# MULTI-OMIC PREDICTION OF SEVERE ACUTE TREATMENT-INDUCED ORAL MUCOSITIS AND DYSPHAGIA IN NASOPHARYNGEAL CARCINOMA PATIENTS

NICOL ALEXANDER JAMES

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

# Multi-omic Prediction of Severe Acute Treatment-Induced Oral Mucositis and Dysphagia in Nasopharyngeal Carcinoma Patients

**NICOL Alexander James**

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

August 2024

# CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduced 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)

Alexander James NICOL_____(Name of student)

# ABSTRACT

**Introduction:** With the improvement in survival rates for nasopharyngeal carcinoma (NPC), it is increasingly important to address the impact of treatment-induced toxicity on patients' quality of life. Acute oral mucositis (OM) and dysphagia are two of the most common toxicities resulting from NPC treatment. Severe cases cause significant suffering and pose a threat to treatment outcome through unexpected hospitalization, weight loss, and treatment interruption. This thesis harnesses high-dimensional multi-omic data to identify patients at risk of severe acute OM and dysphagia to better target preventative interventions and personalized support.

**Methods**: Four hundred and sixty-four NPC patients treated with radiotherapy (RT) at two Hong Kong hospitals were retrospectively recruited for analysis. Radiomic, dosiomic and contouromic features were extracted from planning CT images, RT dose distributions and tumour and organ-at-risk contours respectively. Machine learning models for predicting severe acute OM and dysphagia were developed. Model performance was comprehensively assessed and compared to that of conventional prediction models using clinical and dosimetric features alone.

**Results:** Multi-omic prediction models for severe acute OM and dysphagia outperformed conventional clinical and dosimetric models developed on the same data. Radiomics, by describing pre-treatment tissue characteristics, dosiomics, by describing the spatial distribution of the planned RT dose, and contouromics, by describing the challenges posed by patient geometry, were demonstrated to have unique predictive value, and facilitated

greater model discrimination by supplementing clinical features. Importantly, this study conducted external validation to assess the generalizability of the models, providing a greater level of evidence compared to other prediction models in the literature.

**Conclusion:** Multi-omic features including radiomic, dosiomic and contouromic features enhanced the discrimination performance of models incorporating clinical and dosimetric features and demonstrated independent predictive value. The findings in this project provide an invaluable reference for future work and include important recommendations for future development of multi-omics for toxicity prediction.

# RESEARCH OUTPUT

## Publications

**Nicol AJ**, Ching JCF, Tam VCW, Liu KCK, Leung VWS, Cai J, Lee SWY. Predictive Factors for Chemoradiation-Induced Oral Mucositis and Dysphagia in Head and Neck Cancer: A Scoping Review. *Cancers (Basel)*. 2023 Dec 4;15(23):5705. [**Journal article**]

**Nicol AJ**, Lam SK, Ching JCF, Tam VCW, Teng X, Zhang J, Yip CWY, Lee FKH, Au KH, Chan PLC, Wong KCW, Cai J, Lee SWY. A Multi-Centre, Multi-Organ, Multi-Omic Prediction Model for Treatment-induced Severe Oral Mucositis in Nasopharyngeal Carcinoma. *Radiol med.* 2024. [**Journal article**]

Teng X, Wang Y, **Nicol AJ**, Fung JCF, Wong EKY, Lam K, Zhang J, Lee SWY, Cai J. Enhancing the Clinical Utility of Radiomics: Addressing the Challenges of Repeatability and Reproducibility in CT and MRI. *Diagnostics*. 2024; 14(16):1835. [**Journal article**]

Li W, Shi Y, Li Z, Lam S, Li X, Cheung AHY, Liu C, Liu S, Yang J, Zeng G, Fu J, Lee SWY, **Nicol AJ** et al. Federated Learning-empowered Virtual Contrast-enhanced MRI (FL-VCE-MRI) Enables Toxic-free Clinical Assessments for Highly Infiltrative and Heterogenous Cancer – A Nationwide Study. **[Journal article - submitted for publication]**

## Conference abstracts and presentations

**Nicol AJ**, Ching J, Tam V, Teng X, Zhang J, Cai J, Lee SWY. Developing and Validating a Radiomic Prediction Model for Severe Acute Dysphagia in Nasopharyngeal Carcinoma Patients Undergoing Radiation Therapy AAPM 2024 [**Poster**]

**Nicol AJ,** Lee SWY. Developing and validating a multi-omic prediction model for severe acute oral mucositis in nasopharyngeal carcinoma patients undergoing radiation therapy, ISRRT World Congress 2024 in conjunction with Hong Kong Radiographers and Radiation Therapists Conference [**Conference abstract + presentation**]

**Nicol AJ**, Lee SWY. Radiotherapy-induced oral mucositis prediction using radiomics and dosiomics, HKU Medical AI Imaging Summit 2024 [**Meeting presentation**]

**Nicol AJ**, Zhang J, Teng X, Cheung KM, Cai J, Lee SWY. Multi-omics Prediction of Severe Acute Radiotherapy-Induced Dysphagia using Planning CT and Dose Distribution Data in Locoregionally Advanced Nasopharyngeal Carcinoma Patients receiving Definitive Radiotherapy. PolyU Research Student Conference 2023, May 2023 [**Conference paper + presentation**]

**Nicol AJ**, Lee SWY. Dysphagia in Head and Neck Cancer Patients: A multi-omics prediction model. Paper presented at CUHK ENT Conference 2023, Hong Kong, Hong Kong, May 2023 [**Conference presentation**]

Lee SWY, **Nicol AJ**, Tam VCW, Law HKW, Leung VWS, Cai J. Student Experiential Learning Through Patient Education: Game-Oriented Radiotherapy Simulation for Pediatric Cancer Patients. AAPM (American Association of Physicists in Medicine) 64th Annual Meeting & Exhibition 2022 (July 10-14) Session: Arthur Boyer Award for Innovation in Medical Physics Education [**Conference abstract**]

Lee SWY, Ching JCF, **Nicol AJ**, Tam VCW, Cai J, Leung VWS, Law HKW. Experiential learning through paediatric patient education: game-based radiotherapy rehearsals, ESTRO, May 2023 [**Conference abstract**]

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my chief supervisor, Dr. Shara Wee Yee Lee, for her invaluable guidance, unwavering support, and continuous encouragement throughout my PhD journey. I am also grateful to my co-supervisor, Prof. Jing Cai, for his insightful suggestions and mentorship, which have significantly contributed to the completion of this thesis.

I extend my sincere appreciation to Dr. Au Kwok Hung, Dr. Francis Lee Kar Ho, and Dr. Celia Yip Wai Yi from Queen Elizabeth Hospital Hong Kong, as well as Dr. Kenneth CW Wong, Dr. Catherine Po Ling Chan, and Miss Virginia Kwong from Prince of Wales Hospital Hong Kong, for their invaluable assistance in the collection of clinical and imaging data. Their support has been instrumental to the success of this research.

Special thanks are due to my dedicated research team members, Mr Jerry Chi Fung Ching and Dr Victor Chi Wing Tam, for their unwavering support in data collection, meticulous manuscript revision, and thorough presentation practice. I am incredibly grateful for their contributions.

I am also deeply indebted to the members of Prof. Jing Cai's research team for their regular feedback on the methodology and results of this study. I am particularly grateful to Dr Xinzhi Teng and Dr Jiang Zhang for their exceptional guidance on radiomic model development and for providing the in-house software for feature extraction and perturbation generation.

I would also like to acknowledge the valuable assistance provided by Dr. Sai Kit Lam, Dr. Vincent Leung, and Dr. Helen Law, who have generously shared their expertise and insights throughout my research.

Finally, I extend my heartfelt gratitude to my fiancé, Ms. Haruka Sato, and my family for their unwavering love, support, and encouragement throughout my PhD study. Their presence has been a constant source of inspiration and motivation.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS

| | |
|---|---|
| AI | Artificial intelligence |
| AUC | Area under the ROC curve |
| BMI | Body mass index |
| BW | Body weight |
| CCT | Concurrent chemotherapy |
| CD | Compact disc |
| CECT | Contrast-enhanced CT |
| CI | Confidence interval |
| CLI | Clinical features |
| CON | Contouromic features |
| CRT | Chemoradiotherapy |
| CSV | Comma separated values |
| CT | Computed tomography |
| CTCAE | Common Terminology Criteria for Adverse Events |
| CV | Cross validation |
| DAHANCA | Danish Head and Neck Cancer group |
| DICOM | Digital Imaging and Communications in Medicine |
| DNA | Deoxyribonucleic Acid |
| DOS | Dosiomic features |
| DVF | Deformable vector field |
| DVH | Dose volume histogram |
| EBV | Epstein-Barr virus |
| ECOG | Eastern Cooperative Oncology Group performance status |
| ENT | Ear, Nose and Throat |
| ESR | Erythrocyte sediment rate |
| EUD | Equivalent uniform dose |
| GLCM | Grey-level co-occurrence matrix |
| GLDM | Grey-level difference matrix |
| GLRLM | Grey-level run-length matrix |
| GLSZM | Grey-level size zone matrix |
| GNB | Gaussian Naive Bayes |
| GTV | Gross tumour volume |
| GTVN | Neck nodal gross tumour volume |
| GTVP | Primary gross tumour volume |
| HNC | Head and neck cancer |
| HNSCC | Head and neck squamous cell carcinoma |
| HPV | Human papillomavirus |
| HU | Hounsfield units |
| IBSI | Image biomarker standardization initiative |
| IC | Induction chemotherapy |
| ICC | Intraclass correlation coefficient |
| IMRT | Intensity modulated radiotherapy |
| IPC | Inferior pharyngeal constrictor |
| IPCM | Inferior pharyngeal constrictor muscle |
| IV | Intravenous |
| LLLT | Low level laser therapy |
| MCC | Maximal Correlation Coefficient |
| MLC | Multi-leaf collimator |
| MPC | Middle pharyngeal constrictor |
| MPCM | Middle pharyngeal constrictor muscle |
| MR | Magnetic resonance |
| MRI | Magnetic resonance imaging |
| MRMR | Minimum redundancy, maximum relevance algorithm |
| MVCT | Megavolt CT |
| MWU | Mann-Whitney U test |
| NA | Not applicable / available |
| NCI-CTC | National Cancer Institute Common Toxicity Criteria |
| NGTDM | Neighbouring grey-tone difference matrix |
| NPC | Nasopharyngeal carcinoma |
| NTCP | Normal tissue complication probability |
| OAR | Organ-at-risk |
| OM | Oral mucositis |
| OVH | Overlap volume histogram (contouromics) |
| PBMT | Photobiomodulation therapy |
| PC | Pharyngeal constrictors |
| PCM | Pharyngeal constrictor muscles |
| PEG | Percutaneous endoscopic gastrostomy |
| PET | Positron emission tomography |
| POV | Projection overlap volume (contouromics) |
| PWH | Price of Wales Hospital, Hong Kong |
| QEH | Queen Elizabeth Hospital, Hong Kong |
| QMH | Queen Mary Hospital, Hong Kong |
| QOL | Quality of life |
| RAD | Radiomic features |
| RADAR | Radiotherapy data analysis and reporting toolkit |
| RBF | Radial basis function |
| RF | Random Forest |
| RLN | Retropharyngeal lymph nodes |
| ROC | Receiver operating characteristic |
| ROI | Region of interest |
| RT | Radiotherapy |
| RTOG | Radiation Therapy Oncology Group |
| SHAP | Shapley Additive Explanations |
| SMOTE | Synthetic minority over-sampling technique |
| SNP | Single nucleotide polymorphism |
| SPC | Superior pharyngeal constrictor |
| SPCM | Superior pharyngeal constrictor muscle |
| SVM | Support vector machine |
| TNM | Tumour Node Metastasis staging system |
| TPN | Total parenteral nutrition |
| VIF | Variance inflation factor |
| VMAT | Volumetric modulated arc therapy |
| VOI | Volume of interest |
| WCC | White cell count |
| WHO | World health organization |
| XGB | XGBoost (eXtreme Gradient Boosting) |
| XRCC | X-ray repair cross-complimenting protein |

# OVERVIEW OF THESIS

Nasopharyngeal carcinoma (NPC) has a relatively high prevalence in East and Southeast Asia and is a significant cause of cancer-related deaths in Hong Kong. Survival rates have improved in recent decades with the introduction of intensity-modulated radiotherapy (IMRT) and the use of concurrent chemotherapy for locally advanced cases. However, the prescribed radiation dose is high among head and neck cancers, and combined with chemotherapy, there is a significant burden on patients from treatment-induced toxicity. As survival rates improve, it is increasingly important to address suffering from toxicity. Acute oral mucositis and dysphagia are two of the most common and damaging toxicities experienced by NPC patients, and early identification of patients at risk of severe toxicity is crucial for facilitating targeted preventative intervention and management.

Chapter 1 provides a comprehensive overview of the background of the topic, highlighting the unique challenges of head and neck cancer, nasopharyngeal carcinoma, and treatment-induced toxicity. The remainder of this chapter consists of two literature reviews. The first literature review, published in *Cancers*, investigated predictive factors for treatment-induced OM and dysphagia, identifying the most well-supported risk factors and the conventional predictive models in the literature [1]. The second literature review focuses specifically on the existing research on multi-omics-based toxicity prediction in head and neck cancer, placing the thesis in context and providing evidence for the research gap. Chapter 2 states the aim of the thesis and identifies the research gap, based on the findings from the literature reviews in the previous chapter. The project objectives are then stated. Chapter 3 outlines the core methodology for multi-omic studies, reporting the key considerations and

steps which are common to all three objectives such as image acquisition, preprocessing, feature extraction, and feature selection. There is significant variation in the methodology for radiomics studies, and so the methodological considerations were discussed, and justifications were provided for the chosen approach. Chapter 4 reports the work conducted to complete objective 1, the development of a multi-omic model for severe acute OM in NPC patients undergoing RT. This chapter is supported by a manuscript submitted for publication. Chapter 5 reports the work conducted to complete objective 2, the development of a multi-omic model for severe acute dysphagia in NPC patients undergoing RT. Chapter 6 reports the work conducted to complete objective 3, the development of a multi-label model for severe acute OM and dysphagia in NPC patients undergoing RT. This chapter is exploratory, reporting the results of different approaches to multi-label modelling and identifying the strengths and weaknesses of each, to guide future work. Chapter 7 consists of a comprehensive discussion of the practical considerations involved in future development of multi-omic models for prediction of OM and dysphagia. It addresses wide-ranging aspects of from study design to data analysis, synthesizing the experiences learned during this project and strengths and limitations of the research to provide recommendations for future work. The clinical practicability of the developed models is discussed along with the current barriers to clinical use and recommendations for moving towards clinical implementation.

In summary, findings show that the inclusion of multi-omic features was able to improve on conventional approaches to prediction of OM and dysphagia. The resulting model signatures had independent predictive value when compared to conventional clinical and DVH features. A wide range of recommendations for the future development of multi-omic

prediction models is provided to tackle the many challenges of achieving generalizable and accurate predictions of these multifaceted and complex toxicities.

# CHAPTER 1 BACKGROUND

## 1.1. Introduction

This project focuses on nasopharyngeal carcinoma (NPC), a form of head and neck cancer (HNC). Before describing the unique challenges of NPC, it is important to introduce the context of HNC more generally, providing a foundational understanding of the broader category of cancers that share many of the same treatments and treatment-related toxicities.

### *1.1.1. Head and neck cancer*

*Epidemiology and types of HNC*

In 2020, HNC accounted for over 900,000 new cases globally, representing the seventh most prevalent type of cancer [2]. HNC was in the top eight cancers by mortality, with over 400,000 deaths globally [2]. The worldwide five year survival rate for HNC is approximately 50%, though survival depends on geographical location, tumour site, stage at diagnosis and other factors such as human papillomavirus (HPV) status [3]. HNC encompasses malignancies that originate in the oral cavity, lips, tongue, pharynx, larynx, salivary glands, as well as in the nasal cavity and sinuses, with some studies also including oesophageal malignancies [3]. Around 90% of these malignancies are head and neck squamous cell carcinoma (HNSCC), beginning in the mucosal epithelium [3]. The incidence of HNC is rising in both developed and developing countries, with a 30% increase predicted by 2030 [3]. The increase has been attributed to increasing rates of HPV infection in the United States and Europe, and also chewing of areca nut or smokeless tobacco in Southeast Asia [3].

## *Risk factors for HNC*

HNSCC is two to four times more common in men than in women. The age at which a person is diagnosed with HNSCC can vary depending on whether the cancer is linked to the HPV or the Epstein-Barr virus (EBV). For HNSCC that is not linked to a virus, the age at diagnosis is 66 years old, but it drops to about 50 years old for HNSCC that is linked to HPV or EBV [4]. These viruses, along with tobacco and alcohol use, represent the main risk factors for HNC, though the specific associations vary depending on the subsite of HNC [3]. There is also evidence for occupational exposure and socio-economic risk factors, as well as family history and dietary factors [3]. Genetics have also been investigated in relation to HNC incidence. Genetic loci have been associated with the risk of HNC after HPV infection, and also with affecting metabolism of alcohol to increase HNC risk [3].

## *Clinical presentation of HNC*

The clinical presentation of HNC varies by subsite, with sites like the oral cavity exhibiting more obvious masses and symptoms. Symptoms for HNC are varied, including difficulty eating or speaking, voice changes, painful swallowing, ear pain, hearing loss, nosebleeds, nasal obstruction, or even difficulty breathing [4]. If neck lymph nodes become involved, patients may present with neck masses.

## *Diagnosis and staging of HNC*

A biopsy of the primary tumour or neck nodal mass is necessary to confirm the diagnosis of HNC and allow for the analysis of the histology. Staging is conducted according to the Union for International Cancer Control (UICC) and the American Joint Commission on Cancer (AJCC) cancer staging manual [5], which provides criteria for the TNM staging

including tumour stage (T1-T4), lymph node stage (N0-N3) and overall stage (I-IV). The stage of diagnosis varies by subsite, due to the differences in presenting symptoms and prominence of masses.

### *Treatment for HNC*

The management of HNC is affected by subsite, stage, and patient preference. If the tumour is locally or locoregionally confined then surgical resection is the main line curative treatment, provided that the tumour is accessible [4]. Advances in surgery have expanded the indications of resection as a primary treatment. However, HNCs, particularly laryngeal or pharyngeal cancers, are located close to many vital structures. For this reason, radiation is often employed as the primary treatment. For more advanced disease, postoperative radiation or chemoradiation is employed in order to reduce the risk of recurrence or treatment failure [4]. Studies have also investigated the potential of immunotherapy for HNC. While combinations of different treatments can prove advantageous for survival, they can also pose the increased risk of toxicity.

### *Consequences of HNC treatment*

Head and neck cancer frequently has a significant impact on daily activities and continues to do so even after treatment. Impairments and disabilities often persist for years after treatment, representing long term burdens. As a result, quality of life (QOL) is an important consideration in HNC treatment. Physical, functional, emotional, and social aspects should be considered for the patient and their caregivers. The impact of HNC and its treatment is sadly reflected by the statistic that survivors of HNC are almost twice as likely to die from suicide as other cancers [6].

### *1.1.2. Nasopharyngeal carcinoma*

*Epidemiology of NPC*

Nasopharyngeal carcinoma (NPC) is a form of HNC that develops in the epithelial tissues of the nasopharynx, which is the top part of the throat behind the nasal cavity [7]. It has several important differences from other HNC in terms of the involved anatomy, epidemiology, risk factors, and treatment. In terms of geographic distribution, NPC has a relatively high prevalence in East and Southeast Asia, particularly in southern China [8]. Interestingly, the high incidence in that population remains present in migrant populations in other geographic areas, but with reduced incidence in subsequent generations [8]. It is hypothesized that the pathogenesis of NPC is therefore affected by genetic, cultural, and environmental factors [8].

*Risk factors for NPC*

EBV infection, family history of NPC, smoking, consumption of certain preserved foods, alcohol, and poor oral hygiene have been identified as risk factors [8]. The mechanism by which EBV or environmental carcinogens lead to NPC is through DNA damage and the proliferation of mutated cells. Consequently, EBV-related antibodies may be used as biomarkers for screening and monitoring NPC. Symptoms of NPC correspond to the pattern of spread of the disease and include nosebleeds, nasal obstruction, hearing loss, cranial nerve palsies and swelling of the cervical lymph nodes [8].

*Diagnosis and staging of NPC*

Definitive diagnosis of NPC requires a biopsy, which is typically performed after head and neck evaluation using nasopharyngoscopy and magnetic resonance imaging (MRI), computed tomography (CT) imaging or positron-emission tomography (PET/CT) imaging.

MRI and PET/CT are particularly relevant for the diagnosis of distant metastasis. As with other HNC, staging of NPC is performed according to the UICC/AJCC TNM staging system [5]. Prognostic factors typically include TNM staging, EBV-related biomarkers and PET/CT metabolic indexes [8]. In terms of survival, the Hong Kong Cancer Registry reported a relative 5-year survival rate of 68.7% for NPC in the period 2010-2018 [9]. An analysis on NPC survival trends in a United States population revealed an upward trend in 5-year survival rate over recent decades, from 36% in 1973-1979 to 54.7% in 2000-2007 [10]. The improvement in survival rates over recent decades is illustrated by **Figure 1**, which compares data from several studies in different countries. A combination of factors, including improved screening and advancements in treatment techniques, may be responsible for the improved survival of NPC. With improved survival rates, the need to address patients' quality of life becomes increasingly urgent.

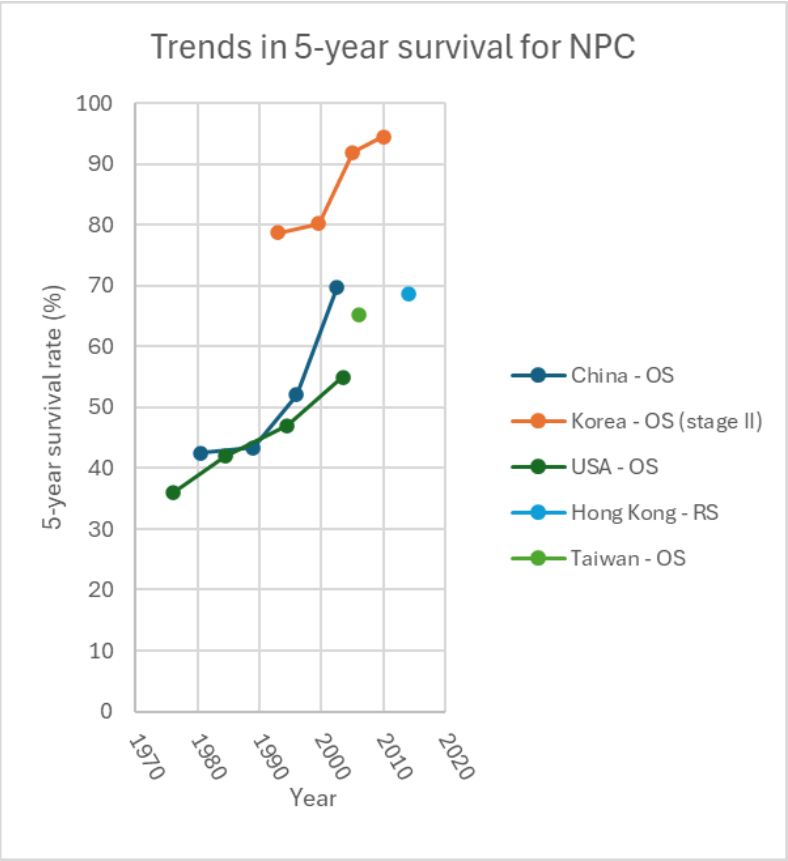



**Figure 1: Trends in 5-year survival for NPC. Data was taken from the following publications: China [11], Korea [12], USA [10], Hong Kong [9], Taiwan [13]. OS = overall survival, RS = relative survival**

## *Radiotherapy for NPC*

The primary treatment for NPC is radiotherapy (RT), due to its sensitivity to ionising radiation and the close proximity of critical structures. Surgery is not typically employed as a treatment, except for certain recurrent cases. Radiation techniques have evolved from conventional two-dimensional RT to 3D conformal RT, to intensity-modulated RT (IMRT) , allowing more accurate targeting of the tumour and better dose sparing to the surrounding tissues [8]. The use of IMRT resulted in a reduction of the 5-year locoregional failure rate of non-metastatic NPC to 7.4% [14]. Future advancements in NPC treatment may include proton

or carbon ion RT, though this is currently not in widespread use. Under the current treatment guidelines, radiation is delivered using a total dose of 66-70Gy applied in 2Gy per fraction [7].

### *Chemotherapy for NPC*

While early-stage NPC is treated with RT alone, advanced stage cases receive chemotherapy concurrently with RT. Additionally, induction (neoadjuvant) chemotherapy may be provided before RT, and adjuvant chemotherapy may be provided after RT. The specific guidelines for delivery of neoadjuvant or adjuvant chemotherapy vary by institution, since the evidence for the benefit of one approach over the other is mixed [8]. The specific chemotherapy drugs and dosing schedules vary, and decisions can depend on the patient's kidney function, any existing co-morbidities, and the severity of side effects experienced. More aggressive treatment can mitigate the risk of local, regional, or distant failure; however, this can take a greater toll on the patient during treatment and can result in more significant toxicity in both the short and long term.

### *Nasopharyngeal carcinoma in Hong Kong*

NPC is particularly significant in Hong Kong, where it was among the top ten cancers by mortality in men in 2021 [15]. Specifically, there were 558 new diagnoses among a population of 7.4 million [15, 16]. This thesis focuses on NPC in Hong Kong, due to its high prevalence and unique challenges within the local population. Consequently, this section summarizes the specific context of NPC in Hong Kong, including its treatment protocols.

There are six public oncology centres which cover more than 90% of the population in the region, managed by the Hong Kong Hospital Authority [17]. A retrospective, multi-centre

study on NPC patients from all of these centres reported statistics for patients who underwent definitive IMRT between 2001 and 2010 [17], which provides an overview of NPC in the region. Seventy-three percent of patients were male, with a median age of 50 years. The ratio of non-smokers to former or current smokers was about 1.35:1. Most (75%) of patients were diagnosed at stage III or higher. Regarding treatment, 72% of patients received chemotherapy in combination with IMRT. This included 28% with concurrent chemotherapy, 14% with concurrent-adjuvant chemotherapy, and 28% with induction-concurrent chemotherapy.

In Hong Kong, the typical process for NPC patients is as follows: First, patients suspected of NPC undergo a nasal endoscopy in the Ear, Nose and Throat (ENT) department. If abnormalities are found, then the patient will receive an MRI and/or PET/CT to determine the extent and stage of cancer, including identification of distant metastases or a different primary cancer. The diagnosis of NPC is then verified by conducting a biopsy and performing an analysis of the tumour histology. Patients eligible for curative treatment are then scheduled for RT. This involves booking the use of the treatment machine and scheduling a planning CT scan. Prompt treatment is desirable; however, in practice, there may be a wait of several weeks between diagnosis and RT, during which some patients at advanced stage may receive induction chemotherapy. Patients receive a planning CT scan while lying in the treatment position, fitted with immobilization devices. Typically, a contrast medium is injected to facilitate discrimination of the tumour. The medical physics team will then contour the tumour and organs-at-risk (OARs) on the CT, referring to any available MRI or PET/CT previously acquired. Then, treatment planning software is used to create a radiation plan, ensuring that the required dose to the target is met and minimizing the dose to critical structures. The patient

then attends the RT clinic 5 times a week, receiving a treatment fraction of 2 – 2.12 Gy for each of the 33 fractions. Concurrent chemotherapy is also delivered during these visits for eligible patients, typically involving cisplatin or carboplatin. After the 66 – 70 Gy dose is delivered, some patients may receive adjuvant chemotherapy. Patients then receive follow-up nasal endoscopy to assess the condition of the nasopharynx post-treatment, along with a biopsy to confirm treatment outcome. If residual disease is detected, patients may be prescribed a radiation boost or other treatment. Patients generally receive weekly follow-up during and shortly after treatment. Patients are then monitored at regular intervals in the subsequent months and years. Generally, patients will remain with their local hospital or have records transferred between hospitals. Occasionally, patients receive imaging or treatment at private hospitals before returning to their local public hospital. Some patients may relocate out of the region, for instance, to mainland China, resulting in some loss of follow-up.

### *1.1.3. Treatment-induced toxicity*

*Mechanism of toxicity*

Radiotherapy is a major component of treatment for HNC, administered to nearly 75% of patients [18]. It is employed to damage tumour cells, stopping their growth and division. This is achieved using high-energy rays or radioactive substances to directly damage the DNA, or other critical cellular molecules of the tumour cells [19]. However, the same radiation will also damage normal cells, and so optimal delivery of RT is dependent on balancing tumour dose and dose to normal tissue [20]. The degree of tissue damage depends on total radiation dose, fractionation, and tissue properties [20]. Certain tissues are more resistant to radiation damage, while others may suffer from dysfunction more quickly. Radiation is frequently

accompanied by chemotherapy, a treatment that likewise aims to inhibit cell multiplication. As a systemic treatment affecting the whole body, it can mitigate the risk of tumour invasion and metastasis. Combination chemotherapy is often employed to minimize the risk of tumour resistance [21]. However, the side effects of chemotherapy are extensive, including reduced blood cell production, kidney toxicity and nausea and vomiting. By inhibiting cell multiplication, chemotherapy is particularly damaging for rapidly dividing tissues. These include tumour cells, but also normal tissues, such as those in the mucosa, resulting in toxicity [22]. Inflammation, healing rates and susceptibility to infection are also affected by chemotherapy.

### *Grading and timeframe of toxicity*

Most patients receiving chemoradiation in the head and neck region will experience moderate to severe toxicity. This may occur during treatment or after treatment, with chronic side effects developing in some patients. The Radiation Therapy Oncology Group (RTOG) and National Cancer Institute Common Toxicity Criteria (NCI-CTC) define acute toxicity as occurring within 90 days of the commencement of RT [23]. Damage to normal tissue triggers inflammatory responses in the acute phase, followed by tissue repair and remodelling after treatment completion. Acute toxicities typically include OM, dysphagia, xerostomia, dysgeusia, odynophagia, and dermatitis. Most acute toxicities resolve within weeks after treatment, though significant suffering and risks to treatment outcome can occur in this period. Late toxicity, which may surface months or years later, results from irreversible damage to tissues and structures. Examples of late toxicities include dysphagia, xerostomia, osteoradionecrosis, myositis, dental caries, oral cavity necrosis, fibrosis, impaired wound

healing and lymphedema [24]. These are typically chronic conditions that require lifelong management. The severity of toxicities can range from mild to life-threatening and may require immediate medical intervention. Two of the most prevalent toxicities are oral mucositis (OM) and dysphagia, each of which poses a significant impact on patient quality of life.

### *1.1.4. Oral mucositis*

#### *Introduction to OM*

Oral mucositis refers to erythema (redness), inflammation and ulceration occurring in the mucosal lining of the mouth and pharynx because of chemotherapy or RT. It is a painful condition which affects the ability of the patient to eat and increases the risk of infection [25]. Almost all HNC patients treated with chemoradiotherapy experience OM [25], and a meta-analysis by Li et al. on NPC patients found that 99% experienced OM, and 52% experienced severe OM [26]. Severe cases can involve haemorrhage and necrosis and can become life-threatening, requiring hospitalization and immediate intervention.

#### *Pathogenesis of OM*

The pathogenesis of OM involves five phases [27]. Firstly, tissue injury results from damage to basal epithelial cells caused by radiation or chemotherapy, and the resultant generation of reactive oxygen species, or free radicals. An inflammatory phase follows, resulting from the increased production of pro-inflammatory cytokines caused by the reactive oxygen species. This further exacerbates tissue injury and cell death. The third phase involves the activation of molecular pathways that further amplify damage to the mucosa. The fourth phase is characterized by ulceration, which is partly due to the role of microorganisms that

colonize the damaged tissues. The final phase is a healing phase, where the production of epithelial cells increases, restoring the mucosa.

## *Clinical presentation of OM*

The clinical presentation of OM broadly matches the five phases of pathogenesis [27]. The mucosa first presents with erythema, then with ulcerations, which may become colonized by microorganisms naturally present in the mucosa, further exacerbating the pain and inflammation [25]. Fungal infections, such as candidiasis, may also compound the effects of OM. Symptoms typically arise within two weeks of the start of treatment, and predominantly affect non-keratinized surfaces such as the tongue, buccal mucosa, and soft palate, as shown in **Figure 2** [27]. Lesions typically heal within 2-4 weeks from the end of treatment [27].

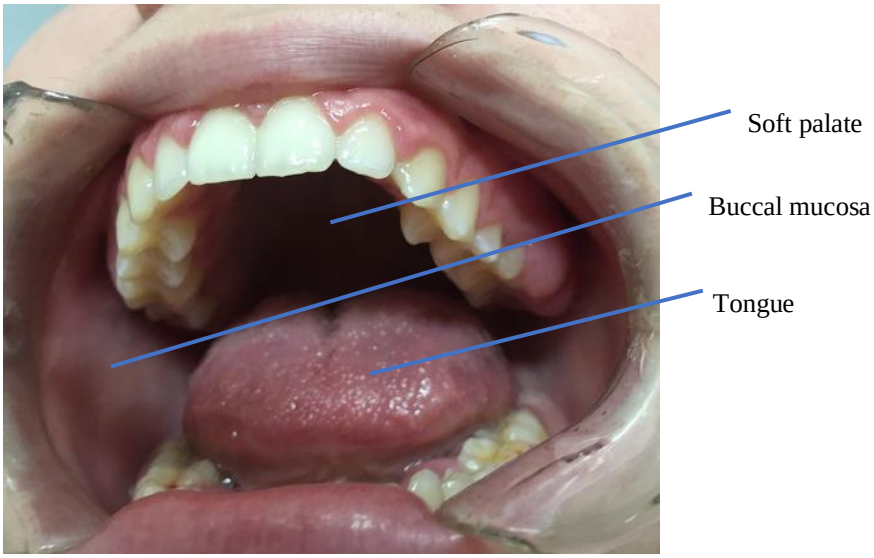



**Figure 2: Anatomical structures involved with oral mucositis [28]**

## *Impact of OM*

Oral mucositis can have a detrimental effect on patients in several ways. Treatment outcomes may be jeopardized by treatment interruptions necessitated by severe OM [27]. The

increased risk of infection, especially for immunosuppressed patients, is associated with a higher rate of infection-related deaths [27]. Patients are also more likely to be hospitalized and spend more time in hospital [27]. OM is associated with higher weight loss and difficulty eating. The impact of the pain from OM on quality of life is also significant. Furthermore, the requirements for intensified care and hospitalization also pose an economic burden to the patient and/or hospital [27].

### *Grading of OM*

**Table 1** lists the grading criteria for OM under the RTOG, CTCAEv2 and CTCAEv5 grading systems. These criteria are broadly similar regarding the severity of each grade, however, the emphasis on physical presentation versus functional impact versus subjective pain level varies. The criteria for severe grades 3 and 4 show a significant impact on patient quality of life, where the pain is sufficient to interfere with oral intake and requires narcotic grade painkillers. Patient-reported scales for the assessment of OM have also been used to measure subjective pain and impact on quality of life in certain studies. In Hong Kong, assessment of OM for NPC patients is conducted as follows: OM is assessed using the latest CTCAE grading system by the doctor or nurse as part of the consultations during RT. The grade is entered into the typed clinical notes, for example, as “G2-3 OM” or “G3 mucosa”. During the seven weeks of RT, the consultations take place approximately once a week. After completion of RT, consultations take place less frequently, with follow up reducing to 6-monthly or yearly consultations. Patient-reported scales for self-reporting of OM are not in widespread use in Hong Kong.

## *Interventions for prevention of OM*

Preventative measures for OM are an active area of research, but the effectiveness of the proposed interventions remains uncertain [25]. One area of research is the use of antioxidant agents, which aim to limit the damage from reactive oxygen species. This direct approach is well justified, but studies have reported mixed effectiveness and some adverse effects [25]. Another approach is the use of drugs that inhibit the production of pro-inflammatory cytokines. While promising, evidence for these approaches is in a preliminary phase [25]. Several natural agents have also been proposed for the prevention of OM, including glutamine, vitamin E, zinc, essential oils, herbal drugs, and topical honey. There have been a few proposed mechanisms of action for natural agents, and these frequently have the advantage of being well-tolerated by patients and posing little risk. However, further evidence is required to support their use [25]. Low-level laser therapy, or photobiomodulation, employs monochromatic light in the red band of the spectrum to promote healing and inhibit inflammation. Research into this treatment is ongoing, but it is considered safe, and is recommended by MASCC/ISOO clinical guidelines to keep patients who are getting hematopoietic stem cell transplants from getting OM [29]. The use of oral cryotherapy, where ice is held in the mouth for a period of time before treatment, has also been explored and likewise recommended for patients undergoing certain treatments [25].

## *Interventions for management of OM*

Management of oral mucositis (OM) symptoms primarily involves the use of analgesics, with narcotics prescribed for severe cases. Pain relief can be administered through mouthwashes containing analgesics or anaesthetics, among other ingredients. Additionally,

topical treatments that form a protective barrier over ulcerations are also explored. Maintaining oral hygiene is crucial to minimize infection risk. The MASCC/ISOO recommends dental hygiene practices and the use of non-medicated mouth rinses [29]. Certain alcohol-based rinses or rinses with active ingredients may be difficult to tolerate and have not been found to be sufficiently effective [27].

### *Necessity for predictive models for OM*

While most HNC patients undergoing chemoradiation will experience OM, identifying those at high risk for severe OM is crucial. Early identification facilitates personalized management strategies, which can reduce the risk of hospitalization and treatment interruptions. This approach not only improves the patient's quality of life but also ensures better adherence to the treatment. Additionally, developing predictive models for OM can provide valuable insights into its risk factors and etiology, further enhancing preventive and therapeutic measures. Current research on the predictive factors and models for severe OM is outlined in **Section 1.2**.

## *1.1.5. Dysphagia*

### *Introduction to dysphagia*

Dysphagia, or difficulty swallowing, refers to physical or functional impairment rather than pain involved in swallowing (odynophagia). Dysphagia results in difficulty with solid food, progressing in severity until even the intake of liquids is restricted. The risk of dehydration and weight loss require medical intervention, which is typically in the form of nasogastric tube feeding. Dysphagia also poses the risk of aspiration and choking.

*Swallowing mechanisms*

The swallowing mechanism involves a complex coordination of different muscles and nerves that facilitate the transport of a bolus of food from the oral cavity into the stomach via the oesophagus. Matsuo et al. describe the swallowing mechanism for a liquid bolus (mass of a substance) and for a solid bolus [30]. In the case of liquids, the bolus is initially confined to the oral cavity by a seal formed by the soft palate and tongue. Next, the bolus is transported into the pharynx by the motion of the tongue, whereupon the pharyngeal stage of swallowing begins. For solid food, a bolus may accumulate in the oropharynx while chewing continues, and there is overlap between the oral and pharyngeal stages. Nevertheless, in the case of both liquids and solids, the next stage is the pharyngeal swallow. This involves the airway being sealed off, preventing the bolus from entering the larynx and trachea, while simultaneously transporting the bolus down into the oesophagus through a coordinated set of muscle contractions [30]. The final component of swallowing involves the transport of the bolus along the oesophagus into the stomach. Muscle contraction behind the bolus and relaxation in front of the bolus work together to achieve this [30]. The swallowing mechanism is illustrated in **Figure 3**.

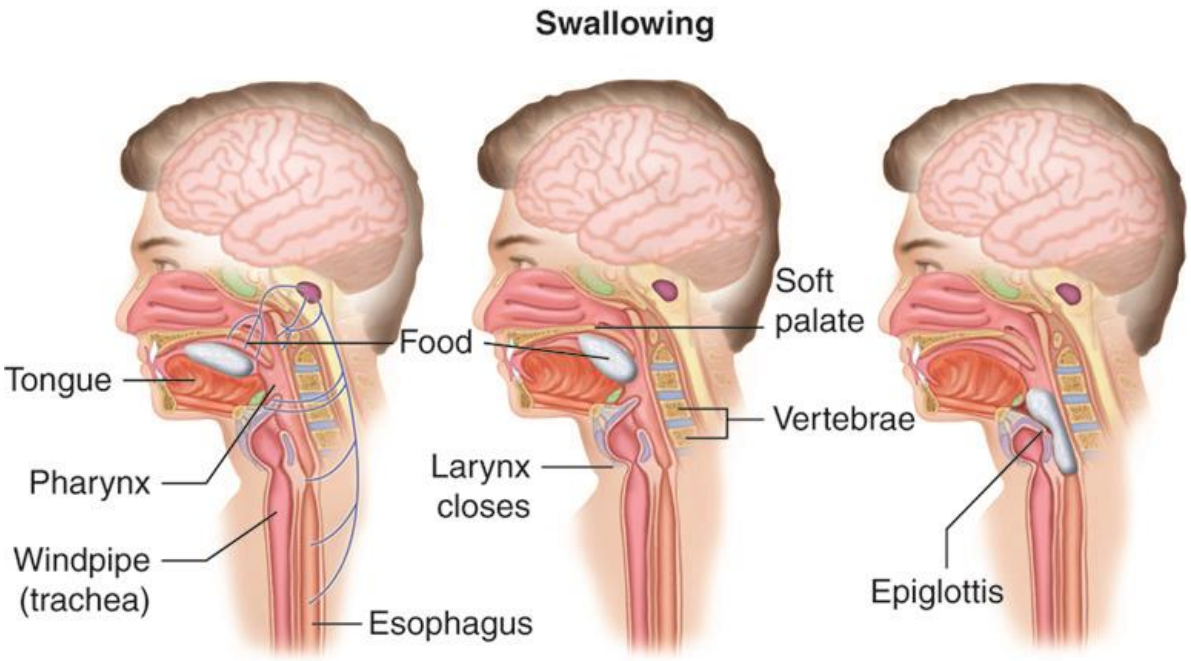



**Figure 3: Anatomical structures involved during swallowing [31]**

## *Causes of dysphagia*

Dysphagia can arise from dysfunction in various components of the swallowing mechanism. Pre-treatment dysphagia may result from obstruction by the tumour or infiltration of swallowing-related structures [32]. Dysphagia induced by RT is typically due to neural and soft tissue damage, along with inflammation, swelling, pain, and altered saliva production [32]. Acute dysphagia, which occurs during and shortly after treatment, often resolves over time. However, late dysphagia can develop from fibrosis and permanent neurological impairment [32].

## *Diagnosis of dysphagia*

Dysphagia may be diagnosed using video fluoroscopy, where a contrast medium is swallowed and visualized on real-time X-ray. Endoscopy can also be used to inspect swallowing-related structures and action. In clinical practice, HNC patients who develop dysphagia during treatment are typically graded (**Table 2**) according to the functional impact.

## *Impact of dysphagia*

A significant proportion of head and neck cancer (HNC) patients undergoing RT develop acute or late dysphagia [33-35]. This difficulty in swallowing can lead to serious complications, including an increased risk of aspiration-related infections and choking. Additionally, dysphagia can adversely affect treatment outcomes by impairing oral intake and leading to weight loss. The condition profoundly impacts patients' daily activities, causing significant discomfort and emotional distress. Identifying patients at high risk for severe dysphagia is crucial for implementing targeted support and preventative measures. Such interventions can help minimize weight loss, reduce the risk of aspiration or choking, and improve overall quality of life. The current research on the predictive factors and predictive models for dysphagia is outlined in **Section 1.2.**

## *Grading of dysphagia*

**Table 2** lists the grading criteria for the RTOG, CTCAEv2, and CTCAEv5 grading systems. All three grading systems broadly agree on the criteria for each level. Moderate dysphagia is indicated by diet modification to soft or liquid diet, while severe dysphagia is identified by difficulty consuming liquids. Consequently, dehydration, significant weight loss, and indication for nutrition support in the form of tube feeding all form part of the criteria for

severe dysphagia, which demands urgent intervention. In Hong Kong, assessment of dysphagia is infrequently recorded in the clinical notes from the consultations with doctors or nurses. However, indication for tube feeding is commonly recorded in the weekly consultation notes during RT. This intervention requires patient consent, and so clinical notes contain statements such as "tube feeding was discussed – patient strongly refuses", or "patient agreed to tube feeding – scheduled for DD/MM/YYYY". Severe dysphagia can be inferred from such statements which provide evidence for the indication for tube feeding. During the seven weeks of RT, the consultations take place approximately once a week. After completion of RT, consultations take place less frequently, with follow up reducing to 6-monthly or yearly consultations. Patient-reported scales for self-reporting of dysphagia are not in widespread use in Hong Kong.

### *Tube feeding*

During treatment, dysphagia results in a reduction in oral intake, causing weight loss. This weight loss is responsible for changes in patient geometry which risk deviation from the radiation treatment plan. In the worst case, the tumour may receive insufficient dose and normal tissues may suffer increased toxicity. Additionally, poor nutrition contributes to fatigue and weakness, whereupon patients may be unable to attend treatment fractions or may opt to discontinue part of their treatment, such as chemotherapy. To prevent severe weight loss, clinicians often offer patients liquid nutrition supplements for easier oral intake or recommend tube feeding. This is typically done using a nasogastric feeding tube, which bypasses the swallowing mechanism. However, the insertion of such a tube is uncomfortable, and patients are often reluctant to use it. Even for those who opt for tube feeding, the tube can be

accidentally pulled out. Alternatives to nasogastric tube include percutaneous gastrostomy, where a tube is passed directly into the stomach through the abdominal wall, or parenteral nutrition, where nutrition is provided intravenously.

### *Interventions to prevent dysphagia*

To prevent RT-induced dysphagia, it is desirable to minimize the radiation dose to swallowing-related structures. The introduction of IMRT facilitates such dose reduction, along with a reduced exposure to the salivary glands [32]. Further prevention of dysphagia in HNC patients treated by IMRT may be possible using exercises for speech and swallowing therapy, though results are inconclusive [32]. Another consideration in preventing dysphagia is that disuse of the swallowing mechanism, such as during prolonged tube feeding, increases the risk of late dysphagia [32]. Severe dysphagia may also be prevented by providing interventions for related toxicities such as OM and xerostomia, which can exacerbate the severity of dysphagia.

