI, et al. Intensity-modulated radiation therapy for oropharyngeal carcinoma: impact of tumor volume. *International Journal of*

*Radiation Oncology*Biology*Physics*. 2004;59(1):43-50. doi:10.1016/j.ijrobp.2003.08.004

91. Basaki K, Abe Y, Aoki M, Kondo H, Hatayama Y, Nakaji S. Prognostic factors for survival in stage III non–small-cell lung cancer treated with definitive radiation therapy: Impact of tumor volume. *International Journal of Radiation Oncology*Biology*Physics*. 2006;64(2):449-454. doi:10.1016/j.ijrobp.2005.07.967

92. Dehing-Oberije C, De Ruysscher D, van der Weide H, et al. Tumor Volume Combined With Number of Positive Lymph Node Stations Is a More Important Prognostic Factor Than TNM Stage for Survival of Non–Small-Cell Lung Cancer Patients Treated With (Chemo)radiotherapy. *International Journal of Radiation Oncology*Biology*Physics*. 2008;70(4):1039-1044. doi:10.1016/j.ijrobp.2007.07.2323

93. Fave X, Zhang L, Yang J, et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. *Translational Cancer Research*. 2016;5(4):349-363. doi:10.21037/8709

94. Zhang L, Dong D, Li H, et al. Development and validation of a magnetic resonance imaging-based model for the prediction of distant metastasis before initial treatment of nasopharyngeal carcinoma: A retrospective cohort study. *EBioMedicine*. 2019;40:327-335. doi:10.1016/j.ebiom.2019.01.013

95. Zhang B, Tian J, Dong D, et al. Radiomics Features of Multiparametric MRI as Novel Prognostic Factors in Advanced Nasopharyngeal Carcinoma. *Clin Cancer Res*. 2017;23(15):4259-4269. doi:10.1158/1078-0432.CCR-16-2910

96. Wang G, He L, Yuan C, Huang Y, Liu Z, Liang C. Pretreatment MR imaging radiomics signatures for response prediction to induction chemotherapy in patients with nasopharyngeal carcinoma. *European Journal of Radiology*. 2018;98:100-106. doi:10.1016/j.ejrad.2017.11.007

97. Sheikh K, Lee SH, Cheng Z, et al. Predicting acute radiation induced xerostomia in head and neck Cancer using MR and CT Radiomics of parotid and submandibular glands. *Radiation Oncology*. 2019;14(1):131. doi:10.1186/s13014-019-1339-4

98. Zhang L, Zhou H, Gu D, et al. Radiomic Nomogram: Pretreatment Evaluation of Local Recurrence in Nasopharyngeal Carcinoma based on MR Imaging. *J Cancer*. 2019;10(18):4217-4225. doi:10.7150/jca.33345

99. Ming X, Oei RW, Zhai R, et al. MRI-based radiomics signature is a quantitative prognostic biomarker for nasopharyngeal carcinoma. *Scientific Reports*. 2019;9(1):10412. doi:10.1038/s41598-019-46985-0

100. Li S, Wang K, Hou Z, et al. Use of Radiomics Combined With Machine Learning Method in the Recurrence Patterns After Intensity-Modulated Radiotherapy for

Nasopharyngeal Carcinoma: A Preliminary Study. *Front Oncol*. 2018;8. doi:10.3389/fonc.2018.00648

101. Ouyang FS, Guo BL, Zhang B, et al. Exploration and validation of radiomics signature as an independent prognostic biomarker in stage III-IVb nasopharyngeal carcinoma. *Oncotarget*. 2017;8(43):74869-74879. doi:10.18632/oncotarget.20423

102. Zhang B, Ouyang F, Gu D, et al. Advanced nasopharyngeal carcinoma: pre-treatment prediction of progression based on multi-parametric MRI radiomics. *Oncotarget*. 2017;8(42):72457-72465. doi:10.18632/oncotarget.19799