## *1.1.6. Grading systems for OM and dysphagia*

**Table 1: Grading systems for OM**

| Grading system | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 |
|---|---|---|---|---|---|---|
| RTOG 1995 | No change over baseline | Injection/may experience mild pain not requiring analgesic | Patchy mucositis that may produce an inflammatory serosanguinous discharge / may experience moderate pain requiring analgesia | Confluent fibrinous mucositis / may include severe pain requiring narcotic | Ulceration, haemorrhage, necrosis | Death |
| CTCAEv2 | None | Erythema of the mucosa | Patchy pseudomembranous reaction (patches generally ≤ 1.5 cm in diameter and non-contiguous) | Confluent pseudomembranous reaction (contiguous patches generally >1.5 cm in diameter) | Necrosis or deep ulceration; may include bleeding not induced by minor trauma or abrasion | Death |
| CTCAEv5 | None | Asymptomatic or mild symptoms; intervention not indicated | Moderate pain or ulcer that does not interfere with oral intake; modified diet indicated | Severe pain; interfering with oral intake | Life-threatening consequences; urgent intervention indicated | Death |

**Table 2: Grading systems for (acute) dysphagia**

| Grading system | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 5 |
|---|---|---|---|---|---|---|
| RTOG 1995 | No change over baseline | Mild dysphagia / may require soft diet | Moderate dysphagia / may require puree or liquid diet | Severe dysphagia with dehydration or weight loss > 15% from pretreatment baseline requiring N-G feeding tube, iv. fluids or hyperalimentation | Complete obstruction, ulceration, perforation, fistula | Death |
| CTCAEv2 | None | Mild dysphagia, but can eat regular diet | Dysphagia, requiring predominantly pureed, soft, or liquid diet | Dysphagia, requiring feeding tube, IV hydration or hyperalimentation | Complete obstruction (cannot swallow saliva); ulceration with bleeding not induced by minor trauma or abrasion or perforation | Death |
| CTCAEv5 | None | Symptomatic, able to eat regular diet | Symptomatic and altered eating/swallowing | Severely altered eating / swallowing; tube feeding, TPN, or hospitalization indicated | Life-threatening consequences; urgent intervention indicated | Death |

### *1.1.7. Multi-omics*

With the advent of DNA sequencing, it became possible to map the entire genome of an organism. Global analysis of a genome was termed ‘genomics’, which along with proteomics, transcriptomics metabolomics, lipidomics and epigenomics, represent global analyses of biological data [36]. The suffix ‘omics’ is therefore associated with high-throughput analysis of large quantities of data that describe a biological system [37]. In 2012, a new form of omics derived from medical imaging was termed ‘radiomics’, under the premise that these images contain information about biological processes [38, 39]. More recently, similar kinds of quantitative data was extracted from the RT dose plan in ‘dosiomics’ [40] and from the geometric relationships between tumour and organ contours in ‘contouromics’ [41]. In this thesis, the role of radiomics, dosiomic and contouromics in toxicity prediction is explored, representing a ‘multi-omic’ analysis.

#### *Radiomics*

Radiomics is a quantitative approach to medical imaging, which aims at enhancing the existing data available to clinicians by means of advanced mathematical analysis [39]. Through this analysis, patterns not readily visible to the human eye may be identified. Specifically, radiomics involves the extraction of numerical features from an area or volume-of-interest within a medical image. The distribution of voxel intensities is described in a set of first order ‘intensity’ features. The geometry of the volume-of-interest is described by a class of ‘shape’ features. The texture within the volume is quantified using a set of matrices: grey-level co-occurrence matrix (GLCM), grey-level difference matrix (GLDM), grey-level run-length matrix (GLRLM), grey-level size zone matrix (GLSZM), and neighbouring grey-tone

difference matrix (NGTDM). Additionally, the medical image may be filtered prior to feature extraction for better discrimination of patterns. Filters include smoothing, edge detection and high and low pass filters. There are over 100 standard radiomic features, which can amount to thousands of individual features once combined with different image filters, volumes-of-interest, and image modalities. Together, the set of radiomic features can provide a comprehensive description of the medical imaging data pertaining to a biological volume. Importantly, each radiomic feature has an established mathematical definition to facilitate accurate reproduction across patients, scans, and centres. To ensure reproducibility of these features, the Image Biomarker Standardization Initiative (IBSI) published a protocol to which the various suppliers of software for radiomic feature extraction can comply [42]. A review by Gul et al. in 2021 identified several applications of radiomics in HNC, including prediction of survival, HPV status and treatment response, as well as models for diagnosis and staging [43]. Radiomic features have also been reported in connection with toxicity prediction models. This is discussed further in **Section 1.3**.

### *Dosiomics*

The development of dosiomics was motivated by the desire to comprehensively characterize the planned RT dose in order to predict xerostomia in HNC patients [40]. Similarly to radiomics, features describing the intensity and texture of the dose map may be extracted, as well as features describing the dose gradient and dose moments. Together, these features can describe the intensity and spatial distribution of the planned radiation dose. Dosiomics expands on the simple dose statistics which are commonly calculated during RT planning. During planning, a dose-volume histogram (DVH) is often plotted as a plan evaluation tool. The DVH

summarizes a complex 3D dose distribution by plotting the relative or absolute volume of a tumour or organ receiving a particular dose over a range of dose bins. While the DVH can show the intensity and overall homogeneity of the dose to a volume, it does not provide information on the spatial distribution of the dose. Conventional toxicity prediction models often utilize DVH parameters, however the interest in dosiomics is motivated by the potential to better characterize the dose distribution and thereby obtain more personalized and accurate prediction of toxicity.

### *Contouromics*

Contouromics was introduced in a study that developed a prediction model for adaptive RT eligibility for NPC [41]. Pre-treatment radiomic and dosiomic features were combined with the newly developed contouromic features, which characterized the complex geometric relationships between pairs of organ-at-risk (OAR) and tumour volumes. Each patient has unique geometry, which poses a challenge in radiation planning, where the radiation must be targeted at the tumour while minimizing the dose to organs-at-risk (OARs). Contouromic features quantify the distance and angular relationships between these pairs. Description of distance between two volumes is not as simple as for two points. Therefore, a distance overlap volume histogram (OVH) was computed for the pair of volumes. This histogram ranged from the minimum separation between the volumes, to the maximum separation between any two points in the volumes. It represents the distance by which the tumour would have to be uniformly expanded to overlap a given fraction of the organ volume. Therefore, the OVH contains comprehensive distance information, including the minimum, maximum and median separation, as well as all points in between. Additionally, the angular relationship was

characterized by a projection overlap volume histogram (POV) which indicates the fraction of the organ which is masked by the tumour when viewed from a given angle about the selected rotation axis. This concept relates to the beam's eye view of a linear accelerator treatment machine, where the tumour and organ contours will appear to be overlapping at certain gantry rotation angles. Together, these features describing distance and angular relationships offer the potential to characterize the patient's geometry and the consequent difficulty of dose sparing. Inclusion of contouromic features is motivated by the hypothesis that patients whose geometry poses a greater challenge may consequently receive more radiation to normal tissues and may therefore experience more severe toxicity.

### *Artificial intelligence, machine learning, and deep learning*

Artificial intelligence (AI) is a topic of profound importance and widespread attention in the present day. However, the term is very broad, as reflected by the definition by European Commission High-level Expert Group on Artificial Intelligence [44]:

*"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."*

Machine learning, commonly considered a subdomain of AI, is concerned with algorithms that learn from data and can generalize to unseen data. As with AI, the term 'learning' is prone to anthropomorphism. A lay person may relate this to a "thinking machine", however a simple definition is that the algorithm "learns" to perform a task if its performance in conducting the task improves with experience [45]. Generally, machine learning can be classified into three approaches: supervised learning, where the algorithm learns by comparing its output to a reference 'ground truth', unsupervised learning, where the algorithm only has access to the unlabelled training data and attempts to learn patterns in the data without human input, and reinforcement learning, where feedback is used to guide the learning process rather than a specific ground truth. Generally, the input data may be represented in the form of a $p$ x $n$ array of $p$ predictors or "features" and $n$ samples. In this way, tasks from a wide range of fields of study can be expressed as a machine learning task. Many 'omics' utilize machine learning for model development, and this is especially true for the selected multi-omics in this thesis: radiomics, dosiomics and contouromics. Machine learning is used to learn patterns in the large number of features in a training set and develop a model which can generalize to the test set. Many different algorithms have been explored in the context of radiomics and related omics, including traditional logistic regression, decision tree, support vector machine (SVM) and many more.

Deep learning is a subdomain of machine learning which is increasingly familiar to the public through its applications in natural language processing and generative AI. Generated images and video are increasingly prevalent in digital media, while tools such as ChatGPT and Google Gemini have been groundbreaking developments. The term 'deep learning' refers to

the use of a neural network with three or more layers [46]. Neural networks are algorithms which draw inspiration from the networks of neurons in the brain by having interconnected nodes which transmit signals based on the input from connected nodes. Developments in technology have facilitated the development of highly complex neural networks with vast numbers of parameters. Deep learning is typically more computationally expensive and involve many more parameters than other machine learning models. Consequently, they tend to require much larger sample sizes. The structure of deep learning models allows highly complex representations of the input to be learned, permitting the analysis of images or text. However, unlike in more conventional machine learning models, each parameter does not have a predefined meaning. Combined with the huge number of parameters, interpretation of deep learning models is not possible in the conventional way. For this reason, deep learning models are often considered as 'black box' methods. While methods for the interpretation and explainability of deep learning models have been explored in the literature, these tend to be qualitative. Deep learning can achieve excellent results but is limited by available sample size and interpretability. In relation to multi-omics, the medical image data, radiation dose data and contour data can be put into deep learning models directly, however the learned parameters will not be interpretable as the pre-defined mathematical features included in radiomics, dosiomics and contouromics. Alternatively, it is possible to develop deep learning models using the aforementioned omics as inputs. In this case, the deep learning model can be used much like any other machine learning model. However, the limitations on sample size and may limit the complexity of such a model and restrict its performance. In this situation, there is limited justification for using a deep learning model. For these reasons, this work focused on

machine learning applications of multi-omics in the prediction of severe acute OM and dysphagia

# 1.2. Predictive factors for chemoradiotherapy-induced OM and dysphagia in HNC

This chapter reports a systematic scoping review on predictive factors for chemoradiation-induced OM and dysphagia in HNC that was published in *Cancers* [1].

## *1.2.1. Introduction*

OM and dysphagia are two of the most prevalent toxicities experienced by HNC patients undergoing RT and have a systemic impact on patients, hampering treatment outcome and harming quality of life. A study on 212 HNC patients undergoing IMRT reported that 50% of patients suffered from moderate-to-severe OM, while 75% faced moderate-to-severe dysphagia [47]. The consequences of severe toxicity can be life-threatening: another study found that 9% of HNC patients were hospitalized or sought emergency care due to acute OM toxicity, and an even higher 19% for dysphagia-related issues [48]. The pain and discomfort which directly result from these toxicities are combined with the frustration over impairments to everyday actions such as eating and drinking. Interventions to maintain nutrition, such as nasogastric tube feeding, can also induce discomfort. The impact on QoL should not be underestimated, with these conditions inflicting a toll on the patient's physical, emotional, and psychosocial health. Accurate prediction of patients at risk of severe toxicity is crucial for improving management strategies and ultimately, patient outcomes.

This scoping review aimed to systematically map current literature on predictive factors for chemoradiation-induced OM and dysphagia among HNC patients, providing quantitative as well as qualitative synthesis. This study sought to address two primary research questions: 1) Which factors are recognized as predictors for treatment-induced OM and dysphagia in HNC

patients? 2) How efficacious are the prevailing predictive models in forecasting the severity of these toxicities? This review synthesized current evidence to offer clinicians and researchers insights for enhancing predictive model development.

## *1.2.2. Materials and methods*

To identify studies on predictive factors for chemoradiation-induced OM and dysphagia in HNC patients, a systematic literature search was conducted in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols Extension for Scoping Reviews (PRISMA-ScR) guidelines [49]. Data were collected through Embase, PubMed, Scopus, and Web of Science from year 2000 to September 2023. The detailed search strategy for each database is provided in **Search strategy** for literature review

**Table 58**. A flowchart of the study search and screening process is illustrated in **Figure 4**.

Eligible studies included those that reported predictive factors for OM or dysphagia. If both toxicities were reported, separate records were created for each. The outcome measures related to the severity, incidence or duration of OM or dysphagia. Measures of dysphagia included objective and subjective severity scales and indicators including diagnoses of stricture or aspiration based on video fluoroscopy or modified barium swallow. Aspiration pneumonia, as a secondary consequence of dysphagia, was not included as an outcome measure. Included studies reported statistically significant factors or predictive models for the outcome. Study subjects were restricted to patients with head and neck cancer who received RT and/or chemotherapy. Consequently, studies using animal or in vitro models were excluded. Studies

with 10 or fewer subjects were excluded, as were those not available in English, or those published before the year 2000.

After removing duplicates, screening was performed in two phases. The first phase involved screening by the title and abstract, followed by full text screening in the second phase. Any non-full-length articles were excluded, due to limited detail and lower quality of evidence. A critical appraisal of individual sources of evidence was not conducted, considering the diversity in study designs and the volume of literature included. Moreover, the primary purpose of the review was to map the existing research rather than assess the level of evidence of each study.

Data charting was conducted separately for OM and dysphagia. For each outcome, details of each study were tabulated in a spreadsheet. Data items included sample size, treatment regimen, outcome measure, outcome incidence, and timeframe. Factors reported to be significantly correlated with the outcome (p-value $< 0.05$) were recorded under the appropriate category, along with an indication of whether the factor was significant in univariate analysis, multivariate analysis, or was reported as a model feature. To quantify the amount of evidence for each factor, the number of studies reporting it as significant in multivariate or model analysis was calculated for each toxicity outcome. Univariate analyses were not included because of concerns over multiple testing bias and confounding variables. Factors for each toxicity outcome were grouped by factor type and ranked by the number of studies reporting it as significant in multivariate or model analysis. The number of studies reporting each factor as significant in any analysis (univariate, multivariate or model) was also included for comparison.

Additionally, a subset of the included studies that reported predictive models for OM or dysphagia were investigated. Studies were included in this subset if they provided some form of validation performance score. The time frame, endpoint definition, model features, sample size and validation type were recorded, along with the test performance score. This was typically reported in the form of the area under the receiver operating characteristic curve (AUC). An AUC close to 0.5 indicated that the prediction was equivalent to random chance, while a value close to 1.0 indicated a perfect prediction.

## *1.2.3. Results*

*Identification and selection of studies*

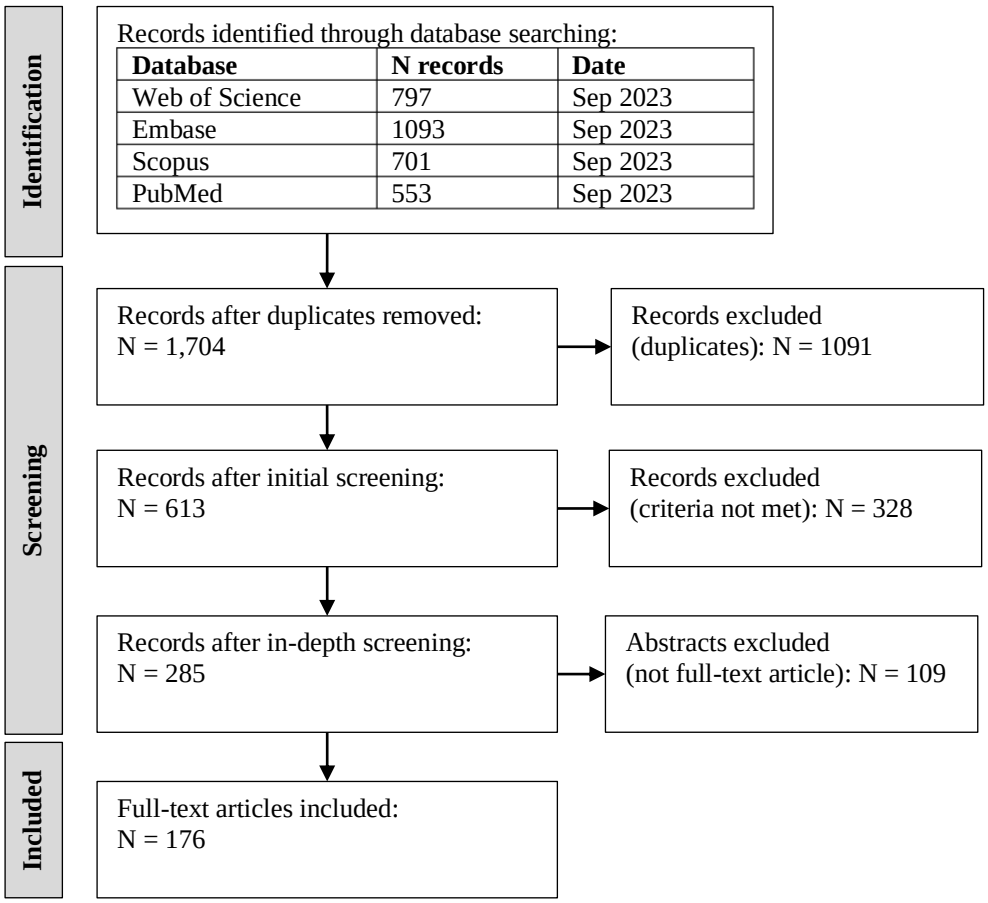



**Figure 4: Flow diagram of selection of sources of evidence.**

The literature search resulted in 1704 unique records, of which 1528 were excluded during the screening process. A total of 176 full-text articles were included for this review. Seventy-three articles reported predictors for OM [50-122], five reported predictors for both OM and dysphagia [123-126], and ninety-eight reported predictors for dysphagia alone [33, 54, 127-222].

## *Overview of included studies*

**Table 3** summarizes the 176 full-text articles included in this review. Some articles reported predictors for both OM and dysphagia, so were treated as separate records in each analysis. The median number of HNC patients in each study were similar for OM (n=91) and dysphagia (n=100). The overall frequency of RT, chemotherapy, and surgery as treatments is also listed. Almost all the patients received RT, and most patients received chemotherapy. It should be noted that 81% of studies did not report the incidence of surgery as a treatment, so the proportion of patients with surgical history for OM studies (54%) and dysphagia studies (24%) may be inaccurate. Clinician-rated outcomes were most reported for both OM and dysphagia. OM was almost exclusively investigated in the acute period using clinician-rated gradings. Dysphagia was investigated at both the acute and late periods, with the majority focused on late dysphagia. Note that some studies reported predictors for both acute and late toxicities, and some without the timeframe specified. Among studies on OM, 59% included only univariate analysis, while 41% utilized multivariate analysis or developed a predictive model. Among studies on dysphagia, 37% included only univariate analysis, while 63% utilized multivariate analysis or developed a predictive model.

**Table 3: Summary statistics of included studies**

| Summary statistic | OM | Dysphagia |
|---|---|---|
| Full-text articles (N) | 78 | 103 |
| Patient cohort size (median) | 91 | 100 |
| Patient cohort with RT history (%) | 100 | 100 |
| Patient cohort with chemotherapy history (%) | 80 | 88 |
| Patient cohort with surgical history (%) | 54 | 24 |
| Clinician rated outcome (%) | 85 | 84 |
| Patient reported outcome* (%) | 4 | 21 |
| Investigated acute toxicity (%) | 80 | 40 |
| Investigated late toxicity (%) | 1 | 62 |
| Studies with univariate analysis only (%) | 59 | 37 |

* Patient reported outcome refers to toxicity outcomes defined by results of questionnaires or scales completed by the patient, such as pain rating, oral intake, swallowing ability or quality of life scales.

**Table 4** describes the incidence of the top-reported OM and dysphagia outcomes. The most frequently reported OM outcome was RTOG/CTCAE/WHO grade 3+ (severe). Approximately 56% and 42% of patients endured grade 2+ (moderate-or-higher) and grade 3+ (severe-or-higher) OM during their treatment respectively. Among dysphagia outcomes, severe dysphagia (as indicated by tube feeding or RTOG/CTCAE grade 3+) was very common in the acute period.

**Table 4: Toxicity outcome incidences**

| Toxicity | Timeframe | OM outcome | Incidence | Reported by N studies |
|---|---|---|---|---|
| OM | Acute | RTOG/CTCAE/WHO grade 3+ | 42% | 47 |
| | | RTOG/CTCAE/WHO grade 2+ | 56% | 8 |
| Dysphagia | Acute | Tube feeding(use/indication/dependence) | 37% | 13 |
| | | RTOG/CTCAE grade 3+ | 37% | 11 |
| | | RTOG/CTCAE grade 2+ | 49% | 2 |
| | Late | Tube feeding(use/indication/dependence) | 17% | 14 |
| | | RTOG/CTCAE grade 3+ | 23% | 7 |
| | | RTOG/CTCAE grade 2+ | 28% | 4 |

**Table 5**, **Table 6**, and **Table 7** detail the analysis of predictive factors for acute OM, acute dysphagia, and late dysphagia. Univariate analysis does not account for relationships between factors and may be influenced by confounding. Therefore, factors were ranked by the number of studies that identified them as significant in multivariate analysis or predictive models. The factors can be classified into seven categories: patient, tumour, treatment, organ-

at-risk (OAR) dose, clinical laboratory test results, genetic expression, and "other". Patient factors include demographics; tumour factors include TNM staging and tumour site; treatment factors include treatment modalities and regimen. Dose factors pertain to the radiation delivered to the organ-at-risk (OAR); genetic factors refer to single nucleotide polymorphisms of genes which were found to be correlated with toxicity; clinical laboratory test results include the results of blood, saliva, or stool tests.

### *Predictors of oral mucositis*

For acute oral mucositis (OM, **Table 5**), smoking was the patient factor most frequently reported as significant in multivariate analysis, followed by sex, body mass index and age. Other factors included weight loss, performance status score and number of teeth. Alcohol was reported by two studies in univariate analysis. The tumour factors most reported in multivariate analysis were tumour site, T-stage, and N-stage. Interestingly, the primary tumour volume was reported by one study in univariate analysis. Treatment factors were led by use of concurrent chemotherapy, followed by chemotherapy drug, RT fractionation, , neoadjuvant chemotherapy, retropharyngeal lymph node irradiation, RT delivery time, RT field size, RT modality and surgery related factors. The dose factor most reported in multivariate analysis was the radiation dose to the oral cavity or extended oral cavity, followed by the dose to the oral mucosal surface, parotid glands, and pharyngeal space. Additionally, the dose to the tongue and pharyngeal constrictor muscles were identified in univariate analysis. Many studies investigated the role of clinical laboratory test results, including white blood cell lymphocyte count, erythrocyte sediment rate (ESR) and γ-H2AX (protein marker) and presence of candida fungus. The role of RADIODTECT blood assay, epidermal growth factor and neutrophil-to-lymphocyte ratio

were also reported. Fourteen studies reported genetic factors as predictors of OM. However, only four adopted multivariate analysis. Tumour necrosis factor alpha was reported by two studies, but with different genotypes reported by each study (TT and GG) [67, 104]. Single nucleotide polymorphisms of XRCC1, a gene involved with DNA repair, were reported by two studies [85, 115]. The remaining ten studies that reported genetic factors each returned a different factor. Beyond the previously mentioned categories, one study crafted a prediction model employing radiomic and dosiomic features derived from the primary tumour volume [57]. Two studies highlighted a significant correction between bioelectrical impedance measurements and OM in univariate analyses [68, 106], while another study identified perfusion parameters as a significant determinant in a univariate analysis [103]. The most robust factors, significant in multivariate or model analysis, include RT dose to the oral cavity / oral mucosa, concurrent chemotherapy, smoking, tumour site, gender, and RT dose to the parotid glands.

**Table 5: Number of studies demonstrating significant correlations between acute OM and various factors.**

| Factor Type | Factor | Multivariate or model | All analyses |
|---|---|---|---|
| Clinical laboratory tests | Blood, saliva, or stool properties | 14 | 26 |
| Dose | RT dose to oral cavity (entire volume) | 10 | 11 |
| | RT dose to oral mucosa (surface only) | 6 | 7 |
| | RT dose to parotid glands | 4 | 4 |
| | RT dose to pharyngeal space | 1 | 1 |
| | RT dose to constrictor muscle | 0 | 1 |
| | RT dose to tongue | 0 | 1 |
| Treatment | Concurrent chemotherapy | 8 | 13 |
| | Chemotherapy drug | 2 | 2 |
| | RT fractionation | 2 | 3 |
| | Neoadjuvant chemotherapy | 2 | 2 |
| | Retropharyngeal lymph node irradiation | 1 | 1 |
| | RT delivery time | 1 | 3 |
| | RT field size | 1 | 1 |
| | RT modality | 1 | 1 |
| | Surgery related factors | 1 | 2 |
| | Number of chemotherapy cycles | 0 | 1 |
| | Use of tongue immobilizer | 0 | 1 |
| Patient | Smoking | 6 | 9 |
| | Sex | 4 | 6 |
| | Body mass index | 3 | 6 |
| | Age | 3 | 9 |
| | Baseline weight loss | 2 | 3 |
| | Performance status score | 1 | 3 |
| | Number of teeth | 1 | 1 |
| | Alcohol-related | 0 | 2 |
| Tumor | Tumor site | 5 | 7 |
| | T-stage | 3 | 5 |
| | N-stage | 3 | 3 |
| | Primary tumor volume | 0 | 1 |
| Genetic | Genetic factors | 4 | 14 |
| Other | Radiomic / dosiomic features | 1 | 1 |
| | Bioelectrical impedance measurement | 0 | 2 |
| | Perfusion / blood flow measurement | 0 | 1 |

**Table 6: Number of studies demonstrating significant correlations between acute dysphagia and various factors.**

| Factor type | Factor | Multivariate or model | All analyses |
|---|---|---|---|
| Tumor | T-stage | 9 | 11 |
| | Tumor site | 8 | 14 |
| | N-stage | 6 | 8 |
| Treatment | Concurrent chemotherapy | 9 | 11 |
| | RT fractionation | 4 | 7 |
| | Chemotherapy drug type | 3 | 4 |
| | Neck irradiation regimen | 3 | 6 |
| | RT field size | 2 | 3 |
| | Surgery related factors | 2 | 2 |
| | Adjuvant chemotherapy | 1 | 1 |
| | Brachytherapy | 1 | 1 |
| | Neoadjuvant chemotherapy | 1 | 1 |
| | RT modality | 1 | 2 |
| Dose | RT dose to constrictor muscles | 8 | 13 |
| | RT dose to inferior pharyngeal constrictor (IPC) | 6 | 7 |
| | RT dose to superior pharyngeal constrictor (SPC) | 6 | 9 |
| | RT dose to middle pharyngeal constrictor (MPC) | 4 | 6 |
| | RT dose to oral cavity volume / oral mucosa surface | 4 | 5 |
| | RT dose to parotids | 3 | 3 |
| | RT dose to larynx | 3 | 6 |
| | RT dose to esophageal inlet / cricopharnygeus | 2 | 3 |
| | RT dose to esophagus | 1 | 2 |
| | RT dose to pharyngeal mucosa | 1 | 1 |
| | RT dose to pharynx | 1 | 1 |
| | RT dose to submandibular glands | 1 | 1 |
| | RT dose to primary tumor | 1 | 1 |
| Patient | Age | 6 | 8 |
| | Body mass index | 4 | 4 |
| | Performance status score | 4 | 5 |
| | Baseline weight loss | 3 | 5 |
| | Sex | 3 | 5 |
| | Smoking history | 3 | 5 |
| | Pretreatment dysphagia | 2 | 4 |
| | Constrictor muscle geometry | 1 | 1 |
| Clinical laboratory tests | Blood or saliva properties | 3 | 3 |
| Genetic | Genetic factors | 3 | 3 |

**Table 7: Number of studies demonstrating significant correlations between late dysphagia and various factors.**

| Factor type | Factor | Multivariate or model | All analyses |
|---|---|---|---|
| Dose | RT dose to constrictor muscles | 16 | 26 |
| | RT dose to superior pharyngeal constrictor (SPC) | 16 | 18 |
| | RT dose to larynx | 10 | 16 |
| | RT dose to middle pharyngeal constrictor (MPC) | 10 | 12 |
| | RT dose to inferior pharyngeal constrictor (IPC) | 9 | 12 |
| | RT dose to esophageal inlet / cricopharnygeus | 5 | 8 |
| | RT dose to oral cavity volume / oral mucosa surface | 4 | 5 |
| | RT dose to parotids | 3 | 7 |
| | RT dose to tongue or base of tongue | 3 | 6 |
| | RT dose to esophagus | 2 | 4 |
| | RT dose to inferior brain stem | 1 | 1 |
| | RT dose to submandibular glands | 0 | 2 |
| Tumor | T-stage | 13 | 21 |
| | Tumor site | 11 | 18 |
| | N-stage | 8 | 12 |
| Patient | Age | 12 | 14 |
| | Smoking history | 6 | 6 |
| | Baseline / acute weight loss | 5 | 8 |
| | Pretreatment or acute dysphagia | 3 | 7 |
| | Body mass index | 1 | 2 |
| | Performance status score | 1 | 4 |
| | Sex | 1 | 3 |
| | Constrictor muscle geometry | 1 | 1 |
| | Alcohol use | 0 | 1 |
| Treatment | Concurrent chemotherapy | 9 | 9 |
| | Surgery related factors | 4 | 6 |
| | RT fractionation | 3 | 11 |
| | Neck irradiation regimen | 2 | 8 |
| | Chemotherapy drug type | 1 | 3 |
| | Neoadjuvant chemotherapy | 1 | 3 |
| | Adjuvant chemotherapy | 1 | 2 |
| | RT modality | 1 | 2 |
| | Brachytherapy | 1 | 1 |
| | RT field size | 0 | 2 |
| Clinical laboratory tests | Blood or saliva properties | 1 | 1 |

## *Predictors of dysphagia*

Regarding acute dysphagia (**Table 4**), the most reported patient factor in multivariate analysis was age, followed by body mass index and performance status score. In terms of tumour factors, T-stage was the most reported, followed by tumour site and N-stage. For treatment factors, most reported was the use of concurrent chemotherapy, followed by RT

fractionation, chemotherapy drug and neck irradiation regimen. The most reported dose factor was the accumulated radiation dose to the pharyngeal constrictor muscles, specifically the superior and inferior pharyngeal constrictors. This was followed by the dose to the medial constrictor, oral cavity or oral mucosa, parotid glands, larynx, and oesophageal inlet or cricopharnygeus. Three studies [126, 127, 217] reported genetic factors regarding single nucleotide polymorphisms. Two studies [123, 203] reported clinical laboratory test results, including the presence of oral candidiasis and the result of the RADIODTECT blood assay. The most well-supported factors, significant in multivariate or model analysis, were T-stage, concurrent chemotherapy, tumour site, RT dose to constrictor muscles, N-stage, and patient age.

With respect to late dysphagia (**Table 5**), the most common patient factor in multivariate analysis was age, followed by smoking history, and baseline or acute weight loss. The most reported tumour factor was T-stage, followed by tumour site and N-stage. The most reported treatment factor was the use of concurrent chemotherapy, followed by surgery related factors, RT fractionation, and neck irradiation regimen. The most reported dose factor was the radiation dose to the pharyngeal constrictor muscles, specifically the superior pharyngeal constrictor, dose to the larynx, medial and inferior constrictors, oesophageal inlet or cricopharnygeus, oral cavity or oral mucosa, parotid glands, tongue or base of tongue, and oesophagus. The doses received by the inferior brain stem and the submandibular glands were also reported. HPV status was the only clinical laboratory test result factor identified [192]. No genetic factors were identified as significantly correlated with late dysphagia. The most well-supported factors, significant in multivariate or model analysis, were RT dose to constrictor

muscles, T-stage, patient age, tumour site, RT dose to larynx, concurrent chemotherapy, and N-stage.

**Table 8: Predictive models for OM.**

| Ref | Time frame | Endpoint | Model features* | Sample size | Validation type | Test AUC |
|---|---|---|---|---|---|---|
| [50] | Acute | Increase from RTOG G1-G2 | Oral bacteria genetic information | 41 | Internal | 0.646 |
| [51] | Acute | CTCAE G3+ OM | BMI, Combined parotid glands EUD, Oral cavity EUD | 132 | Internal | 0.67 |
| [52] | Acute | CTCAE G3+ OM | Oral cavity $D_{mean}$, Mean RT dose at which 50% of patients experience toxicity (51 Gy), Slope of dose response curve | 169 | External | 0.67 |
| [53] | Acute | CTCAE G3+ OM | Definitive RT, Male, Age, Chemotherapy modality, Chemotherapy drug, Tumour site, Volumes of oral cavity receiving 20-260cGy per fraction in 20cGy/fraction increments | 351 | Internal | 0.71 |
| [54] | Acute | WHO G3+ OM | Age, N-stage, # of cycles of neoadjuvant chemotherapy, V40 (oral cavity) | 190 | Internal | 0.759 |
| [55] | Acute | RTOG G3+ OM | BMI, RLN irradiation, Mucosa surface contour V55 | 270 | Internal | 0.782 |
| [56] | Acute | DAHANCA G3+ OM | Extended oral cavity DVH parameters converted into 2 Principal Components, Treatment acceleration | 802 | Internal | 0.808 |
| [57] | Acute | CTCAE G3+ OM | 4 cT1-w MR and 1 CECT radiomic texture features extracted from gross tumour volume (primary and nodal tumour) | 242 | Internal | 0.81 |

* Dmean = mean dose, BMI = body mass index, DVH = dose volume histogram, Vx = volume receiving x Gy dose, RLN = retropharyngeal lymph nodes, EUD = equivalent uniform dose, cT1-w MR = contrast T1 weighted magnetic resonance image, CECT = contrast enhanced CT image.

## *Predictive models for oral mucositis and dysphagia*

**Table 8, Table 9** and **Table 10** present predictive models from the included studies that underwent either internal or external validation, providing insights into the current predictive performance of published models. Internally validated models were evaluated using hold-out test sets, cross-validation or bootstrapping comprised of samples from the same centre as the training set. External validated models were evaluated on a hold-out test set taken from a

separate centre to assess the generalizability of the model. The details of the outcome measure, model features and type of validation are also tabulated.

All the validated predictive models for OM were specific to the acute period. Severe-or-higher OM (≥ grade 3), as scored by RTOG, CTCAE or WHO, was used as an out-come by 6 out of 8 models [51-55, 57]. Alternatively, an increase in RTOG grade from mild (grade 1) to moderate (grade 2) was used by Zhu et al. [50] and DAHANCA grade 3+ was used by Hansen et al. [56]. The validation performance, as measured by AUC, ranged from 0.65 to 0.81. Clinical features used in the models include sex, age, BMI, tumour site, N-stage, use of chemotherapy, chemotherapy drug, number of cycles of neoadjuvant chemotherapy, treatment acceleration, retropharyngeal lymph node irradiation, and treatment dose parameters. Dose volume histogram (DVH) parameters used in the models included dose to the oral cavity and dose to the mucosa surface contour. Zhu et al. reported a model using genetic information from oral bacteria [50]. Dong et al. reported a model using MR and CECT radiomic features extracted from the gross tumour volume [57]. The model size ranged between 2 to 19 features.

Among the validated predictive models for dysphagia, five predicted acute dysphagia and nine predicted late dysphagia. For acute dysphagia, outcomes were defined as tube feeding dependence [128-130] or CTCAE grading severe or higher (≥ grade 3) [127, 131]. The validation performance, as measured by the AUC, ranged from 0.60 to 0.82. Clinical features used in the models included sex, age, BMI, texture modified diet, tumour site, T-stage, N-stage, performance status, pre-treatment weight loss, use of chemotherapy versus RT alone, use of concurrent chemotherapy, use of induction chemotherapy, and chemotherapy drug. DVH parameters used in the models included dose to the superior and inferior pharyngeal constrictor

muscles, dose to the pharyngeal mucosa, dose to the contralateral parotid gland, dose to the oral cavity, and dose to the contralateral submandibular gland. De Ruyck et al. also incorporated a genetic polymorphism feature into their model [127]. The model size ranged from a single feature up to 20 features.

For late dysphagia, outcomes were defined as tube feeding dependence [137, 140], occurrence of a dysphagia criteria (including tube feeding, aspiration, stricture, aspiration pneumonia) [133, 138, 139], RTOG/CTCAE moderate-or-higher dysphagia (≥ grade 2) [134-136], or improvement in dysphagia grading [132]. The validation performance, as measured by the AUC, ranged from 0.70 to 0.85. Clinical features used in the models included age, T-stage, N-stage, tumour site, HPV status, smoking status, baseline weight loss, baseline dysphagia score, treatment acceleration, use of chemotherapy versus RT alone, neck dissection, and total dose to tumour. DVH parameters used in the models included dose to the pharyngeal constrictor muscles, dose to the larynx, dose to the contralateral parotid, dose to the cricopharyngeal muscle, dose to the mylogeniohyoid, and dose to the oral cavity. The model size ranged from a single feature up to 9 features.

**Table 9: Predictive models for acute dysphagia**

| Ref | Time frame | Endpoint | Model features* | Sample size | Validation type | Test AUC |
|---|---|---|---|---|---|---|
| [127] | Acute | CTCAE G3+ dysphagia | CCT, D2 SPCM, Rs321345_TC(XRCC1) polymorphism | 189 | Internal | 0.6 |
| [128] | Acute | Tube feeding use ≥ 4 weeks | Pre-treatment weight change %, Texture modified diet., ECOG > 0, Tumor site, N-stage ≥ 2, $D_{mean}$ contralateral parotid, $D_{mean}$ oral cavity | 334 | External | 0.624 |
| [129] | Acute | Tube feeding use ≥ 4 weeks | Tumor site, T-stage ≥ 3, Chemotherapy (vs RT alone), $D_{mean}$ contralateral parotid | 225 | Internal | 0.708 |
| [130] | Acute | Tube feeding use ≥ 4 weeks | BMI, Texture modified diet, WHO performance scale > 0, Tumor site, T-stage ≥ 2, N-stage ≥ 2, CCT (vs RT alone), $D_{mean}$ contralateral submandibular gland, $D_{mean}$ contralateral parotid | 450 | Internal | 0.723 |
| [131] | Acute | CTCAE G3+ dysphagia | Definitive RT, Male, Age, IC, No CCT, Chemotherapy drug, Tumor site, Volumes of Pharyngeal mucosa receiving 20-260cGy per fraction in 20cGy/fraction increments | 90 | External | 0.82 |

* ECOG = Eastern Cooperative Oncology Group performance status, $D_{mean}$ = mean dose, $D_x$ = dose to x% of volume, BMI = body mass index, IC = induction chemotherapy, CRT = chemoradiation, CCT = concurrent chemotherapy, [S/M/I]PCM = superior/medial/inferior constrictor muscle, Vx = volume receiving x Gy dose.

**Table 10: Predictive models for late dysphagia.**

| Ref | Time frame | Endpoint | Model features* | Sample size | Validation type | Test AUC |
|---|---|---|---|---|---|---|
| [132] | Late | Dysphagia improvement (reduction of at least one grade from CTCAE grade ≥3) | $D_{min}$ larynx | 90 | Internal | 0.697 |
| [133] | Late | Aspiration (>25 months) | Age, Neck dissection, $D_{mean}$ MPCM | 107 | Internal | 0.73 |
| [134] | Late | RTOG G2+ dysphagia (6 months) | $D_{mean}$ SPCM, $D_{mean}$ supraglottic larynx | 186 | External | 0.75 |
| [135] | Late | CTCAE G2+ dysphagia (6 months) | $D_{mean}$ oral cavity, $D_{mean}$ SPCM, $D_{mean}$ MPCM, $D_{mean}$ IPCM, Tumor site, Baseline dysphagia score | 277 | External | 0.8 |
| [136] | Late | RTOG G2+ dysphagia (6 months) | $D_{mean}$ SPCM, $D_{mean}$ supraglottic larynx | 354 | Internal | 0.8 |
| [137] | Late | Tube feeding dependence (6 months) | T-stage ≥ 3, N-stage > 0, Baseline weight loss, Accelerated RT, CRT, Neck irradiation | 183 | External | 0.82 |
| [138] | Late | Aspiration or stricture or tube feeding or aspiration pneumonia (> 12 months) | Age, V69 Mylo/geniohyoid complex | 300 | Internal | 0.835 |
| [139] | Late | Feeding tube insertion or aspiration (6 months) | Tumor-organ distances for superior, inferior and medial pharyngeal constrictors, plus mylogeniohyoid, cricopharyngeal muscle and supraglottic larynx, Clinical feature clusters comprised of smoking status, T-stage, N-stage, HPV status, Pathological grade, tumor site, CRT combination, tumor laterality, age, total dose to tumor | 200 | Internal | 0.84 |
| [140] | Late | Tube feeding dependence (6 months) | T-stage ≥ 3, Baseline weight loss > 10%, RT + cetuximab, Accelerated RT, CCT, $D_{mean}$ SPCM, $D_{mean}$ IPCM, $D_{mean}$ contralateral parotid, $D_{mean}$ cricopharyngeal muscle | 355 | Internal | 0.85 |

* ECOG = Eastern Cooperative Oncology Group performance status, $D_{mean}$ = mean dose, $D_x$ = dose to x% of volume, BMI = body mass index, IC = induction chemotherapy, CRT = chemoradiation, CCT = concurrent chemotherapy, [S/M/I]PCM = superior/medial/inferior constrictor muscle, Vx = volume receiving x Gy dose.

### *1.2.4. Discussion*

To our knowledge, this was the first scoping review to map current literature on predictors of and predictive models for the severity of OM and dysphagia in HNC patients. One hundred and seventy-six studies were included in this review. The reported predictors were categorized, grouped by toxicity and timeframe, and the numbers reported in univariate and multivariate analysis were analysed (**Table 5**, **Table 6**, **Table 7**). Additionally, eight, five and nine studies that reported predictive models for the severity of acute OM, acute dysphagia and late dysphagia were analysed (**Table 8**, **Table 9**, **Table 10**).

#### *Predictors of OM and dysphagia*

A broad range of predictors for the severity of OM and dysphagia have been identified, indicating the multifactorial and complex aetiology of these conditions. Ranking predictors by the number of studies where they were significant in multivariate analysis is indicative of the quantity of evidence per predictor. Some predictor types, such as genetic factors for OM, were frequently reported as significant in univariate analyses, but not in multivariate analyses. For example, genome-wide association studies reported genetic variants associated with acute OM [108, 120]. Even in other studies where multivariate analysis was performed, a limited set of predictor types were included. Further investigation is required to confirm the independent value of predictors and identify combined or interactive effects among a comprehensive range of predictor types.

#### *Performance of predictive models*

Predictive models mostly focused on severe toxicity, indicating their intended use for identifying high-risk patients who can be targeted for closer monitoring and more aggressive

preventative measures. For acute OM, eight models were identified with AUCs ranging from 0.65 to 0.81 [50-57]. Five models emerged for acute dysphagia (AUC: 0.60-0.82) [127-131], and nine for late dysphagia (AUC: 0.70-0.85) [132-140]. Considerable variability in model performance was evident, suggesting opportunities for further improvement. The best performing models tended to incorporate multiple predictor types. For example, for OM, Hansen et al. included treatment acceleration alongside DVH parameters of the extended oral cavity [56], and Dong et al. included texture features from multiple imaging modalities [57]. For dysphagia, Dean et al. included patient, tumour, treatment and DVH parameters of the pharyngeal mucosa [131]. Wopken et al. included tumour, treatment and DVH parameters from multiple organs at risk (OARs) [140]. Investigation of a broader range of factor types offers potential to capture more of the multifactorial nature of these toxicities and further improve performance, provided that the challenges of increased dimensionality and interactive effects can be addressed.

### *Limitations of predictive models*

External validation on data from a separate centre provides a higher level of evidence. However, less than one third of the models were externally validated. Many of the studies reporting these models highlighted the lack of external validation and small sample size as limitations.

One of the main challenges involved in developing a predictive model for OM or dysphagia is in its generalizability to other clinical centres. The differences between centres may explain why only 27% of the studies utilized external validation. The grading system used for assessing toxicity can vary, as can the criteria for interventions such as tube feeding, a

common outcome measure for dysphagia. For example, Dean et al. observed a difference in the scoring system for dysphagia between their training data and their validation data [131], and Willemsen et al noted that individual and in-situational preferences in feeding tube insertion policy may affect the apparent incidence of dysphagia [128]. Furthermore, the treatment regimen can also vary between centres. For example, there may be different guidelines for the use of neoadjuvant, concurrent or adjuvant chemotherapy depending on tumour site and staging, and different guidelines for choice of chemotherapy drug or radiation delivery. Sharabiani et. al. recognized that the normal tissue complication probability (NTCP) models they used were not fully up to date with current treatment regimens, and so the generalizability of the models would be reduced [52]. The contouring of OARs may also vary between centres, especially for organs which are not commonly delineated during standard practice such as the oral mucosa surface [53].

Models generally did not incorporate all types of predictors. For example, clinical laboratory test results and genetic factors were underrepresented in the models and may offer potential to enhance prediction. For example, blood tests have an established role in patient monitoring. Information such as blood cell counts, and presence of inflammatory markers have been highlighted as potential predictors of severe toxicity. Incorporation of factors such as blood group type and its relationship with head and neck cancer subtypes may also offer potential for more personalized models [223]. Regarding genetic factors, Hansen et al. suggest that characterizing normal tissue radiosensitivity through genomic or microbiomic data might improve prediction of OM [56]. Regarding the role of DVH parameters, some studies did not include all relevant OARs in their analysis, such as the oral cavity for OM [57]. In some models,

social factors such as smoking status and alcohol use were omitted [131]. Additionally, Dean et al. suggested that subjective patient-reported factors such as pain tolerance should also be investigated [131].

Another limitation was in terms of the reporting. Most studies did not display the receiver-operator curve for the validation set, preventing the comparison of sensitivity and specificity across different prediction thresholds. Where performance metrics were re-ported, it was sometimes unclear whether the value applied to a training set or validation set. Moreover, often insufficient information was provided to independently validate the findings. Such information might include definitions of all model features, coefficient values and model hyperparameters.

### *Recommendations for future model development*

Based on the limitations identified in the studies reporting predictive models, some recommendations are warranted. Studies should endeavour to recruit sufficiently large sample sizes to better identify patterns and reduce the impact of overfitting. Models should be externally validated to achieve a higher level of evidence, though the differences between centres should also be identified and discussed. Study methodology should be reported comprehensively, including details on patient selection criteria, variable and outcome definitions, preprocessing, feature selection and the validation approach. Guidelines for OAR delineation should be followed wherever possible to facilitate reproducibility. Likewise, greater standardization in the reporting of results is also desirable. Sufficient information should be provided to reproduce the model for validation purposes.

Certain types of predictors merit further investigation, particularly the role of clinical laboratory tests and genetic factors. Furthermore, exploration of radiomic and dosiomic features may be beneficial through their ability to quantify textural properties and spatial dose distribution within OARs. It should be noted that toxicity has a subjective component. While most of the studies in this review have investigated clinician-rated toxicity, further exploration of patient-reported toxicity outcomes and psychosocial factors as predictors of severe toxicity is warranted.

Future development of predictive models for OM and dysphagia should include prospective studies. These may allow for a more comprehensive range of predictors to be measured and would improve the level of evidence by reducing the risk of selection bias. However, cross-institutional prospective studies would still face issues from differences in toxicity grading and treatment regimen between centres. This presents a bottleneck in the further development of models to predict severe toxicity.

### *Limitations of this review*

A limitation of this review is the broad definition of dysphagia, which includes not just dysfunction in the swallowing mechanism but also impaired oral intake and indication for tube feeding, which in turn is often determined by weight loss. Analysing predictors for each aspect separately might yield more specific findings. However, collecting a range of specific dysphagia outcomes is not typical in clinical practice, so the quantity of results would be reduced.

### *1.2.5. Conclusion*

After reviewing 176 studies on OM and dysphagia, predictors were systematically assessed. Discrepancies observed between the findings from univariate and multivariate analyses suggest the need for deeper investigation into the relationships between different predictors. While several predictive models for the severity of OM or dysphagia have been proposed, the variability in their performance indicates potential for enhancement. This review identified several areas for improvement. Future studies should prioritize larger sample sizes, external validation, standardized predictor and outcome definitions, and comprehensive reporting to facilitate reproducibility. A broad range of predictor types should be collected to capture the multifactorial aetiology of OM and dysphagia. Careful design of prospective studies will mitigate selection bias and allow some of the challenges of obtaining standardized and comprehensive predictor data to be overcome.

Only one prediction model for acute OM was externally validated, indicating the limited level of evidence among existing models and their unknown generalizability. Among the models for acute dysphagia, two models were externally validated. However, both of these were trained on multi-site HNC data and were not specific to NPC patients. While a general HNC toxicity prediction model is useful, NPC has key differences in its treatment which make further exploration in NPC-only cohorts desirable. For late dysphagia, three models were externally validated. Interestingly, discrimination scores were generally higher for late dysphagia than for acute OM or acute dysphagia.

## 1.3. Multi-omics for toxicity prediction in HNC

### *1.3.1. Introduction*

The integration of radiomics, dosiomics and contouromics into the clinical management of HNC promises advancements in precision oncology by offering enhanced toxicity prediction and personalized preventative management. Treatment of HNC is complicated by the presence of various vital structures in proximity to the target, and by the differences in tumour site and histology. Traditional approaches to toxicity prediction can struggle to fully capture this complexity, limiting their use in personalized medicine and underscoring the need for more sophisticated predictive tools. Radiomics provides a non-invasive method for characterizing phenotypic characteristics of tumours or OARs which are associated with higher risk of toxicity or more severe toxicity. Similarly, dosiomics provides a means to quantify the spatial distribution of the planned radiation dose, allowing characterization of the dose delivery beyond the aggregate values of DVH parameters. Contouromics, a newly emerging field, may represent a way to describe the difficulty of dose sparing by quantifying the geometric relationships between tumour and OAR. Together, these tools can harness three-dimensional medical imaging and RT data and machine learning to utilize tumour and treatment characteristics which are invisible to the naked eye.

A growing number of studies have reported multi-omics-based toxicity prediction models for HNC. A systematic review by Carbonara et al. reported eight studies that predicted HNC toxicities including xerostomia, radiation-induced brain injury, parotid shrinkage, trismus and hearing loss using radiomics [224]. A subsequent systematic review by Araújo reported sixteen studies that utilized radiomic features in combination with clinical, dosimetric, or

dosiomic features for toxicity prediction in HNC [225]. Additionally, a systematic review by Tan et al. reported thirteen studies that utilized delta radiomics for toxicity prediction [226]. Limitations of the studies were discussed. In particular, Araújo reported that a majority of the included studies were at high risk of bias according to the TRIPOD checklist and failed to find an overall benefit of imaging biomarkers over conventional approaches in a meta-analysis on a subset of three studies. Furthermore, only three of the models utilized external validation. The authors also noted the need for more comprehensive reporting.

The aim of this section was to conduct a systematic literature search for studies which reported multi-omic models for HNC toxicity prediction, and perform a scoping review in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols Extension for Scoping Reviews (PRISMA-ScR) guidelines [49]. These guidelines were selected in order to ensure a systematic approach with comprehensive reporting.

## *1.3.2. Methods*

### *Search strategy*

Four databases, Web of Science, Embase, Scopus, and PubMed, were searched in November 2023 using the search terms outlined in **Table 11.**

**Table 11: Search strategy**

| Fields | Search String |
|---|---|
| Title, abstract, keywords | (“radiomic*" OR "dosiomic*" OR "textur* analy*" OR "textur* feat*") AND (“head and neck" OR "HNC" OR "nasopharyn*" OR "esophag*" OR "oesophag*" OR "lip" OR "lips" OR "tongue*" OR "pharyn*" OR "hypopharyn*" OR "laryn*" OR "salivar*" OR "nasal" OR "sinus*") |
| Title | "toxi*" OR "morbi*" OR "side effect*" OR "mucositis" OR "dysphagia" OR "xerostomia" OR "saliva" OR "dysgeusia" OR "necro" OR "hearing" OR "caries" OR "weight loss" OR "thyroid" OR "feeding tube" OR "tube feeding" OR "Ryle" OR "stricture" OR "aspiration" OR "osteo*" OR "hypothyroid*" OR "trismus" OR "fibrosis" OR "stenosis" OR "edema" OR "oedema" |
| Document type | Article / full text |

## *Inclusion and exclusion criteria*

The inclusion criteria comprised full-text English-language articles reporting radiomics, dosiomics or contouromics-based toxicity prediction models for head and neck cancer. Abstracts and reviews were excluded.

## *Selection of studies based on PRISMA*

**Figure 5** shows the PRISMA flow diagram for the selection of sources of evidence. Sixty-two studies were identified after removing duplicates. After screening, twenty-seven studies were initially included for analysis.

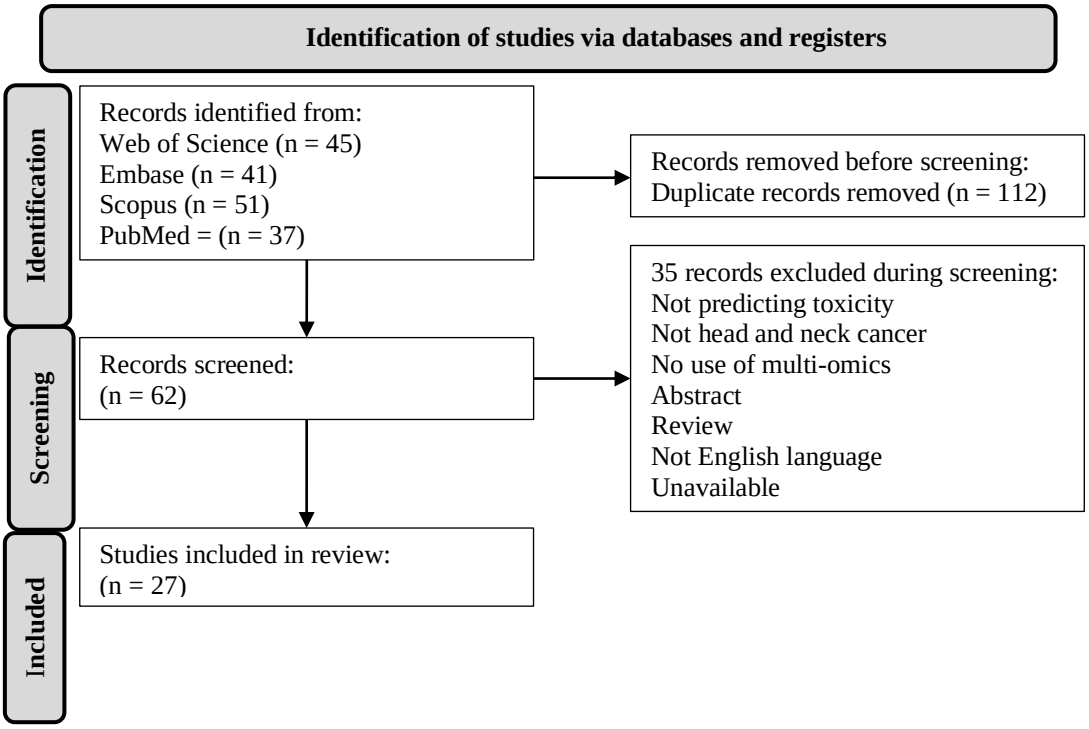



**Figure 5: PRISMA flow diagram of selection of sources of evidence.**

## *Data charting and analysis*

Data charting was conducted by tabulating key details from each study. This included the toxicity outcome, timeframe, feature types investigated, image modality, sample size, use of external validation, feature extraction settings, feature selection method, model type, model size, outcome incidence, and details of any feature stability assessment. Additionally, the included studies were also assessed using items from the CheckList for EvaluAtion of Radiomics research (CLEAR) [227]. This checklist identifies many important aspects of radiomics research which will facilitate reproducibility if reported. Items 1-7, referring to the title, abstract, keywords and introduction, were excluded in order to focus on the methodology. Items 49-58, referring to the discussion and data availability, were also excluded for the same reason. Analysis of these aspects was performed in order to identify strengths, weaknesses, and

standard practice of studies in this research area, which in turn would provide recommendations for the rest of this project.

### *1.3.3. Results*

The twenty-seven included studies reported prediction models for a broad range of radiation-induced toxicities, particularly focusing on xerostomia. The number of studies predicting each toxicity is listed in **Table 12**. Fourteen studies reported prediction models for xerostomia and three reported prediction models for hypothyroidism. The remaining studies each reported prediction models for different toxicities: saliva amount reduction, parotid shrinkage, hearing loss, trismus, dysgeusia, weight loss, fibrosis, stenosis, osteoradionecrosis, and OM. Many toxicities were the focus of only a single study, indicating potential for further exploration.

**Table 12: Toxicities predicted by included studies**

| Toxicity | Number of studies |
|---|---|
| Xerostomia | 14 |
| Hypothyroidism | 3 |
| Saliva amount reduction | 1 |
| Parotid shrinkage | 1 |
| Hearing loss | 1 |
| Trismus | 1 |
| Dysgeusia | 1 |
| Weight loss | 1 |
| Fibrosis | 1 |
| Stenosis | 1 |
| Osteoradionecrosis | 1 |
| OM | 1 |

**Table 13** shows that the included studies focused on radiomic features, often in combination with clinical and DVH features. Studies extracted radiomic features from CT images (70%), MRI (22%) and PET (7%). Dosiomics, a more recently developed field, was explored by 19% of the studies.

**Table 13: Feature types analysed by included studies**

| Feature type | Number of studies |
|---|---|
| Clinical | 19 (70%) |
| Radiomics | 25 (93%) |
| DVH | 19 (70%) |
| Dosiomics | 5 (19%) |

**Table 14** describes the characteristics of the included studies. The median sample size was small, at 93 patients, and only 4 studies were externally validated (15%). The majority of studies focused on late toxicity (70%), with 10 studies focusing on acute toxicity (37%).

**Table 14: Characteristics of included studies**

| Characteristic | |
|---|---|
| Sample size | 20-337 (median = 93) |
| External validation | 4 (15%) |
| Acute toxicity | 10 (37%) |
| Late toxicity | 19 (70%) |

Details of the included studies are provided in **Table 15**. Most of the models used pre-treatment features (56%), however some models used imaging data taken during treatment to predict the subsequent evolution of toxicity, for example by using the change in feature values over time. The feature selection method varied extensively across studies. Some used logistic regression with Least Absolute Shrinkage and Selection Operator (LASSO), some used model-based wrapper methods such as forward selection or recursive feature elimination, some used filter methods such as Maximum Relevance Minimum Redundancy (MRMR) or filtering by statistical tests for relevance and pairwise correlation between features, while others used tree-based models such as Random Forest or XGBoost to perform feature selection. The type of model reported also varied significantly. Logistic regression was the most frequently reported, but other models included Random Forest, XGBoost, Likelihood Fuzzy Analysis, SVM, K-nearest neighbours, extra-trees, and Gaussian Naïve Bayes. The model complexity, as

indicated by the model size, varied significantly across models, ranging from single-feature models to a model consisting of 30 features.

The number of studies reporting each aspect of selected items from the CLEAR guidelines is displayed in **Table 16**. Only 19% of studies reported adherence to guidelines or checklists, and among those that did, such guidelines were not necessarily specific to radiomics-related studies. Instead, guidelines such as TRIPOD were referenced, which are general to diagnostic or prognostic studies [228]. While ethics or institutional approval was generally well-reported alongside study nature as retrospective or prospective, the sample size was rarely justified with a calculation based on statistical power. It can be assumed that most sample sizes were chosen based on convenience or availability. Eligibility criteria was reported by all studies, and most identified the origin of the data, for example by providing the name of the institution. Data overlap refers to use of data from a previous publication, which was specified by four studies. Most studies reported the data split methodology, with the exception of exploratory studies that did not perform validation. Reporting of imaging protocol and definition of predictors and outcomes was also included by most studies. While the segmentation strategy was usually outlined, 26% of studies did not identify who performed the delineation. Image pre-processing and feature extraction settings were not well reported. Key settings such as discretization, normalization and resampling were omitted from the main text and also from the supplementary data. Image filters were also rarely reported. The feature extraction method refers to reporting the software used for feature extraction. Studies did not always specify the exact software version, and insufficient information was usually provided for reproduction of features if in-house or custom software was used.

**Table 15: Details of included studies**

| | Pre-treatment features | Toxicity | Acute vs Late | Clinical features | Radiomic features | DVH features | Dosiomic features | Image modality | Sample size | External validation | Model size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Nardone, 2018 [229] | Y | Xerostomia | Late | Y | Y | Y | N | CT | 78 | No | 4 |
| Zhou, 2022 [230] | N | Saliva reduction | Acute | Y | Y | Y | N | CT | 52 | No | 17 |
| Pota, 2017 [231] | N | Parotid shrinkage | unclear | Y | Y | N | N | CT | 37 | No | 9 |
| Abdollahi, 2018 [232] | Y | Hearing loss | Either | N | Y | N | N | CT | 47 | No | 10 |
| Ren, 2021 [233] | Y | Hypothyroidism | Late | Y | N | Y | Y | Dose | 145 | No | 3 |
| Wu, 2018 [234] | N | Xerostomia | Acute | N | Y | Y | N | CT | 59 | Yes | 2 |
| Thor, 2017 [235] | N | Trismus | Either | N | Y | Y | N | MRI | 20 | No | - |
| Berger, 2022 [236] | N | Xerostomia | Late | N | Y | Y | N | Daily MVCT | 337 | No | 3 |
| Qin, 2023 [237] | N | Xerostomia | Late | Y | Y | N | N | MRI | 123 | No | 20 |
| van Dijk, 2018 [238] | Y | Xerostomia | Late | Y | Y | Y | N | PET | 161 | No | 3 |
| van Dijk, 2018 [239] | Y | Xerostomia | Late | Y | Y | Y | N | MRI | 93 | Yes | 4 |
| Li, 2023 [240] | Y | Xerostomia | Late | Y | Y | Y | N | PET | 137 | Yes | 3 |
| Busato, 2023 [241] | Y | Dysgeusia | Late | N | N | Y | Y | Dose | 80 | No | 5 |
| van Dijk, 2019 [242] | N | Xerostomia | Late | Y | Y | Y | N | CT | 68 | No | 3 |
| Ritlumlert, 2023 [243] | Y | Hypothyroidism | Late | Y | Y | Y | N | CT | 220 | No | 30 |
| Berger, 2023 [244] | N | Xerostomia | Late | Y | Y | Y | N | Daily MVCT | 117 | No | 1 |
| Abdollahi, 2023 [245] | N | Xerostomia | Late | Y | Y | Y | Y | CT | 31 | No | varies |
| Sheikh, 2019 [246] | Y | Xerostomia | Acute | Y | Y | Y | N | CT, MRI | 266 | No | ≤ 17 |
| Cheng, 2019 [247] | Y | Weight loss | Acute | Y | Y | Y | N | CT | 163 | No | 18 |
| Liu, 2019 [248] | N | Xerostomia | Acute | Y | Y | N | N | CT | 35 | No | 14 |
| Wang, 2020 [249] | Y | Fibrosis | Late | N | Y | N | N | CT, MRI | 186 | No | NA |
| Gabrys, 2018 [40] | Y | Xerostomia | Either | Y | Y | Y | Y | CT | 153 | No | ≤ 5 |
| Liu, 2022 [250] | Y | Stenosis | Late | Y | Y | N | N | CT | 65 | No | 7 |
| Barua, 2021 [251] | N | Osteoradionecrosis | Late | N | Y | N | N | CT | 21 | No | 6 |
| Calamandrei, 2023 [252] | N | Xerostomia | Acute | N | Y | N | N | MRI | 27 | No | NA |
| Smyczynska, 2021 [253] | Y | Hypothyroidism | Late | Y | Y | Y | N | CT | 98 | Yes | ≤ 6 |
| Dong, 2023 [57] | Y | OM | Acute | Y | Y | Y | Y | CT, MRI | 242 | No | ≤ 19 |

**Table 16: Reporting of CLEAR items across included studies.**

| CLEAR item number | Description | Number of studies | Percent |
|---|---|---|---|
| 7 | Adherence to guidelines or checklists (e.g. CLEAR checklist) | 5 | 19% |
| 8 | Ethical details (e.g., approval, consent, data protection) | 22 | 81% |
| 9 | Sample size calculation | 2 | 7% |
| 10 | Study nature (e.g., retrospective, prospective) | 23 | 85% |
| 11 | Eligibility criteria | 27 | 100% |
| 12 | Flowchart for technical pipeline | 12 | 44% |
| 13 | Data source (e.g., private, public) | 22 | 81% |
| 14 | Data overlap | 4 | 15% |
| 15 | Data split methodology | 24 | 89% |
| 16 | Imaging protocol (i.e., image acquisition and processing) | 27 | 100% |
| 17 | Definition of non-radiomic predictor variables | 25 | 93% |
| 18 | Definition of the reference standard (i.e., outcome variable) | 26 | 96% |
| 19 | Segmentation strategy | 24 | 89% |
| 20 | Details of operators performing segmentation | 20 | 74% |
| 21 | Image pre-processing details | 6 | 22% |
| 22 | Resampling method and its parameters | 10 | 37% |
| 23 | Discretization method and its parameters | 5 | 19% |
| 24 | Image types (e.g., original, filtered, transformed) | 9 | 33% |
| 25 | Feature extraction method | 20 | 74% |
| 26 | Feature classes | 27 | 100% |
| 27 | Number of features | 22 | 81% |
| 28 | Default configuration statement for remaining parameters | 2 | 7% |
| 29 | Handling of missing data | 4 | 15% |
| 30 | Details of class imbalance | 23 | 85% |
| 31 | Details of segmentation reliability analysis | 3 | 11% |
| 32 | Feature scaling details (e.g., normalization, standardization) | 8 | 30% |
| 33 | Dimension reduction details | 23 | 85% |
| 34 | Algorithm details | 25 | 93% |
| 35 | Training and tuning details | 25 | 93% |
| 36 | Handling of confounders | 2 | 7% |
| 37 | Model selection strategy (defined as choosing between model types) | 12 | 44% |
| 38 | Testing technique (e.g., internal, external) | 24 | 89% |
| 39 | Performance metrics and rationale for choosing | 5 | 19% |
| 40 | Uncertainty evaluation and measures (e.g., confidence intervals) | 16 | 59% |
| 41 | Statistical performance comparison (e.g., DeLong's test) | 5 | 19% |
| 42 | Comparison with non-radiomic and combined methods | 15 | 56% |
| 43 | Interpretability and explainability methods | 3 | 11% |
| 44 | Baseline demographic and clinical characteristics | 26 | 96% |
| 45 | Flowchart for eligibility criteria | 6 | 22% |
| 46 | Feature statistics (e.g., reproducibility, feature selection) | 3 | 11% |
| 47 | Model performance evaluation | 25 | 93% |
| 48 | Comparison with non-radiomic and combined approaches | 16 | 59% |

### *1.3.4. Discussion*

Over half of the included studies focused on xerostomia; other toxicities were less well explored. Most studies focused on late toxicity, though the impact of acute toxicity should not be ignored. The focus on CT radiomics may have been because the planning CT is typically used for tumour and organ-at-risk (OAR) segmentation and so would provide the most accurate VOIs. Other imaging modalities require careful registration to align with the planning CT, and even then, differences in geometry remain due to differences in patient position and from morphological changes during the time between acquisitions. The role of dosimetric data such as dose-volume-histogram parameters in toxicity prediction is well established. However, relatively few studies explored the role of dosiomics, which can further characterise the dose information with more sophisticated intensity features and second-order texture features. The potential role of contouromic features, which may be able to characterize the difficulty of dose sparing, remains unexplored.

The level of evidence across the studies was generally low, reflecting their exploratory nature. The median sample size was small, at less than 100 cases, and over 80% of the studies were not externally validated. The results demonstrate a lack of standardization in the methodology and reporting of multi-omics models. This is also reflected by the 81% of studies which did not report adherence to any guidelines or checklists for reporting. While there are differences in the items included by different checklists, and authors may not agree with the necessity to include all items, future studies should refer to reporting guidelines wherever possible in order to avoid omission of information and improve the transparency and reproducibility of research.

Sample size calculation was often omitted. Justification of the sample size is less important for a retrospective study where the risk to the recruited patients is low, though such a calculation can ensure that the sample size is large enough to achieve the desired power. However, there are often practical limitations which ultimately determine the sample size. Clear reporting of the inclusion and exclusion criteria, along with the method of identifying patients to include, should be provided even if a sample size calculation was not performed.

In terms of the analysis of feature data, studies rarely reported how missing data was handled, even though this can introduce bias and merits discussion. Analysis of segmentation reliability was also rarely performed, likely due to the resource cost of delineating multiple sets of contours. However, simulating perturbations to contours can be used to achieve the same effect [254]. Normalization or scaling of feature data was only reported by a minority of studies, even though this is an important step which can bias the results if performed incorrectly. All preprocessing and feature extraction settings must be reported to ensure reproducible research. Additionally, the software used for feature extraction should be compliant with the IBSI to ensure reproducibility [42]. Most studies reported their methodology for dimension reduction and model optimization, but handling of confounders was rarely addressed. Studies stated the metrics used for performance evaluation, but rarely justified their choice of metric. Given that the use of AUC for classification is so widespread, this may have been seen as unnecessary. Most, but not all studies reported confidence intervals for their performance metrics. Often these were obtained from cross validation or bootstrapping. However, only rarely was a statistical test used to compare between the performance of different models. Many studies did include a comparison with conventional

(clinical or DVH-based) approaches. Such baseline clinical characteristics were generally well reported.

Omission of image pre-processing and feature extraction settings makes accurate reproduction of the models very difficult. It may be assumed that only original features were used if no image filters were mentioned, however it would be better to clarify this explicitly. Given the lack of reporting of feature extraction settings, it is difficult to identify any trends or standards in terms of gray-level discretization, normalization, or resampling. Since most studies investigated CT radiomics, normalization was likely not applied to the images, because the voxel values have physical meaning as Hounsfield Units.

Another point of difference between studies is from variations in the outcome definition. This must be clearly stated, and it should be noted that the interpretation of the findings of a study should consider the incidence and clinical relevance of the outcome definition. For example, prediction of moderate-or-higher toxicity may be of less clinical relevance, since most patients will experience such toxicity, and any intervention would need to be applied to most patients.

When analysing the results, statistical performance comparison was rarely performed. Instead, comparison tended to be based on direct comparison of the performance metric without considering the variability or confidence interval. Methods for interpreting and explaining the resulting models were limited, with the discussion mostly focusing on the discrimination performance. Comparisons with non-radiomic and combined approaches was not always conducted. Studies should compare their model with conventional approaches and any related literature, to better clarify the novelty and advantages of their approach.

A summary of the strengths and weaknesses of existing prediction models for acute OM and dysphagia are provided in Table 17.

**Table 17: Summary of strengths and weaknesses of existing prediction models for acute OM and dysphagia**

| Strengths | Weaknesses |
|---|---|
| • Utilize readily available features: clinical and DVH<br>• Several models utilize multi-center training data – more generalizable<br>• Broad range of clinical and treatment features included<br>• Some use custom VOIs to attempt to better capture the region relevant to the toxicity | • Often lacking external validation<br>• Rarely include multiple organs at risk in model<br>• Variations in contouring between institutions<br>• Rarely adhere to a reporting guideline e.g., TRIPOD<br>• Limited interpretation of model features<br>• Limited comparison with conventional approaches |

## *1.3.5. Conclusion*

In conclusion, the use of radiomics and dosiomics for toxicity prediction in HNC is a developing field. Prediction of toxicities other than xerostomia remain relatively unexplored, and research tends to focus on late toxicity. Studies focused particularly on CT radiomics, and the role of dosiomics remains to be further explored. The level of evidence of these studies is generally low, with small sample sizes and a lack of external validation. There is significant variation in the methodology for model development. Comprehensiveness of reporting is mixed, with studies rarely following reporting guidelines. This consequently limits the reproducibility of the results since insufficient information is available to extract equivalent features and build equivalent models. There is also a need to include stability assessment of features to ensure more reproducible models. Study findings should be compared with related literature and conventional approaches, and statistical performance comparison should be utilized rather than comparing single valued measures of the discrimination performance.

Inclusion of the items in the CLEAR guidelines should greatly improve the standardization of reporting and result in more reproducible and transparent research.

# CHAPTER 2 RESEARCH AIMS & OBJECTIVES

## 2.1. Research aim

*'With improvements in the survival rates of NPC patients, it is increasingly important to develop holistic clinical decision-making strategies that address quality of life for patients suffering from treatment-induced toxicities. This project harnesses high-dimensional multi-omic data to identify patients at risk of severe acute OM and dysphagia, two of the most common and debilitating toxicities, thereby facilitating the targeting of preventative interventions and support.'*

## 2.2. Research gap

Literature reviews in **Section 1.2** and **1.3** provided a comprehensive overview of the published prediction models for OM and dysphagia in HNC patients. Among these studies, very few investigated radiomics, dosiomics or contouromics in this context. Specifically, no full text articles reported externally validated prediction models for OM and dysphagia using radiomic, dosiomic or contouromic features. No full-text articles investigating multi-omics for dysphagia were found. Furthermore, there was a general lack of external validation across all prediction models for OM and dysphagia, with studies mostly based on single-centre cohorts. Many studies investigated mixed cohorts consisting of multiple HNC subsites. The development of NPC-specific prediction models represents another under-explored area of research. Given the distinct challenges and treatment guidelines associated with NPC, and its significant prevalence in Hong Kong where it is one of the most prevalent cancers in men, the development of tailored prediction models for NPC is particularly important. Furthermore,

treatment for NPC has a particularly high burden on patients, with a relatively high radiation dose, close proximity of critical structures in the tumour region, and the simultaneous effects of concurrent chemotherapy.

## 2.3. Research objectives

### *2.3.1. Objective 1*

***To develop and externally validate a multi-omic prediction model for severe acute OM in NPC patients undergoing RT, analysing clinical, DVH, radiomic, dosiomic and contouromic features.***

Severe OM has a major impact on patients' quality of life and poses the risk of treatment interruption and unplanned hospitalization. Accurate prediction of severe OM is important for the delivery of personalized prevention and management strategies. As identified in the literature review (**Sections 1.2** and **1.3**), prediction of OM using radiomic, dosiomic or contouromic features is underexplored. Two studies reported models utilizing radiomic features, however neither were externally validated. Furthermore, conventional clinical and DVH-based models for OM were also lacking external validation. Multi-centre data from two hospitals was therefore used to develop and externally validate a multi-omic prediction model for severe acute OM in NPC patients undergoing RT. This was motivated by the need for generalizable models with a higher level of evidence. To the best of our knowledge, this represented the first externally validated model for treatment-induced OM using multi-omic features.

### *2.3.2. Objective 2*

***To develop and externally validate a multi-omic prediction model for severe acute dysphagia in NPC patients undergoing RT, analysing clinical, DVH, radiomic, dosiomic and contouromic features.***

Severe acute dysphagia harms patients' quality of life and poses the risk of weight loss which can result in deviations from the RT plan, threatening worsened treatment outcome and additional toxicity. Accurate prediction of severe acute dysphagia would allow for earlier intervention to mitigate these risks. The literature review in **Sections 1.2** and **1.3** highlighted the absence of radiomic, dosiomic or contouromic models for this toxicity. Multi-centre data from two hospitals was therefore used to develop and externally validate a multi-omic prediction model for severe acute dysphagia in NPC patients undergoing RT. This was motivated by the need for generalizable models with a higher level of evidence. To the best of our knowledge, this represented the first full-length article reporting a model for treatment-induced severe acute dysphagia using multi-omic features.

### *2.3.3. Objective 3*

***To develop and externally validate a multi-omic, multi-label combined model to predict both severe acute OM and dysphagia, utilizing the interaction between the two toxicities to further improve performance.***

OM and dysphagia are related conditions which both impact oral intake. Severe OM is correlated with severe acute dysphagia and is thought to contribute to difficulty swallowing through the pain that it causes. A multi-label model may be able to improve on the accuracy of predictions of each toxicity by incorporating information about the relationship or interaction between the two toxicities. Different approaches for multi-label modelling were evaluated on

the combined data for OM and dysphagia. To the best of our knowledge, no such multi-label model for acute OM and dysphagia had been published.

# CHAPTER 3 CORE METHODOLOGY IN MULTI-OMIC STUDIES

Predictive models for OM and dysphagia are reported in subsequent chapters. This chapter reports the methodology common to those chapters, including the steps listed in the flowchart in **Figure 6.** The technical workflow is illustrated visually in **Figure 7**.

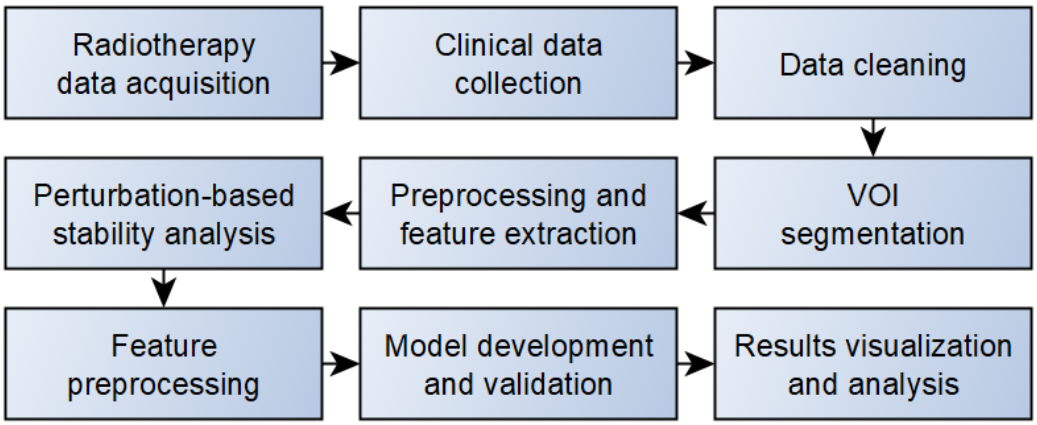



**Figure 6: Core methodology flowchart**

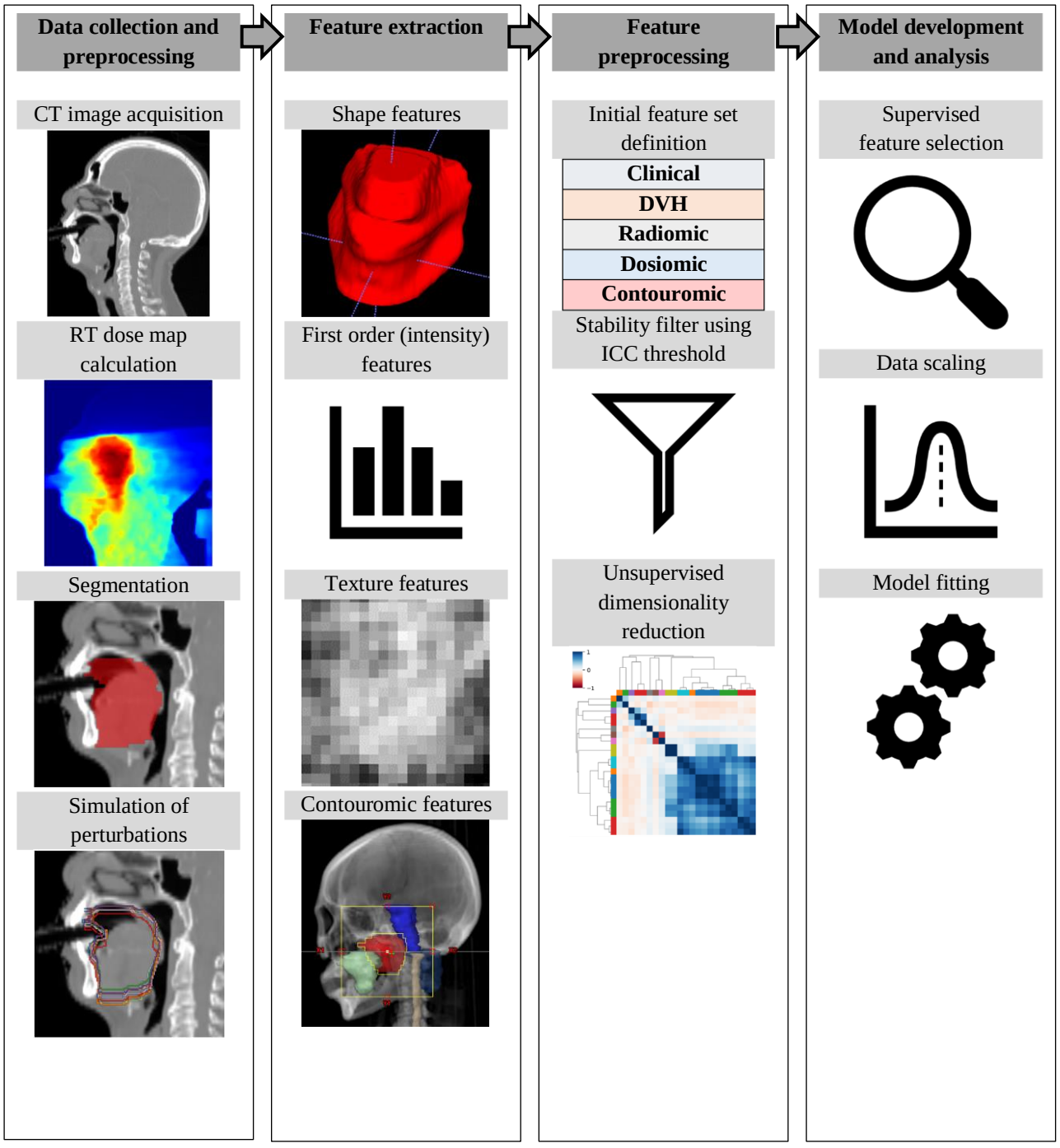



**Figure 7: Technical workflow**

## 3.1. Study population

Patients with biopsy-proven primary NPC treated with RT were retrospectively collected from two public hospitals in Hong Kong. Data from 397 patients who received RT at Hong Kong Queen Elizabeth Hospital (QEH) between 2008 and 2018 had previously been collected for the investigation of adaptive RT eligibility prediction [41]. In order to perform external validation, data from 109 patients who received RT at Hong Kong Prince of Wales Hospital (PWH) between 2020 and 2021 were collected as part of this project. Patients were recruited consecutively by scheduled start date of RT. The sample size for the external validation set was determined using MedCalc v22.018, to detect an AUC of 0.7 versus a null hypothesis value of 0.5 with 80% power and 0.05 significance level, assuming an incidence of severe OM of 40% as observed from the literature search [1, 255]. The expected incidence of severe acute dysphagia was similar or higher than that for severe OM.

The patients from both datasets were screened for study eligibility. Exclusion criteria consisted of distant metastasis at diagnosis, missing planning CT image, missing RT dose map, and missing primary tumour contour. The patient recruitment diagram is shown in **Figure 8.** Institutional review board ethics approval was obtained from each institution, and patient informed consent was waived due to the retrospective nature of the study.

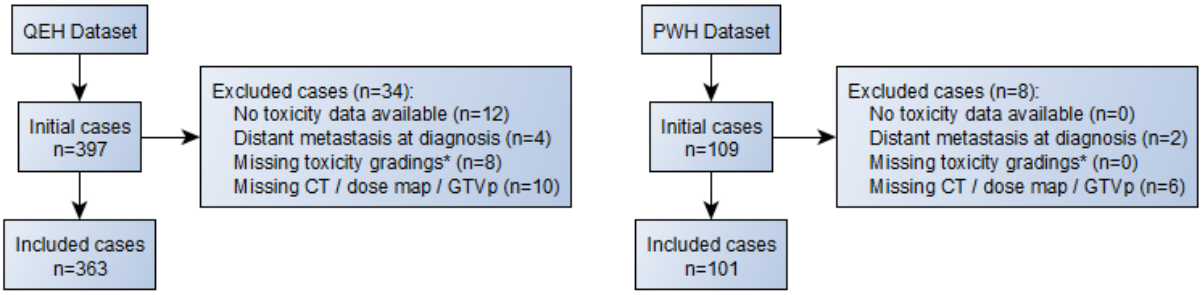



**Figure 8: Eligibility flowchart. See Section 3.8.1 for missing data handling**

# 3.2. Radiotherapy data acquisition

*Imaging acquisition parameters*

For both datasets, planning contrast-enhanced CT (CECT) images were acquired with 16-slice Brilliance Big Bore CT scanners (Philips Medical Systems, Cleveland, OH). Acquisition parameters were as follows: scan mode = helical, voltage = 120 kVp, pixel spacing = 1.2x1.2mm, slice thickness = 3mm, matrix = 512x512px. The X-ray tube current was typically 264mA for QEH patients, and typically 165 mA for PWH patients. Information on the reconstruction kernel was not collected. Patients were scanned in a supine position, wearing a thermoplastic immobilization mask. Intravenous contrast agents were injected 30 seconds prior to scanning.

*Radiation dose distribution*

All patients received IMRT. In the QEH dataset, 92% of patients received helical tomotherapy, while in the PWH dataset, all patients were treated with volumetric modulated arc therapy (VMAT) on a linear accelerator. The standard treatment protocol was for 70Gy to GTVp and GTVn delivered in 33 fractions of 2.12 Gy. After treatment planning on the

respective software, the planned radiation dose distribution was calculated and saved as a DICOM file.

Helical tomotherapy is a form of IMRT in which radiation is delivered continuously in a fan-shaped beam moving in a helical pattern around the patient. The shape of the beam is adjusted by a binary (open or closed) multi-leaf collimator (MLC). VMAT, another form of IMRT, also involves continuous delivery of radiation as the gantry rotates, however the multi-leaf collimator allows for intermediate positions and more complex beam shaping. VMAT can be administered either in step-and-shoot mode, where the beam is only delivered after the MLC has been adjusted to a desired, stationary position, or in sliding window, where the radiation beam is maintained while the MLC leaves move across the aperture at varied rates. Additionally, unlike helical tomotherapy, VMAT involves multiple arcs around the patient, rather than following a helical path.

It would be desirable to compare the dose features extracted from patients treated with VMAT and tomotherapy to examine the effect of RT modality. However, because almost all (92%) QEH patients were treated with tomotherapy, and all PWH patients were treated with VMAT, such a comparison would be confounded by other differences between the institutions. Comparison of the performance of models trained on patients treated with each modality would also be confounded by inter-institutional differences, and the number of samples for model development would have to be limited such that both datasets had the same sample size, for a fair comparison. This would likely reduce model performance.

*Contours from radiotherapy planning*

All of the contours created during the RT planning process were extracted in the form of a DICOM RT structure set. There were typically a large number of contours per patient, since these included some pseudo structure such as margin expansion for planning purposes. Usually these included the GTVp, GTVn and several OARs, but naming conventions varied and data cleaning was required.

*Collection of DICOM files*

DICOM files containing the CT images, radiation dose distributions and VOI contours were provided on compact discs (CDs) by hospital staff. All data was anonymized to remove identifying information.

## 3.3. Clinical data collection

Clinical data for the QEH dataset had previously been manually extracted from the patient records and digitised into a spreadsheet by other members of the research group. The data had previously been included in a publication on prediction of adaptive RT in NPC patients [41]. The data included age, gender, height, weight at CT simulation, TNM staging according to 8th Edition of UICC/AJCC [5, 256], chemotherapy regimen, and details of RT delivery. Additionally, the raw text from the mid-treatment consultation notes had been extracted and stored in the spreadsheet, organized into the 7 weeks of RT per patient. This raw text was analysed by the author to extract the toxicity outcome labels for QEH. Specifically, the text was manually searched for OM gradings and for information about tube feeding.

The PWH dataset was collected for the purposes of this project. Informed by the existing QEH dataset, clinical data was recorded onto a spreadsheet during visits to the hospital. Data was organized into two tabs: one including the per-patient data such as age at RT start, gender, weight at CT simulation, and another including separate records for each consultation note. Data from collection notes were collected from the start of RT to the end of the acute time period (90 days after the start of RT) for each patient. Data collection involved careful inspection of the entire RT patient folder. Consultation notes, which had been typed by the clinician, included data such as weight, blood test results, toxicity grades, tube feeding information and also key events such as changes in treatment regimen and results of biopsies. Relevant information about each factor was recorded within separate columns in the consultation notes tab in the spreadsheet.

Details of the toxicity outcome definitions are provided in the respective chapters on OM and dysphagia.

## 3.4. Data cleaning

Planning contrast-enhanced CT images, radiation dose distributions and contours were collected in Digital Imaging and Communications in Medicine (DICOM) format. Data cleaning was required to identify the corresponding sets of DICOM files for each patient and identify the relevant contours. GTV and organ contours were identified by their name field, however the naming convention varied significantly across patients and multiple contours with similar names were present for several patients. Therefore, manual checking of the contours was

required along with VOI name standardization. Having performed this step, the number of patients with each contour could be tabulated.

The cleaned CT, dose and contour files were extracted into .mha files, a format based on the open-source Insight Toolkit (ITK) which allowed for more flexible analysis and visualization. Available contours included the GTVp, GTVn, parotid glands, larynx, and oesophagus. However, there were significant differences in the contouring guidelines for OARs across the two centres, especially for the larynx and oesophagus, with different anatomy included in each centre.

## 3.5. VOI segmentation

A deep learning auto-segmentation model was used to generate segmentations for the extended oral cavity and the pharyngeal constrictor muscles from the contrast-enhanced planning CT image. The open-source model, nnU-Net, had demonstrated good performance on several organ segmentation tasks and so was selected for this purpose [257]. The source code enabled a segmentation pipeline to be automatically configured for a dataset, provided that the data was structured in a specified manner. The software identified the optimal model architecture and parameters based on five-fold cross validation and returned a trained model for inference.

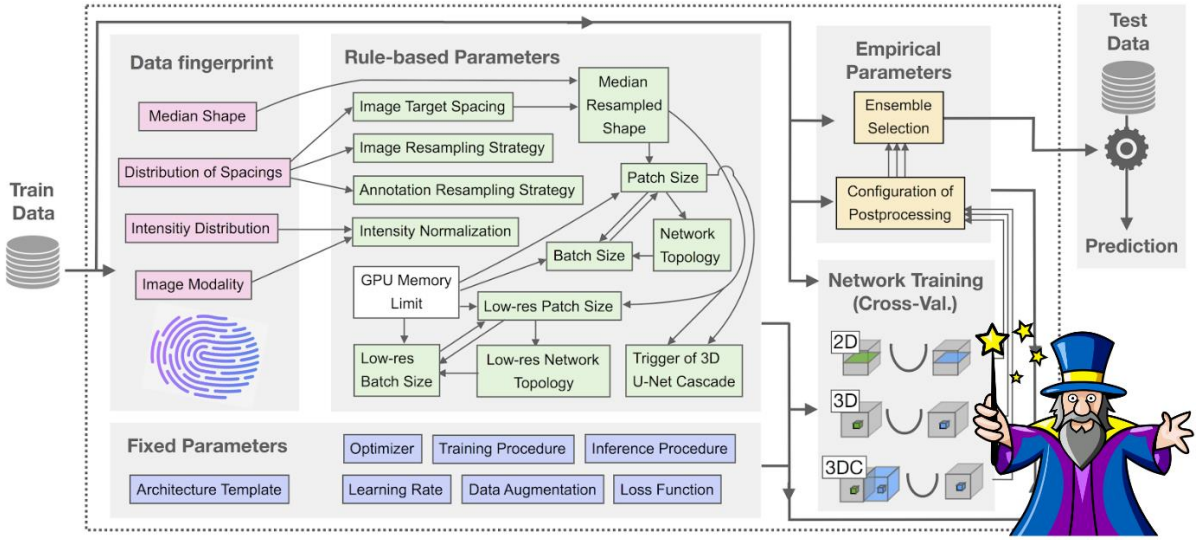



**Figure 9: nnU-Net workflow [257]**

### *3.5.1. Extended oral cavity*

The oral cavity contour was only clinically available for a few patients in the QEH and PWH cohorts. Furthermore, the clinical segmentation of the oral cavity varied within and across cohorts. For some cases, it only covered a few axial slices or only covered the oral immobilization device held in the patient's mouth. Based on the literature search detailed in **Section 1.2**, the oral cavity and tongue are relevant to OM prediction, and these can be contoured according to different guidelines. The extended oral cavity, as defined by Brouwer et al., was selected as a relevant VOI, since it included many of the areas where OM can present [258].