103. Zhang Y, Lam S, Yu T, et al. Integration of an imbalance framework with novel high-generalizable classifiers for radiomics-based distant metastases prediction of advanced nasopharyngeal carcinoma. *Knowledge-Based Systems*. 2022;235:107649. doi:10.1016/j.knosys.2021.107649

104. Lam SK, Zhang J, Zhang YP, et al. A Multi-Center Study of CT-Based Neck Nodal Radiomics for Predicting an Adaptive Radiotherapy Trigger of Ill-Fitted Thermoplastic Masks in Patients with Nasopharyngeal Carcinoma. *Life*. 2022;12(2):241. doi:10.3390/life12020241

105. Lam SK, Zhang Y, Zhang J, et al. Multi-Organ Omics-Based Prediction for Adaptive Radiation Therapy Eligibility in Nasopharyngeal Carcinoma Patients Undergoing Concurrent Chemoradiotherapy. *Front Oncol*. 2022;11:792024. doi:10.3389/fonc.2021.792024

106. Yu TT, Lam SK, To LH, et al. Pretreatment Prediction of Adaptive Radiation Therapy Eligibility Using MRI-Based Radiomics for Advanced Nasopharyngeal Carcinoma Patients. *Front Oncol*. 2019;9:1050. doi:10.3389/fonc.2019.01050

107. Hou J, Li H, Zeng B, et al. MRI-based radiomics nomogram for predicting temporal lobe injury after radiotherapy in nasopharyngeal carcinoma. *European Radiology*. 2021;32:1106-1114. doi:10.1007/s00330-021-08254-5

108. Hu C, Zheng D, Cao X, et al. Application Value of Magnetic Resonance Radiomics and Clinical Nomograms in Evaluating the Sensitivity of Neoadjuvant Chemotherapy for Nasopharyngeal Carcinoma. *Frontiers in Oncology*. 2021;11. doi:10.3389/fonc.2021.740776

109. Zhu C, Huang H, Liu X, et al. A Clinical-Radiomics Nomogram Based on Computed Tomography for Predicting Risk of Local Recurrence After Radiotherapy in Nasopharyngeal Carcinoma. *Frontiers in Oncology*. 2021;11. doi:10.3389/fonc.2021.637687

110. Shen H, Wang Y, Liu D, et al. Predicting Progression-Free Survival Using MRI-Based Radiomics for Patients With Nonmetastatic Nasopharyngeal Carcinoma. *Frontiers in Oncology*. 2020;10. doi:10.3389/fonc.2020.00618

111. Zhao L, Gong J, Xi Y, et al. MRI-based radiomics nomogram may predict the response to induction chemotherapy and survival in locally advanced nasopharyngeal carcinoma. *Eur Radiol*. 2020;30(1):537-546. doi:10.1007/s00330-019-06211-x

112. Zhuo EH, Zhang WJ, Li HJ, et al. Radiomics on multi-modalities MR sequences can subtype patients with non-metastatic nasopharyngeal carcinoma (NPC) into distinct survival subgroups. *Eur Radiol*. 2019;29(10):5590-5599. doi:10.1007/s00330-019-06075-1

113. Peng L, Hong X, Yuan Q, Lu L, Wang Q, Chen W. Prediction of local recurrence and distant metastasis using radiomics analysis of pretreatment nasopharyngeal [18F]FDG PET/CT images. *Annals of Nuclear Medicine*. 2021;35:458-468. doi:10.1007/s12149-021-01585-9

114. Li S, Deng YQ, Zhu Z, Hua HL, Tao ZZ. A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging. *Diagnostics*. 2021;11. doi:10.3390/diagnostics11091523

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

116. Jia X, Ren L, Cai J. Clinical implementation of AI technologies will require interpretable AI models. *Medical Physics*. 2020;47(1):1-4. doi:10.1002/mp.13891

117. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*. 2018;102(4):1143-1158. doi:10.1016/j.ijrobp.2018.05.053

118. Park SH, Lim H, Bae BK, et al. Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer. *Cancer Imaging*. 2021;21(1):19. doi:10.1186/s40644-021-00388-5

119. Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol*. 2021;31(3):1460-1470. doi:10.1007/s00330-020-07174-0

120. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*. 2020;11(1):91. doi:10.1186/s13244-020-00887-2

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

122. Jha AK, Mithun S, Jaiswar V, et al. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Sci Rep*. 2021;11(1):2055. doi:10.1038/s41598-021-81526-8

123. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. *Int J Cancer*. 2015;136(5):E359-E386. doi:10.1002/ijc.29210

124. Chua MLK, Wee JTS, Hui EP, Chan ATC. Nasopharyngeal carcinoma. *The Lancet*. 2016;387(10022):1012-1024. doi:10.1016/S0140-6736(15)00055-0

125. Zhang MX, Li J, Shen GP, et al. Intensity-modulated radiotherapy prolongs the survival of patients with nasopharyngeal carcinoma compared with conventional two-dimensional radiotherapy: A 10-year experience with a large cohort and long follow-up. *European Journal of Cancer*. 2015;51(17):2587-2595. doi:10.1016/j.ejca.2015.08.006

126. Qu W, Li S, Zhang M, Qiao Q. Pattern and prognosis of distant metastases in nasopharyngeal carcinoma: A large-population retrospective analysis. *Cancer Med*. 2020;9(17):6147-6158. doi:10.1002/cam4.3301

127. Sun X, Su S, Chen C, et al. Long-term outcomes of intensity-modulated radiotherapy for 868 patients with nasopharyngeal carcinoma: An analysis of survival and treatment toxicities. *Radiotherapy and Oncology*. 2014;110(3):398-403. doi:10.1016/j.radonc.2013.10.020

128. Kam MKM, Teo PML, Chau RMC, et al. Treatment of nasopharyngeal carcinoma with intensity-modulated radiotherapy: The Hong Kong experience. *International Journal of Radiation Oncology*Biology*Physics*. 2004;60(5):1440-1450. doi:10.1016/j.ijrobp.2004.05.022

129. Xu Y, Huang T, Fan L, Jin W, Chen X, Chen J. Patterns and Prognostic Value of Lymph Node Metastasis on Distant Metastasis and Survival in Nasopharyngeal Carcinoma: A Surveillance, Epidemiology, and End Results Study, 2006–2015. *Journal of Oncology*. 2019;2019:1-8. doi:10.1155/2019/4094395

130. Amin MB, Edge SB, eds. *AJCC Cancer Staging Manual*. 8th ed. Springer; 2017.

131. Zhou X, Ou X, Yang Y, et al. Quantitative Metastatic Lymph Node Regions on Magnetic Resonance Imaging Are Superior to AJCC N Classification for the Prognosis of Nasopharyngeal Carcinoma. *Journal of Oncology*. 2018;2018:e9172585. doi:10.1155/2018/9172585

132. Chiang CL, Guo Q, Ng WT, et al. Prognostic Factors for Overall Survival in Nasopharyngeal Cancer and Implication for TNM Staging by UICC: A Systematic Review of the Literature. *Front Oncol*. 2021;11:703995. doi:10.3389/fonc.2021.703995

133. Xu Y, Chen X, Zhang M, et al. Prognostic effect of parotid area lymph node metastases after preliminary diagnosis of nasopharyngeal carcinoma: a propensity score matching study. *Cancer Medicine*. 2017;6(10):2213-2221. doi:10.1002/cam4.1154

134. Zhang Y, Zhang ZC, Li WF, Liu X, Liu Q, Ma J. Prognosis and staging of parotid lymph node metastasis in nasopharyngeal carcinoma: An analysis in 10,126 patients. *Oral Oncology*. 2019;95:150-156. doi:10.1016/j.oraloncology.2019.06.013

135. Huang L, Zhang Y, Liu Y, et al. Prognostic value of retropharyngeal lymph node metastasis laterality in nasopharyngeal carcinoma and a proposed modification to the UICC/AJCC N staging system. *Radiotherapy and Oncology*. 2019;140:90-97. doi:10.1016/j.radonc.2019.04.024