Initially, an approximate VOI for the extended oral cavity was obtained by performing binary morphological operations on the contour of the mandible, which was present for almost all cases. A convex hull operation was used to find the minimum convex volume that covered the whole mandible. This volume would contain large sections of the oral cavity and tongue, as in the extended oral cavity definition. Initial exploratory analysis utilized this approximate

contour. However, to obtain a more accurate set of contours, the nnU-Net AI segmentation model was employed [257]. An additional set of imaging and contour data was available for NPC patients from Hong Kong Queen Mary Hospital (QMH). This dataset included a larger quantity of oral cavity contours, with some variation in the definitions used. A subset of forty-seven contours which covered the oral cavity and tongue, extending towards the back of the pharynx but not including the pharynx, were selected for training the nnU-Net model. These contours appeared consistent with the definition of the extended oral cavity by Brouwer et al. [258]. The nnU-Net software evaluated a 2D model architecture, a 3D model architecture and a 3D cascade architecture, comparing the cross-validation performance. Ensembles of these models are also evaluated to determine the best configuration (see Figure 9). The optimal model configuration was an ensemble of the 2D and the 3D full resolution models. The mean performance metrics across the five-fold cross validation were as follows: 0.843 (Dice), 0.732 (IoU). An example of this VOI is shown in **Figure 10**.

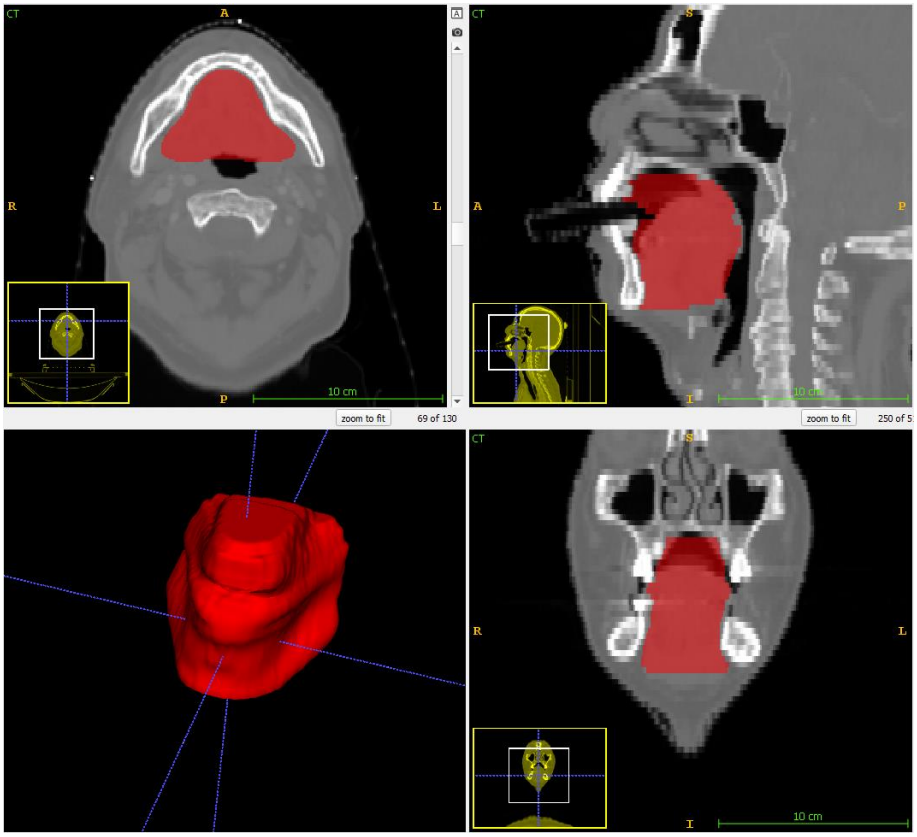



**Figure 10: Example of extended oral cavity segmentation in axial (upper left), sagittal (upper right), and coronal (bottom right) planes, with 3D projection (bottom left).**

### *3.5.2. Pharyngeal constrictor muscles*

As identified in the literature search, the radiation dose to the pharyngeal constrictor muscles has been identified as a predictor of dysphagia. However, many of the included cases lacked contours of the pharyngeal constrictor muscles. The nnU-Net model was also used to generate contours for those cases. The model was trained on the thirty-nine cases from QEH which had separate contours of the superior, middle, and inferior pharyngeal constrictors. The separate VOIs produced by the model were later combined into a single pharyngeal constrictor muscle VOI for subsequent analysis. This was done because the individual muscles constituted VOIs with small volumes, which would increase their susceptibility to perturbations and any

inaccuracies in the automatic segmentation. This hypothesis was confirmed by the observation that the stability of the independent muscles in the perturbations analysis was lower than that for the combined VOI. Furthermore, exploration of model development using the separate VOIs did not suggest that using any of the individual muscles would give better results than using the combined VOI.

The optimal model configuration was a 3D full resolution model. The mean performance metrics across the five-fold cross validation were 0.742 (Dice) / 0.595 (IoU), 0.773 (Dice) / 0.636 (IoU), 0.766 (Dice) / 0.625 (IoU) for the superior, middle, and inferior constrictor muscles respectively. An example of this VOI is shown in **Figure 11.**

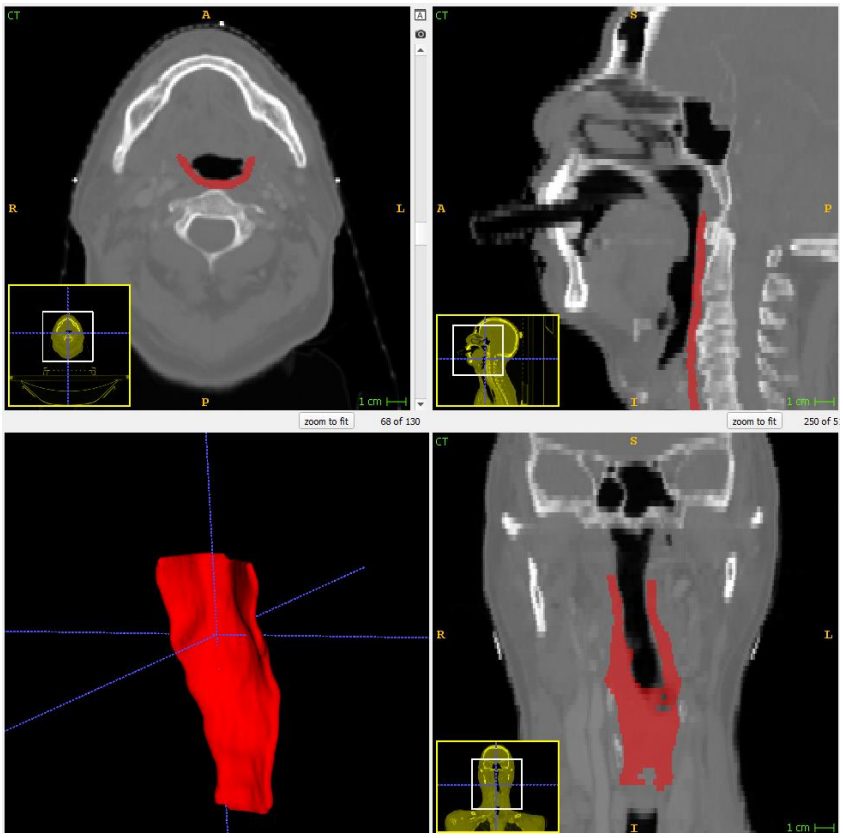



**Figure 11: Example of pharyngeal constrictor (PC) segmentation in axial (upper left), sagittal (upper right), and coronal (bottom right) planes, with 3D projection (bottom left).**

### *3.5.3. Parotid glands and larynx*

Statistically significant differences in the volume of the parotids contour and the larynx contour were observed between the QEH and PWH datasets. The contouring guidelines for the larynx were noticeably different for each centre. These differences would severely affect the generalizability of models and made inclusion of these OARs inadvisable. However, given that the radiation doses to these OARs were reported as predictors of OM and dysphagia, the AI segmentation model was utilized to standardize the contours between datasets. The contours from the PWH dataset were selected as training data. These contours, being performed more recently, would conform more closely to the latest contouring guidelines. Additionally, the

definition of the larynx was more consistent across cases. The model was trained on 111 cases from the PWH dataset with contours of the larynx and parotid glands available. The model configuration was a 3D full resolution model. The mean performance metrics across the five-fold cross validation were 0.874 (Dice) / 0.778 (IoU) and 0.896 (Dice) / 0.818 (IoU) for the parotid glands and larynx respectively. Examples of the parotid glands and larynx VOIs are shown in **Figure 12**.

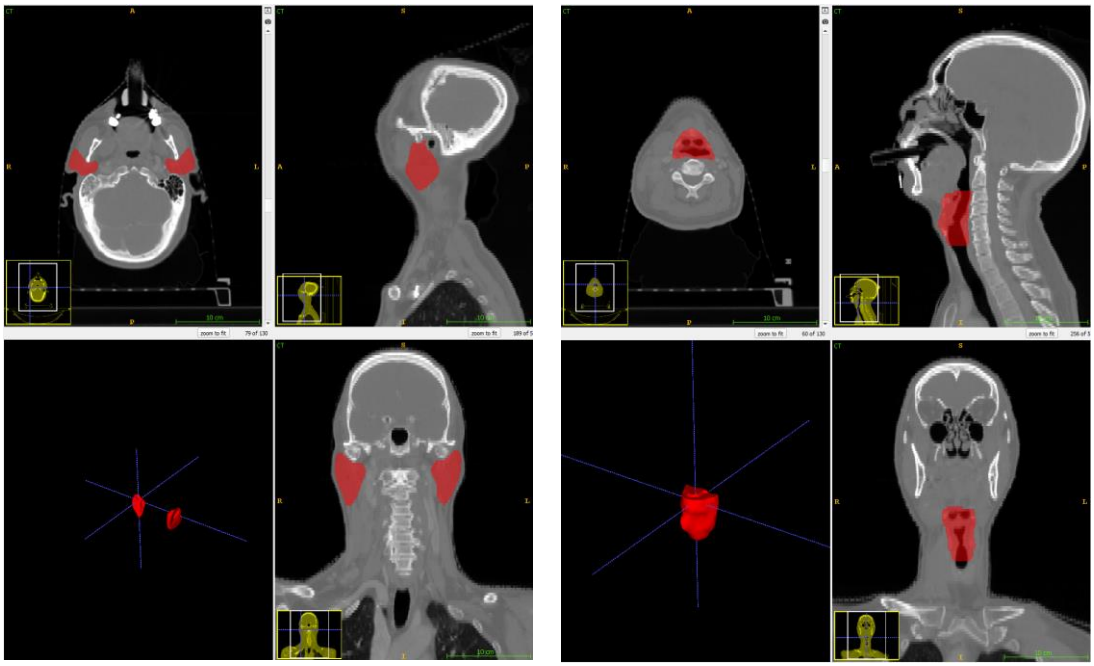


**Figure 12: Examples of parotid glands segmentation (left) and larynx segmentation (right).**

### *3.5.4. Oesophagus*

The collected RT data included contours of other OARs beyond those highlighted in the previous sections. Some of these, like the spinal cord, were not considered to be sufficiently related to OM or dysphagia for inclusion in the analysis. However, the cervical oesophagus had been contoured for many patients, and is part of the swallowing anatomy that could affect dysphagia. Despite this, there were substantial variations in the contouring of this OAR across

patients and institutions that precluded its use. Particularly, the axial limits of the contour, and its diameter and margin varied extensively. The exclusion of this VOI was likely to be less impactful for acute dysphagia than for late dysphagia, since narrowing and stiffness of the oesophagus result from fibrosis, which is recognised as a late treatment toxicity that worsens with time [259].

### *3.5.5. Tumour volumes*

The primary gross tumour volume (GTVp) and neck nodal gross tumour volume (GTVn) were extracted from the DICOM files collected from the treatment planning system. These contours were delineated by experienced radiation oncologists on the contrast-enhanced planning CT with reference to registered planning MRI.

Aside from the gross tumour volume, the peritumoral region has been proposed as an important VOI for NPC prognosis prediction [260, 261]. This may be defined as a shell of 3mm thickness expanding uniformly from the gross tumour volume [260]. Given the simplicity of obtaining such peritumoral VOIs, their connection with toxicity was also investigated. Initial findings were not suggestive of predictive value of these VOIs, and without strong justification for their subsequent inclusion, these peritumoral VOIs were not incorporated into the model development in subsequent chapters.

## 3.6. Preprocessing and feature extraction

Feature extraction was performed using in-house software “Radiotherapy data analysis and reporting (RADAR) toolkit” as developed by Zhang [262]. This software utilized the PyRadiomics v3.0.1 [263] package for Python v3.8.15 , which is compliant with the Image Biomarker Standardization Initiative (IBSI) [42]. The software extracted features from a database of CT, radiation dose and contour .mha files according to a set of user-defined feature extraction settings.

CT images were resampled to isotropic 1mm x 1mm x 1mm resolution during preprocessing. No normalization was applied to the image intensity, whose values represented Hounsfield Units (HU). A resegmentation range of -150 HU to 180 HU was selected to restrict the VOI to relevant soft tissues and exclude air and bone. Features were extracted from the original CT images, as well as from Laplacian of Gaussian filtered images, using radius parameters of 1mm, 2mm and 3mm. These filters had been previously found to give reasonably stable features compared to other filters such as Wavelet and offered additional information from their edge-detection effect. The HU values were discretized using a fixed bin count of 50 bins. Within the wider research group, the effect of bin count on feature stability had been investigated for counts between 16 and 128. For original and Laplacian-of-Gaussian filtered images, the stability was not strongly affected by bin count. Preliminary experiments were conducted with different bin counts and no clear optimal bin count was apparent, therefore a bin count of 50 was selected, representing a mid-range value that had been previously used in other projects.

Radiation dose distribution maps were resampled to isotropic 2.5mm x 2.5mm x 2.5mm resolution to match the original pixel spacing of the dose distribution map. Again, no normalization was applied to the intensity, whose values represented the planned dose in gray (Gy). For the dose features, a fixed bin width of 1.00 Gy was used, in accordance with previous dosiomics studies [57, 264]. A resegmentation range of 0 to 100 Gy was selected to exclude any erroneous dose values. Features were extracted from the original dose maps, as well as from Laplacian of Gaussian filtered maps, using the same parameters as the CT image filters. All dosiomic and DVH features were extracted from the standardized contours obtained from the nnU-Net model, except for the contours selected for training the model, and the contours of the GTVp and GTVn, which were contoured by clinicians.

Image pre-processing and feature extraction was performed by in-house software which utilized PyRadiomics v3.0.1 and SimpleITK v2.2.0 [36, 37]. Feature extraction was compliant with a well-established protocol of the Image Biomarker Standardization Initiative (IBSI) [38]. Radiomic features were extracted from the planning CECT, including shape, first order, and texture features. The texture features included grey-level co-occurrence matrix (GLCM), grey-level difference matrix (GLDM), grey-level run-length matrix (GLRLM), grey-level size zone matrix (GLSZM), and neighbouring grey-tone difference matrix (NGTDM) features. Dosiomic features were extracted from the planned radiation dose, including first order and texture features. Features were extracted from the original and Laplacian-of-Gaussian filtered CECT image and dose distribution. Original first order mean, median, minimum, and maximum dose features were categorized as DVH parameter features in subsequent analysis. Additional DVH features were calculated, including Dx%, the dose in Gy received by x% of the VOI and VxGy,

the fractional volume receiving at least x Gy, as defined by Gabryś et al. [21]. Vx%, the fractional volume receiving at least x% of the maximum dose to the volume, was also calculated.

Contouromic features were calculated using the method outlined by Lam et al. [41]. Here, the pairs of VOIs were GTVp and extended oral cavity, GTVp and parotid glands, GTVp and pharyngeal constrictor muscles, GTvn and extended oral cavity, GTVn and parotid glands, GTVn and pharyngeal constrictor muscles. As detailed in the study by Lam et al., overlap-volume histogram (OVH) and projection-overlap-volume (POV) features were extracted. The OVH features describe the distance between the pair of contours, specifically the fractional volume of the OAR within a certain distance of the GTV. The POV features describe the fractional volume of the OAR which is masked by the GTV at a certain angle about a given rotation axis. Additionally, the integrals of the OVH and POV curves were included as features since the area under each curve is descriptive of the overall separation or masking.

Two types of contouromic features were extracted: overlap volume histogram (OVH) and projection overlap volume (POV), as defined in the paper by Zhang et al. [265]. The OVH features described the distance between an OAR and a GTV, by quantifying how far the GTV had to be expanded to overlap with a given proportion of the OAR volume. Mean OVH curves for the included VOIs are shown in **Figure 13**.

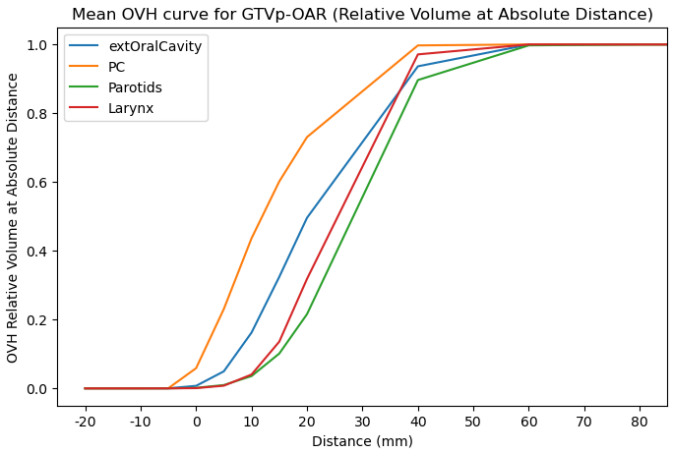



**Figure 13: Mean OVH curves for GTVp-OAR pairs in the development dataset**

POV features describe the angular relationship between tumour volumes and OARs. Specifically, they quantify the proportion of the OAR that is masked by the GTV from a given projection angle about the rotation axis. In this study, features were extracted for two different rotation axes: rotation in the sagittal plane (dim = 0) and rotation in the axial plane (dim = 2), which corresponds to the rotation axis of the RT gantry. For the rotation in the axial plane, the POV features can be illustrated by the beam's eye view **Figure 15**. The value of the POV feature indicates the fraction of the OAR masked by the GTV in the beam's eye view at that angle. **Figure 16** shows the mean POV features for GTVp-OAR pairs for rotation in the sagittal and axial planes. As an example, consider the POV curve for the larynx. The curve for rotation in the sagittal plane has a single peak, representing the range of angles where the GTVp will mask the larynx. However, the corresponding curve for rotation in the axial plane is zero everywhere, indicating that the GTVp does not mask the larynx at any rotation angle because it is located superior to the larynx.

**Figure 14: Rotation in sagittal plane (dim=0, left) and rotation in axial plane (dim = 2, right) for POV features**

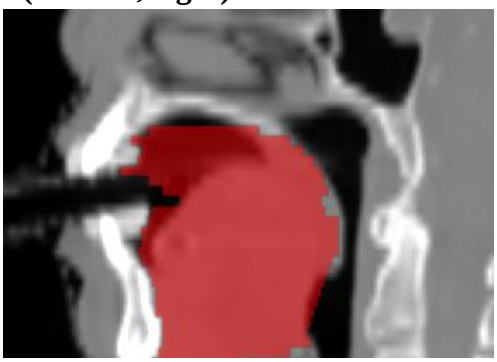


**Figure 15: Example of beam' eye view showing the masking of organs-at-risk (OARs) by the tumour volume**

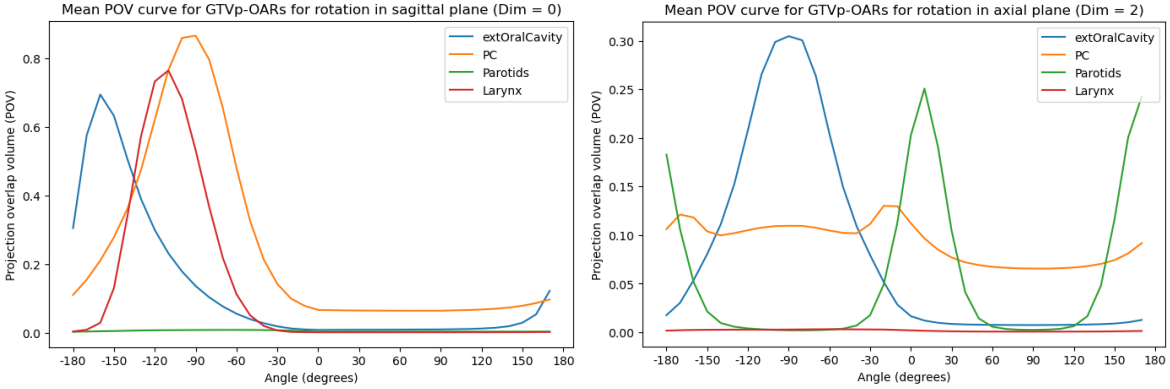



**Figure 16: POV curves for GTVp-OAR pairs for rotation in the sagittal plane (left) and axial plane (right)**

Additionally, the sum of each set of OVH and POV features was calculated and included as OVH integral and POV integral features, representing the area under the curves. These integral features were theorized to contain additional information about the distance and angular relationship between GTV and OAR.

## 3.7. Perturbation-based stability assessment

To develop a reproducible, repeatable and robust toxicity model, it is important to ensure that the model features are stable and robust to common sources of variations. Feature stability was assessed using the perturbation-based approach outlined by Zwanenburg et al. [266]. The approach involves generating synthetic perturbations to the image and contour which replicates the effects of test-retest scanning or inter-observer variation in contouring. The aim of this step is to develop reproducible models by ensuring that each feature included in model development is stable and reproducible with respect to random changes or noise. Random translations, rotations and contour deformation using a deformation vector field (DVF) were applied, and all features were re-calculated. Forty random perturbations were applied, resulting in 40 sets of perturbed features. These perturbations were applied to a subset of the patients in the development (QEH) dataset. The perturbation settings were identical to those in the publication by Zhang et al. [254].

The perturbations should replicate the effect of inter-observer variability. To verify that the selected perturbation settings resulted in a similar level of variability, the Dice similarity coefficient was calculated between each perturbation and the original segmentation for an example patient. The mean and standard deviation in the Dice coefficient across the forty perturbations was calculated and is shown in **Table 18**. The similarity varied by VOI, being highest for the extended oral cavity and larynx, and lowest for the pharyngeal constrictors. The size and surface-to-volume ratio were calculated to see how these factors impact the similarity of the perturbations. A linear trend between higher surface to volume ratio and higher Dice similarity was observed ($R^2$=0.86), as well as a linear trend between higher volume and higher

Dice similarity ($R^2$=0.58). However, insufficient data was available to tailor the perturbation settings to each VOI. The mean Dice coefficient for the GTVp was 0.78, which compares to a value of 0.72 ± 0.15 as reported in a thesis by Panyura, in which five radiation oncologists contoured the GTV for each of 30 NPC patients, and the mean Dice was computed [267]. The Dice score for the selected perturbation settings is therefore approximately in line with the similarity observed between 5 radiation oncologists' contours.

**Table 18: Dice similarity coefficient across original-perturbed pairs of contours for an example case**

| | Mean Dice | Standard deviation | Volume (cc) | Surface Volume Ratio |
|---|---|---|---|---|
| GTVp | 0.78 | 0.07 | 30 | 0.34 |
| GTVn | 0.80 | 0.03 | 75 | 0.36 |
| extOralCavity | 0.92 | 0.02 | 152 | 0.14 |
| PC | 0.75 | 0.06 | 20 | 0.59 |
| Larynx | 0.91 | 0.02 | 67 | 0.17 |
| Parotids | 0.89 | 0.02 | 71 | 0.23 |

Having computed the sets of perturbed features, the stability was assessed by calculating the intraclass correlation coefficient (ICC) for each feature. The stability was assessed by calculating the one-way, random, absolute, single rater intraclass correlation coefficient (ICC) for each feature using the Python package Pingouin v0.5.3 [268]. Features with poor stability against the effect of perturbations were removed, using an ICC threshold of 0.7 in the development dataset. Publications investigating repeatable or reproducible radiomic features previously utilized various ICC thresholds between 0.5 and 0.9 to identify stable features [269-271].

**Figure 17** shows 10 example perturbations to each VOI for a single patient, shown in the axial and lateral planes. The differences between contours should approximate typical inter-observer variations in contouring. The regions where perturbed contours disagree form a

smaller proportion of the total volume for larger, more uniform VOIs such as the extended oral cavity, larynx, and parotids, agreeing with the observations from the Dice scores. Conversely, VOIs such as the pharyngeal constrictors and GTVn were more significantly affected by the perturbations.

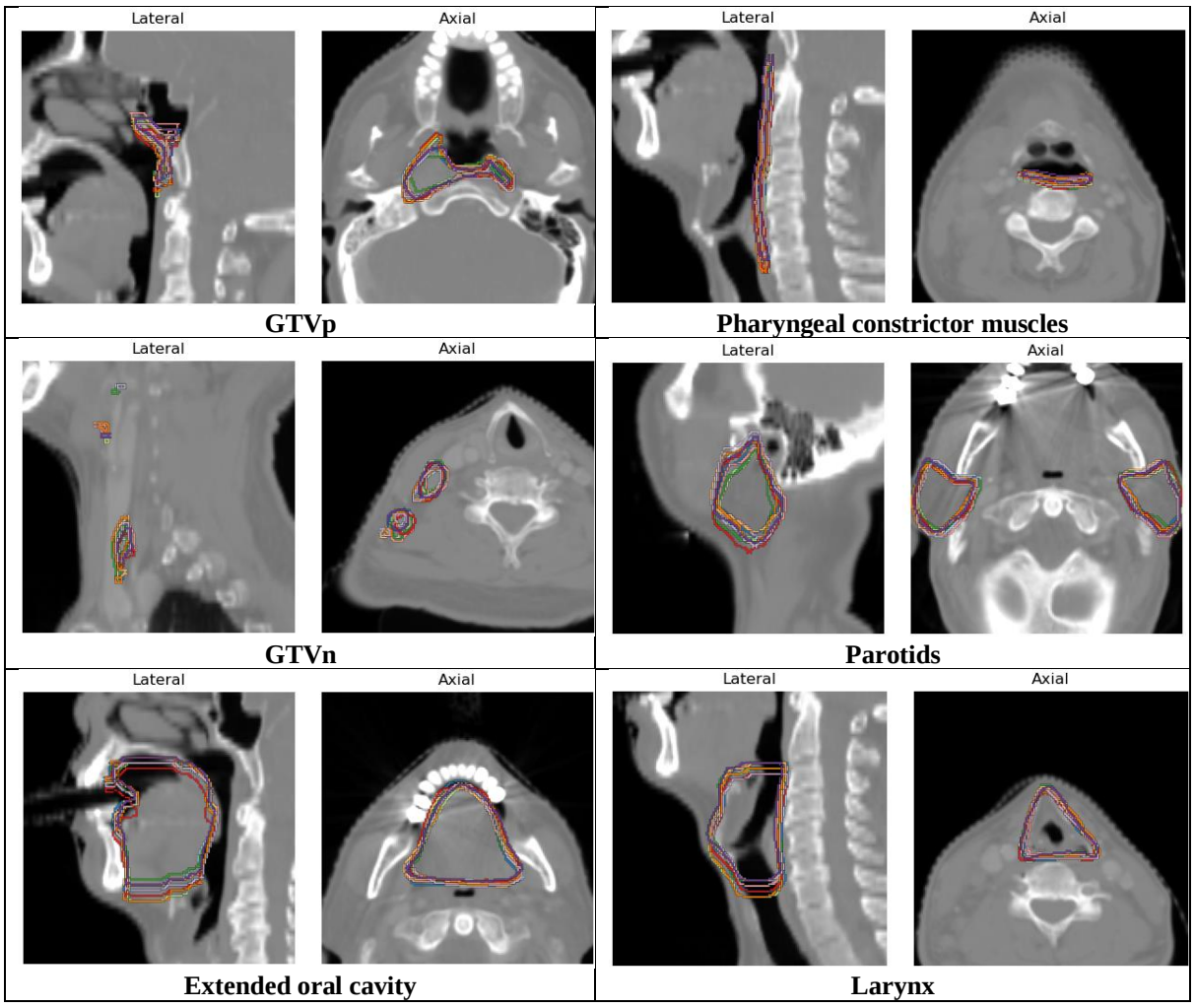



**Figure 17: Examples of perturbations to VOI contours**

**Figure 18** shows the feature stability in ICC as calculated from the image perturbation features. This includes the original features and the Laplacian-of-Gaussian filtered features. Contouromic features had high stability because the geometric changes from the perturbations had a small effect on the geometric relationships between tumour and OAR. The DVH features were mostly stable, especially for the OAR VOIs. Dosiomic features were moderately stable, and radiomic features were the least stable, particularly for OARs like the PC muscles, parotid glands and extended oral cavity. The reduced stability of DVH features for GTVn and GTVp may have been due to the steep dose fall-off around these volumes. Likewise, the low stability

of features for the PC muscles, parotid glands and extended oral cavity may have been due to the close presence of air/bone or other tissue with significantly different radiodensity to the OAR. The shape of each OAR may also have played a role, for example the PC muscles had a high surface to volume ratio and low sphericity compared to the GTVp.

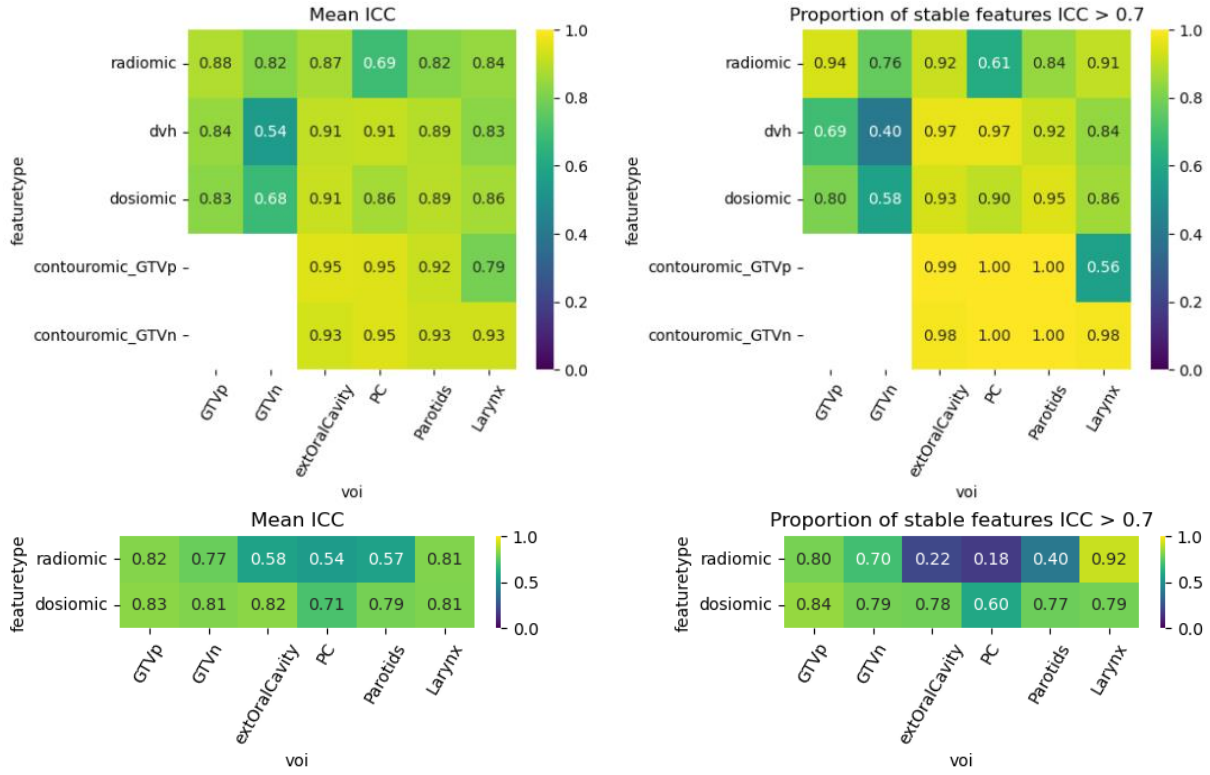



**Figure 18: Feature stability by feature type and VOI, for original features (top pair) and Laplacian-of-Gaussian filtered features (bottom pair)**

## 3.8. Model development and validation

Model development was conducted using JupyterLab, a browser-based interactive notebook interface for Python programming [272]. The entire workflow from data loading and preprocessing to visualization of the final model performance was conducted within this interface. All JupyterLab interactive notebook scripts for model development, evaluation and visualization were developed by the author. **Figure 19** outlines the key steps in model development and validation. Each step will be discussed in this section.

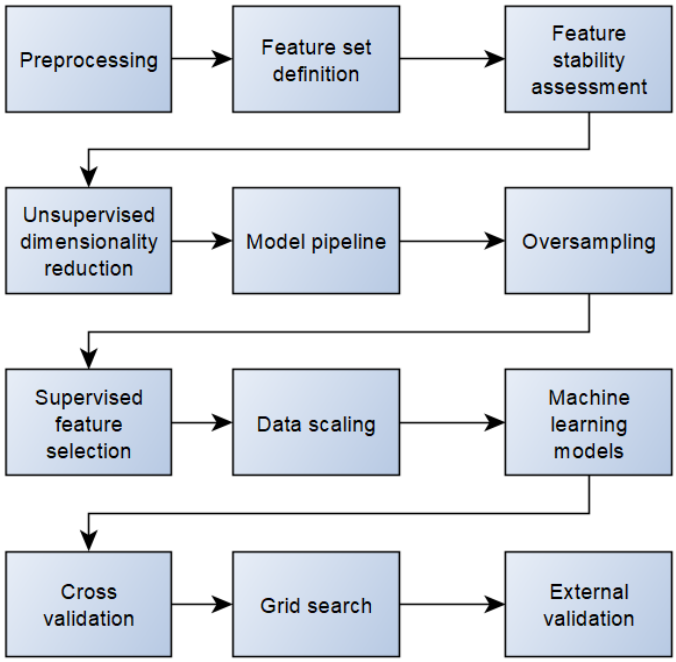



**Figure 19: Flowchart for model development and validation**

### *3.8.1. Preprocessing*

The pre-extracted features, stored as CSV files for the development (QEH) and external validation (PWH) datasets, were loaded into the notebook in preparation for preprocessing. Data tables were stored and manipulated using the Pandas package for Python. The data for the QEH and PWH datasets were stored as Pandas DataFrames, and a MultiIndex was created for

the table columns in order to allow for grouping by feature type, feature class, VOI, and individual feature name. The DataFrame rows each represented a unique patient.

The first preprocessing step was to exclude any cases with metastases, in accordance with the exclusion criteria. Next, cases were excluded if consultation notes were available for fewer than three of the seven weeks of RT and if no severe toxicity was observed. This approach is similar to the missing data handling approach detailed by Dean et al., designed to compensate for the under-reporting effect of missing weekly gradings [131]. The three-week threshold was chosen to balance statistical power against mitigation of the under-reporting effect.

Strategies were also employed for handling missing feature data. Patients who were missing any categorical clinical features (gender, chemotherapy regimen, TNM stage) were removed. The small number of remaining missing values, such as missing BMI values, were imputed by the median feature value to minimize the impact of outliers and be suitable for various value distributions. Constant or quasi-constant continuous-valued features which were more than 20% single-valued were removed since their low variance would limit their predictive value.

Not all features were normally distributed, or even symmetrically distributed. Features with high skewness may distort the results of many machine learning models. The resulting predictions may be unduly influenced by highly skewed features, which may obtain an inflated importance in the model. Therefore, experiments were performed where features were excluded if their Pearson skewness exceeded a given threshold. Some benefit was observed;

however, this was partly due to the reduction in the number of features and the dimensionality of the prediction task. After reviewing related radiomics literature, this approach did not appear to be in widespread use, and risks discarding important toxicity-related information. Therefore, the skewness filter was not utilized in subsequent chapters.

Another aspect considered was the variance of features. Features with low variance may be considered to have less information and can therefore be discarded prior to model building. However, due to the wide range of feature magnitudes, and the variation in the shape of feature value distributions, using the mathematical variance (the square of the standard deviation), does not permit valid comparisons between features. Scaling of features prior to computing the variance should also be undertaken with great care, since the scaling can be affected by outliers or skewness in the feature values. For this reason, a more robust measure of the 'spread' in the feature values was selected. The quartile coefficient of variance, defined in terms of the median and interquartile range, was investigated as a way to filter features. Different threshold values were applied, and the resulting performance of the developed models were compared. However, as with the skewness filter, this approach could not be observed in other radiomic-related studies. Furthermore, there was insufficient evidence for the improvement of performance from this approach to justify including it in subsequent chapters.

### *3.8.2. Feature set definition*

When developing a model, the starting feature set must first be defined. This includes the type of features: clinical, DVH, radiomics, dosiomics or contouromics, and also the choice of which VOIs to include. Additionally, one can choose to include features only from the

original image / dose map, or from images with specific filters applied. The decisions about each of these aspects can be informed by evidence from the literature or from a hypothesis about the potential value of these features. Including all possible filters, VOIs, feature types and feature classes for exploratory purposes is inadvisable because of the very large number of features that will result, compared to the relatively small number of samples. This results in high complexity and high redundancy / intercorrelation, while not necessarily increasing the useful information. Certain feature types or feature classes may be quantifying similar image characteristics, and so a balance had to be struck between including useful information and excluding features which were likely to be redundant. Some studies may conduct a pre-screening of features using a statistical test of association with the outcome label. However, if this is performed before splitting the data into training and validation sets then information about the validation set will effectively be leaked. This will result in an optimistic bias in the performance on the validation set. Similarly, fitting separate models for each feature type and then subsequently performing data/model fusion will also result in information leakage unless the validation set is separate. Given the limited number of samples, the model development for this project utilized cross validation as the internal validation method. Data fusion was performed prior to cross validation, in order to avoid these issues. A benchmarking study on feature selection methods likewise recommended the data fusion approach, selecting features concurrently from all feature types rather than performing feature selection separately for each [273].

To evaluate the relative performance of different combinations of VOIs and feature types, the model development process was repeated using the combinations in **Table 19**. The

table shows the combinations of VOIs and feature types for the initial feature sets. These were not an exhaustive list since this would amount to thousands of unique combinations. Rather, these include the combinations of individual VOIs and individual feature types, plus pairs of VOIs and feature types, plus some selected cases such as inclusion of all VOIs and all feature types. Together, this resulted in 330 combinations of initial feature sets. For each initial feature set, the dimensionality reduction approach could involve hierarchical clustering or VIF, and in each case, 6 different machine models were evaluated using a cross-validated grid search. Analysis of all initial feature set combinations for a given dimensionality reduction approach and resampling approach took between 8 – 24 hours, after employing the efficiency methods outlined in the other sections in this chapter. These aspects may also be referred to as model development parameters.

**Table 19: Initial feature set combinations: CLI = clinical features, RAD = radiomic features, DOS = dosiomic features, CON_GTVP = contouromic features based on GTVp, CON_GTVN = contouromic features based on GTVn**

| VOI combinations (N = 22) | Feature type combinations (N = 15) |
|---|---|
| extOralCavity | CLI + DVH |
| PC | CLI + RAD |
| GTVp | CLI + DOS |
| GTVn | CLI + CON_GTVP |
| Larynx | CLI + CON_GTVN |
| Parotids | CLI + CON_GTVP + CON_GTVN |
| extOralCavity + PC | CLI + DVH + RAD |
| extOralCavity + GTVp | CLI + DVH + DOS |
| extOralCavity + GTVn | CLI + DVH + CON_GTVP + CON_GTVN |
| extOralCavity + Larynx | CLI + RAD + DOS |
| extOralCavity + Parotids | CLI + RAD + CON_GTVP + CON_GTVN |
| PC + GTVp | CLI + DOS + CON_GTVP + CON_GTVN |
| PC + GTVn | CLI + RAD + CON_GTVP + CON_GTVN |
| PC + Larynx | CLI + RAD + DOS + CON_GTVP + CON_GTVN |
| PC + Parotids | CLI + DVH + RAD + CON_GTVP + CON_GTVN |
| GTVp + GTVn | |
| GTVp + Larynx | |
| GTVp + Parotids | |
| GTVn + Larynx | |
| GTVn + Parotids | |
| Larynx + Parotids | |
| extOralCavity + PC + GTVp + GTVn + Larynx + Parotids | |

### *3.8.3. Feature stability assessment*

After defining the initial feature set, unstable features were excluded. This served two purposes. Firstly, it was intended to increase the repeatability of the resulting models, and secondly, it would reduce the dimensionality of the prediction problem. Unstable features were identified by their robustness to the simulated perturbations, as measured by the ICC. A threshold of 0.7 was used to remove unstable features. This value was one of the frequently reported thresholds identified in a systematic review by Xue et al. [274].

### *3.8.4. Unsupervised dimensionality reduction*

The datasets in this project are representative of other datasets in radiomics and related fields which tackle high-dimensional data. Specifically, there are a large number of extracted

features and comparatively few samples. By contrast, many applications of deep learning utilize vast amounts of data consisting of many orders of magnitude more samples. With many more features than samples, machine learning models may struggle to discern patterns amidst the noise. For this reason, feature selection and dimensionality reduction methods are employed. Supervised methods which utilize the outcome variable to determine relevance should be applied to the training data only, to avoid introducing bias in the results from information leakage**.** However, unsupervised methods can remove redundant or low-variance features prior to data partitioning.

Multicollinearity is the phenomenon where several features are linearly correlated with each other. If they are also correlated with the outcome variable, then during model building, it is unclear which feature to select. This redundant information can result in multiple correlated features being selected, resulting in an overly complex model that may also risk overfitting. Pairwise tests such as the Pearson correlation coefficient can determine how strongly correlated two features are, and a matrix of these coefficients can be computed for any feature set. However, identifying and removing redundant features can be performed in many different ways. One approach explored in this project was to identify clusters of correlated features using these coefficients. Then, the most stable and conceptually simplest feature was selected from each cluster. An alternative approach to detecting multicollinearity was performed by calculating Variance Inflation Factors (VIFs) [275]. VIFs are calculated by fitting a regression model for the selected feature using only that feature as input, then fitting a regression model for the same feature using all of the available features as inputs. The VIF is the factor by which the variance in the estimate is inflated in the multi-feature model versus the single-feature

model. This is indicative of the amount of collinearity or redundancy between the features in the set. The VIF can be calculated for each feature in a feature set. The higher the VIF, the greater the multicollinearity. A cutoff value for VIF can be applied to remove features with high multicollinearity. A cutoff value of 10 has previously been proposed [275].

Two different approaches to dimensionality reduction were employed, each with the aim of reducing the amount of multicollinearity and redundant features. The first approach was based on hierarchical clustering using correlation coefficient. The second approach used a combination of correlation coefficients and VIF calculations. The motivation of this second approach was to build up a set of non-redundant, non collinear features using VIF that would make subsequent model building easier. The logic is similar to that described in the paper by Cheng et al., except that it is used for unsupervised feature selection instead of supervised feature selection [276]. The initial feature set is sorted in order of perturbation stability by ICC, then by complexity of features, from clinical to DVH, radiomic, dosiomic, and contouromic. Then, features are iteratively added to the reduced feature set on two conditions: 1) The maximum pairwise Pearson correlation coefficient between any two features in the resulting set is less than a given threshold, and 2) The maximum VIF of the features in the resulting set is less than the VIF cutoff value. This results in a reduced feature set with low redundancy. No information is provided about the outcome variable, however, so this can be applied to the whole development dataset.

To summarize, two different unsupervised dimensionality reduction methods were investigated in this project: **1)** a hierarchical clustering approach using Pearson correlation, and

2) a VIF-based approach: iteratively selecting features while Pearson correlation and VIF thresholds were met. The pseudocode for these two approaches is displayed in **Table 20**.

**Table 20: Pseudocode for unsupervised dimensionality reduction approaches**

| Hierarchical clustering | VIF approach |
|---|---|
| 1. Calculate Pearson correlation matrix *R* for input features<br>2. Define distance as 1 - \|*R*\|<br>3. Calculate feature clusters based on distance and distance threshold (0.9)<br>4. Sort each cluster by feature type (clinical-DVH-radiomic-dosiomic-contouromic) then by feature stability (ICC)<br>5. Select first feature from each cluster<br>6. Return reduced feature set | 1. Sort features by feature type (clinical-DVH-radiomic-dosiomic-contouromic) then by feature stability (ICC)<br>2. Add first feature to set of reduced features<br>3. Iterate over remaining features<br>4. Check whether each feature meets the following criteria:<br>a. The feature is not highly correlated (R > 0.9) with any feature in the reduced set<br>b. The resulting maximum VIF does not exceed the threshold (10)<br>5. If the criteria are met, add the feature to the reduced feature set |

## *3.8.5. Model pipeline*

Supervised feature selection, scaling and model fitting steps were incorporated into a Pipeline object from the Scikit-learn package for Python [277]. The Pipeline object was selected because it provided a convenient encapsulation of the series of steps needed to fit or conduct inference on a model, while also avoiding information leakage by ensuring the correct treatment of training and test data. The Pipeline object was called with the 'fit' method during training, which resulted in the feature selection, scaling and model fitting being applied in a supervised manner on labelled data. Then, during validation (either within cross-validation or

on the external validation set), the object was called with the ‘predict’ method. This method evaluated each step of the pipeline on the input data without changing the fitted coefficients, as should be the case for inference, thereby preventing information leakage to the test set.

Another important use of the Pipeline object was the ‘memory’ argument. This would store the results of each step in the pipeline in temporary memory, so that it could be recalled later if an identical step was performed. The model pipeline would be fitted and evaluated thousands of times during the optimization process, and therefore it was critical to employ this functionality to reduce the computational burden and increase the time efficiency. Using the memory argument allowed these results to be recalled, saving computation time.