136. Ai QY, King AD, Poon DMC, et al. Extranodal extension is a criterion for poor outcome in patients with metastatic nodes from cancer of the nasopharynx. *Oral Oncology*. 2019;88:124-130. doi:10.1016/j.oraloncology.2018.11.007

137. Lu T, Hu Y, Xiao Y, et al. Prognostic value of radiologic extranodal extension and its potential role in future N classification for nasopharyngeal carcinoma. *Oral Oncology*. 2019;99:104438. doi:10.1016/j.oraloncology.2019.09.030

138. Mao Y, Wang S, Lydiatt W, et al. Unambiguous advanced radiologic extranodal extension determined by MRI predicts worse outcomes in nasopharyngeal carcinoma: Potential improvement for future editions of N category systems. *Radiotherapy and Oncology*. 2021;157:114-121. doi:10.1016/j.radonc.2021.01.015

139. Ma H, Liang S, Cui C, et al. Prognostic significance of quantitative metastatic lymph node burden on magnetic resonance imaging in nasopharyngeal carcinoma: A retrospective study of 1224 patients from two centers. *Radiotherapy and Oncology*. 2020;151:40-46. doi:10.1016/j.radonc.2020.07.023

140. Sun Y, Yu XL, Luo W, et al. Recommendation for a contouring method and atlas of organs at risk in nasopharyngeal carcinoma patients receiving intensity-modulated radiotherapy. *Radiotherapy and Oncology*. 2014;110(3):390-397. doi:10.1016/j.radonc.2013.10.035

141. Peng Y, Chen S, Qin A, et al. Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. *Radiotherapy and Oncology*. 2020;150:217-224. doi:10.1016/j.radonc.2020.06.049

142. Luo X, Liao W, Chen J, et al. Efficient Semi-supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency. In: de Bruijne M, Cattin PC, Cotin S, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Vol 12902. Lecture Notes in Computer Science. Springer International Publishing; 2021:318-329. doi:10.1007/978-3-030-87196-3_30

143. Davidson-Pilon C. lifelines: survival analysis in Python. *JOSS*. 2019;4(40):1317. doi:10.21105/joss.01317

144. Tang LL, Guo R, Zhou G, et al. Prognostic Value and Staging Classification of Retropharyngeal Lymph Node Metastasis in Nasopharyngeal Carcinoma Patients Treated with Intensity-modulated Radiotherapy. *PLOS ONE*. 2014;9(10):e108375. doi:10.1371/journal.pone.0108375

145. Li YZ, Xie CM, Wu YP, et al. Nasopharyngeal Carcinoma Patients With Retropharyngeal Lymph Node Metastases: A Minimum Axial Diameter of 6 mm Is a More Accurate Prognostic Predictor Than 5 mm. *American Journal of Roentgenology*. 2015;204(1):20-23. doi:10.2214/AJR.14.12936

146. Chen J, Luo J, He X, Zhu C. Evaluation of Contrast-Enhanced Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) in the Detection of Retropharyngeal Lymph Node Metastases in Nasopharyngeal Carcinoma Patients. *Cancer Manag Res*. 2020;12:1733-1739. doi:10.2147/CMAR.S244034

147. Guo R, Mao YP, Tang LL, Chen L, Sun Y, Ma J. The evolution of nasopharyngeal carcinoma staging. *BJR*. 2019;92(1102):20190244. doi:10.1259/bjr.20190244

148. Lee AW, Ng WT, Pan JJ, et al. International guideline for the delineation of the clinical target volumes (CTV) for nasopharyngeal carcinoma. *Radiotherapy and Oncology*. 2018;126(1):25-36. doi:10.1016/j.radonc.2017.10.032

149. Lin L, Dou Q, Jin YM, et al. Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. *Radiology*. 2019;291(3):677-686. doi:10.1148/radiol.2019182012

150. Salzano G, Perri F, Maglitto F, et al. Pre-Treatment Neutrophil-to-Lymphocyte and Platelet-to-Lymphocyte Ratios as Predictors of Occult Cervical Metastasis in Clinically Negative Neck Supraglottic and Glottic Cancer. *JPM*. 2021;11(12):1252. doi:10.3390/jpm11121252

151. Abbate V, Barone S, Troise S, et al. The Combination of Inflammatory Biomarkers as Prognostic Indicator in Salivary Gland Malignancy. *Cancers*. 2022;14(23):5934. doi:10.3390/cancers14235934

152. Zhong L, Dong D, Fang X, et al. A deep learning-based radiomic nomogram for prognosis and treatment decision in advanced nasopharyngeal carcinoma: A multicentre study. *EBioMedicine*. 2021;70:103522. doi:10.1016/j.ebiom.2021.103522

153. Shen H, Yin J, Niu R, 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. *Radiotherapy and Oncology*. 2022;171:107-113. doi:10.1016/j.radonc.2022.04.017

154. Severn C, Suresh K, Görg C, Choi YS, Jain R, Ghosh D. A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features. *Sensors*. 2022;22(14):5205. doi:10.3390/s22145205

155. Oba K, Paoletti X, Alberts S, et al. Disease-Free Survival as a Surrogate for Overall Survival in Adjuvant Trials of Gastric Cancer: A Meta-Analysis. *JNCI: Journal of the National Cancer Institute*. 2013;105(21):1600-1607. doi:10.1093/jnci/djt270

156. Chen L, Hu CS, Chen XZ, et al. Adjuvant chemotherapy in patients with locoregionally advanced nasopharyngeal carcinoma: Long-term results of a phase 3 multicentre randomised controlled trial. *European Journal of Cancer*. 2017;75:150-158. doi:10.1016/j.ejca.2017.01.002

157. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012

158. Al-Sarraf M, LeBlanc M, Giri PG, et al. Chemoradiotherapy versus radiotherapy in patients with advanced nasopharyngeal cancer: phase III randomized Intergroup study 0099. *JCO*. 1998;16(4):1310-1317. doi:10.1200/JCO.1998.16.4.1310

159. You R, Cao YS, Huang PY, et al. The Changing Therapeutic Role of Chemo-radiotherapy for Loco-regionally Advanced Nasopharyngeal Carcinoma from Two/Three-Dimensional Radiotherapy to Intensity-Modulated Radiotherapy: A Network Meta-Analysis. *Theranostics*. 2017;7(19):4825-4835. doi:10.7150/thno.21815

160. Chen YP, Wang ZX, Chen L, et al. A Bayesian network meta-analysis comparing concurrent chemoradiotherapy followed by adjuvant chemotherapy, concurrent chemoradiotherapy alone and radiotherapy alone in patients with locoregionally advanced nasopharyngeal carcinoma. *Annals of Oncology*. 2015;26(1):205-211. doi:10.1093/annonc/mdu507

161. Ribassin-Majed L, Marguet S, Lee AWM, et al. What Is the Best Treatment of Locally Advanced Nasopharyngeal Carcinoma? An Individual Patient Data Network Meta-Analysis. *JCO*. 2017;35(5):498-505. doi:10.1200/JCO.2016.67.4119

162. Chan ATC, Hui EP, Ngan RKC, et al. Analysis of Plasma Epstein-Barr Virus DNA in Nasopharyngeal Cancer After Chemoradiation to Identify High-Risk Patients for Adjuvant Chemotherapy: A Randomized Controlled Trial. *JCO*. 2018;36(31):3091-3100. doi:10.1200/JCO.2018.77.7847

163. Mao YP, Tang LL, Chen L, et al. Prognostic factors and failure patterns in non-metastatic nasopharyngeal carcinoma after intensity-modulated radiotherapy. *Chin J Cancer*. 2016;35(1):103. doi:10.1186/s40880-016-0167-2

164. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*. 1983;23(3):341-352. doi:10.1016/0734-189X(83)90032-4

165. Elhalawani H, Mohamed ASR, White AL, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Scientific Data*. 2017;4(1):170077. doi:10.1038/sdata.2017.77