### *3.8.6. Oversampling*

Data imbalance, referring to differences in the incidence of each output label, is a common consideration in machine learning. As class imbalance increases, models may be less able to learn to identify or separate classes, since they learn from fewer examples compared to the majority class. If the sample size is small, there may also be very few samples for the minority class, risking overfitting. There are several ways to address this issue. Firstly, most machine learning models allow for a weighting parameter for each sample, which can be adjusted to better balance the impact of each sample based on the rarity of the class it belongs to. In Sci-kit Learn, this is performed using the ‘class_weight’ parameter. By default, this weights samples proportionally to their incidence. However, adjusting weightings alone may be insufficient, because there may be too few samples for the minority class for the model to learn meaningful patterns. Another approach is to use sampling.

Imbalanced datasets may be under-sampled in order to remove samples of the majority class, or over-sampled to increase the number of samples of the minority class. Typically, under-sampling is more appropriate for a large dataset, where the loss of data is less impactful. Over-sampling may be performed using bootstrapping, or using more sophisticated techniques which generate synthetic samples of the minority class. Synthetic Minority Over-sampling (SMOTE) is one such approach [278]. SMOTE identifies *k* nearest neighbours to each minority sample and uses them to generate a synthetic sample, which yields more generalizable results than over-sampling with bootstrapping.

Model development explored two different approaches: no under/oversampling, and SMOTE. The performance of each approach was compared. This oversampling step was incorporated into the front of the model pipeline, so that it would be applied before any of the other steps. By including it in the model pipeline, the correct application of oversampling was ensured. Namely, oversampling would only be applied to the training set or training folds during cross validation. SMOTE was implemented using the Imbalanced-Learn package for Python [279]. Additionally, the Imbalanced-Learn version of the Pipeline object was used to ensure compatibility. Initial experiments explored performing SMOTE oversampling as the first step of the model pipeline, however no benefit was observed. Therefore, no oversampling was used for subsequent modelling.

### *3.8.7. Supervised feature selection*

Supervised feature selection generally refers to the identification of a subset of relevant features for model building. This serves several purposes. It identifies the features which have

strong associations with the outcome, removes irrelevant features and reduces the dimensionality of the problem. This results in simpler, more interpretable models, reduces the computational burden, and even mitigates the risk of overfitting. Within multi-omics literature, a wide range of feature selection approaches have been conducted, and no consensus has yet been reached on a standard feature selection approach. Feature selection approaches can broadly be categorized into filter-based, embedded, and wrapper-based methods [280]:

Filter-based feature selection methods are conducted prior to model fitting and are therefore independent of the model algorithm [280]. Examples include selecting features based on tests of association with the outcome, screening out redundant features, information gain, and minimum redundancy maximum relevance (MRMR).

Embedded methods undergo feature selection as part of the model training [280]. Typically, the trained model will learn coefficients or feature importance values which can go to zero or be compared against a threshold in order to exclude features. Examples include logistic regression with using the least absolute shrinkage and selection operator (LASSO) and linear support vector machine (SVM) with recursive feature elimination.

Wrapper-based feature selection methods are based on optimizing the feature set based on a performance score from the prediction model [280]. The performance score would typically be based on the average score from cross validation of the model using a given subset of features. Forward selection approaches involve iteratively adding to the subset of selected features until the improvement in performance falls below some threshold. Backward selection approaches involve iteratively removing features until an optimal feature set is reached [280].

A review on feature selection methods for machine learning by Bolón-Canedo recommended the use of filter-based methods, having conducted experiments on synthesized data [280]. They argued that they have good generalization ability and are faster than embedded or wrapper-based methods. Additionally, because they are independent of the model algorithm, they may be more interpretable and more robust. This project utilizes MRMR, a filter-based feature selection method, following the recommendations of a benchmark study [273].

MRMR is a feature selection framework first proposed by Ding et al [281]. Rather than ranking features by their individual correlation with the target outcome, MRMR seeks to minimize the redundant features that would be selected using this approach, while still maximizing the ability of the features to separate the target outcome labels. The framework can be used for categorical or continuous variables.

The relevance is scored by performing a statistical test on the feature data and target. For discrete-valued features, the mutual information can be used as a statistic. For continuous features, the F-statistic can be used. The redundancy is scored by quantifying the correlation between features. For discrete-valued features, this can be found using the mutual information between features. For continuous features, this can be done using the Pearson correlation coefficients, or Euclidean distances between features. The overall MRMR score can then be assigned by combining the relevance and redundancy scores. They can be combined using either the difference or the quotient of the relevance and the redundancy score. The choice of difference or quotient, and the type of data (discrete vs continuous) results in a set of possible MRMR calculations which are assigned the following names: Mutual Information Difference,

Mutual Information Quotient, F-test correlation difference and F-test correlation quotient [281].

The MRMR method can also be modified to use different measures of relevance such as the Kolmogorov-Smirnov statistic or Random Forest feature importance. The Python module 'mrmr-selection' was used to implement MRMR in this project [282]. It uses the F-statistic as a default relevance scorer and Pearson correlation as a default redundancy scorer. Given that the feature data was not necessarily normally distributed, the use of this function with non-parametric tests was investigated, with the Kolmogorov-Smirnov test for relevance and the Spearman correlation for redundancy. However, performance was reduced with these scorers, and therefore the default settings were used.

To improve the efficiency of the model pipeline during the grid search optimization, rather than computing MRMR separately for different numbers of features to select *k*, MRMR was computed for the maximum number of features to select, and then the top *k* chosen features were simply selected from that according to the grid search. This is because the MRMR object returned the top k features in order of importance.

### *3.8.8. Data scaling*

The extracted multi-omic features had extensive variation in the scale and distribution of values. Scaling features is a standard preprocessing step for the development of machine learning models, ensuring that the coefficients for each feature are more directly comparable and can be optimized more easily. Different scaling methods are available, including normalization to a range of 0 to 1 or scaling to zero mean and unit variance. The advantage of

the latter approach is that it will centre normally distributed data. However, many features are not normally distributed, and may contain outliers that could distort the mean and standard deviation. For these reasons, features were instead scaled by the median and interquartile range, which would be less susceptible to the effect of outliers [283]. This was implemented using the RobustScaler object from the Scikit-learn package.

### *3.8.9. Machine learning models*

A number of different machine learning algorithms are commonly employed for radiomic models. Typically, studies compare performance and select an optimal algorithm for their application, rather than selecting the algorithm in advance. The exception to this would be where the algorithm is selected for simplicity or for its use in a specific embedded feature selection approach.

Six types of model algorithms were investigated in this project:

1. Ridge logistic regression
2. Support vector machine (SVM) classifier with linear kernel
3. Support vector machine (SVM) classifier with radial basis function kernel
4. Random Forest classifier (RF)
5. XGBoost classifier (XGB)
6. Gaussian Naïve Bayes classifier (GNB)

#### *Ridge logistic regression*

This algorithm refers to logistic regression with a regularization penalty proportional to the square of the model coefficients [284]. Regularization is desirable in order to limit model

complexity and control overfitting. Ridge regression differs from Lasso logistic regression, where the penalty is proportional to the absolute values of the model coefficients. Lasso regression is commonly used for feature selection since it commonly sets coefficients to zero. Ridge regression was selected for this project in order to restrict the feature selection to the MRMR algorithm.

### *Support Vector Machine*

This algorithm works by identifying a hyperplane in the feature space which best separates between classes. The hyperplane parameters are optimized by maximizing the distance between the hyperplane and the nearest data points from each class. A kernel function may be used to transform the feature space to a higher-dimensional space, so that a non-linear decision surface can be represented [285].

### *Random Forest*

A decision tree is a method for classifying data, where each sample begins at the root node, a test is performed in which a feature value is compared to a threshold, and the sample passes to the next node, where further tests are performed. This continues until each sample has reached a leaf node with an associated class label. The tests at each node are optimized by using a statistical measure such as information gain or Gini index. Random forest is an ensemble method which generates a number of decision trees each with a random subset of the data [285]. The output is determined from the combination of the output of all the decision trees in the random forest.

*XGBoost*

Extreme Gradient Boosting algorithm, or XGBoost, is an ensemble method similar to Random Forest, in that it also uses decision trees as base learners [286]. The algorithm sequentially adds weak learners to the ensemble, with subsequent learners further minimizing the loss from previous learners. The minimization of the loss function is performed using a gradient descent approach.

*Gaussian Naïve Bayes*

This algorithm uses Bayes' Theorem of probability, where posterior probabilities of an event are calculated according to prior probabilities of other events [285]. Gaussian Naïve Bayes assumes that each class follows a normal distribution, and that each feature is independent.

### *3.8.10. Cross validation*

Determining the optimum MRMR K and model hyperparameters for a given model was performed using a cross validated grid search. For each combination of parameters, the model pipeline was fitted and evaluated using cross validation. The mean performance across validation folds was then used to identify the best set of parameters. Cross validation was used instead of a single train test split in order to consider the variability within the development dataset and better utilize the limited number of samples.

When performing the grid search, 20-fold stratified cross validation was used. It was observed that increasing the number of folds gave improved performance, which was likely due to increasing the effective size of the training data. Some studies perform leave-one-out

cross validation, which is the extreme case of using all but one sample for training for each fold, however this is computationally expensive and mostly applied to studies with very small datasets. 20-fold cross validation was selected as a balance between computation time and maximizing the size of training folds. The random state was fixed in order to ensure reproducibility of results.

### *Nested cross validation*

Some studies utilize nested cross validation. This involves conducting a second cross validation procedure within the training folds of an outer cross validation structure, and then taking the mean performance of the resulting models on the outer validation folds. It can further reduce the optimism bias in the internal validation score, however performing the grid search within a nested CV is extremely computationally expensive and was unnecessary due to the availability of an external validation set. The bias from having non-nested cross validation is minimal, since scaling, feature selection and model optimization all take place within the cross-validation training folds.

### *Leave-one-out cross validation*

There was significant variability in the model development process across different training folds when 5, 10 or 20-fold cross validation was used. If the sample size were very large, for example in the tens of thousands, then such differences would be much less significant. One approach which could address this issue is leave-one-out cross-validation, where all but one sample are used as training folds, and a single sample is reserved for validation. There are consequently as many folds as there are samples. This approach results in less variability between training folds and has been employed by several radiomics-related studies with small

sample sizes. However, this method results in a substantial increase in computational burden, by a factor of 10 or more. When comparing models constructed using different combinations of VOIs and feature types, the multiplicative effect of this computational burden prevented the use of this approach. Moreover, a simulation study by Geroldinger et al. found that the leave-one-out strategy was strongly negatively biased for measures of discrimination [287].

### *3.8.11. Grid search optimization*

*Hyperparameter grid*

A hyperparameter grid was defined for each model, providing a range of values for performance optimization. For example, the logistic regression models had a hyperparameter *C* which controlled the strength of the regularization, and Random Forest models had hyperparameters which controlled the depth of the trees and the minimum leaf size. The range of values defined for each model is shown in **Table 21**. There are potentially a large number of hyperparameters to choose from in models such as Random Forest, but the ranges were chosen to give reasonable coverage from simple / highly penalized models to more complex models.

**Table 21: Model hyperparameter grid ranges**

| Model type | Hyperparameter grid |
|---|---|
| Ridge logistic regression | MRMR k : [1, 2, 3, 4, 5, … $k_{max}$]<br>C : [0.001, 0.01, 0.1, 1, 10, 100, 1000]<br>class weight: [equal, balanced] |
| SVM linear | MRMR k : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>C : [0.001, 0.01, 0.1, 1, 10, 100, 1000]<br>class weight: [equal, balanced] |
| SVM RBF | MRMR k : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>C : [0.001, 0.01, 0.1, 1, 10, 100, 1000]<br>class weight: [equal, balanced] |
| Random Forest | MRMR k : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>n estimators : [50]<br>max depth : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>max features: [sqrt, log2, none]<br>class weight: [equal, balanced] |
| XGBoost | MRMR k : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>n estimators : [50]<br>max depth : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>learning rate: [0.01, 0.1, 0.3] |
| Gaussian Naïve Bayes | MRMR k : [1, 2, 3, 4, 5, 6, 7, 8, 9]<br>var smoothing : [1e-9,1e-7,1e-11] |

## *Obtaining a final model*

The best-performing set of hyperparameters in the cross validated grid search were applied to the model pipeline. It should be noted that each of the models fitted using these hyperparameters within each training fold of the cross validation were different. To obtain a final model with a single set of coefficients, the pipeline was fit to the entire development dataset (QEH). The training or "apparent" performance was indicated by comparing the predictions of the model on the development dataset with the corresponding labels. This performance score is not a validation score but indicates the degree of overfitting of the model when compared to the internal and external validation scores.

The internal validation performance was calculated from the mean performance across the cross-validation folds. The development dataset was used for cross validation, and therefore the validation folds and training folds were from the same centre. This represents an estimate

of the generalization performance; however, it does not consider any structural differences between centres. Alternative approaches to internal validation include a single train-test split or bootstrapping. Using a single train-test split does not capture the variability within the dataset and the results can be highly dependent on which samples are included in the training and test sets. Bootstrapping involves resampling from the dataset with replacement and is more often used for estimating the standard error on a performance estimate.

### *3.8.12. External validation*

The external validation performance was calculated using the data from PWH and was completely separate from all of the previous model development steps: stability filtering, unsupervised dimensionality reduction, and grid search optimization of the model pipeline.

## 3.9. Results visualization and analysis

*Bootstrapped confidence intervals*

Confidence intervals on the training and external validation scores were obtained by bootstrapping. 1000 bootstrapped samples were generated for each dataset, along with the corresponding model predictions. The resulting performance metrics were computed, and the confidence intervals were calculated from the standard error.

*Statistical tests*

To test for statistically significant differences in the performance of two models on the same data, the DeLong test was employed. For example, this allowed comparison of multi-omic models against conventional clinical and DVH-based models. Additionally, multivariate analysis of the model signatures, that is, the predicted probabilities from the model, against the

outcome variable was also employed to determine whether the model signatures were independently associated with the outcome.

### *Feature importance assessment*

Feature importance was assessed in two ways: Firstly the Shapley Additive exPlanations (SHAP) approach was used [288]. This method quantifies the impact of each feature on the model output. Secondly, the impact of each model feature on model performance was also assessed using a model-agnostic permutation variable importance procedure [289]. The effect of removing each feature from the model was assessed by calculating the AUC on 1000 sets of bootstrapped samples after shuffling the values of the selected feature. The greater the impact of the feature, the larger the difference between the original AUC and the AUC after shuffling the feature.

### *Receiver Operating Characteristic curves*

Receiver Operating Characteristic (ROC) curves were plotted, indicating the true positive rate and false positive rate for different probability thresholds. The true positive rate indicates the sensitivity, and the false positive rate is equivalent to 1 – specificity. For the internal validation performance from cross-validation, the mean false positive rates and true positive rates from each fold were used to plot the curve, and the standard deviations were used to plot a boundary indicating the variability in the ROC curve in cross-validation. The curves for training and external validation were also plotted.

## *Calibration curve*

The machine learning models developed in this project were optimized for discrimination performance, which does not guarantee good calibration. Different machine learning algorithms are known to have different distortions in the distribution of predicted probabilities, necessitating calibration [290]. Therefore, models were re-calibrated using the Scikit-learn CalibratedClassifierCV object. Prior to assessing the calibration and performing decision curve analysis, a dataset-specific logistic probability mapping was applied to the model. The feature coefficients of the underlying model were unchanged, and therefore the discrimination performance in AUC was also unchanged by this calibration.

Model calibration was assessed by plotting calibration curves and calculating the Brier score. For the calibration curves, the raw model predictions were binned into 5 quantiles, averaged, and plotted against the ratio of positive cases in each bin. The slope and intercept of the resulting curves could then be compared to the ideally calibrated line. The Brier score ranged from 1, indicating a completely incorrect calibration, to 0, indicating a perfect calibration. Considering both the curve and the Brier score allowed for assessment of the overall calibration. Predicted probability bins were defined using quantiles rather than being uniformly distributed, in order to ensure an equal number of samples per bin and an equal significance of each point on the curve.

## *Optimal threshold, sensitivity, and specificity*

An optimum prediction threshold was determined from the calibrated models using the Youden index, that is, the point which maximizes the sensitivity and specificity [291]. The resulting confusion matrices were calculated, along with the resulting sensitivity, specificity,

and overall accuracy. While the threshold can be adjusted depending on clinical requirements, these metrics serve as an additional means of comparison between models for a reasonable prediction threshold.

*Decision curve analysis*

Decision curve analysis was conducted for the calibrated models. The net benefit, as defined in equation (1), was plotted against the threshold probability ($p_t$) [292]. This indicates the clinical usefulness of the model in comparison to assigning all patients to be at risk of severe toxicity or assigning no patients to be at risk of severe toxicity, for different thresholds of the predicted probabilities. Clinicians may use the decision curve to choose the best model for a given threshold probability determined by the utility of the intervention [293].

$$Net\ benefit = \frac{True\ positives - False\ positives \times \frac{p_t}{1-p_t}}{N} \quad (1)$$

# 3.10. Data description

**Table 22** displays the baseline characteristics for the two datasets. The mean value of each characteristic is reported, along with the p-value significance according to either the Mann-Whitney U test or Fisher's Exact Test, depending on whether the variable was continuous or categorical. Cases in QEH had a higher proportion of advanced disease (T3, T4, N2). The contour of the gross tumour volume was smaller on average, however. The voxel volumes of the larynx, parotids, extended oral cavity and PC muscles were not significantly different between institutions, due to the use of AI segmentation. Different contouring guidelines in each institution meant that use of the original contours would have resulted in significant differences between institutions. The mean HU in the GTVp, larynx, extended oral

cavity and PC muscles were significantly different between datasets. This may have been a result of differences relating to the CT contrast and its timing in relation to CT acquisition. Statistically significant differences in the mean dose to the GTVp and OARs between institutions was likely a result of treatment planning differences. Different dose sparing guidelines were likely used, as evidenced by the differences in the original OAR contours prior to automatic segmentation. Additionally, there were likely differences in dose distribution due to the different RT modality used in each institution: QEH used helical tomotherapy while PWH used VMAT. The differences in dose distribution within institutions was less than that across institutions, explaining the statistical significance. These differences may not be clinically significant though since the difference in mean dose was less than 10% for each VOI.

**Table 22: Baseline characteristics**

| Characteristic | QEH | PWH | Sig |
|---|---|---|---|
| Sex_Male | 0.738 | 0.782 | 0.437 |
| Chemotherapy (vs RT only) | 0.848 | 0.851 | 1.000 |
| Neoadjuvant chemotherapy | 0.127 | 0.287 | <0.001* |
| Adjuvant chemotherapy | 0.179 | 0.020 | <0.001* |
| T4 | 0.201 | 0.168 | 0.569 |
| T3 | 0.667 | 0.426 | <0.001* |
| N2 | 0.736 | 0.347 | <0.001* |
| AgeAtRTStart | 54.317 | 55.505 | 0.358 |
| BMI_CTsim | 23.709 | 24.552 | 0.073 |
| BW_CTsim | 63.824 | 68.018 | 0.013 |
| GTVn Voxel Volume (cc) | 25.166 | 19.097 | 0.257 |
| GTVp Voxel Volume (cc) | 46.760 | 55.215 | 0.005* |
| Larynx Voxel Volume (cc) | 67.663 | 69.496 | 0.360 |
| Parotids Voxel Volume (cc) | 68.382 | 73.349 | 0.054 |
| extOralCavity Voxel Volume (cc) | 137.601 | 137.237 | 0.684 |
| PC Voxel Volume (cc) | 19.547 | 20.269 | 0.051 |
| GTVn mean HU | 45.025 | 46.017 | 0.171 |
| GTVp mean HU | 50.482 | 53.794 | 0.032* |
| Larynx mean HU | 39.938 | 46.103 | <0.001* |
| Parotids mean HU | 8.751 | 13.701 | 0.219 |
| extOralCavity mean HU | 46.854 | 57.065 | <0.001* |
| PC mean HU | 46.487 | 55.605 | <0.001* |
| GTVn mean dose (Gy) | 72.169 | 71.894 | 0.104 |
| GTVp mean dose (Gy) | 73.315 | 72.044 | <0.001* |
| Larynx mean dose (Gy) | 46.526 | 43.999 | <0.001* |
| Parotids mean dose (Gy) | 41.524 | 37.918 | <0.001* |
| extOralCavity mean dose (Gy) | 51.413 | 48.640 | <0.001* |
| PC mean dose (Gy) | 56.526 | 60.076 | <0.001* |

Some clinical features were excluded from the analysis due to sparsity of data. Smoking status and alcohol consumption were only available for a minority of patients in the QEH dataset, and so could not be included in model development. Other social factors such as marital status, financial status, education status, and accommodation co-inhabitants were likewise available only for a minority of cases. Performance status score, as measured by the Eastern Cooperative Oncology Group (ECOG) scale, indicated patients' level of function for daily life. However, this information was missing for a substantial proportion of patients, and so could not be included. Blood tests results from the consultation notes were available for a subset of

patients in the PWH dataset. These were not included in model development but correlations between pre-treatment blood test results and toxicity were also explored.

# CHAPTER 4 MULTI-OMIC PREDICTION MODELS FOR SEVERE ACUTE ORAL MUCOSITIS

## 4.1. Introduction

Oral mucositis (OM) is one of the most prevalent and crippling toxicities experienced by NPC patients receiving RT, posing a tremendous adverse impact on quality of life. Severe cases threaten treatment outcome by causing unplanned hospitalization or treatment interruption. Accurate pre-treatment prediction of severe OM is highly desirable, offering the potential for more targeted care and enhanced clinical decision-making. Published prediction models for severe OM in HNC generally use conventional clinical and DVH features. This chapter expands on the results from the submitted research article [294] in order to address Objective 1. Specifically, multi-omic prediction models for severe acute oral mucositis (OM) in NPC patients undergoing RT were developed and externally validated. Clinical, DVH, radiomic, dosiomic and contouromic features were investigated, along with a range of VOIs covering the GTVs and OARs. Investigation of the optimal combination of VOIs and feature types was conducted to further improve the discrimination performance. Correlations with pre-treatment blood tests, which had been reported as potential predictors of severe OM in the literature, were also investigated prior to model development. To the best of our knowledge, this represented the first externally validated model to use the specified multi-omics for OM prediction in HNC.

## 4.2. Methodology

### *4.2.1. Definition of severe acute OM*

As discussed in **Section 1.1.4**, OM can be graded according to different severity scales. In both institutions (QEH and PWH), the CTCAE grading system was used for assessing the severity of OM, with grade 3 or higher indicating severe toxicity. The severe acute OM outcome label was defined as the occurrence of CTCAE grade 3 or higher during RT.

### *4.2.2. Statistical analysis of baseline characteristics*

Patients were grouped by severe and non-severe OM outcome label within each dataset. Differences between baseline characteristics in severe and non-severe OM groups were assessed for statistical significance using Fisher's Exact test for categorical features, and Mann Whitney U test for continuous features. The non-parametric Mann Whitney U test was selected because a significant proportion of the features were not normally distributed, with Shapiro test p-value $< 0.05$. The univariate analysis was conducted separately from model development, and served to verify whether there were any clinical or DVH features significantly correlated with severe acute OM in both datasets.

### *4.2.3. Analysis of pre-treatment blood tests as predictors of severe OM*

As discussed in **Section 1.2**, blood test results such as white blood cell lymphocyte count, erythrocyte sediment rate (ESR), and neutrophil-to-lymphocyte ratio were reported in association with OM. Since blood tests are routinely performed for NPC patients, particularly those undergoing chemotherapy, they would potentially represent valuable and convenient biomarkers for estimating the risk of severe OM if their predictive value was confirmed. Laboratory test results were identified from the clinical records of patients in the PWH dataset

during the data collection process. These included the results of blood tests conducted before the start of RT. Measures of blood pressure and pulse rate were also included for completeness. The number of patients with pre-RT blood test data available varied, depending on the test result. Most patients with blood test results received concurrent chemotherapy, and approximately half received neoadjuvant chemotherapy.

Pre-radiotherapy blood test results were only available for a minority of PWH patients, and were not available for the QEH dataset, therefore these factors were not included in model development. Indeed, during the literature review in **Section 1.2**, no validated prediction models for OM included blood test results (other than genetic properties of oral bacteria). Nevertheless, statistical analysis of the blood test results was conducted in order to provide recommendations for future studies. To identify correlations between these pre-treatment blood test results and severe OM, the median pre-treatment test result was calculated for each patient. The median value was selected to minimize the impact of outliers across pre-treatment test results. The mean blood test result in the severe OM and non-severe OM groups were computed, along with the effect size from Cohen's D. The statistical significance was reported using the Mann Whitney U test.

### *4.2.4. Model development*

Predictive models for severe OM were developed from two feature sets: 1) conventional clinical and DVH features only, and 2) multi-omic features including clinical, DVH, radiomic, dosiomic and contouromic features.

In the work submitted for publication, the extended oral cavity and pharyngeal constrictor (PC) muscles were selected as VOIs [294]. Several studies have previously

investigated the extended oral cavity for predicting OM [54, 56, 295]. This VOI, as defined by the guidelines by Brouwer et al. [258], contained several areas that typically exhibit the most severe mucosal changes, including the soft palate, tongue, and floor of the mouth [296]. The PC VOI, consisting of the superior, middle, and inferior muscles, was frequently contoured as part of the RT planning process, and included part of the mucosa at risk of severe reaction. Specifically, the hypopharyngeal mucosa was reported as the region experiencing the most severe OM after the soft palate [296]. Moreover, Tao et al. reported the radiation dose to the pharyngeal space as a significant predictor of OM [113].

In this chapter, a broader range of VOIs were explored to determine the optimal combination for OM prediction. Several VOIs were available for the included patients: GTVp, GTVn, extended oral cavity, PC muscles, parotid glands, and larynx. It was hypothesized that some of these VOIs could further enhance the model discrimination performance. Particularly, the GTVp and GTVn VOIs had been utilized by Dong et al. for OM prediction [57], while the parotid glands' role in saliva production could influence OM through its impact on oral health. The larynx VOI, located inferior to the oral cavity, was also included to explore whether its geometry or dose profile could be linked to higher severity of OM.

Data preprocessing, model development and performance evaluation was conducted as described in **CHAPTER 3**. As described in **Section 3.8.2**, different combinations of VOIs and feature types were investigated. Six different machine learning models were fitted for each combination: Ridge regression, Support Vector Machine (SVM) with linear and radial basis function kernels, Random Forest, XGBoost and Gaussian Naïve Bayes classifier. Additionally, all combinations were explored using two different dimensionality reduction approaches: hierarchical clustering and VIF-based dimensionality reduction. **Figure 20** displays an

overview of the model development parameters. For each combination, the model pipeline was optimized using cross-validation on the QEH dataset, then externally validated on the PWH dataset. This resulted in AUC scores for training, internal validation and external validation for each combination of model development parameters.

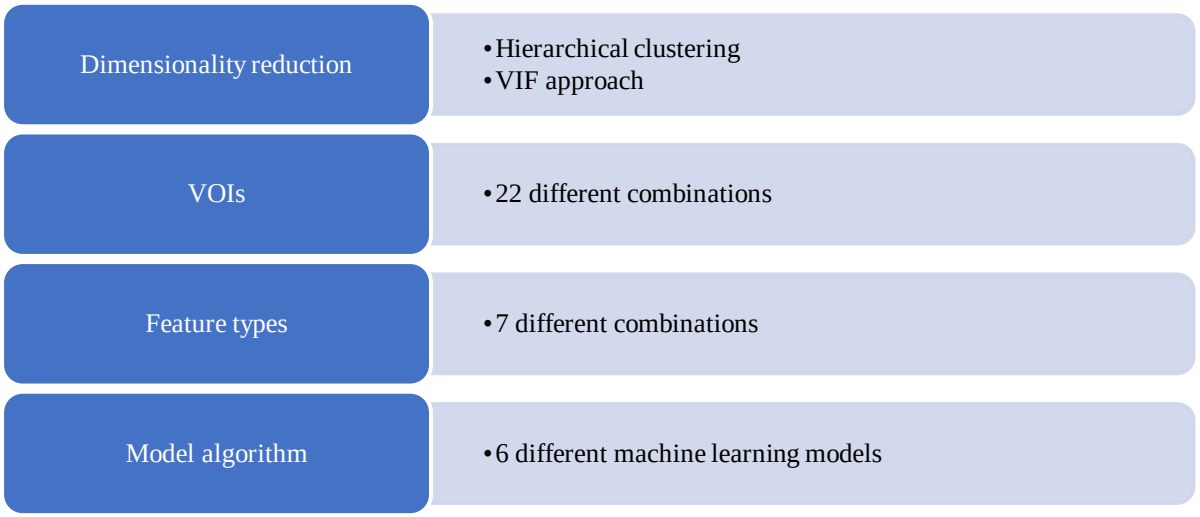



**Figure 20: Overview of model development parameters**

The optimization of each model pipeline was performed using a cross-validated grid search across the range of hyperparameters defined in **Section 3.8.11**. For the MRMR feature selection algorithm, the maximum number of features to select, $k_{max}$, was set to 9, since this would correspond to an event-per-variable rate of approximately 10, in line with the rule of thumb [131]. The optimal number of features to select was varied in the grid search from 1 up to a maximum of $k_{max}$.

To determine the most generalizable combination of VOI, omics and model algorithm, the training, internal validation and external validation AUCs were compared. For each combination, the specific models had been optimized using the cross-validated grid search on the internal validation dataset. Low training scores indicated that the model had insufficient complexity to capture the patterns in the data. Low internal validation scores indicated that the

model was overfitting to the training set. Low external validation scores indicated that the model had poor generalizability and was highly institution specific. The optimal model development parameters were identified by finding the models with the highest AUC across training, internal validation and external validation. An aggregate of the three scores was used for comparison. The mean or maximum were unsuitable as aggregate measures, because a high training score or internal validation score could obscure overfitted or poorly generalizing models. Therefore, the minimum of the three scores was selected as an aggregate score. The aggregate score was used to rank the model performances and identify the top-performing model. It is common for studies to compare training, internal and external validation scores to select the best model. The aggregate score provides a means to quantify this comparison.

Having identified the best-performing conventional and multi-omic models, the DeLong test was used to test for statistically significant differences in AUC, and a multivariate analysis was conducted with both model signatures against the severe OM label to determine independent predictive value. Feature importance within each model was assessed using the SHAP approach and the permutation feature importance approach. The calibration of each model was assessed, and the clinical utility was compared using decision curve analysis.

Additionally, the developed models were compared to the only externally validated prediction model for severe OM in the literature. This literature model was also evaluated on the QEH and PWH datasets to assess its generalizability.

# 4.3. Results

## *4.3.1. Baseline demographic and clinical characteristics*

Severe OM occurred in 90 (25%) of patients in the QEH dataset and in 30 (30%) of patients in the PWH dataset. Comparison of features between severe OM groups in each dataset is shown in **Table 23**, along with univariate analysis. It should be noted that this was not performed for the purpose of feature selection and did not influence model development. Instead, this was conducted as a separate verification of the role of clinical and mean dose DVH features. Effect size was calculated from $\frac{1}{1.81}\ln(\text{odds ratio})$ for categorical features [297], and from Cohen's d for continuous features. P values were calculated from Fisher's Exact test for categorical features, and from the Mann Whitney U test for continuous features. No correction for multiple comparisons was applied. Chemotherapy and the mean dose to the GTVn were the only significant features in the development dataset. Only chemotherapy remained significant in the external validation dataset.

**Table 23:Univariate analysis of clinical and mean dose DVH features. Incidence is shown for categorical features, and median value is shown for continuous features.**

| | Development (QEH) | | | | External validation (PWH) | | | |
|---|---|---|---|---|---|---|---|---|
| **Feature** | **Severe OM** | **No Severe OM** | **Effect size** | **P value** | **Severe OM** | **No Severe OM** | **Effect size** | **P value** |
| Age at start of RT | 54.5 | 54.0 | -0.21 | 0.180 | 55.5 | 58.0 | -0.40 | 0.132 |
| BMI at CT simulation | 23.528 | 23.528 | -0.13 | 0.548 | 24.351 | 23.932 | 0.22 | 0.169 |
| Weight at CT simulation (kg) | 62.5 | 62.5 | -0.01 | 0.728 | 71.7 | 65.9 | 0.20 | 0.167 |
| Chemotherapy (vs RT only) | 87 (97%) | 221 (81%) | 0.44 | **<0.001*** | 29 (97%) | 57 (80%) | 0.46 | **0.036*** |
| N stage = 2 | 67 (74%) | 200 (73%) | 0.03 | 0.891 | 10 (33%) | 25 (35%) | -0.04 | 1.000 |
| Male sex | 73 (81%) | 195 (71%) | 0.22 | 0.074 | 24 (80%) | 55 (77%) | 0.06 | 1.000 |
| T stage = 3 | 55 (61%) | 187 (68%) | -0.16 | 0.200 | 16 (53%) | 27 (38%) | 0.31 | 0.189 |
| T stage = 4 | 21 (23%) | 52 (19%) | 0.11 | 0.368 | 1 (3%) | 16 (23%) | -0.52 | **0.019*** |
| GTVn $D_{mean}$ (Gy) | 72.4 | 72.1 | 0.21 | **0.018*** | 72.3 | 72.3 | 0.01 | 0.266 |
| GTVp $D_{mean}$ (Gy) | 73.6 | 73.4 | 0.10 | 0.279 | 72.3 | 72.4 | 0.07 | 0.540 |
| Larynx $D_{mean}$ (Gy) | 47.0 | 47.2 | -0.03 | 0.892 | 42.8 | 43.2 | 0.05 | 0.856 |
| PC $D_{mean}$ (Gy) | 56.2 | 56.7 | -0.12 | 0.462 | 60.7 | 60.3 | 0.23 | 0.359 |
| Parotid glands $D_{mean}$ (Gy) | 40.8 | 41.5 | 0.03 | 0.673 | 37.8 | 37.8 | 0.04 | 0.685 |
| Ext. oral cavity $D_{mean}$ (Gy) | 51.9 | 51.6 | 0.10 | 0.403 | 50.4 | 47.8 | 0.55 | **0.010*** |

### *4.3.2. Correlations with pre-treatment blood tests*

The median pre-RT value of each blood test result was computed for patients in the PWH dataset. The Mann-Whitney U test was used to check for statistically significant differences in these blood test results between severe and non-severe OM groups. The effect size was also calculated using Cohen's d, and the resulting achieved statistical power was computed. The results are shown in **Table 24**. Only blood test results collected before the start of RT were included. Significant correlations were found with potassium level, white blood cell count, and creating clearance. The achieved power was low for all tests, at less than 80%, indicating that the likelihood of the test detecting a true effect was not high, and therefore that real correlations might be missed by these tests, due to the small sample size.

**Table 24: Correlations between pre-treatment blood tests and severe OM**

| Blood test result | Severe OM | No severe OM | Effect size | Power | Sig. | Sample size |
|---|---|---|---|---|---|---|
| Pulse pressure | 47.2 | 53.8 | -0.48 | 0.18 | 0.288 | 38 |
| Mean arterial pressure | 106.4 | 101.2 | 0.34 | 0.11 | 0.317 | 38 |
| Rate pressure product | 11229.5 | 11003.3 | 0.10 | 0.06 | 0.983 | 35 |
| Systolic blood pressure | 137.8 | 137.2 | 0.03 | 0.05 | 0.984 | 38 |
| Diastolic blood pressure | 90.7 | 83.1 | 0.52 | 0.21 | 0.149 | 38 |
| Pulse | 80.7 | 80.4 | 0.02 | 0.05 | 0.930 | 35 |
| Sodium (Na) | 139.9 | 138.0 | 0.17 | 0.08 | 0.500 | 43 |
| **Potassium (K)** | **4.4** | **4.1** | 0.77 | 0.58 | **0.029*** | **44** |
| Urea | 5.2 | 5.6 | -0.23 | 0.10 | 0.385 | 41 |
| Creatinine (Cr) | 74.3 | 83.1 | -0.56 | 0.39 | 0.084 | 54 |
| **White cell count (WCC)** | **6.5** | **5.2** | 0.63 | 0.51 | **0.026*** | **53** |
| Platelet (plt) | 268.3 | 299.9 | -0.38 | 0.23 | 0.309 | 54 |
| **Creatinine clearance (CrCl)** | **98.4** | **80.5** | 0.77 | 0.61 | **0.024*** | **48** |
| Haemoglobin (Hb) | 12.7 | 12.8 | -0.06 | 0.05 | 0.588 | 55 |

The univariate correlations did not adjust for confounding factors. Neoadjuvant chemotherapy was likely the most impactful factor on pre-treatment blood test results. A multivariate logistic regression analysis of the significant blood test results and neoadjuvant chemotherapy was conducted. Missing values were removed, resulting in a sample size of 24. Thirteen of these patients received neoadjuvant chemotherapy, and seven experienced severe

OM. The results of this analysis are shown in **Table 25**. None of the blood test results remained significant when accounting for the impact of neoadjuvant chemotherapy.

**Table 25: Multivariate logistic regression for blood test results against severe OM**

| Factor | z | P>\|z\| |
|---|---|---|
| const | -1.142 | 0.254 |
| Neoadjuvant | -0.002 | 0.999 |
| preRT_median_Blood_K | 1.188 | 0.235 |
| preRT_median_Blood_WCC | 0.913 | 0.361 |
| preRT_median_Blood_CrCl | -1.088 | 0.277 |

### *4.3.3. Conventional and multi-omic prediction models for severe OM*

**Table 26** lists the top 5 conventional prediction models developed for severe OM, consisting only of clinical and DVH features. The ranking was determined by the minimum of the training, internal validation and external validation scores, to ensure that the most internally valid and most generalizable models were selected.

**Table 26: Top 5 conventional prediction models for severe OM**

| Rank | Dimensionality reduction | VOIs | Algorithm | Model size | Train AUC | Int. AUC | Ext. AUC |
|---|---|---|---|---|---|---|---|
| 1 | Hierarchical | GTVp, Larynx | SVM Linear | 4 | 0.662 | 0.655 | 0.656 |
| 2 | VIF | Ext. oral cavity | Random Forest | 3 | 0.676 | 0.646 | 0.740 |
| 3 | Hierarchical | GTVp, Larynx | Ridge | 4 | 0.655 | 0.643 | 0.646 |
| 4 | VIF | Ext. oral cavity | XGBoost | 3 | 0.808 | 0.653 | 0.642 |
| 5 | VIF | Ext. oral cavity | Random Forest | 3 | 0.676 | 0.639 | 0.740 |

**Table 27** lists the top 5 multi-omic prediction models for severe OM, with the same ranking mechanism as the previous table. Notably, models containing radiomic, dosiomic and contouromic features were present in the top 5 models. All of the top 5 models outperformed the best conventional model in internal and external validation. The extended oral cavity and larynx were the most frequently selected VOIs in the top-performing models.

**Table 27: Top 5 multi-omic prediction models for severe OM. CLI = clinical, RAD = radiomic, DOS = dosiomic, CON = contouromic**

| Rank | Dimensionality reduction | VOIs | CLI | DVH | RAD | DOS | CON | Algorithm | Model size | Train AUC | Int. AUC | Ext. AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Hierarchical | Ext. oral cavity, Larynx | ✓ | | ✓ | | | XGBoost | 5 | 0.954 | 0.684 | 0.688 |
| 2 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Random Forest | 4 | 0.751 | 0.678 | 0.699 |
| 3 | VIF | GTVn, Larynx | ✓ | | | ✓ | | XGBoost | 9 | 0.849 | 0.691 | 0.677 |
| 4 | VIF | Ext. oral cavity, PC | ✓ | | | | ✓ | Ridge | 8 | 0.673 | 0.674 | 0.682 |
| 5 | Hierarchical | Ext. oral cavity, Larynx | ✓ | | ✓ | ✓ | | XGBoost | 8 | 0.788 | 0.690 | 0.673 |

## *Best-performing conventional prediction model for severe OM*

The highest performing model using only clinical and DVH features was a SVM model with linear kernel fitted after using the hierarchical clustering approach. It consisted of features from the GTVp and the larynx. The model achieved a training AUC of 0.662 (95% CI: 0.602, 0.725), internal validation AUC of 0.655 (95% CI: 0.593, 0.718) and external validation AUC of 0.656 (95% CI: 0.539, 0.768). It consisted of four features: male sex, chemotherapy, the fractional volume receiving at least 90% of the maximum dose to the GTVp, and the fractional volume receiving at least 20% of the maximum dose to the larynx. **Figure 21** shows the ROC curves for the conventional model, along with the impact of each feature on the model output, calculated using SHAP analysis.

**Table 28: Conventional model for severe acute OM**

| | |
|---|---|
| **Initial feature set** | Clinical, DVH |
| **VOIs** | GTVp<br>Larynx |
| **N features after ICC filter and hierarchical clustering** | 37 |
| **MRMR K** | 4 |
| **Model** | SVM linear<br>balanced class weights<br>C = 1 |

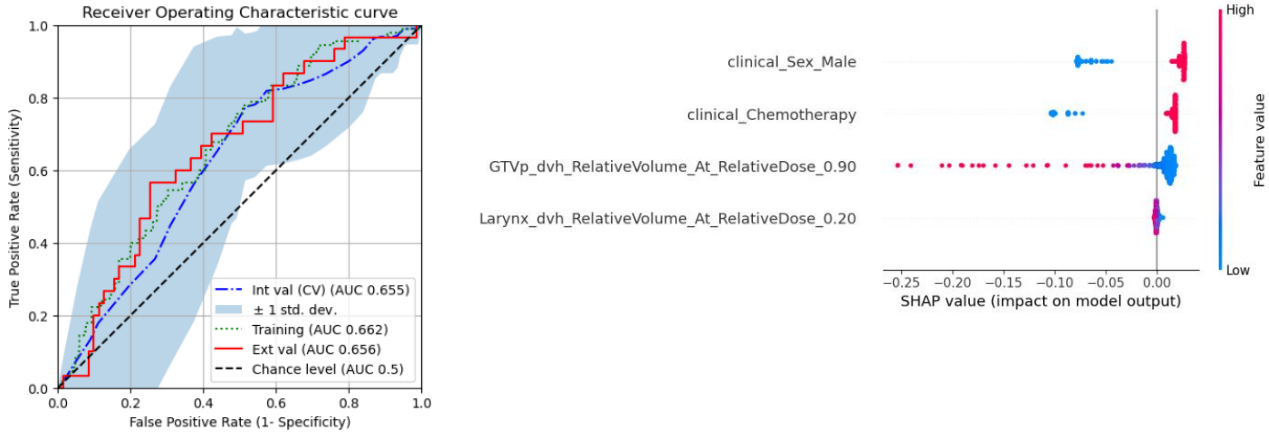



**Figure 21: ROC curve and SHAP feature analysis for conventional model**

## *Best-performing multi-omic model for severe OM*

The highest multi-omic model performance was achieved using clinical and radiomic features extracted from the extended oral cavity and larynx, filtered using the hierarchical clustering approach. The XGBoost model consisted of features from the extended oral cavity and the larynx. The model achieved a training AUC of 0.954 (95% CI: 0.929, 0.974), internal validation AUC of 0.684 (95% CI: 0.613, 0.754) and external validation AUC of 0.688 (95% CI: 0.580, 0.791). It consisted of 5 features: the sphericity of the larynx VOI, the HU intensity within the extended oral cavity, two features describing CT texture within the extended oral cavity, and chemotherapy. **Figure 22** shows the ROC curves for multi-omic model A, along with the impact of each feature on the model output, calculated using SHAP analysis. None of

the radiomic features in the multi-omic model were highly correlated (Pearson |R| > 0.7) with any clinical or DVH feature in either dataset.

**Table 29: Multi-omic model for severe acute OM**

| | |
|---|---|
| **Initial feature set** | Clinical, radiomic |
| **VOIs** | Extended oral cavity<br>Larynx |
| **N features after ICC filter and hierarchical clustering** | 243 |
| **MRMR K** | 5 |
| **Model** | XGBoost<br>balanced class weights<br>learning rate = 0.3<br>max depth = 2<br>n estimators = 50 |

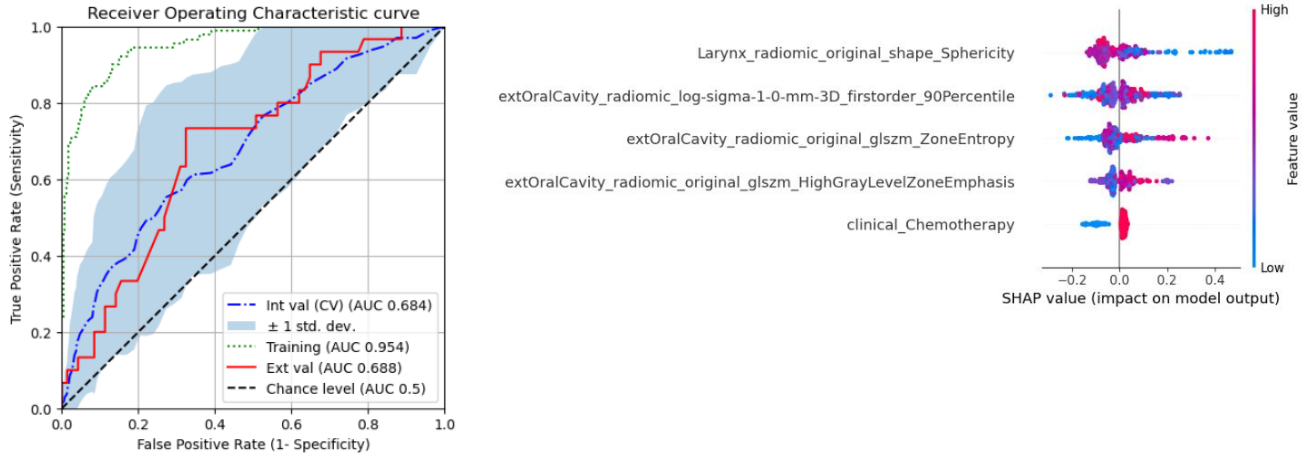



**Figure 22: ROC curve and SHAP feature analysis for multi-omic model**

## *4.3.4. Model comparisons*

*Feature correlations and model signature correlations*

The Pearson correlation coefficient between the conventional model signature and the multi-omic model signature was 0.28, indicating that the two signatures were not highly correlated. Additionally, none of the radiomic features were highly correlated (Pearson |R| > 0.7) with any of the features in the conventional model, in either dataset.

*Comparison of discrimination performance*

**Table 30** shows the discrimination performance, as measured by the AUC, for training, internal validation, and external validation. The mean AUC and its 95% confidence interval are shown. This information is also visualized in **Figure 23**. A discrepancy between the training score and the internal validation score is apparent for multi-omic model A. The internal and external validation scores for the multi-omic model are greater than those for the conventional model. As evidenced by the 95% confidence intervals, both models significantly outperformed random chance (AUC=0.5) in both internal and external validation ($p < 0.05$).

**Table 30: Comparison of discrimination performance across models for severe OM.**

| | Training (refit) AUC | Internal validation AUC | External validation AUC |
|---|---|---|---|
| | Mean and 95% CI (1000 bootstraps) | Mean and 95% CI (cross validation) | Mean and 95% CI (1000 bootstraps) |
| Conventional model | 0.662 (0.602, 0.725) | 0.655 (0.593, 0.718) | 0.656 (0.539, 0.768) |
| Multi-omic model | 0.954 (0.929, 0.974) | 0.684 (0.613, 0.754) | 0.688 (0.580, 0.791) |

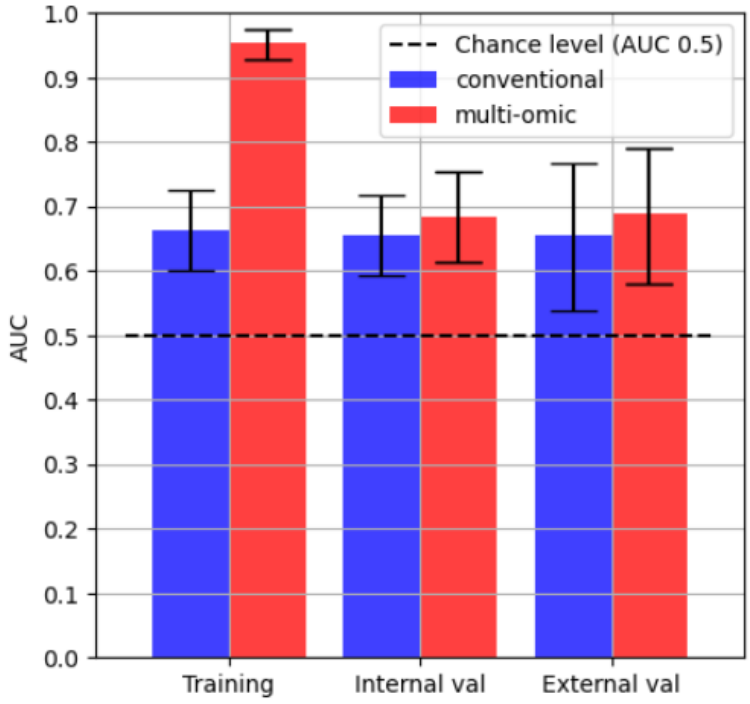



**Figure 23: Comparison of discrimination performance for severe OM models**

**Table 31** shows the results of a multivariate logistic regression analysis of the conventional and multi-omic model signatures against the severe OM outcome label. The multi-omic model signature had a statistically significant association with severe OM in both datasets when controlling for the effects of the conventional model, indicating its independent predictive value. The lack of significance for the conventional model suggests that it does not provide any additional benefit for prediction of severe OM beyond the multi-omic model. However, it should be noted that both the multi-omic model signature and the conventional model signature were significantly associated with the outcome in univariate analysis in both datasets using the Mann Whitney U test ($p < 0.05$).

**Table 31: Multivariate logistic regression of model signatures**

| | Variable | P value |
|---|---|---|
| Development dataset (QEH) | Multi-omic model | <0.001* |
| | Conventional model | 0.964 |
| External validation dataset (PWH) | Multi-omic model | 0.009* |
| | Conventional model | 0.317 |

## *Statistical significance of the results*

The DeLong test was used to calculate the statistical significance of the differences between the conventional and multi-omic models in the training and external validation datasets. The results are shown in Table 32. Neither the training nor external validation differences were statistically significant at the level of 0.05. This may have been partly due to the limited sample size in this study, resulting in insufficient power for the test. Despite the bootstrapped confidence intervals showing that the training AUCs were much higher for the multi-omic model, the p-value was still non-significant. The improvement in external validation would likely require a much larger sample size to reach statistical significance.

**Table 32: DeLong test p-values for top OM models**

| Dataset | P-value |
|---|---|
| Development (QEH) | 0.0547 |
| External validation (PWH) | 0.5729 |

An alternative approach to investigating the statistical significance of the results was performed by taking bootstrapped samples from the training and external validation sets, then calculating the performance of each model on the bootstraps. The difference between the performance of each model was recorded and the distribution over 1000 bootstraps was plotted, as shown in Figure 24. This approach allowed a confidence interval in the improvement in the AUC from the multi-omic model to be calculated. These confidence intervals and resulting p-values are shown in Table 33. This analysis showed a significant improvement in the training score, but not in the external validation score, although the multi-omic model had a net improvement in the performance over the conventional model across bootstraps.

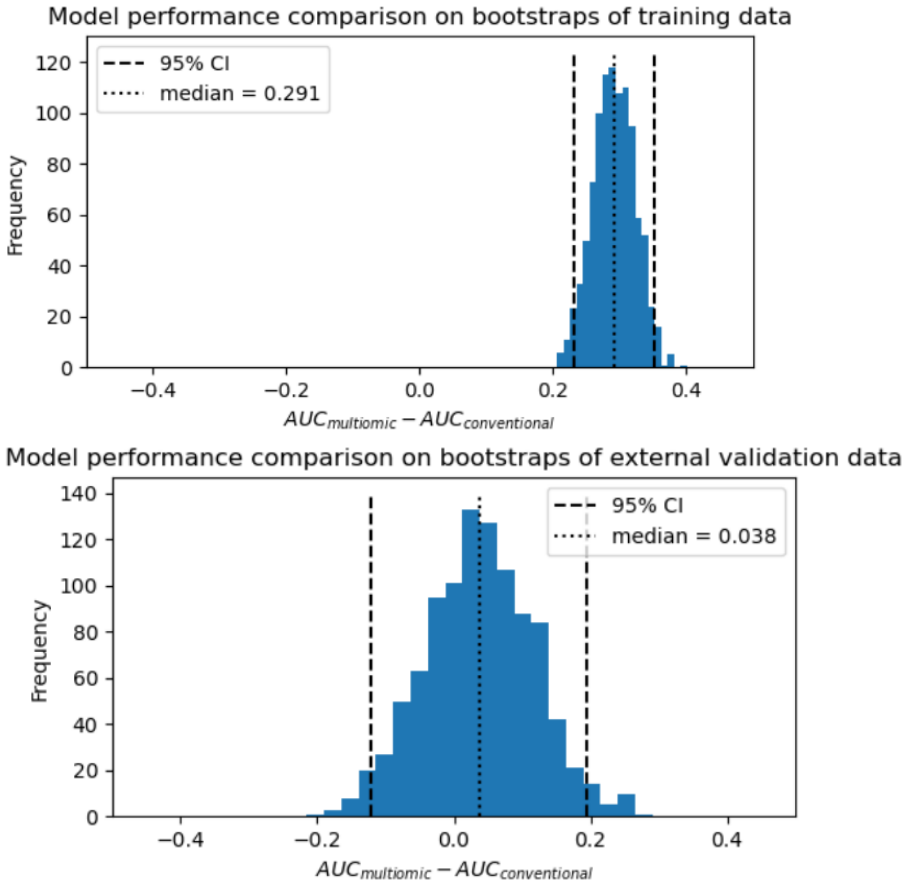



**Figure 24: Distribution of performance improvement across bootstraps for OM**

**Table 33: Performance improvement across bootstraps**

| Dataset | 95% CI | P-value |
|---|---|---|
| Development (QEH) | (0.231, 0.351) | 0.000* |
| External validation (PWH) | (-0.122,0.195) | 0.318 |

## *Comparison with externally validated model from the literature*

The only externally validated prediction model for severe OM in the literature consisted of a single feature: the mean dose to the oral mucosa. The definition of this VOI was similar to that defined for the extended oral cavity in this study, including both the tongue and oral cavity. **Table 34** shows the performance of this logistic regression model by Otter et al. [295] in their original study, in external validation conducted by Sharabiani et al. [52], and in external validation on the two datasets in this study. Relatively good discrimination is observed on the data from Sharabiani and from PWH, however the discrimination on the original dataset and on the QEH dataset is relatively poor. The results exhibit a wide range in AUC, from 0.53 to

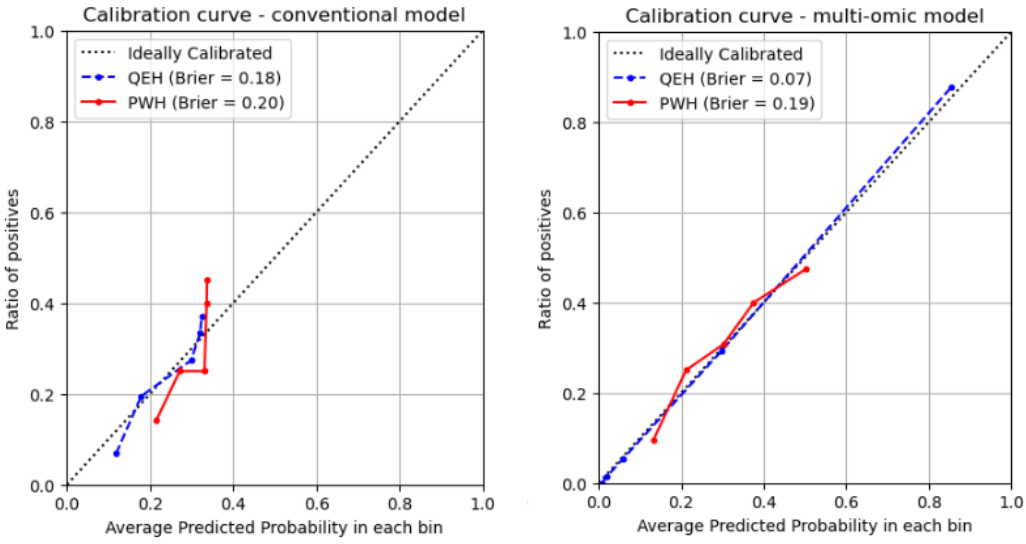
0.67. The multi-omic model outperformed the model by Otter in internal and external validation.

**Table 34: Performance of logistic regression model by Otter et al.**

| | |
|---|---|
| **Internal validation – original study [295]** | 0.62 AUC |
| **External validation by Sharabiani et al. [52]** | 0.67 AUC |
| **External validation – QEH dataset** | 0.53 AUC |
| **External validation – PWH dataset** | 0.66 AUC |

## *Calibration*

**Figure 25** shows the calibration curves for the conventional model and for the multi-omic model. The calibration curve for the multi-omic model on the training dataset was very close to the ideal calibration. The calibration on the external validation set also closely follows the ideal curve, albeit over a shorter range of predicted probabilities. The Brier scores for the multi-omic model were better than for the conventional model, and the range of predicted probabilities was larger in all cases.




Conventional model Multi-omic model A

**Figure 25: Calibration curves for severe OM models**

## *Decision curve analyses*

**Figure 26** shows the decision curves for the multi-omic model and the conventional model in both QEH (development) and PWH (external validation) datasets. The multi-omic model demonstrated a greater net benefit over the conventional model in both datasets. The important range for the threshold probability corresponds to the range of expected incidences of severe OM. These range from around 0.25 to around 0.5, as identified from the datasets included in this study and from the literature.

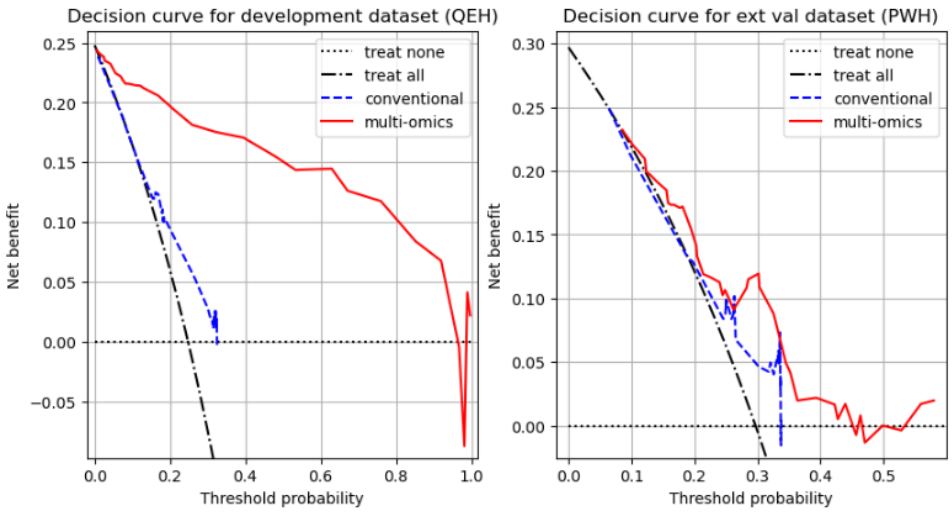
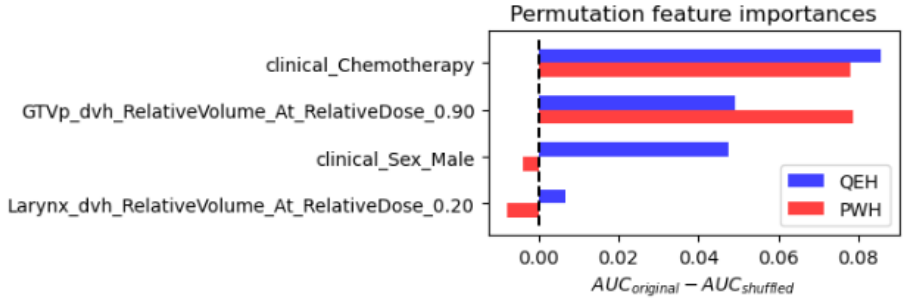



**Figure 26: Decision curve analysis for severe OM models**

## *Permutation feature importance*

**Figure 27** shows the permutation feature importance for the best conventional model and multi-omic model. Unlike for the conventional model , all of the features in the multi-omic model had positive permutation feature importance for both training and external validation datasets, indicating that the model discrimination score was reduced when that feature was removed from the model.

**Conventional model**




**Multi-omic model**

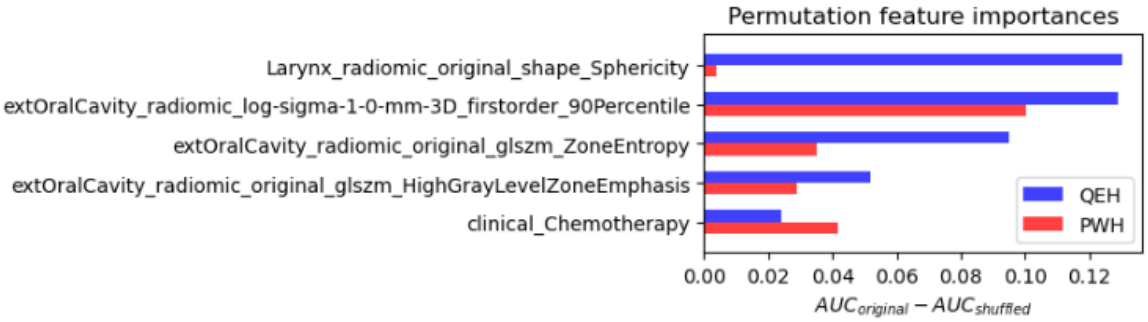



**Figure 27: Permutation feature importance for severe OM models**

### *4.3.5. Frequently selected features in top 5% of models*

The conventional and multi-omic models with the highest discrimination performance were identified in the previous section. Some models with different VOI, feature type, dimensionality reduction approach and algorithm combinations achieved reasonable internal and external validation performance but selected different features. To identify trends in the types of features selected, an analysis of the model features across the top 5% of developed models was conducted. The top 5% was chosen in order to focus only on the models with the highest discrimination performance. This analysis was conducted in two ways: firstly, counting the number of models where each feature was selected (**Table 35**), and secondly, by weighting each feature by the aggregate AUC (minimum of training, internal, external validation) of its corresponding model and finding the sum for each feature (**Table 36**). The Pearson correlation matrices for the frequently selected features are shown in APPENDIX (**Figure 41** and **Figure 42**).

**Table 35: Feature counts across top 5% of models for severe OM**

| Feature | Number of models |
|---|---|
| clinical_Chemotherapy | 189 |
| clinical_Sex_Male | 142 |
| clinical_AgeAtRTStart | 56 |
| clinical_T3 | 47 |
| extOralCavity_radiomic_log-sigma-3-0-mm-3D_firstorder_Mean | 38 |
| clinical_T4 | 21 |
| extOralCavity_dosiomic_original_glcm_MaximumProbability | 20 |
| extOralCavity_radiomic_original_glszm_ZoneEntropy | 16 |
| extOralCavity_dvh_RelativeVolume_At_RelativeDose_0.94 | 16 |
| extOralCavity_dosiomic_original_gldm_LargeDependenceLowGrayLevelEmphasis | 15 |
| PC_radiomic_original_glszm_ZoneEntropy | 15 |
| Larynx_radiomic_original_shape_Sphericity | 14 |
| extOralCavity_dvh_RelativeVolume_At_RelativeDose_0.97 | 14 |
| PC_radiomic_original_glcm_Autocorrelation | 13 |
| extOralCavity_contouromic_GTVn_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-20.00 | 13 |
| GTVp_dvh_RelativeVolume_At_RelativeDose_0.95 | 12 |
| extOralCavity_dosiomic_log-sigma-2-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis | 10 |
| GTVn_dvh_RelativeVolume_At_AbsoluteDose_72.00 | 10 |
| GTVp_radiomic_original_glszm_ZoneEntropy | 9 |
| Larynx_dosiomic_original_glszm_LargeAreaEmphasis | 9 |

**Table 36: Weighted feature counts across top 5% of models for severe OM**

| Feature | Weighted sum |
|---|---|
| clinical_Chemotherapy | 121.4 |
| clinical_Sex_Male | 91.1 |
| clinical_AgeAtRTStart | 36.0 |
| clinical_T3 | 30.4 |
| extOralCavity_radiomic_log-sigma-3-0-mm-3D_firstorder_Mean | 24.3 |
| clinical_T4 | 13.6 |
| extOralCavity_dosiomic_original_glcm_MaximumProbability | 13.0 |
| extOralCavity_radiomic_original_glszm_ZoneEntropy | 10.4 |
| extOralCavity_dvh_RelativeVolume_At_RelativeDose_0.94 | 10.2 |
| extOralCavity_dosiomic_original_gldm_LargeDependenceLowGrayLevelEmphasis | 9.6 |
| PC_radiomic_original_glszm_ZoneEntropy | 9.6 |
| Larynx_radiomic_original_shape_Sphericity | 9.1 |
| extOralCavity_dvh_RelativeVolume_At_RelativeDose_0.97 | 9.0 |
| extOralCavity_contouromic_GTVn_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-20.00 | 8.6 |
| PC_radiomic_original_glcm_Autocorrelation | 8.3 |
| GTVp_dvh_RelativeVolume_At_RelativeDose_0.95 | 7.6 |
| extOralCavity_dosiomic_log-sigma-2-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis | 6.6 |
| GTVn_dvh_RelativeVolume_At_AbsoluteDose_72.00 | 6.4 |
| Larynx_dosiomic_original_glszm_LargeAreaEmphasis | 5.8 |
| GTVp_radiomic_original_glszm_ZoneEntropy | 5.8 |

## 4.4. Discussion

Severe OM has a significant negative impact on quality life and also threatens treatment outcome by causing unplanned hospitalization or treatment interruptions. Identification of patients at high risk of severe OM is desirable to enable targeted interventions for prevention and management. This study pioneered the development of integrated multi-omic prediction

models for severe OM by identifying the best combinations of feature types and VOIs. Specifically, clinical, DVH, radiomic, dosiomic and contouromic features were investigated. To our best knowledge, this represented the first externally validated model to use these omics for severe OM prediction in NPC or other HNCs.

Prior to model development, univariate analysis of baseline characteristics was conducted in each dataset. Among the conventional clinical and mean dose DVH features, only chemotherapy status was significantly associated with severe OM in both datasets, highlighting the inadequacy of these conventional features for severe OM prediction.

Pre-treatment blood tests were investigated for their role as potential predictors of severe OM. Sparsity of pre-treatment blood tests prevented their inclusion in models and also limited the statistical power for identifying correlations with OM. Pre-treatment potassium, white blood cell count and creatinine clearance were significant in univariate analysis, however none of these factors remained statistically significant when included in multivariate analysis alongside neoadjuvant chemotherapy status. It was likely that neoadjuvant chemotherapy was a confounding factor affecting these pre-treatment blood test results. Further investigation would be required to determine the independent association of each blood test result with severe OM. The lack of available blood test data precluded the inclusion of these features in the model development. It should be noted that the sample size for all of the blood test analyses was small, and the associated statistical power was often low, therefore some true correlations between pre-treatment blood tests and severe OM may not have been detected. Neoadjuvant chemotherapy was not included as a feature in model development due to the significant differences in usage between the two institutions, preventing models from learning a generalizable association between feature and outcome.

Models were developed and evaluated for different combinations of VOIs, feature types, dimensionality reduction approaches and machine learning algorithms. For each combination, the model hyperparameters were optimized using cross-validation. It was hypothesized that comparing these models would uncover the most relevant VOI and feature type combination for severe OM prediction. The discrimination performance of the models was compared by ranking an aggregate AUC, calculated by taking the minimum of the training, internal validation and external validation score for a given model. This would ensure that only models with good performance on both datasets would be ranked highly.

Comparison of the discrimination scores reveals that all of the top 5 multi-omic models outperformed the top 5 conventional models in internal and external validation. Interestingly, clinical, radiomic, dosiomic and contouromic features were represented in these top 5 models, suggesting that each of these omics holds predictive value for severe OM.

The top-performing conventional model consisted of male sex, chemotherapy status, and DVH parameters for the GTVp and larynx. The identification of chemotherapy as a critical factor increasing the risk of severe OM aligned with prior research [298]. Male sex was also associated with higher incidence of severe OM, as identified in other literature [53, 98]. Interestingly, no DVH features for the extended oral cavity were present, unlike the mean dose feature used in the model by Otter et al. [295]. The GTVp DVH feature indicated that patients with a higher fraction of the GTVp volume with high relative dose were assigned lower predicted probabilities for severe OM. This could represent greater dose conformity within the

GTVp. Lower values of this feature could represent cases with lower conformity, where more dose was delivered to the oral mucosa tissues. The larynx DVH feature had a smaller impact on the model output, as indicated by the SHAP analysis, and its interpretation is less clear.

The top-performing multi-omic model consisted of chemotherapy status and four radiomic features from the extended oral cavity and larynx. The training AUC was significantly higher than the internal or external validation AUCs. This would typically be indicative of over-fitting. In this case, the optimal hyperparameters from the cross-validated grid search likely resulted in excess model complexity. Such discrepancies between training and validation scores were typical of the XGBoost algorithm in the results. However, it should be noted that the grid search evaluated a range model hyperparameters and model sizes ranging from 1 feature to 9 features, and the selected settings yielded the best internal validation score. The discrepancy between the training score and internal validation score should not invalidate the model, instead, clinicians should be aware that the training score is not representative of the expected performance on data from the same or different institutions. A possible solution that could be employed in future model development would be to manually restrict the range of model hyperparameters to enforce simpler models.

The feature in the multi-omic model with the greatest impact on the model output, as identified by the SHAP analysis, was the sphericity of the larynx VOI. Higher sphericity was associated with lower probability of severe OM. There are many factors which could affect this property, including patient anatomy, patient neck position, and even variability within the auto-segmentation model, and it is difficult to explain the specific connection with severe OM. The 90th percentile of the Laplacian-of-Gaussian-filtered CT intensity was the feature with the next greatest impact on the model output. The image filter is an edge-detection filter, indicating the

rate-of-change in HU values. It indicates that patients with greater rate-of-change in HU were allocated a higher probability of severe OM. The GLSZM zone entropy feature indicates that patients with greater textural heterogeneity in the CT volume had greater probability of experiencing severe OM. The GLSZM high gray level zone emphasis feature represents the number of zones with higher CT intensity in the volume. Higher values of this feature were associated with greater probability of severe OM. Together, these three features describe the pre-treatment tissue characteristics in the oral cavity. Aside from the mucosa itself, these features could be influenced by the number of teeth, presence of dental fillings, and the corresponding scattering artifacts which were visually apparent on the CT images. In this way, these features could indicate poorer oral health in correlation with severe OM. Finally, chemotherapy status was included in the model, in agreement with the conventional model.

The multi-omic model signature was not strongly correlated with the conventional model signature, and the radiomic features included in the model were not strongly correlated with any clinical or DVH features. This indicates that the multi-omic model had independent predictive value for severe OM and was not simply an alternative representation of conventional features. While the multi-omic model outperformed the conventional model in discrimination performance, the DeLong test did not find a statistically significant difference in AUC. However, due to the relatively small increase in AUC and the limited sample size for external validation, this test had low statistical power. Retrospective sample size calculation indicated that thousands of patients would be required to achieve 80% power in detecting the observed difference in AUC. Multivariate analysis of the conventional and multi-omic signatures against the severe OM label found that only the multi-omic model achieved significance in the external validation dataset, further confirming its independent predictive

value separate from the conventional model. The multi-omic model also outperformed the externally validated model by Otter et al. [295], which utilized a similar VOI for the oral cavity, in internal and external validation. Independent validation of the multi-omic signature by other studies would be desirable to confirm this finding. Particularly noteworthy is the low performance of the Otter model on the QEH dataset, where the mean dose to the extended oral cavity does not appear to be a good predictor of severe OM, unlike in other institutions. This may result from the differences in the RT modality and dose sparing guidelines.

The calibration and clinical utility of the conventional and multi-omic models were also analysed. The calibration of both models was comparable in the external validation set, though the range of predicted probabilities was quite limited. Good calibration is desirable for clinical implementation, since it provides predicted probabilities that accurately match the observed incidence. However, the discrimination performance for severe OM prediction was still relatively low, and combined with the limited sample size, this makes achieving a good calibration across the full range of predicted probabilities quite difficult. However, for future clinical implementation, models may be calibrated to specific institutions with the use of additional data. Decision curve analysis indicated the increased utility of the multi-omic model in both training and external validation datasets. The overfitting to the training data makes comparison of the net benefit in the QEH dataset less informative, however in the external validation dataset the multi-omic model consistently had higher net benefit, making it the preferable model out of the two options.

Permutation feature importance analysis revealed that all of the features in the multi-omic model had a positive contribution to the AUC in both datasets, reinforcing their relevance. Conversely the analysis for the conventional model suggested that male sex and the DVH

feature for the larynx were less relevant in the PWH dataset, resulting in improved AUC when negating the contribution of these features.

Apart from comparing the highest-scoring conventional model and multi-omic model, the features selected across the top 5% of models were also analysed. Given the large number of evaluated models, this would still capture much of the variability of feature selection, while restricting the results to the highest-scoring models. This provided a qualitative impression of the most relevant features. However, since clinical features were included in all of the evaluated feature sets, their importance may be over-estimated in this analysis. Features were ranked by their frequency in the top 5% of models, and an alternative analysis also weighted these results by the aggregate AUC of each model. The results for each approach were similar. The most relevant clinical features broadly aligned with those reported in the literature: chemotherapy status, sex, age and T-stage. The extended oral cavity was the most frequently selected VOI, as is expected for OM. Interestingly, DVH, radiomic, dosiomic, and contouromic features were represented in the top 20. When the pairwise Pearson correlations were measured for these top 20 features (see APPENDIX **Figure 41 & Figure 42**), very few features were strongly correlated ($|R| > 0.7$), indicating that these features each have independent predictive value for severe OM and that the selected radiomic, dosiomic and contouromic features were not interchangeable.

Subsequent development of models using these frequently selected features is inadvisable without collecting additional external validation data, since information leakage would result in over-inflated performance estimates. These results emphasize the variability in feature selection, and the significant impact of the choice of VOIs and feature types. However,

these features can serve as a starting point for future studies to decide on which VOIs and features to include.

Published prediction models for severe OM suffered from variable performance and a lack of external validation, with the exception of a single conventional model developed by Otter et al., [295] and validated by Sharabiani et al., [52] which consisted of only one DVH feature. The performance of this model on the QEH dataset was poor, demonstrating the difficulty of developing a generalizable model. Other prediction models for severe OM have achieved high internal validation scores but have a low level of evidence and unknown generalizability without validation in external datasets [53-57]. Nevertheless, studies conducted by Dean, Liu and Hansen incorporated additional treatment-related variables such as chemotherapy modality, specific chemotherapeutic agents, number of chemotherapy cycles, and treatment acceleration, which may have enhanced the predictive value of their models [53, 54, 56]. However, it is crucial to acknowledge that treatment protocols often differ substantially between institutions and across HNCs, potentially affecting the generalizability of these findings.

Despite the promising outcomes of this analysis, it is important to recognize several limitations. The severe OM label used in this study was confined to the first seven weeks from the onset of RT and did not encompass the entire 90-day period typically used to evaluate acute toxicity. However, as OM generally peaks during the fourth to fifth weeks of therapy, extending the observation period beyond seven weeks is unlikely to significantly affect the accuracy of the severe OM label [299]. This temporal boundary ensures that the critical peak of mucositis is captured, minimizing the impact of this limitation on the study's findings. Another limitation is the exclusion of social determinants like smoking and alcohol consumption from the

analysis, attributed to the limited availability of such in the dataset. These factors have previously been identified as predictors of OM [56, 67, 70, 77, 300]. The data imbalance observed restricted the use of these features in our model development. While the external validation approach employed in this study allowed for some assessment of model generalizability, the lack of multi-centre training meant that the generalizability of the models could not be optimized. Future studies employing multi-centre cohorts should aim to construct models that more effectively generalize across the inherent structural variations between centres, thereby enhancing predictive accuracy and clinical applicability. This approach will be vital in advancing the field and improving patient outcomes.

## 4.5. Conclusion

This study pioneered the development of integrated multi-omic prediction models for severe OM, identifying the most effective combinations of clinical, DVH, radiomic, dosiomic, and contouromic features. Multi-omic models outperformed conventional models developed on the same dataset and also outperformed the only externally validated conventional model from the literature. The limited performance of conventional models demonstrated the inadequacy of clinical and DVH features to fully capture the complex correlations with severe OM. Radiomics, by describing pre-treatment tissue characteristics, dosiomics, by describing the spatial distribution of the planned RT dose, and contouromics, by describing the challenges posed by patient geometry, were shown to achieve greater discrimination by supplementing clinical features. Importantly, this study conducted external validation to assess the generalizability of the models to another local institution, providing a greater level of evidence compared to other prediction models in the literature. To the best of our knowledge, this represented the first externally validated model for treatment-induced OM using multi-omic

features. Future studies can build on these results to further enhance the stability, generalizability and discriminative performance of prediction models, leading the way towards clinical implementation for achieving early intervention and personalized management of OM.

# CHAPTER 5 MULTI-OMIC PREDICTION MODELS FOR SEVERE ACUTE DYSPHAGIA

## 5.1. Introduction

Acute dysphagia is a common and debilitating toxicity among NPC patients undergoing RT. In addition to the major detrimental impact on quality of life, severe acute dysphagia threatens treatment outcome through weight loss and treatment interruption. Pre-treatment prediction of severe acute dysphagia offers the potential to deliver more targeted care for the prevention and management of the condition. This study sought to harness high-dimensional multi-omic data (radiomics, dosiomics and contouromics) for enhanced prediction of severe acute dysphagia. Published prediction models for dysphagia, utilizing conventional clinical and DVH features, have frequently focused on mixed cohorts of HNC patients, rather than being specific to NPC. Additionally, to the best of our knowledge, there are no full-length articles published on the prediction of severe acute dysphagia using radiomics, dosiomics or contouromics.

## 5.2. Methodology

### *5.2.1. Definition of severe acute dysphagia*

As discussed in **Section 1.1.5**, dysphagia can be graded according to different severity scales. In this retrospective study, patient-reported outcomes were not available, nor were videofluoroscopy assessments. CTCAE and RTOG are two commonly used grading scales for clinician assessment of dysphagia, focusing on the functional impact of the condition. Both scales include the indication for tube feeding as part of the criteria for severe dysphagia. While

specific gradings of dysphagia were rarely included in the clinical notes, tube feeding was well-reported. Clinicians noted when tube feeding was initially offered to the patient, based on their observations of reduced dietary intake, weight loss or other dysphagia-related symptoms. The patient's response was also noted – with many patients initially refusing tube feeding and choosing to continue with milk supplements. A minority of patients offered tube feeding did go on to have a nasogastric feeding tube fitted, and this was recorded in the clinical notes. Since the indication for tube feeding was in the criteria for severe dysphagia, rather than the delivery of tube feeding, this indication was selected as the outcome definition for severe dysphagia in this study. Specifically, the severe acute dysphagia label was defined as the indication for tube feeding during RT, as identified from the clinical consultation notes. The evidence for an indication of tube feeding was determined from statements such as "patient agreed to receive tube feeding", "patient declined tube feeding", or "feeding tube inserted on DD/MM/YYY". The specific wording varied extensively, and great care had to be taken during data collection to check for all synonyms of tube feeding, including: "R/T feeding", "Ryle's tube", "N-G tube", "nasogastric tube", "enteral feeding", "PEG insertion", and "Entriflex". To ensure that all tube feeding indication events were captured, all consultation notes, discharge summaries and nursing consultation notes were inspected from diagnosis up to the end of RT. Data post-RT was not available for the QEH dataset. While events after the end of RT were not included in the outcome label, the frequency of consultations was significantly reduced post-RT. Additionally, tube feeding was not typically offered after RT, or even towards the end of RT, because the motivation for tube feeding was primarily to mitigate the detrimental impact of

weight loss on treatment outcome from deviations to the radiation plan from changes in geometry.

### *5.2.2. Statistical analysis of baseline characteristics*

Patients were grouped by severe and non-severe acute dysphagia outcome label within each dataset. Differences between baseline characteristics in severe and non-severe acute dysphagia groups were assessed for statistical significance using Fisher's Exact test for categorical features, and Mann Whitney U test for continuous features. The non-parametric Mann Whitney U test was selected because a significant proportion of the features were not normally distributed, with a Shapiro test p-value $< 0.05$. The univariate analysis was conducted separately from model development, and served to verify whether there were any clinical or DVH features significantly correlated with severe acute dysphagia in both datasets.

### *5.2.3. Model development*

Predictive models for severe acute dysphagia were developed from two feature sets: 1) conventional clinical and DVH features only, and 2) multi-omic features including clinical, DVH, radiomic, dosiomic and contouromic features.

The model development process outlined in CHAPTER 3 was conducted for different combinations of VOIs and feature types. The discrimination performance in AUC was compared for each feature set and model type. The VOI and feature type combination that gave the highest discrimination performance across training, internal and external validation was selected for further analysis. Conventional models and multi-omic models were developed using this set of VOIs and compared.

In this chapter, several VOIs were explored to determine the optimal combination for severe acute dysphagia prediction: GTVp, GTVn, extended oral cavity, PC muscles, parotid glands, and larynx. It was hypothesized that some of these VOIs could further enhance the model discrimination performance. Particularly, DVH features from PC muscles, oral cavity, parotids, and larynx had been reported as predictors in the literature [1]. The PC muscles are directly involved in the swallowing mechanism, so radiation damage could directly impact dysphagia. Radiation dose to the oral cavity has been associated with OM, as in the previous chapter, and the resulting pain can impact dysphagia through its effect on swallowing and oral intake. The parotid glands are the largest of the salivary glands and radiation damage to this OAR could result in xerostomia. A lack of saliva, or changes in its consistency, could contribute to difficulty swallowing. Additionally, the epiglottis, situated at the top of the larynx, moves downward during swallowing to seal off the entrance to the larynx, thereby preventing food or liquid from entering the airway. The inclusion of the larynx VOI may therefore be relevant to prediction of severe dysphagia. The proximity of the GTVp and GTVn to key swallowing anatomy justifies their inclusion in models.

Data preprocessing, model development and performance evaluation was conducted as described in **CHAPTER 3**. Specifically, the model development process was repeated for different combinations of VOI and different combinations of feature types as described in **Section 4.2.4**. The optimization of each model pipeline was performed using a cross-validated grid search across the range of hyperparameters defined in **Section 3.8.11**. For the MRMR feature selection algorithm, the maximum number of features to select, $k_{max}$, was set to 11, since this would correspond to an event-per-variable rate of approximately 10, in line with the rule

of thumb [131]. Within the grid search optimization, the number of features to select was varied from 1 to $k_{max}$. Validation and analysis of the resulting models was conducted in a similar manner to that described in **Section 3.8.11**.

### *5.2.4. Experiment: removing perturbation stability filter*

To explore whether the perturbation stability filter was overly stringent and removing relevant features, model development was repeated with the stability filter removed. As discussed in **Section 3.7**, the settings used for perturbation were not known to be optimal for every VOI. The proportion of stable features varied across VOIs and across feature types (radiomic, dosiomic, contouromic). This suggests that some component of feature stability is dependent on the shape, intensity, and texture of the VOI rather than inherent to the calculation of the feature itself. For example, the extended oral cavity and GTVp had relatively high stability compared to the GTVn and PC muscles. In terms of the geometry, the extended oral cavity and GTVp are larger and more spherical, with a lower surface-to-volume ratio. The GTVn typically consisted of multiple small sub-volumes, and the PC VOI was a longer, flatter volume. Certain features may be inherently more sensitive to VOIs with smaller volumes or larger surface-to-volume ratios. Additionally, the choice of perturbation parameters may have a greater impact on such volumes. Whether this is commensurate with the impact of inter-observer variation from real experts' segmentations remains to be seen. However, this section addresses the hypothesis that the feature stability filter based on perturbation ICC may be too punishing, particularly for VOIs such as the GTVn and PC muscles. By removing the filter, the predictive potential of these VOIs might be observed more clearly. To explore this hypothesis, the model development process outlined in the previous **Section 5.2.3** was repeated with the

ICC stability threshold set to 0. That is, no features were excluded on the basis of apparent stability. The best performing conventional and multi-omic models were identified and compared to those obtained under the previous model development process.

## 5.3. Results

### *5.3.1. Baseline demographic and clinical characteristics*

Severe acute dysphagia occurred in 122 (34%) of patients in the QEH dataset and in 53 (52%) of patients in the PWH dataset. Comparison of clinical and mean dose DVH features against severe acute dysphagia in each dataset is shown in **Table 37**, along with univariate analysis. Effect size was calculated from $\frac{1}{1.81}\ln(\text{odds ratio})$ for categorical features [297], and from Cohen's d for continuous features. P values were calculated from Fisher's Exact test for categorical features, and from the Mann-Whitney U test for continuous features. Body weight at CT simulation, male sex, and mean dose to the neck nodal GTV were the only significant features in the development dataset. None of these features were significant in the external validation dataset.

**Table 37: Univariate analysis of clinical and mean dose DVH features against severe acute dysphagia. Incidence is shown for binary features, and median value is shown for categorical features.**

| | Development (QEH) | | | | External validation (PWH) | | | |
|---|---|---|---|---|---|---|---|---|
| | **Severe acute dysphagia** | **No severe acute dysphagia** | **Effect size** | **P value** | **Severe acute dysphagia** | **No severe acute dysphagia** | **Effect size** | **P value** |
| Age at RT start | 53 | 55 | -0.05 | 0.617 | 53 | 61 | -0.44 | 0.050 |
| BMI at CT simulation | 23.8 | 23.5 | 0.13 | 0.132 | 24.2 | 23.8 | 0.03 | 0.629 |
| Weight at CT simulation (kg) | 64.4 | 61.8 | 0.21 | **0.037*** | 67.0 | 66.3 | 0.14 | 0.426 |
| Chemotherapy | 109 (89%) | 199 (83%) | 0.19 | 0.120 | 52 (98%) | 34 (71%) | 0.82 | **<0.001*** |
| N stage = 2 | 93 (76%) | 174 (72%) | 0.09 | 0.451 | 23 (43%) | 12 (25%) | 0.39 | 0.062 |
| Male sex | 100 (82%) | 168 (70%) | 0.28 | **0.012*** | 41 (77%) | 38 (79%) | -0.04 | 1.000 |
| T stage = 3 | 76 (62%) | 166 (69%) | -0.14 | 0.239 | 30 (57%) | 13 (27%) | 0.61 | **0.005*** |
| T stage = 4 | 29 (24%) | 44 (18%) | 0.14 | 0.216 | 6 (11%) | 11 (23%) | -0.31 | 0.182 |
| GTVn $D_{mean}$ (Gy) | 72.3 | 72.1 | 0.30 | **0.001*** | 72.3 | 72.4 | -0.19 | 0.152 |
| GTVp $D_{mean}$ (Gy) | 73.7 | 73.4 | 0.22 | 0.106 | 72.4 | 72.4 | -0.23 | 0.731 |
| Larynx $D_{mean}$ (Gy) | 47.1 | 47.2 | 0.16 | 0.559 | 43.5 | 42.9 | 0.08 | 0.783 |
| PC $D_{mean}$ (Gy) | 56.8 | 56.6 | 0.11 | 0.981 | 60.6 | 60.3 | 0.04 | 0.525 |
| Parotids $D_{mean}$(Gy) | 40.8 | 41.2 | -0.04 | 0.603 | 38.7 | 37.1 | 0.18 | 0.353 |
| extOralCavity $D_{mean}$ (Gy) | 51.3 | 51.7 | 0.14 | 0.724 | 48.8 | 48.1 | 0.33 | 0.134 |

### *5.3.2. Correlations with pre-treatment blood tests*

Pre-treatment blood test results were analysed in a similar manner to **Section 4.2.3**. The Mann-Whitney U test was used to check for statistically significant differences in the median pre-treatment value of each blood test result against severe dysphagia incidence. The results are shown in **Table 38.** Only blood test results collected before the start of RT were included. No significant correlations were observed between blood test results and dysphagia.

**Table 38: Correlations between pre-treatment blood tests and severe dysphagia**

| Blood test result | Severe dysphagia | No severe dysphagia | MWU p | Number of cases |
|---|---|---|---|---|
| Pulse pressure | 51.9 | 53.9 | 0.848 | 38 |
| Mean arterial pressure | 103.8 | 99.8 | 0.437 | 38 |
| Rate pressure product | 11002.8 | 11088.7 | 0.908 | 35 |
| Systolic blood pressure | 138.5 | 135.9 | 0.567 | 38 |
| Diastolic blood pressure | 86.4 | 81.8 | 0.547 | 38 |
| Pulse | 80.2 | 80.7 | 0.868 | 35 |
| Sodium (Na) | 139.4 | 137.7 | 0.703 | 43 |
| Potassium (K) | 4.1 | 4.2 | 0.436 | 44 |
| Urea | 5.7 | 5.3 | 0.574 | 41 |
| Creatinine (Cr) | 79.6 | 82.4 | 0.639 | 54 |
| White cell count (WCC) | 5.7 | 5.3 | 0.438 | 53 |
| Platelet (plt) | 287.7 | 296.7 | 0.560 | 54 |
| Creatinine clearance (CrCl) | 87.9 | 81.5 | 0.357 | 48 |
| Haemoglobin (Hb) | 12.5 | 13.0 | 0.357 | 55 |

### *5.3.3. Conventional and multi-omic prediction models*

**Table 39** lists the top 5 conventional prediction models for severe acute dysphagia, consisting only of clinical and DVH features. The ranking was determined by the minimum of the training, internal validation and external validation scores, to ensure that the most internally valid and most generalizable models were selected.

**Table 39: Top 5 conventional prediction models for severe acute dysphagia**

| Rank | Dimensionality reduction | VOIs | Algorithm | Model size | Train AUC | Int. AUC | Ext. AUC |
|---|---|---|---|---|---|---|---|
| 1 | VIF | Larynx | SVM RBF | 6 | 0.733 | 0.598 | 0.597 |
| 2 | VIF | Ext. oral cavity, PC | Gaussian Naive Bayes | 8 | 0.617 | 0.599 | 0.593 |
| 3 | VIF | GTVn, Larynx | SVM RBF | 9 | 0.897 | 0.638 | 0.592 |
| 4 | VIF | NA | SVM Linear | 5 | 0.594 | 0.588 | 0.610 |
| 5 | Hierarchical | Larynx | SVM RBF | 8 | 0.742 | 0.587 | 0.624 |

**Table 40** lists the top 5 multi-omic prediction models for severe acute dysphagia, with the same ranking mechanism as the previous table. Notably, models containing radiomic, dosiomic and contouromic features were all present in the top 5 models. All of the top 5 models

outperformed the best conventional model in internal and external validation. The GTVn was the most frequently selected VOI in the top-performing models.

**Table 40: Top 5 multi-omic prediction models for severe acute dysphagia**

| Rank | Dimensionality reduction | VOIs | CLI | DVH | RAD | DOS | CON | Algorithm | Model size | Train AUC | Int. AUC | Ext. AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Hierarchical | GTVn, Parotids | | | ✓ | ✓ | | Random Forest | 4 | 0.970 | 0.633 | 0.625 |
| 2 | Hierarchical | GTVn | | | ✓ | | | Random Forest | 2 | 0.644 | 0.621 | 0.614 |
| 3 | VIF | Ext. oral cavity | ✓ | | | | ✓ | SVM RBF | 3 | 0.661 | 0.613 | 0.649 |
| 4 | Hierarchical | GTVn | | | ✓ | | | Ridge | 2 | 0.612 | 0.637 | 0.613 |
| 5 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Gaussian Naive Bayes | 5 | 0.622 | 0.612 | 0.648 |

## *Best-performing conventional prediction model for severe acute dysphagia*

The best-performing conventional model for severe acute dysphagia, using only clinical and DVH features, was a SVM model with RBF kernel, developed using the VIF dimensionality reduction approach. The starting feature set included clinical features and DVH features from the larynx VOI. The model achieved a training AUC of 0.733 (95% CI: 0.682, 0.789), internal validation AUC of 0.585 (95% CI: 0.546, 0.650), and external validation AUC of 0.597 (95% CI: 0.482, 0.705). It consisted of 6 features: chemotherapy status, T-stage = 3, T-stage = 4, N-stage = 2, male sex, and the fractional volume of the larynx receiving 63Gy or higher. **Figure 28** shows the ROC curves for the conventional model along with the impact of each feature on the model output, calculated using SHAP analysis.

**Table 41: Conventional model for severe acute dysphagia**

| | |
|---|---|
| **Initial feature set** | Clinical, DVH |
| **VOIs** | Larynx |
| **N features after ICC filter and VIF clustering** | 9 |
| **MRMR K** | 6 |
| **Model** | SVM with RBF kernel<br>No class weighting<br>C = 1000 |

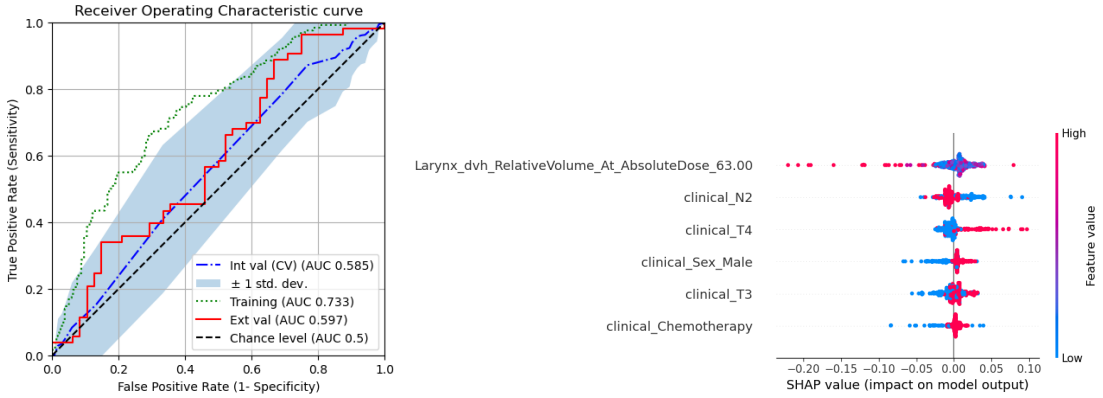



**Figure 28: ROC curve and SHAP feature analysis for conventional model for severe acute dysphagia**

## *Best-performing multi-omic model for severe acute dysphagia*

The best-performing multi-omic model was a Random Forest model developed using the hierarchical clustering dimensionality reduction approach. The initial feature set consisted of clinical, DVH, radiomic and dosiomic features from the GTVn and parotid glands VOIs. The model achieved a training AUC of 0.970 (95% CI: 0.954, 0.982), internal validation AUC of 0.633 (95% CI: 0.577, 0.688), and external validation AUC of 0.625 (95% CI:0.512, 0.725). The final model consisted of 4 features: 2 radiomic features and 2 dosiomic features. **Figure 29** shows the ROC curves for the conventional model along with the impact of each feature on the model output, calculated using SHAP analysis.

**Table 42: Multi-omic model for severe acute dysphagia**

| | |
|---|---|
| **Initial feature set** | Clinical (N = 8)<br>DVH (N = 126)<br>Radiomic (N = 784)<br>Dosiomic (N = 712) |
| **VOIs** | GTVn<br>Parotids |
| **N features after ICC filter and hierarchical clustering** | 460 |
| **MRMR K** | 4 |
| **Model** | Random Forest Classifier<br>balanced class weights<br>max depth = 6<br>n_estimators = 50 |

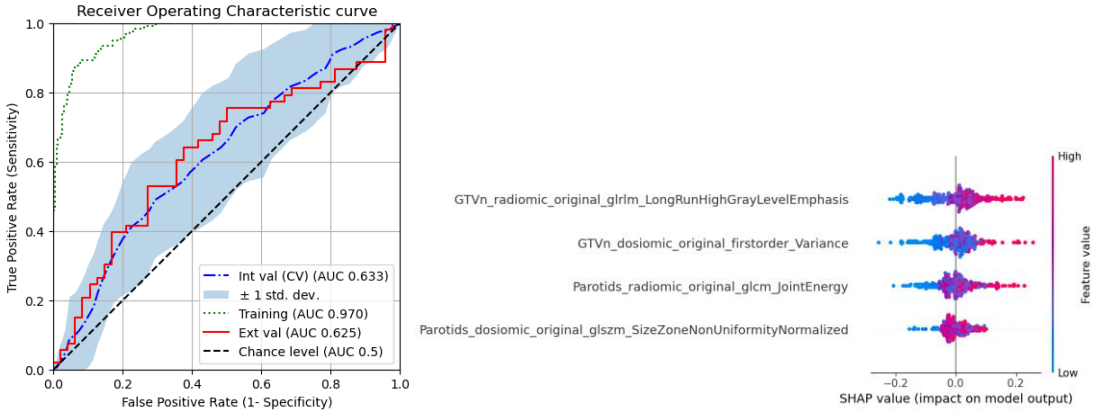



**Figure 29: ROC curve and SHAP feature analysis for multi-omic model for severe acute dysphagia**

## *5.3.4. Model comparisons*

*Feature correlations and model signature correlations*

The Pearson correlation coefficient between the conventional model signature and the multi-omic model signature was -0.07, indicating that the two signatures were not correlated. Additionally, none of the radiomic or dosiomic features in the multi-omic model were highly

correlated (Pearson |R| > 0.7) with any of the features in the conventional model, in either dataset.

## *Comparison of discrimination performance*

**Table 30** shows the discrimination performance, as measured by the AUC, for training, internal validation, and external validation. The mean AUC and its 95% confidence interval are shown. This information is also visualized in **Figure 30**. The internal and external validation scores for the multi-omic model are greater than those for the conventional model, however this difference was not statistically significant according to the DeLong test. As evidenced by the 95% confidence intervals, both models significantly outperformed random chance (AUC=0.5) in internal validation ($p < 0.05$). However, only the multi-omic model significantly outperformed random chance in external validation ($p < 0.05$). The training scores for both models were significantly higher than the validation scores.

**Table 43: Comparison of discrimination performance across models for severe acute dysphagia.**

| | Training (refit) AUC | Internal validation AUC | External validation AUC |
|---|---|---|---|
| | Mean and 95% CI (1000 bootstraps) | Mean and 95% CI (cross validation) | Mean and 95% CI (1000 bootstraps) |
| Conventional model | 0.733 (0.682, 0.789) | 0.585 (0.546, 0.650) | 0.597 (0.482, 0.705) |
| Multi-omic model | 0.970 (0.954, 0.982) | 0.633 (0.577, 0.688) | 0.625 (0.512, 0.725) |

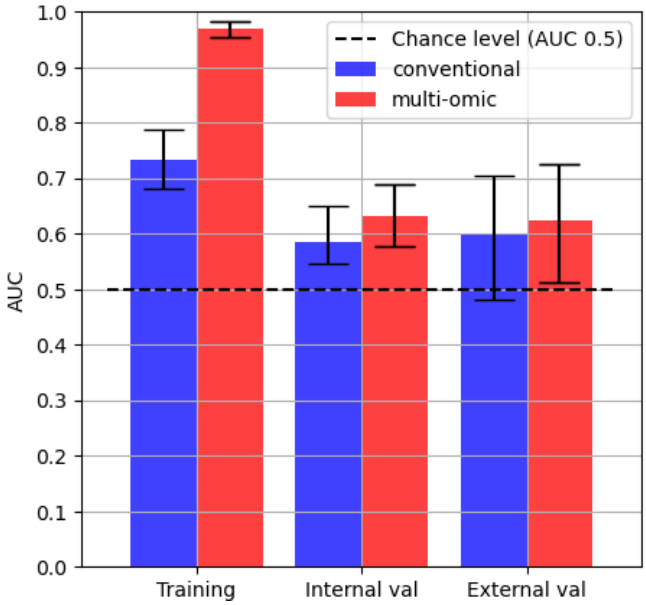



**Figure 30: Comparison of discrimination performance for severe acute dysphagia models**

**Table 44** shows the results of a multivariate logistic regression analysis of the conventional and multi-omic model signatures against the severe acute dysphagia outcome label. Both models were significantly associated with severe acute dysphagia in both datasets, indicating that the multi-omic model had independent predictive value separate from the conventional model.

**Table 44: Multivariate logistic regression of model signatures**

| Dataset | Variable | P value |
|---|---|---|
| Development (QEH) | Multi-omic model | <0.001* |
| | Conventional model | <0.001* |
| External validation (PWH) | Multi-omic model | 0.040* |
| | Conventional model | 0.034* |

## *Statistical significance of the results*

As in the chapter on OM, the DeLong test was used to calculate the statistical significance of the differences between the conventional and multi-omic models in the training and external validation datasets. The results are shown in Table 45. Only the difference in

training score was statistically significant at the level of 0.05. The improvement in external validation would likely require a much larger sample size to reach statistical significance.

**Table 45: DeLong test p-values for top dysphagia models**

| Dataset | P-value |
|---|---|
| Development (QEH) | 0.0000* |
| External validation (PWH) | 0.8783 |

An alternative approach to investigating the statistical significance of the results was performed by taking bootstrapped samples from the training and external validation sets, then calculating the performance of each model on the bootstraps. The difference between the performance of each model was recorded and the distribution over 1000 bootstraps was plotted, as shown in Figure 31. This approach allowed a confidence interval in the improvement in the AUC from the multi-omic model to be calculated. These confidence intervals and resulting p-values are shown in Table 46. This analysis showed a significant improvement in the training score, but not in the external validation score, although the multi-omic model had a net improvement in the performance over the conventional model across bootstraps.

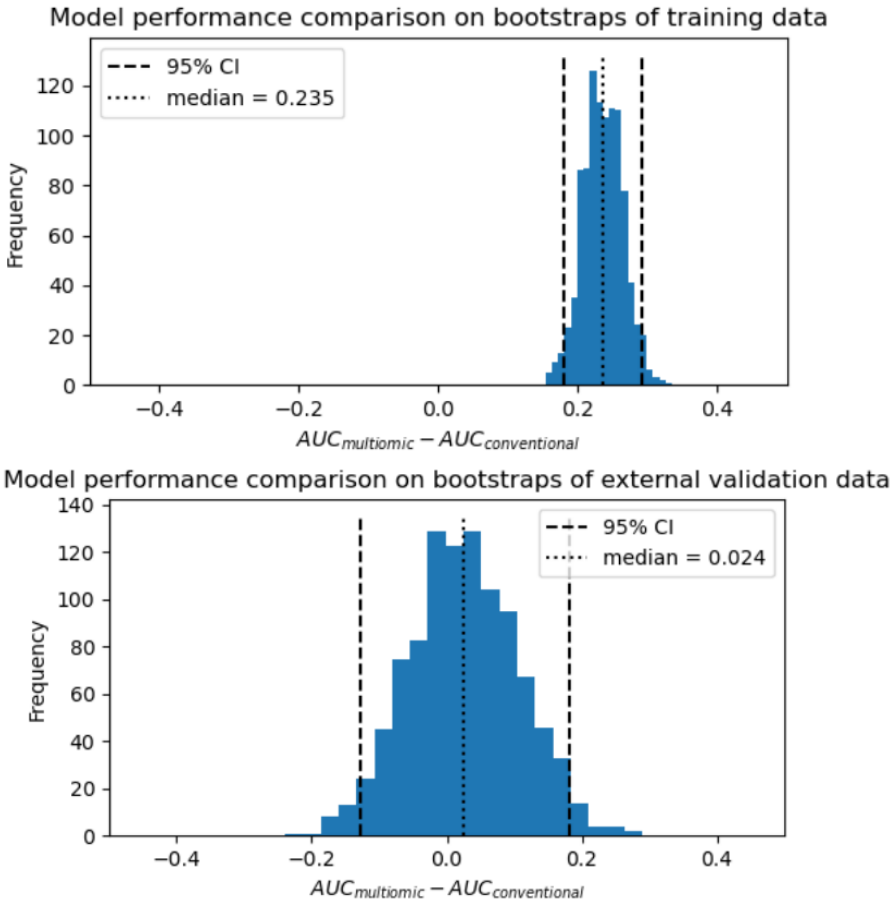



**Figure 31: Distribution of performance improvement across bootstraps for dysphagia**

**Table 46: Performance improvement across bootstraps for dysphagia top models**

| Dataset | 95% CI | P-value |
|---|---|---|
| Development (QEH) | (0.179, 0.293) | 0.000* |
| External validation (PWH) | (-0.128,0.183) | 0.390 |

## *Calibration*

**Figure 32** shows the calibration curves for the conventional model and for the multi-omic model. Calibration on the training dataset was better than on the external validation dataset. The Brier scores for the multi-omic model and conventional model were comparable for the external validation dataset.

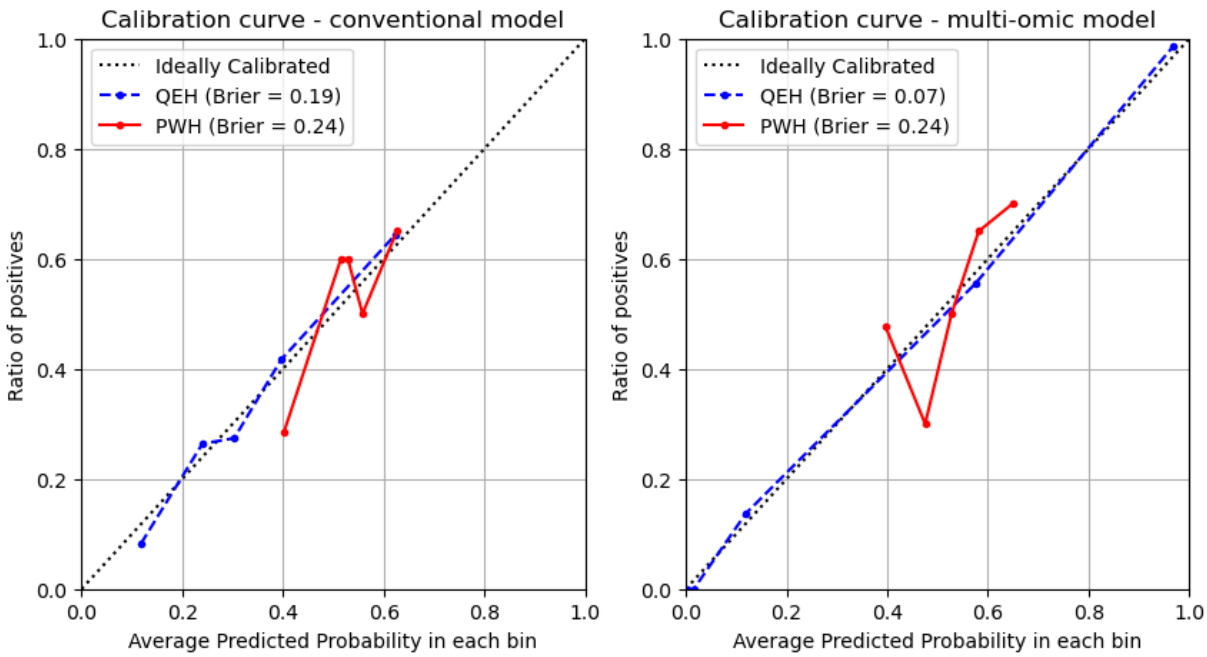



**Figure 32: Calibration curves for models for severe acute dysphagia**

## *Decision curve analyses*

**Figure 33** shows the decision curves for the multi-omic model and the conventional model in both QEH (development) and PWH (external validation) datasets. The multi-omic model demonstrated a greater net benefit over the conventional model in the development dataset, however the net benefit was similar for both models in the external validation dataset. The net benefit in the external validation dataset was greater for the conventional model in the range of threshold probabilities between 0.3 and 0.5 and was greater for the multi-omic model in the range 0.5-0.65. The incidence of dysphagia in the external validation set was 0.52, therefore the multi-omic model had superior clinical utility at this incidence level.

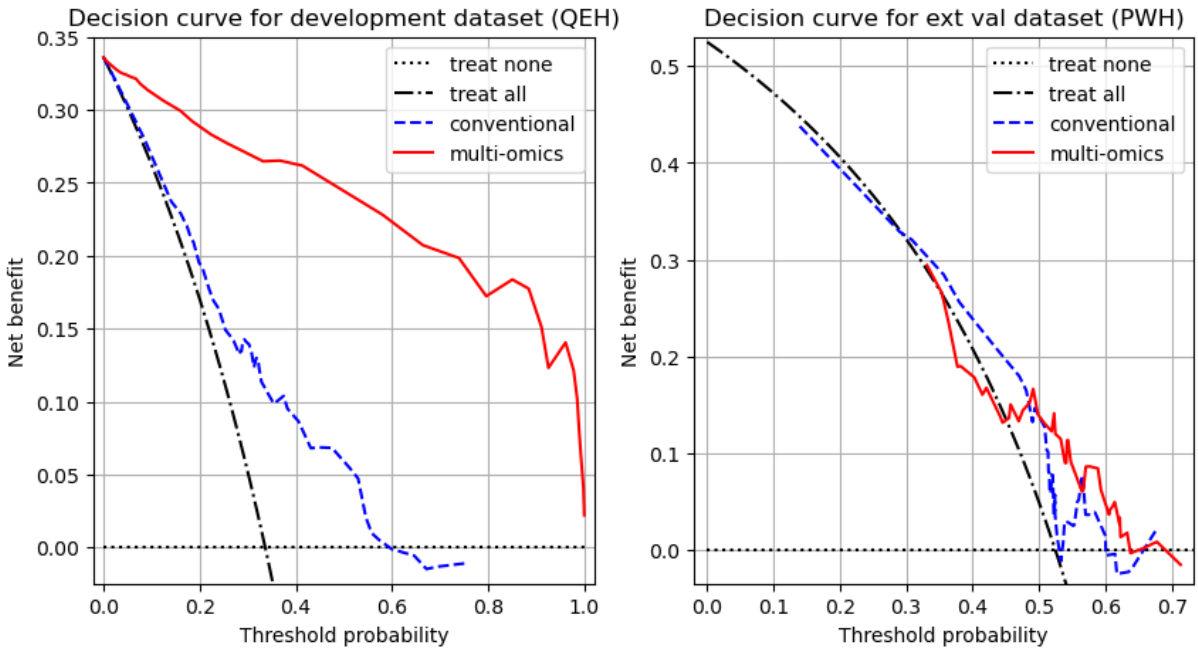



**Figure 33: Decision curves for the training dataset (left) and external validation dataset (right)**

## *Permutation feature importance*

**Figure 34** shows the SHAP feature importance for each model on the training dataset. The feature with greatest impact on the multi-omic model was the radiomic GLSZM zone entropy for the pharyngeal constrictor muscle. The plot indicates that higher heterogeneity in the texture in this volume resulted in a higher predicted probability for severe OM.

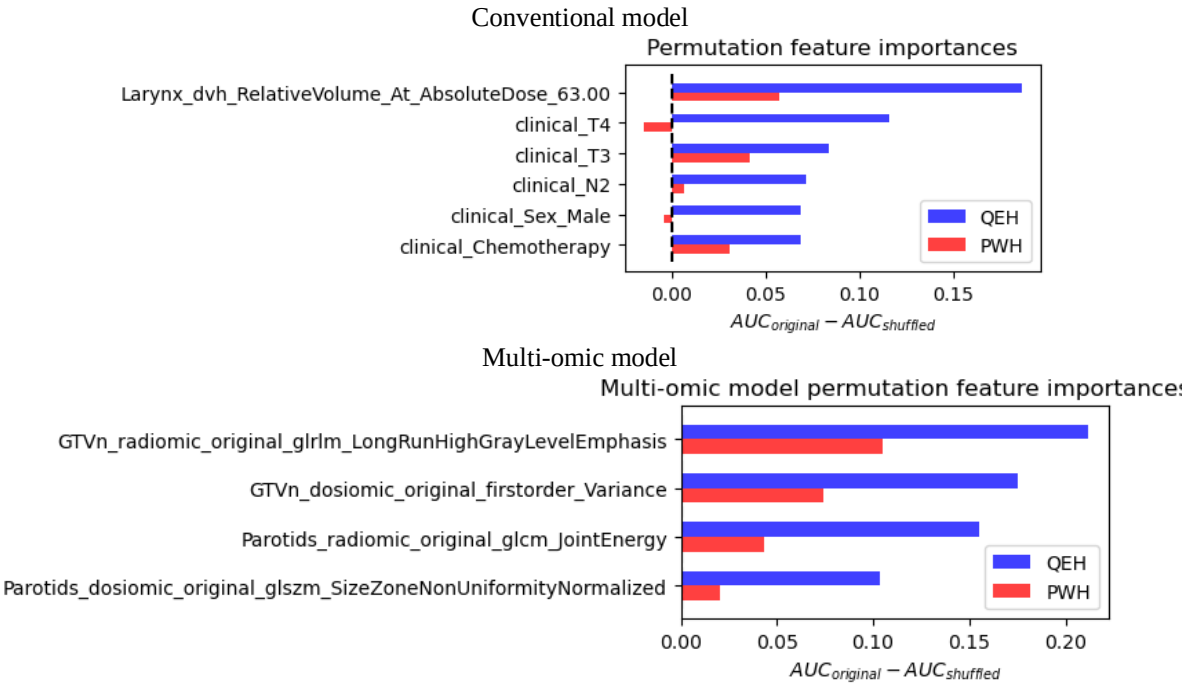



**Figure 34: Permutation feature importance for the multi-omic model**

### *5.3.5. Frequently selected features in top 5% of models*

The conventional and multi-omic models with the highest discrimination performance were identified in the previous section. Some models with different VOI, feature type, dimensionality reduction approach and algorithm combinations achieved reasonable internal and external validation performance but selected different features. To identify trends in the types of features selected, an analysis of the model features across the top 5% of developed models was conducted. The top 5% was chosen in order to focus only on the models with the highest discrimination performance. This analysis was conducted in two ways: firstly, counting the number of models where each feature was selected (**Table 47**), and secondly, by weighting each feature by the aggregate AUC (minimum of training, internal, external validation) of its corresponding model and finding the sum for each feature (**Table 48**). The Pearson correlation

matrices for the frequently selected features are shown in APPENDIX (**Figure 43 & Figure 44**). These matrices show that the most frequently selected features were not highly correlated, indicating that the selected radiomic, dosiomic and contouromic features held independent predictive value and were not interchangeable.

**Table 47: Feature counts across top 5% of models for severe acute dysphagia**

| Feature | Number of models |
|---|---|
| clinical_Sex_Male | 129 |
| clinical_Chemotherapy | 120 |
| GTVn_radiomic_original_glrlm_LongRunHighGrayLevelEmphasis | 77 |
| clinical_T4 | 52 |
| clinical_T3 | 52 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-150.00 | 35 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_0_AbsoluteDegree_-100.00 | 34 |
| clinical_N2 | 33 |
| clinical_AgeAtRTStart | 27 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-170.00 | 16 |
| Parotids_radiomic_original_glszm_LargeAreaLowGrayLevelEmphasis | 16 |
| GTVn_radiomic_log-sigma-3-0-mm-3D_firstorder_Skewness | 14 |
| Parotids_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_30.00 | 8 |
| Parotids_dosiomic_original_glszm_SizeZoneNonUniformityNormalized | 6 |
| PC_dvh_RelativeVolume_At_RelativeDose_0.97 | 5 |
| GTVp_radiomic_log-sigma-2-0-mm-3D_gldm_LargeDependenceEmphasis | 5 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-20.00 | 5 |
| GTVp_radiomic_original_glszm_SizeZoneNonUniformity | 5 |
| extOralCavity_radiomic_log-sigma-3-0-mm-3D_firstorder_Maximum | 4 |
| extOralCavity_dvh_RelativeVolume_At_AbsoluteDose_69.00 | 4 |

**Table 48: Weighted feature counts across top 5% of models for severe acute dysphagia**

| Feature | Weighted sum |
|---|---|
| clinical_Sex_Male | 77.5 |
| clinical_Chemotherapy | 72.1 |
| GTVn_radiomic_original_glrlm_LongRunHighGrayLevelEmphasis | 46.6 |
| clinical_T3 | 31.3 |
| clinical_T4 | 31.2 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-150.00 | 21.1 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_0_AbsoluteDegree_-100.00 | 20.5 |
| clinical_N2 | 19.9 |
| clinical_AgeAtRTStart | 16.3 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-170.00 | 9.6 |
| Parotids_radiomic_original_glszm_LargeAreaLowGrayLevelEmphasis | 9.6 |
| GTVn_radiomic_log-sigma-3-0-mm-3D_firstorder_Skewness | 8.5 |
| Parotids_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_30.00 | 4.8 |
| Parotids_dosiomic_original_glszm_SizeZoneNonUniformityNormalized | 3.6 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-20.00 | 3.0 |
| GTVp_radiomic_original_glszm_SizeZoneNonUniformity | 3.0 |
| GTVp_radiomic_log-sigma-2-0-mm-3D_gldm_LargeDependenceEmphasis | 3.0 |
| PC_dvh_RelativeVolume_At_RelativeDose_0.97 | 3.0 |
| GTVn_dosiomic_log-sigma-1-0-mm-3D_glcm_Idmn | 2.4 |
| GTVn_dosiomic_original_glcm_MCC | 2.4 |

### *5.3.6. Experiment: removing perturbation stability filter*

After removing the perturbation stability filter and evaluating the performance of models with different combinations of VOIs and feature types, the top-performing multi-omic model was identified. The best-performing multi-omic model was a SVM model with RBF kernel consisting of 5 features, including 4 radiomic features and 1 dosiomic feature. The selected VOIs were the GTVn and parotid glands. The model achieved a training AUC of 0.915 (0.879, 0.945), internal validation of 0.658 (0.586, 0.730) and external validation of 0.639 (0.209, 0.355). The receiver operating characteristic curve and SHAP feature analysis for this model are shown in . The features with ICC < 0.7 were: GTVn_radiomic_log-sigma-2-0-mm-3D_glcm_Autocorrelation (ICC = 0.61), Parotids_radiomic_log-sigma-2-0-mm-

3D_firstorder_Range (ICC = 0.21), Parotids_radiomic_log-sigma-1-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis (ICC = 0.11).

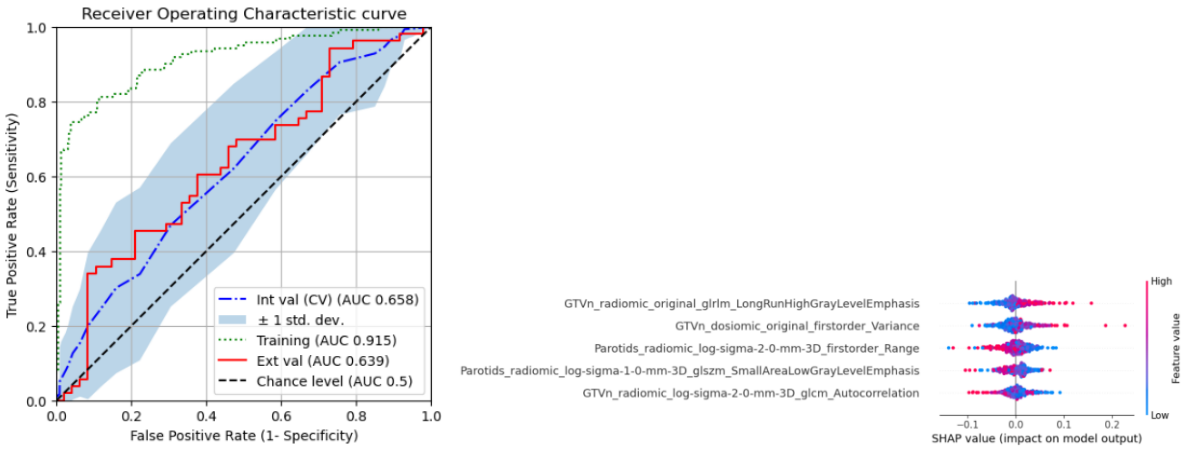



**Figure 35: ROC curve and SHAP feature analysis for best-performing multi-omic model after removing perturbation stability filter**

## 5.4. Discussion

Severe acute dysphagia has a devastating impact on patients' quality of life and threatens treatment outcome. Without urgent intervention, it risks significant weight loss and treatment interruptions. Early identification of patients at risk of severe acute dysphagia is crucial for targeted preventative treatment and management. This study pioneered the development of integrated multi-omic prediction models for severe acute dysphagia by identifying the best combinations of feature types and VOIs. Specifically, clinical, DVH, radiomic, dosiomic and contouromic features were investigated. To our best knowledge, this represented the first model to use these omics for the prediction of severe acute dysphagia in NPC or other HNCs.

Prior to model development, univariate analysis of baseline characteristics was conducted in each dataset. No clinical or mean dose DVH features were significantly correlated with severe acute dysphagia in both datasets. This contrasts with the findings from the literature, where associations with concurrent chemotherapy, T-stage, N-stage, age, sex, BMI and RT dose to OARs such as the PC muscles have been reported [1]. The statistical power for each univariate test varied depending on the sample size of each dataset, the incidence of severe dysphagia in each dataset and the effect size for correlations with each feature, therefore the lack of significance for some of these features may have been due to small effect size and insufficient power. However, comparison of the magnitude and direction of effect sizes does not suggest any common predictors of severe dysphagia in both datasets either. This suggests that the traditional clinical and mean dose DVH features may be of limited value for the prediction of severe acute dysphagia in NPC.

Correlations between pre-treatment blood tests and severe acute dysphagia were investigated as was conducted for severe OM in the previous chapter. No significant correlations were identified. While the statistical power of these tests was limited by the small number of available samples, these factors were also absent from the set of predictors of severe acute dysphagia listed in the literature. As in the previous chapter, these factors were not included in the model development process due to substantial missing data.

Models were developed and evaluated for different combinations of VOIs, feature types, dimensionality reduction approaches and machine learning algorithms. For each combination, the model hyperparameters were optimized using cross-validation. It was hypothesized that comparing these models would uncover the most relevant VOI and feature

type combination for severe acute dysphagia prediction. The discrimination performance of the models was compared by ranking an aggregate AUC, calculated by taking the minimum of the training, internal validation and external validation score for a given model. This would ensure that only models with good performance on both datasets would be ranked highly.

Comparison of the discrimination scores reveals that all of the top 5 multi-omic models outperformed the top 5 conventional models in internal and external validation. Interestingly, clinical, radiomic, dosiomic and contouromic features were represented in these top 5 models, suggesting that each of these omics holds predictive value for severe acute dysphagia. The GTVn was frequently included as a VOI in the top 5 multi-omic models.

The top-performing conventional model consisted of 1 DVH feature and 5 clinical features. The internal and external validation AUCs were comparable and were significantly lower than the training AUC. This discrepancy likely resulted from the complexity of the optimal model from the cross-validated grid search. This was a SVM model with RBF kernel, which transformed the feature space into a higher-dimensional space, allowing for a more complex decision boundary that was able to fit to the training data model closely. However, these model hyperparameters did result in the best internal validation score. Simpler models with reduced number of features and stronger regularization did not improve the internal validation score. In terms of the model features, higher fractional volume of the larynx receiving 63Gy or higher was associated with lower predicted probabilities for severe acute dysphagia. Aside from this DVH feature, male sex, chemotherapy and higher T-stage were associated with higher predicted probabilities for severe acute dysphagia. The association with male sex agrees with the study by Dean et al. [131], but not with the study by Willemsen et al.

which found a correlation in the opposite direction. The correlation with chemotherapy agrees with that reported by Dean et al., while the association with higher T-stage agrees with that reported by Willemsen et al. The model also included an N-stage feature, highlighting the role of the nodal spread.

The top-performing multi-omic model consisted of 2 radiomic features and 2 dosiomic features, extracted from the GTVn and parotid glands. No clinical or DVH features were selected, despite these features being included in the initial feature set for the best performing model. This reinforces the findings from the analysis of baseline characteristics, suggesting the limited predictive value of the conventional features. Regarding discrimination performance, the training AUC was significantly higher than the internal or external validation AUCs. This would typically be indicative of over-fitting. In this case, the optimal hyperparameters for the Random Forest model selected in the cross-validated grid search included a maximum tree depth of 6, resulting in a complex model that fit too closely to the training data. Nevertheless, these model hyperparameters did yield the best internal validation performance and were selected accordingly. As was the case in the chapter on severe OM, the discrepancy between the training score and internal validation score should not invalidate the model, instead, clinicians should be aware that the training score is not representative of the expected performance on data from the same or different institutions. A possible solution that could be employed in future model development would be to manually restrict the range of model hyperparameters to enforce simpler models.

SHAP analysis indicated the impact of each of the features on the output of the multi-omic model. Higher values for the radiomic long run high gray level emphasis within the GTVn

was associated with higher predicted probabilities for severe acute dysphagia. This feature describes the size of regions of voxels with high CT values within the GTVn. This could potentially describe some tumour properties which are indirectly associated with greater toxicity through their impact on the dose received by organs-at-risk (OARs). The other radiomic feature in the model was the GLCM joint energy for the parotid glands VOI. This feature, associated with higher homogeneity in the radiodensity within the parotid glands, resulted in higher predicted probabilities of severe acute dysphagia. This could associate pre-treatment tissue characteristics of the parotid glands with greater susceptibility to radiation damage, resulting in worsened dysphagia from the impact on saliva production. Aside from these two radiomic features, the dosiomic features describing the variance in the dose within the GTVn and the normalized GLSZM size zone non uniformity for the dose within the parotid glands were each associated with higher predicted probabilities for severe acute dysphagia. These features describe the spread and heterogeneity in the dose within these OARs, suggesting that greater dose uniformity, and possibly greater dose conformity to the GTVs, is associated with reduced risk of swallowing-related toxicity.

The multi-omic model developed in this study outperformed the conventional model in both internal and external validation. While the improvement in AUC was not significant according to the DeLong test, both model signatures were significantly correlated with the outcome in each dataset in multivariate analysis, demonstrating their independent predictive value. The 95% confidence intervals on the discrimination performance indicated that only the multi-omic model significantly outperformed chance level, providing further evidence for the improved performance of this model. Furthermore, the radiomic and dosiomic features

included in the multi-omic model were not correlated with any clinical or DVH parameters, and the two model signatures were not strongly correlated, demonstrating the independent predictive value of the multi-omic model.

The calibration and clinical utility of the conventional and multi-omic models were also analysed. Calibration was comparable for both models. Decision curve analysis demonstrated superior clinical utility of the multi-omic model in both datasets. These findings are encouraging for the further exploration of radiomics and dosiomics in connection with severe acute dysphagia prediction, providing a non-invasive way to better identify patients at risk of severe acute dysphagia before treatment.

Comparison of the discrimination performance of the multi-omic model with the externally validated predictive models in the literature reveals that one model by Willemsen et al. achieved comparable performance in external validation (AUC = 0.624) [128], while another by Dean et al. greatly outperformed the multi-omic model (AUC = 0.82) [131]. Both of the literature models significantly outperformed the conventional model, despite also using clinical and DVH features. However, there are some important differences between the studies to consider.

Firstly, both of the highlighted models from the literature were developed using mixed cohorts of head and neck cancers, which included several different disease subsites. This diversity was incorporated into the models as a tumour site feature, potentially enhancing discrimination performance. This improvement may stem from the ability to capture the variation in severe dysphagia incidence across subsites, which results from the differing

distribution of radiation to healthy tissues depending on tumour location. Additionally, variations in radiation plans across tumour sites likely enhanced the discriminative ability of DVH features compared to those in our study's NPC-only cohorts, where DVH parameters variation is significantly less. This could explain the poor performance of the conventional clinical and DVH features in our datasets. In principle, a prediction model specific to NPC would provide a more precise and specialized solution. It is also important to recognize that other HNCs often involve surgery as part of their treatment, which can have a separate impact on dysphagia.

Secondly, both models used multi-centre data for training, in addition to having separate external validation. This is desirable for developing a generalizable model and may have helped their models to better handle the structural differences between institutions. Both the multi-omic model developed in this study and the model by Willemsen et al. utilized features extracted from the parotid glands, reinforcing the role of this VOI for dysphagia prediction. Conversely, the model by Dean et al. utilized features extracted from a custom pharyngeal mucosa contour. This volume may represent some dysphagia-specific information; however, this volume was not contoured by clinicians during RT planning, and it was not possible enrol experts to retrospectively contour this volume for the patients in this study. Another difference between our study and those in the literature is regarding the outcome definition. Willemsen et al. defined their outcome as the use of tube feeding for 4 weeks or more [128], whereas Dean et al. defined their outcome as CTCAE grade 3 or higher [131]. Interestingly, the incidence of tube feeding duration $\geq$ 4 weeks was 59% in the Willemsen study, whereas in the QEH and PWH data, only a minority of patients received tube feeding of

any duration. This suggests that the impact of hospital policy on tube feeding is a significant factor in study design and definition of dysphagia outcomes. In both of these studies, and also in this study, there were significant differences in the outcome incidence between training and external validation sets, presenting a further challenge for the development of predictive models. Willemsen et al. excluded cases where the patient declined tube feeding despite the clinician's recommendation, whereas in this study such cases were included as severe dysphagia cases. In both the training and development datasets, only a small proportion of patients agreed to receive tube feeding, and so excluding cases where tube feeding was refused would be significantly detrimental to the clinical applicability of the model. One further point of difference compared to the literature models is the number of features in each model. The high-performing model by Dean et al. consisted of 26 different features. This compares to 114 patients with severe dysphagia in their training set, resulting in an events-per-variable ratio of only 4.4, less than the rule of thumb of 10 events per variable [131]. Furthermore, their model was significantly larger than of the other models reported in this chapter. The model by Willemsen et al., by contrast, consisted of only 7 features despite being developed on a much larger dataset, resulting in an events-per-variable ratio of 37 [128]. The multi-omic model developed in this study consisted of 4 features, representing a comparable events-per-variable ratio of 30.5. Having a large model with a low events-per-variable ratio increases the risk of overfitting and hinders interpretability due to the resulting complexity of the model.

There were also differences in the clinical features included in each model. Willemsen et al. included the pre-treatment percentage weight change in their prediction model. Inclusion of this feature in general models for dysphagia is difficult, because some patients may receive

neoadjuvant chemotherapy, which could potentially affect the calculation of baseline weight loss. Additionally, the availability of sufficient pre-treatment weight measurements in our datasets was limited. However, if detailed assessment of pre-treatment weight loss was included in standard practice, then this feature could be quite informative for dysphagia prediction. Dean et al. included features which characterized the chemotherapy regimen in greater detail, specifying neoadjuvant chemotherapy status and chemotherapy drug. Inclusion of these features would be desirable in a larger study with multi-centre training but was not practical in this study due to missing drug data for the QEH dataset and the differences in policy on neoadjuvant chemotherapy between centres.

Ranking of the top 20 most-selected features in the top-scoring 5% of models revealed some trends. Male sex and chemotherapy were frequently selected, along with T-stage. One radiomic feature was ranked higher than other multi-omic features: the GLRLM long run high gray level emphasis feature for the GTVn VOI. This feature was also selected by the top-performing multi-omic model, where higher feature values were associated with higher risk of severe dysphagia. Other top features included contouromic features describing the masking of the extended oral cavity by the GTVP, and radiomic features describing the GTVn, parotids, extended oral cavity and GTVp. Findings were quite similar between weighted and unweighted counts. However, two of the features in the best-performing multi-omic model were not present in the tables of top features: the dosiomic first order variance feature for the GTVn and the radiomic GLCM joint energy feature for the parotid glands were not listed in the top 20 unweighted or weighted features. This shows the high variability in feature selection and the difficulty of identifying relevant features.

Removing the feature stability filter resulted in improvements in the discrimination of the best-performing models. This may be because the previously used stability filter was too strict, and relevant information was lost. Indeed, the article by Wennmann et al. found that using reproducible radiomic features did not always improve the external validation score, and in fact worsened the external validation performance in some cases [301]. After removing the stability filter, the best VOI and feature type combination was the same, while two of the same features were selected. The remaining 3 features in the updated model were similar to the previous model but described texture and intensity of the LoG-filtered image rather than the original image. These findings suggest that the stability filter could be too strict for the GTVn: The radiomic GLRLM long run high gray level emphasis feature and the dosiomic first order variance feature were included in the model despite having ICC values of 0.82 and 0.73 respectively. While these features may be more sensitive than others to the kind of perturbations applied, they appear to provide relevant information regarding severe dysphagia. The lack of a feature stability filter in the model development would mean that future studies would need to confirm whether the features selected by the models are adequately stable and repeatable. Also, future investigation with multiple sets of manually delineated contours by different radiologists would be required to determine the optimal perturbation settings for each VOI in order to replicate inter-observer variation.

Interestingly, the PC muscles VOI did not yield high-performing models even after removal of the ICC filter, despite the dose to the PC muscles being frequently reported as correlated with dysphagia in the literature. However, the externally validated models from the literature did not use the PC muscle VOIs either. Instead, Dean et al. included dose features

from a pharyngeal mucosa VOI, while Willemsen et al. included the mean dose to the parotid glands and oral cavity. The role of the PC muscles in prediction models for severe acute dysphagia remains to be seen.

This study had some limitations. Several of these applied to the definition of the outcome label. Firstly, there was likely to be some degree of under-estimation of the frequency of the severe dysphagia outcome. While consultations were typically conducted weekly during RT, the interval between consultations could be irregular and many patients had fewer than 6 sets of consultation notes over the RT period. This could cause a reduction in the apparent incidence of severe dysphagia, especially if consultations closer to the start of RT were absent. Clinicians may have been less likely to offer tube feeding to patients who experienced weight loss and reduced oral intake towards the end of RT, since fewer fractions of radiation remained to be delivered and the impact of weight loss would be less damaging. Furthermore, dysphagia may have been under-reported as a result of the lack of a standardized assessment according to a grading system. As outlined in **Section 5.2.1**, severe dysphagia was defined as the indication for tube feeding as recorded in the clinical notes. In many cases, clinicians may only have identified the indication for tube feeding based on recognizing a significant drop in body weight or based on feedback from the patient. Some patients may not have actively reported their symptoms, and this could lead to an under-reporting of dysphagia. There was also the risk of missing outcome data due to human error during the data collection process. Identification of severe dysphagia was based on manual inspection of the consultation notes for specific keywords relating to tube feeding. Consultation notes were not structured in a standardized way, and variations in the wording and use of abbreviations could have resulted in the omission

of some events, despite the precautions taken. Under-reporting of severe dysphagia would result in inaccuracy of the outcome label, reducing the validity of the results and increasing the difficulty of achieving a high discrimination score. However, steps were taken to mitigate this factor, such as excluding cases with less than 3 weeks of consultation notes.

The lack of specificity of the dysphagia label is another limitation. Whether severe dysphagia is defined as the indication for tube feeding or as CTCAE grade 3 or higher, the label does not distinguish between impairment of the swallowing mechanism and reduced oral intake due to other reasons, such as pain or xerostomia. However, assessment of the swallowing mechanism was not part of the routine clinical practice and would require a prospective study with a specialized assessment protocol that included videofluoroscopy studies.

The incidence of tube feeding indication was significantly different between the two institutions. This difference could partly result from differences in the record-keeping and frequency of consultations. Additionally, differences in tube feeding indication could have resulted from different conventions or policies regarding the provision of tube feeding. The decision to offer tube feeding depended on several factors: the amount of weight loss, which could be measured as the percentage change from baseline, swallowing difficulty reported by the patient, the recent dietary intake, and the number of RT fractions remaining. Even if both institutions used the same threshold weight loss, such as 10%, the definition of the baseline weight could be affected by many factors. The weight was typically recorded at the time of CT simulation, however the time interval between CT simulation and RT varied between patients. Furthermore, the provision of neoadjuvant chemotherapy would affect this weight measurement.

## 5.5. Conclusion

This study pioneered the development of integrated multi-omic prediction models for severe acute dysphagia, identifying the most effective combinations of radiomic, dosiomic and contouromic features. A multi-omic model consisting of radiomic and dosiomic features demonstrated improved predictive performance for identifying NPC patients at risk of severe acute dysphagia compared a conventional model consisting of clinical and DVH features developed on the same dataset. To the best of our knowledge, the multi-omic model represented the first externally validated model employing radiomic or dosiomic features for the prediction of severe acute dysphagia in NPC. However, the multi-omic model did not outperform the externally validated conventional models from the literature. The reasons for the superior performance of the literature models were explored, including the role of multi-site HNC training data instead of NPC-only cohorts, multi-centre training data, and more comprehensive characterization of chemotherapy regimen. While further improvement for the prediction of severe acute dysphagia is desirable, the findings indicate that multi-omic features have independent predictive value and can supplement clinical features for enhanced discrimination.

# CHAPTER 6 MULTI-OMIC, MULTI-LABEL PREDICTION MODELS FOR ORAL MUCOSITIS AND DYSPHAGIA

## 6.1. Introduction

Treatment-induced OM and dysphagia are interrelated conditions. This chapter explores the hypothesis that combining these toxicities into a multi-label problem, complimentary information about their relationship can enhance the discrimination performance of prediction models. One aspect of the relationship between OM and dysphagia is the impact of OM on increasing the severity of dysphagia. OM can exacerbate dysphagia by making swallowing more painful due to ulcerations occurring across multiple parts of the swallowing anatomy, including oral cavity, tongue, and pharyngeal mucosa. Patients may seek to minimize contact between food and the irritated mucosa, restricting the ability to chew food [302], which can result in greater difficulty in swallowing and the need for dietary modification. Another aspect of the relationship between OM and dysphagia is that they have predictive factors in common, as identified in the literature review in **Section 1.2**. Furthermore, there is an overlap in the CTCAEv5 grading criteria for these toxicities, resulting from the shared impact on oral intake [303]. The association between OM and dysphagia is reported in the literature [304], which was further confirmed by the significant correlation between severe OM and severe dysphagia in the data collected for this thesis ($p < 0.05$ in both datasets under Fisher's Exact Test). Multi-label models may be able to harness the relationship between the severity of each condition to enhance the accuracy of predictions. This chapter investigates

different approaches to multi-label models and explores whether this approach can facilitate more accurate prediction of severe OM and dysphagia.

Traditional supervised learning methods involve learning a mapping from a feature space to a label space, allowing the prediction of a label for each sample. This process assumes that each sample has a single label, and that all labels are mutually exclusive. In contrast, multi-label methods involve samples that each have a set of associated labels [305]. Multi-label classification using machine learning is less common than single-label classification. However, it has been utilized in specific applications, including the classification of colon cancer histological subtypes from histopathological images [306], prediction of cardiovascular events from carotid plaque ultrasound images [307], prediagnosis of cervical cancer from clinical record data [308], automatic categorization of pathology report text [309], prediction of antimicrobial resistances from bacterial genomics [310], and prediction of drug toxicity from assay results [311]. In each of these applications, each sample was associated with multiple non mutually exclusive labels, and machine learning models were trained to predict these labels from the input feature data. To our best knowledge, this study is the first to report multi-label models for predicting severe acute OM and dysphagia in cancer patients treated with radiotherapy.

## 6.2. Methodology

A common approach to multi-label learning involves problem transformation, where the problem is converted into binary or multi-class tasks, making the labels mutually exclusive [305]. For the multi-label prediction of severe acute OM and dysphagia, the simplest approach

would be to develop separate binary classification models for each outcome, as reported in CHAPTER 4 and CHAPTER 5. However, this approach does not consider label correlations or the interaction between labels [305].

The first multi-label approach explored in this study was the Label Powerset approach. This method converts the problem into a multi-class task with mutually exclusive classes corresponding to every possible combination of labels. As the number of labels increases, this approach results in exponentially many classes, resulting in higher complexity, computational cost and overfitting [308]. However, only two labels were investigated in this study, therefore this approach resulted in just four mutually exclusive classes. This approach therefore remained feasible for multi-label modelling.

The second multi-label approach explored in this study was the Classifier Chain approach. This method converts the problem into two binary classification tasks performed sequentially. The key aspect is that the predictions from the first model in the chain are passed into the second model in the chain as an additional input, allowing the model to learn from the correlation between toxicities.

There are other approaches to multi-label models that involve more complex tasks or specialized algorithms [305], however the two selected approaches were chosen as a well-established starting point. Additionally, each approach was compatible with the six machine learning algorithms used for model development in previous chapters.

### *6.2.1. Label powerset approach*

Under this approach, the severe OM and severe dysphagia outcome labels were combined into a single target representing the four possible outcomes, as shown in **Table 49**. These were automatically encoded as integers 0-3 using the Pandas library for Python. Consequently, the problem became a multi-class classification with four mutually exclusive labels. The MRMR feature selection function and machine learning algorithms supported this kind of multi-class task by default, and therefore this constructed label was directly used for model development. For model optimization, the AUC was computed using the 'one-versus-rest' approach for multi-class data, where each label was scored against all other labels, e.g., severe OM and severe dysphagia versus all other combinations. The final score was the mean AUC across all combinations. For evaluation purposes, as well as the multi-class 'one-versus-rest' AUC as defined above, the AUC for each individual toxicity was also computed. Comparison between algorithm, VOI and feature type combinations was performed using the AUC for each individual toxicity, rather than the multi-class 'one-versus-rest' AUC, to facilitate comparison with the previous chapters and ensure that the discrimination was adequate for both toxicities.

**Table 49: Multiclass target values**

| | |
|---|---|
| No severe OM<br>No severe dysphagia | Severe OM<br>No severe dysphagia |
| No severe OM<br>Severe dysphagia | Severe OM<br>Severe dysphagia |

### *6.2.2. Classifier chain approach*

Classifiers for severe acute OM and dysphagia were combined into a classifier chain under this approach, with the chain consisting of the OM model followed by the dysphagia

model. That is, the predicted probabilities from the OM model were fed into the dysphagia model, aligning with the hypothesized causal relationship between the two toxicities. While dysphagia might exacerbate OM indirectly through the consequent reduction in nutrition impacting tissue repair and resilience to infection, the more direct link would be through the impact of pain and discomfort from OM on dysphagia. Discussions with clinicians suggested that dysphagia might be worsened by the shedding of dead tissue or pseudo membrane that occurs in the latter phases of OM development.

The Scikit-Learn library for Python provides a classifier chain function wrapper which was used to evaluate this approach. However, the MRMR feature selection step in the pipeline required further modification. The input was provided in the form of an $N \times 2$ vector, representing the binary outcomes for OM and dysphagia. MRMR feature selection could be approached in two ways: firstly, to treat the outcome labels as being multi-class and find the optimal feature set for this pooled label. Secondly, to perform MRMR separately for each outcome label, then take the set, or union, of features that were selected, to prevent duplicate features. Experiments showed that these two approaches gave similar results, so the first approach was used for model development for simplicity.

An alternative variant of the classifier chain approach was also conducted, where the whole model pipeline was connected in series in the chain – having separate feature selection, scaling and model fitting for each classifier. This variant allowed only the most relevant features, including the predictions from the previous model in the chain, to be selected for each classification task in the chain, rather than sharing the same features in each model.

The classifier chain models were optimized in a similar manner to that employed in previous chapters, with a cross-validated grid search over the previously defined range of model hyperparameters. The maximum value of *k* for the MRMR feature selection was set to 9, as with the model development for severe OM. The optimum hyperparameters were identified by the model that gave the highest mean AUC across both severe OM and severe acute dysphagia labels.

Multi-omic models were previously developed for severe acute OM and dysphagia in CHAPTER 4 and CHAPTER 5. These models could not be combined into a classifier chain for several reasons. Firstly, selecting the final feature set from the two models would result in information leakage about the internal validation set, resulting in bias if the resulting classifier chain were cross validated on the development dataset. Secondly, at least one of the models would have to be re-trained with the predictions of the previous model in the chain as input, otherwise the results would be identical to having two separate binary classifiers. Thirdly, the Scikit-learn classifier chain class did not support freezing the parameters of one of the models in the chain, which would mean that both models in the chain would be re-trained. Also, the model hyperparameters were optimized with respect to the mean of the AUCs for each toxicity, instead of being optimized for a single toxicity, which would result in different model coefficients being fitted compared to the two original models. For these reasons, multi-label models were trained on the development dataset without reference to the models in CHAPTER 4 and CHAPTER 5.

## 6.3. Results

The incidence of each label in the label powerset approach is shown in **Table 50**. These four labels were used for model training and optimization. Patients who experienced neither severe OM nor severe dysphagia were the largest group in both datasets.

**Table 50: Label incidences for label powerset approach**

| Severe OM | Severe acute dysphagia | Development dataset (QEH) | External validation dataset (PWH) |
|---|---|---|---|
| ✓ | ✓ | 44 (12%) | 22 (22%) |
| ✓ | | 46 (13%) | 8 (8%) |
| | ✓ | 78 (21%) | 31 (31%) |
| | | 195 (54%) | 40 (40%) |

For each multi-label approach, models were evaluated for different combinations of VOIs and feature types, as performed in previous chapters. Six different machine learning algorithms were also evaluated. The performance of the different models was ranked by an aggregate AUC score, calculated by taking the minimum of the training, internal validation and external validation AUCs across both severe OM and severe dysphagia. If any one of these scores were low, it would indicate poor reproducibility or generalizability of the model.

**Table 51** displays the top 5 multi-label models developed using the label powerset approach. **Table 52** displays the top 5 multi-label models developed using the classifier chain approach with shared feature selection. **Table 53** displays the top 5 multi-label models developed using the classifier chain approach with separate pipelines. Among the top scoring models, VIF dimensionality reduction was the most frequently selected approach. The extended oral cavity was the most frequently selected VOI, distantly followed by the GTVn. The most frequently selected feature types were clinical features, contouromic features, radiomic features and DVH features. Interestingly, dosiomic features were not selected in the

top performing models. Gaussian Naive Bayes was the most frequently selected algorithm among the top scoring models. The model size ranged from 3 features to 9 features. Validation AUC scores were generally higher for OM than for dysphagia. Overall, there was not any obvious difference in the performance between the different multi-label approaches.

**Table 51: Top 5 multi-label prediction models for label powerset approach**

| Rank | Dimensionality reduction | VOIs | CLI | DVH | RAD | DOS | CON | Algorithm | Model size | Severe OM | | | Severe dysphagia | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Train AUC | Int. AUC | Ext. AUC | Train AUC | Int. AUC | Ext. AUC |
| 1 | VIF | Ext. oral cavity, parotids | ✓ | | ✓ | | ✓ | Gaussian Naïve Bayes | 4 | 0.633 | 0.636 | 0.623 | 0.635 | 0.617 | 0.657 |
| 2 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Gaussian Naïve Bayes | 3 | 0.650 | 0.624 | 0.621 | 0.616 | 0.612 | 0.646 |
| 3 | VIF | Ext. oral cavity, GTVn | ✓ | ✓ | | | ✓ | SVM Linear | 7 | 0.623 | 0.66 | 0.625 | 0.630 | 0.634 | 0.611 |
| 4 | VIF | Ext. oral cavity | ✓ | | ✓ | | ✓ | Random Forest | 4 | 0.671 | 0.64 | 0.666 | 0.661 | 0.61 | 0.645 |
| 5 | VIF | GTVp | ✓ | ✓ | | | | XGBoost | 9 | 0.667 | 0.635 | 0.651 | 0.637 | 0.617 | 0.609 |

**Table 52: Top 5 multi-label prediction models for classifier chain approach with shared feature selection**

| Rank | Dimensionality reduction | VOIs | CLI | DVH | RAD | DOS | CON | Algorithm | Model size | Severe OM | | | Severe dysphagia | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Train AUC | Int. AUC | Ext. AUC | Train AUC | Int. AUC | Ext. AUC |
| 1 | VIF | Ext. oral cavity | ✓ | | ✓ | | ✓ | Gaussian Naïve Bayes | 4 | 0.660 | 0.632 | 0.673 | 0.632 | 0.618 | 0.675 |
| 2 | Hierarchical | Ext. oral cavity, GTVn | ✓ | ✓ | ✓ | | | Gaussian Naïve Bayes | 9 | 0.711 | 0.628 | 0.627 | 0.685 | 0.613 | 0.615 |
| 3 | VIF | Ext. oral cavity, PC | ✓ | ✓ | ✓ | | | Gaussian Naïve Bayes | 7 | 0.657 | 0.634 | 0.638 | 0.631 | 0.613 | 0.615 |
| 4 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Gaussian Naïve Bayes | 6 | 0.644 | 0.629 | 0.628 | 0.629 | 0.611 | 0.647 |
| 5 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Gaussian Naïve Bayes | 3 | 0.652 | 0.65 | 0.610 | 0.615 | 0.615 | 0.647 |

**Table 53: Top 5 multi-label prediction models for classifier chain approach with separate pipelines**

| Rank | Dimensionality reduction | VOIs | CLI | DVH | RAD | DOS | CON | Algorithm | Model size | Severe OM | | | Severe dysphagia | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Train AUC | Int. AUC | Ext. AUC | Train AUC | Int. AUC | Ext. AUC |
| 1 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Ridge | 3 | 0.644 | 0.644 | 0.644 | 0.609 | 0.611 | 0.649 |
| 2 | VIF | Ext. oral cavity, PC | ✓ | | ✓ | | ✓ | Ridge | 7 | 0.673 | 0.652 | 0.608 | 0.622 | 0.616 | 0.636 |
| 3 | VIF | Ext. oral cavity | ✓ | | | | ✓ | Gaussian Naïve Bayes | 5 | 0.648 | 0.635 | 0.625 | 0.626 | 0.607 | 0.643 |
| 4 | VIF | Ext. oral cavity | ✓ | | ✓ | | ✓ | Gaussian Naïve Bayes | 3 | 0.651 | 0.656 | 0.635 | 0.638 | 0.609 | 0.604 |
| 5 | Hierarchical | Ext. oral cavity, larynx | ✓ | | | | ✓ | Gaussian Naïve Bayes | 7 | 0.690 | 0.624 | 0.638 | 0.618 | 0.604 | 0.654 |

### *6.3.1. Label powerset*

*Best label powerset model*

A Gaussian Naïve Bayes model constructed from the set of clinical, radiomic, and contouromic features extracted from the extended oral cavity and parotids glands VOIs after applying the VIF dimensionality reduction approach achieved the best performance. The model consisted of only 4 features: chemotherapy status, male sex, radiomic GLSZM large area low gray level emphasis within the parotid glands, and a contouromic projection overlap volume feature between the extended oral cavity and GTVp. For severe OM, the model achieved AUCs of 0.633 (95% CI: 0.567, 0.694), 0.636 (95% CI: 0.567, 0.705) and 0.623 (95% CI: 0.505, 0.743) in training, internal validation and external validation. For severe dysphagia, the model achieved AUCs of 0.635 (95% CI: 0.572, 0.696), 0.617 (95% CI: 0.568, 0.665) and 0.657 (95% CI: 0.550, 0.759) respectively. Details of the model settings are shown in APPENDIX, **Table 59.**

**Figure 36** shows the SHAP analysis for the label powerset model. SHAP values were calculated for each of the four labels, indicating the impact of each feature on the prediction of that label. To facilitate comparison with other models, the values corresponding to labels representing severe OM were added together, and the values corresponding to labels representing severe acute dysphagia were added together. The resulting analyses indicated the impact of each feature on the model predictions for each toxicity. The direction of correlation for each feature was comparable for each plot. Chemotherapy and male sex increased the risk of severe toxicity, as did the contouromic feature and radiomic feature. The ranking of each feature, indicating the overall importance of that feature, differed between toxicities.

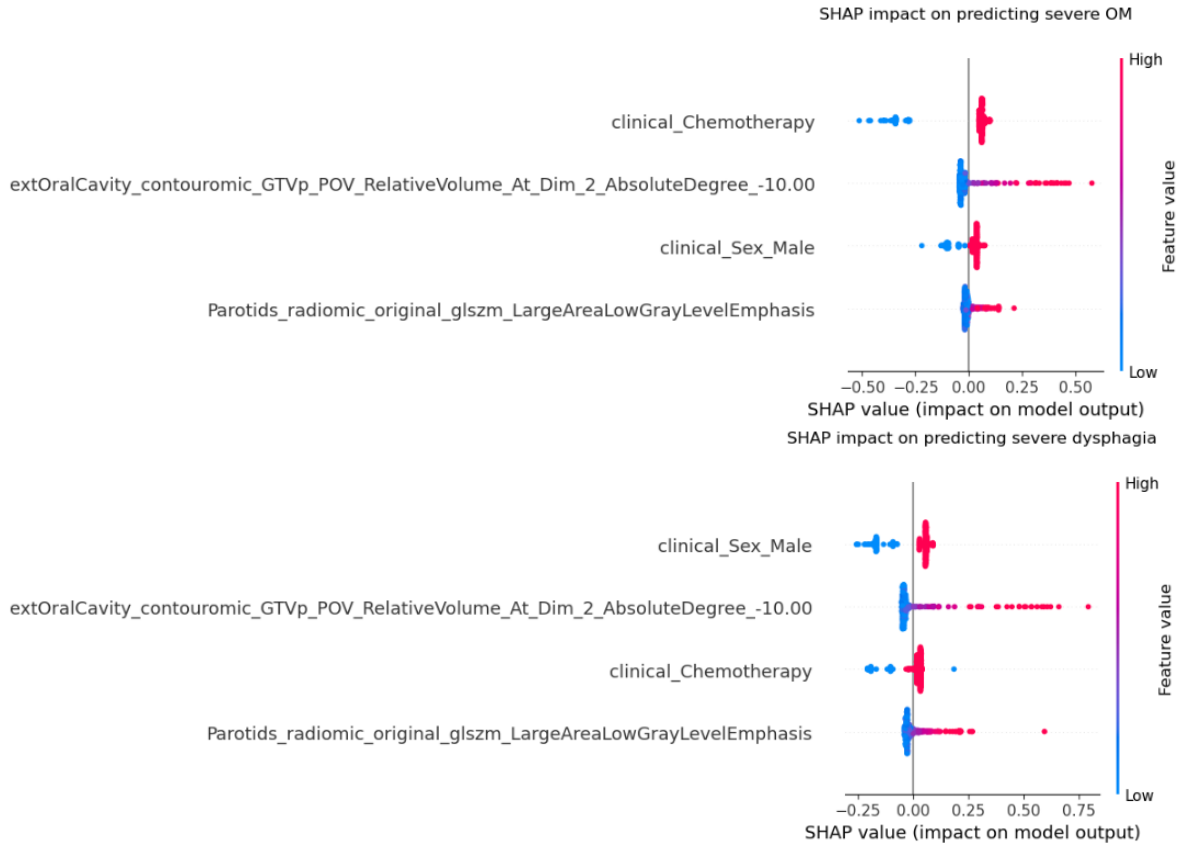



**Figure 36: SHAP analysis for the best multi-label model developed using the label powerset approach, for severe OM (top) and severe acute dysphagia (bottom)**

### *6.3.2. Classifier chain with shared scaling and feature selection*

*Best classifier chain with shared scaling and feature selection*

A Gaussian Naïve Bayes model constructed using clinical, radiomic and contouromic features using the extended oral cavity VOI after applying the VIF dimensionality approach achieved the best performance. The model consisted of 4 features: chemotherapy status, male sex, radiomic LoG-filtered image mean intensity within the extended oral cavity and a contouromic projection overlap volume filter between the extended oral cavity and GTVp. For severe OM, the model achieved AUCs of 0.660 (95% CI: 0.595, 0.719), 0.632 (95% CI: 0.558, 0.707) and 0.673 (95% CI: 0.544, 0.787) in training, internal validation, and external validation.

For severe dysphagia, the model achieved AUCs of 0.632 (95% CI: 0.568, 0.696), 0.618 (0.551, 0.685) and 0.675 (95% CI: 0.565, 0.782) respectively. Details of the model settings are shown in APPENDIX, **Table 60**.

**Figure 37** shows the SHAP analysis for the classifier chain model developed using combined scaling and feature selection. In this case, the SHAP values were calculated for severe OM and for severe acute dysphagia separately. The same set of features was used for the prediction of each toxicity, though the relative importance of features differed, as shown by the different ordering. The direction of correlation between each feature and toxicity was consistent between toxicities.

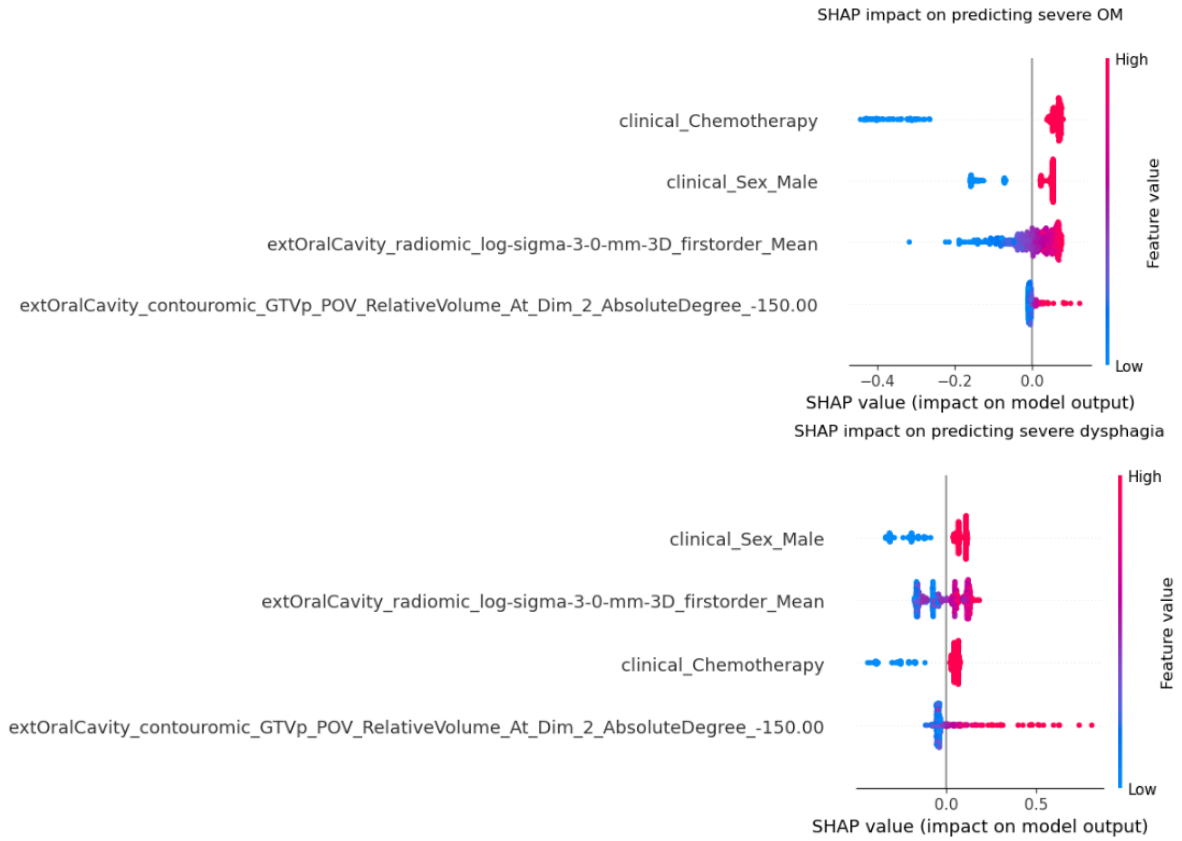



**Figure 37: SHAP analysis for the best multi-label classifier chain developed using shared feature selection, for severe OM (top) and severe acute dysphagia (bottom)**

### *6.3.3. Classifier chain with separate pipelines*

*Best classifier chain model with separate pipelines*

A logistic ridge regression model constructed using clinical and contouromic features using the extended oral cavity VOI after applying the VIF dimensionality approach achieved the best performance. The classifier chain consisted of two models: The OM model consisted of 3 features: chemotherapy status, male sex, and a contouromic overlap volume histogram feature describing the distance between the GTVp and extended oral cavity. The dysphagia model consisted of 3 features: the predicted probabilities from the severe OM model, a contouromic projection overlap volume feature describing the masking of the extended oral cavity by the GTVp, and male sex. For severe OM, the model achieved AUCs of 0.644 (95% CI: 0.579, 0.706), 0.644 (95% CI: 0.581, 0.707) and 0.644 (95% CI: 0.522, 0.759) in training, internal validation, and external validation. For severe dysphagia, the model achieved AUCs of 0.609 (95% CI: 0.546, 0.667), 0.611 (0.552, 0.670) and 0.649 (95% CI: 0.536, 0.756) respectively. Details of the model settings are shown in APPENDIX, **Table 61.**

**Figure 38** shows the SHAP analysis for the classifier chain model developed using separate pipelines. The SHAP values were calculated for severe OM and for severe acute dysphagia separately. Only features with nonzero SHAP values were included in the plots. Since different feature sets were selected for each model in the chain, the set of features in each plot is different. The severe acute dysphagia model utilized the predicted probabilities from the severe OM model, therefore the SHAP values for features in the severe OM model were also nonzero for the dysphagia model.

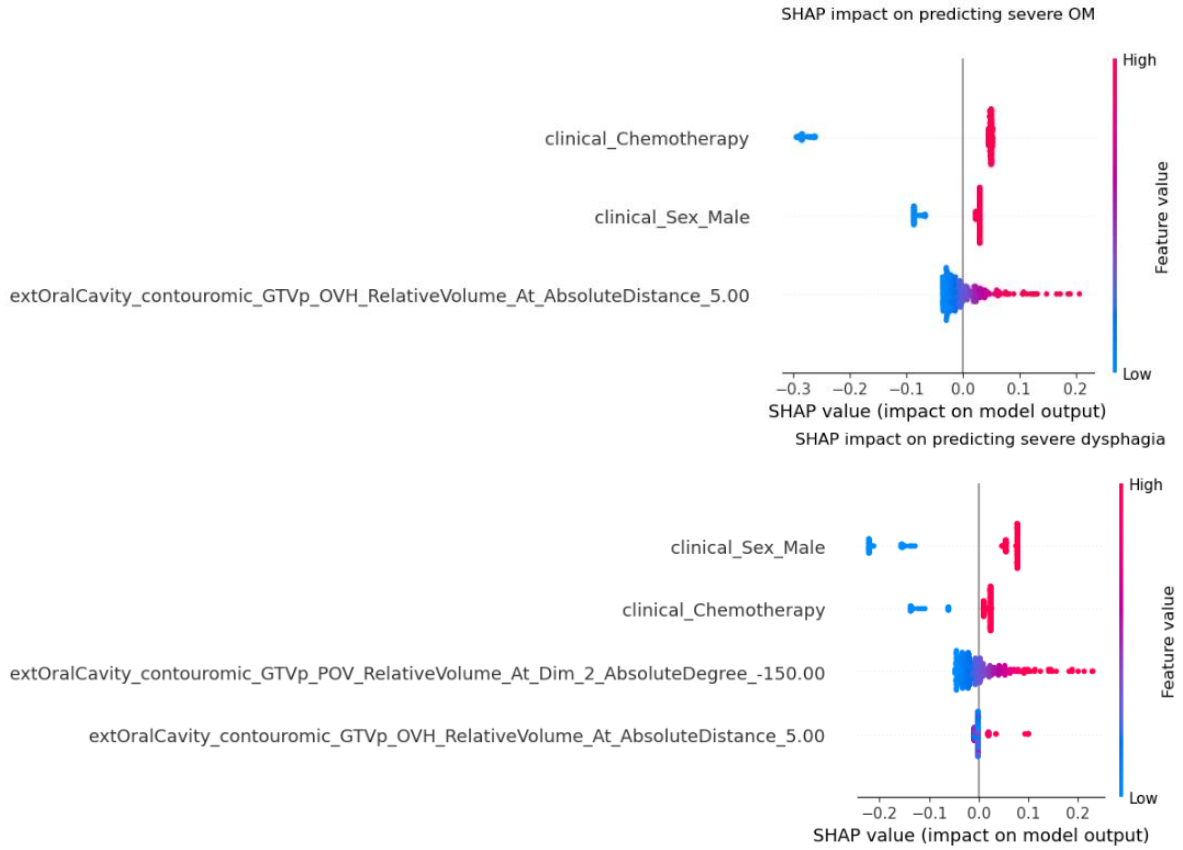



**Figure 38: SHAP analysis for the best classifier chain developed using separate model pipelines, for severe OM (top) and severe acute dysphagia (bottom)**

### *6.3.4. Comparison of top-performing multi-label models*

**Table 54** compares the top-performing multi-label models. For all three approaches, the highest performance was obtained using the VIF dimensionality reduction approach. Discrimination performance in training, internal validation and external validation was generally higher for OM than for dysphagia. However, none of the models out-performed both of the top-performing single-toxicity models from CHAPTER 4 and CHAPTER 5. There were no significant differences between the three models according to the DeLong test.

**Table 54: Comparison of top-performing multi-label models**

| Approach | Dimensionality reduction | Model type | OM train / internal / external AUC | Dysphagia train / internal / external AUC |
|---|---|---|---|---|
| Label powerset | VIF | GNB | 0.633 / 0.636 / 0.623 | 0.635 / 0.617 / 0.657 |
| Classifier chain - shared | VIF | GNB | 0.660 / 0.632 / 0.673 | 0.632 / 0.618 / 0.675 |
| Classifier chain - separate | VIF | Ridge | 0.644 / 0.644 / 0.644 | 0.609 / 0.611 / 0.649 |

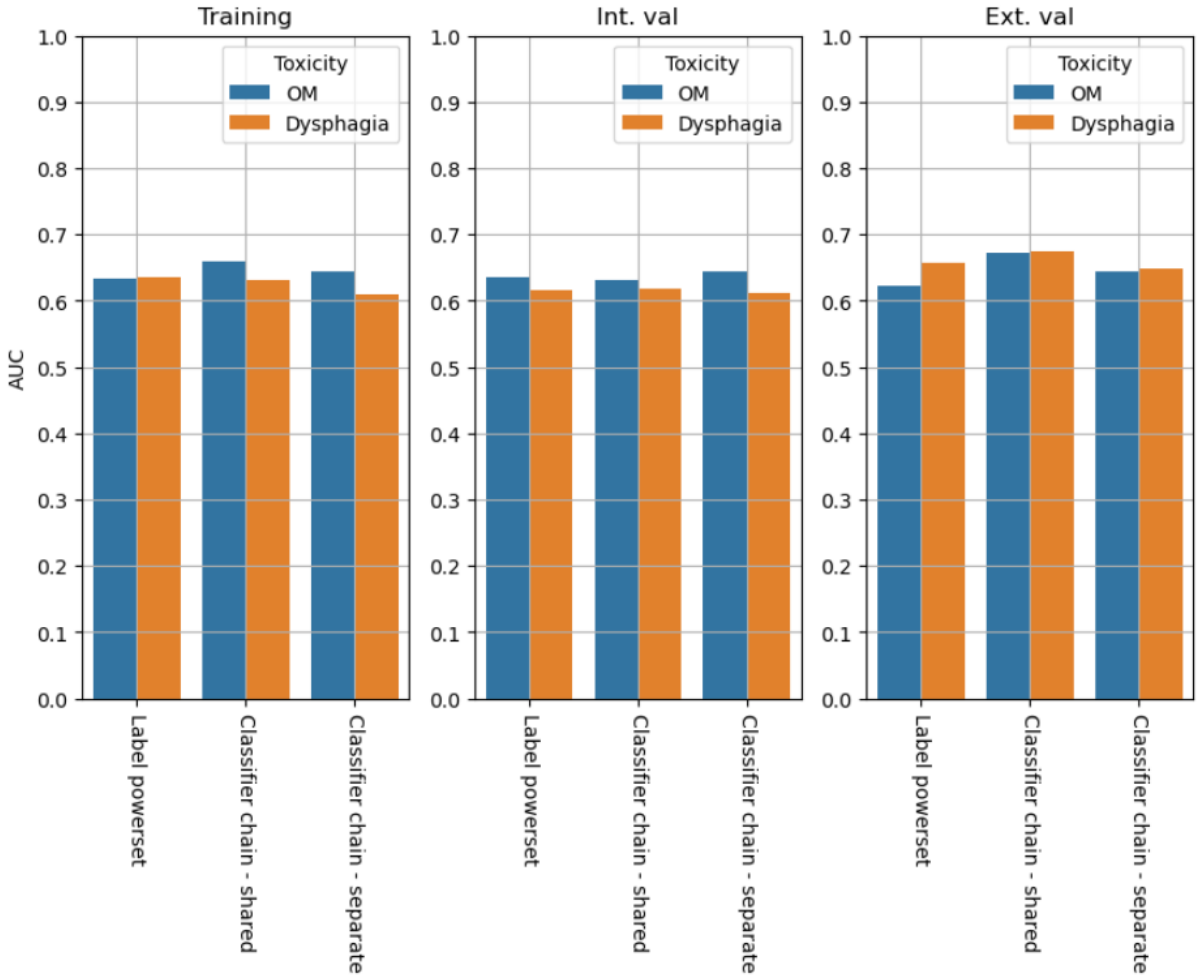



**Figure 39: Graphical comparison of top-performing multi-label models**

**Table 55: Comparison of AUC scores for best multi-label model and best binary classification models for OM and dysphagia**

| | Classifier chain (OM) | Previous OM model | Classifier chain (Dysphagia) | Previous dysphagia model |
|---|---|---|---|---|
| Training | 0.660 | 0.954 | 0.632 | 0.970 |
| Internal validation | 0.632 | 0.684 | 0.618 | 0.633 |
| External validation | 0.673 | 0.688 | 0.675 | 0.625 |

**Table 55** shows the comparison of AUC scores for the best-performing multi-label model and the best-performing binary classification models for OM and dysphagia. The multi-label classifier chain model did not improve on the performance of the OM model, however the external validation score for the dysphagia prediction was markedly improved.

**Table 56** shows the top features in the highest-scoring 5% of classifier chain models using shared feature selection, as this was the approach that gave the highest-scoring model. Apart from chemotherapy status, sex and T-stage, contouromic features were ranked highly. The extended oral cavity was frequently selected as a VOI.

**Table 56: Top features in highest-scoring 5% of Classifier Chain models using shared feature selection**

| Feature | Number of models |
|---|---|
| clinical_Chemotherapy | 191 |
| clinical_Sex_Male | 184 |
| clinical_T3 | 48 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-170.00 | 33 |
| Parotids_contouromic_GTVn_POV_RelativeVolume_At_Dim_0_AbsoluteDegree_20.00 | 30 |
| clinical_AgeAtRTStart | 23 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-150.00 | 23 |
| extOralCavity_dvh_RelativeVolume_At_RelativeDose_0.94 | 17 |
| extOralCavity_dosiomic_original_gldm_LargeDependenceLowGrayLevelEmphasis | 17 |
| extOralCavity_radiomic_log-sigma-3-0-mm-3D_firstorder_Mean | 17 |
| extOralCavity_contouromic_GTVp_OVH_RelativeVolume_At_AbsoluteDistance_5.00 | 12 |
| GTVp_radiomic_original_glszm_ZoneEntropy | 11 |
| Parotids_contouromic_GTVn_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_130.00 | 10 |
| GTVp_radiomic_log-sigma-2-0-mm-3D_glcm_Imc2 | 10 |
| GTVn_dosiomic_log-sigma-1-0-mm-3D_firstorder_10Percentile | 9 |
| GTVn_dvh_RelativeVolume_At_AbsoluteDose_72.00 | 7 |
| extOralCavity_contouromic_GTVp_POV_RelativeVolume_At_Dim_2_AbsoluteDegree_-10.00 | 7 |
| Parotids_radiomic_original_glszm_LargeAreaLowGrayLevelEmphasis | 7 |
| GTVn_dosiomic_log-sigma-1-0-mm-3D_gldm_DependenceNonUniformityNormalized | 6 |
| GTVp_dosiomic_log-sigma-1-0-mm-3D_glcm_ClusterProminence | 6 |

## 6.4. Discussion

Multi-label models were developed for severe OM and severe acute dysphagia, harnessing the complimentary information from the relationship between toxicities. Two types of multi-label models were explored: multi-class models using the label powerset approach, and classifier chains which passed the predicted probabilities of severe OM into the model for severe acute dysphagia. In the classifier chain approach, two variants were evaluated: performing shared scaling and feature selection for both models in the chain and performing separate scaling and feature selection for each model in the chain. As in the previous chapters, combinations of different feature types, VOIs and model algorithms were evaluated to determine the best model.

The top 5 multi-label models developed under each approach were identified. The extended oral cavity was the most frequently selected VOI across these models, suggesting that it contained the most relevant information for both OM and dysphagia, and inclusion of other VOIs provided little additional benefit. Clinical and contouromic features were frequently selected, along with some radiomic features. DVH and dosiomic features were less frequently selected. The role of contouromic features may be to quantify the difficulty of dose sparing of the oral cavity, which could be a shared risk factor for both severe OM and severe dysphagia.

The top-performing model from each of the 3 multi-label approaches was identified. None of the resulting models were able to outperform both best-performing binary classification models from CHAPTER 4 and CHAPTER 5. However, all three of the highlighted multi-label models significantly outperformed random chance, as indicated by their 95% confidence intervals. There were no significant differences between the three models according to the DeLong test, but the classifier chain model with shared scaling and feature selection achieved the best internal and external validation scores, outperforming the label powerset model in external validation and the classifier chain model with separate pipelines in internal validation. It did not outperform the top-scoring model for severe OM from CHAPTER 4, but did outperform the top-scoring model for severe acute dysphagia from CHAPTER 5 in external validation. Additionally, the discrepancy between training score and internal and external validation scores was reduced. Further investigation in future studies would be required to confirm the preferred multi-label approach. Studies could also include a wider range of multi-label approaches, including ensembles of classifier chains or specialized multi-label

algorithms [305], though larger sample sizes would be desirable to avoid overfitting from higher model complexity.

The comparison of the three top-performing multi-label models revealed that chemotherapy status, male sex and a contouromic POV feature for the GTVp - extended oral cavity pair for rotation in the axial plane were independently selected in all three models. The direction of correlation for chemotherapy status and male sex was consistent across the models and agreed with findings from previous chapters. The POV features were also frequently selected in the top-performing models for severe acute dysphagia (**Table 47**). The POV features corresponded to the same rotation plane as that of the LINAC gantry, suggesting its connection to the difficulty of dose sparing. The specific POV feature differed between the models, indicating that the overlap at a different angle was selected. However, inspection of the POV curve for the extended oral cavity in **Figure 14**, **Section 3.7** shows that both features correspond to points near the tails of the distribution rather than points at the peak of the curve. These features suggest that if the GTVp masked the extended oral cavity over a wider angular range, the risk of severe toxicity would be increased. Therefore, patients whose geometry resulted in greater dose sparing difficulty for the extended oral cavity were associated with higher risk of severe OM and severe dysphagia. Interestingly, no DVH features or dosiomic features were selected. These features would be more strongly affected by differences in RT planning and RT modality between institutions. The identified contouromic features may represent an underlying aspect of the patient geometry that is associated with greater risk of toxicity from greater difficulty in dose sparing.

This study had some limitations. One limitation concerns the optimization of classifier chain models. The classifier chains had shared hyperparameters: MRMR k, model algorithm, and model settings. Independent optimization of hyperparameters for each model could further improve performance. However, this would drastically increase computational complexity due to the resulting large number of combinations of hyperparameters to be investigated. Another limitation was that the label powerset approach suffered from lower number of samples per target label and a greater degree of imbalance between labels. Having fewer examples per label would restrict the ability of the model to adequately learn feature correlations without overfitting. However, over-sampling using the SMOTE approach was not observed to improve the results, possibly because there were too few samples in the minority classes for generation of generalizable synthetic samples.

## 6.5. Conclusion

This chapter demonstrated the feasibility of developing multi-label models for severe acute OM and dysphagia. As two interconnected toxicities experienced by NPC patients, having a single model that harnesses predictive factors for both toxicities offers the potential to be more efficient and more accurate. However, having explored different approaches for multi-label modelling, none of the developed models were able to outperform both highest-scoring models for the individual toxicities. Even so, the methods evaluated in this chapter could be useful for future research into multi-label model development. With larger sample sizes, multi-centre training data, and more detailed clinical data about the chemotherapy regimen, it would be possible to construct more complex models that identify more subtle patterns and relationships between features. This study also emphasized the potential role of

contouromic features, being less susceptible to differences between institutions. Contouromic features, like POV features that describe how the GTVp masks the extended oral cavity for rotation around the gantry axis, could identify variations in the underlying difficulty of dose sparing that are caused by the shape of the patient and their tumour. Contouromic features are a relatively recent development, and further exploration and standardization is merited to better understand this geometric information. The potential for these features to be less affected by differences between institutions could be highly desirable for more generalizable models. Further development of multi-label models is warranted to explore their potential for improving the discrimination of severe acute OM and dysphagia.

# CHAPTER 7 PRACTICAL CONSIDERATIONS FOR FUTURE DEVELOPMENT

## 7.1. Introduction

The aim of this project was to investigate the role of multi-omics in facilitating early intervention for severe acute OM and dysphagia in NPC patients receiving RT, to alleviate suffering and improve their quality of life. Three objectives were defined to achieve this aim. Firstly, to develop and externally validate a multi-omic prediction model for severe acute OM in NPC patients undergoing RT, which was reported in CHAPTER 4. Secondly, to develop and externally validate a multi-omic prediction model for severe acute dysphagia, in NPC patients undergoing RT, which was reported in CHAPTER 5. Thirdly, to develop and externally validate a multi-omic, multi-label model to predict severe acute OM and dysphagia, which was reported in CHAPTER 6. These models demonstrated the potential of multi-omic features to improve on the performance of conventional and clinical features in predicting these common and damaging toxicities. However, to alleviate suffering and improve patients' quality of life, models must be implemented in clinical practice. This chapter focuses on two key aspects. Firstly, it provides recommendations for further improvement in the performance of multi-omic models, emphasising the need for greater discrimination ability and a stronger level of evidence to advance towards clinical implementation. Secondly, it directly addresses the challenges associated with clinical implementation, offering insights and potential solutions.

## 7.2. How can the performance of multi-omic models be improved?

### *7.2.1. Defining model performance*

The performance of multi-omic models can be measured in several ways: discrimination, calibration, generalizability, and repeatability.

Discrimination quantifies how well the model can separate the severe and non-severe cases. This is typically measured using the AUC score and is the primary way to compare models because it is independent of the prediction threshold and the incidence of severe cases. A higher AUC is desirable, but the other aspects of model performance should also be considered. Model discrimination can also be measured using the sensitivity, specificity, precision and recall. These aspects are important for assessing the clinical utility of a model but are less useful for comparing the overall performance because they depend on the choice of probability threshold.

Model calibration refers to how well the predicted probabilities match the real incidence of severe toxicity. For the output of a model to be meaningful and useful, the predicted probabilities must be well-calibrated, otherwise the predicted probabilities will be misleading. Calibration is assessed using calibration curves and metrics such as the Brier score.

Generalizability refers to the ability of the model to perform well on new, unseen data. The validity of the model can only be determined by testing it on unseen data, whether from hold-out testing or cross-validation techniques. External validation is the gold standard of evidence for model validity, since it will penalize models that are over-fitted to the training

data or those that cannot cope with structural differences between datasets. With regards to clinical implementation, generalizability refers to the robustness of the model across different institutions with different scanner hardware, treatment regimen, and population demographics. Developing a generalizable model is challenging, while a hospital-specific model can potentially achieve higher discrimination. However, a generalizable model should be the starting point, even for hospital specific models, because a generalizable model should be able to identify underlying, common factors associated with toxicity while also providing a higher level of evidence. Hospital specific models risk overfitting, due to the inherent limitations on sample size and the lack of external validation.

Repeatability refers to how well the predicted probabilities from the model can be reproduced on the same patients using the same hardware for scanning and treatment. A high performing model should provide consistent predictions and not be strongly affected by small changes in the conditions at the time of data acquisition. Examples of factors that can affect repeatability include inter- and intra-observer variations in VOI segmentation, scanner noise, and breathing motion during scanning.

Each aspect must be considered to provide recommendations for future development of multi-omic models for prediction of severe acute OM and dysphagia.

### *7.2.2. Challenges in model development*

*Variability in imaging acquisition parameters*

A generalizable model should be resilient to changes in imaging acquisition across institutions such as different scanner machines, different reconstruction parameters and

contrast media. In some cases, differences in the time-to-scan from diagnosis or the time interval between scanning and start of RT must be considered. The differences in CT acquisition parameters between institutions is shown in Table 57. Differences in X-ray tube current and reconstruction kernel may lead to variations in noise levels and could therefore affect the values of extracted features.

**Table 57: Differences in CT acquisition parameters between institutions**

| Parameter | QEH | PWH |
|---|---|---|
| Scan mode | Helical | Helical |
| Voltage | 120 kVp | 120 kVp |
| Pixel spacing | 1.2 x 1.2 mm | 1.2 x 1.2 mm |
| Slice thickness | 3 mm | 3 mm |
| Matrix | 512 x 512 | 512 x 512 |
| X-ray tube current | 264 mA | 165 mA |
| Reconstruction kernel | Not collected | Not collected |

## *Variability in radiotherapy modalities and planning*

Radiation may be delivered using fixed-field IMRT, volumetric modulated arc therapy (VMAT) or helical tomotherapy. The decision of which modality to provide depends on the availability of treatment machines in the hospital, but there is evidence that the dose sparing for normal tissues differs between each modality [312]. The development of a model using data based on a single modality may struggle to generalize to other modalities, where the dose distribution in organs-at-risk (OARs) may vary. This reinforces the need for multi-centre training, as the RT modality varies between centres. It is desirable to uncover the causes of toxicity that result from these variations, such as differences in dosimetric features.

In terms of variations in contouring between institutions, there are differences in the anatomy included in contoured organs-at-risk (OARs) resulting from following different contouring guidelines. Furthermore, intra-, and inter-observer variation in contouring will

cause significant variability in the VOIs used for feature extraction. This aspect can affect feature repeatability and generalizability and must be investigated by conducting multiple sets of contouring or by utilizing the perturbation approach.

Variation in dose sparing guidelines presents another challenge for generalizability. The Chinese Society of Clinical Oncology reported guidelines for normal tissue delineation and dose limitation for OARs including the parotid glands, oral cavity, PC muscles, and larynx[7]. The mean dose to each of these organs in the QEH and PWH datasets exceeded these limits. Furthermore, there were significant differences in the mean dose to each of these OARs between the two datasets. This may partly be explained by the different RT modalities used in each dataset, and partly from differences in the contouring guidelines used for each OAR. However, there may also be differences in the set of OARs included in the dose sparing guidelines used by each institution. It should be noted that many patients were missing contours of the oral cavity, larynx and PC muscles, which suggests that dose constraints for these OARs were less consistently applied than for tissues such as the brain stem and spinal cord.

Finally, different dose calculation algorithms used by the treatment planning software are another source of variation that could impact the generalizability of DVH and dosiomic features [313].

### *Variability in chemotherapy treatment protocols*

Variability in chemotherapy treatment protocols presents significant challenges for developing generalizable toxicity prediction models. The Chinese Society of Clinical Oncology has reported strong evidence supporting the benefit of concurrent chemotherapy for

locoregionally advanced NPC [7]. However, they did not reach a conclusion on whether concurrent chemotherapy should be combined with neoadjuvant chemotherapy or with adjuvant chemotherapy. Accordingly, the provision of neoadjuvant and adjuvant chemotherapy varied significantly between the two institutions in this study. Furthermore, there were additional variations in the choice of chemotherapy drug, the number of cycles of chemotherapy, and the dose of chemotherapy drug. Some of these variations were required due to the presence of contraindications in patients indicated for chemotherapy. These variations present a challenge for the development of generalizable toxicity prediction models because of the difficulty in characterizing a wide range of possible chemotherapy regimens in the training data and model bias from imbalance in the distribution of chemotherapy-related features across datasets. Furthermore, neoadjuvant chemotherapy is delivered prior to CT simulation and RT planning, therefore it represents a confounding factor that could influence the values of multi-omic features.

### *Variability in clinical assessment and follow-up schedule*

The availability of clinical and toxicity data represents another challenge for model development. Much of the clinical and toxicity data is typically recorded in the clinical notes by clinicians during consultation with patients. The frequency and completeness of these consultation notes can vary. The notes for the institutions in this study were in the form of raw text and did not generally have a fixed structure or fixed data items to collect. While certain measurements were performed routinely, other assessments were only conducted or noted if the patient visibly presented with or complained of issues. There was therefore the risk of under-reporting, particularly of toxicity outcomes. Variations in record keeping, assessment

and grading between clinicians would also affect the clinical and toxicity data. While hospitals may officially utilize a specific grading system, the toxicity grades were often recorded without reference to a grading system and may have been influenced by the clinician's experience with past or alternative grading systems. This was evidenced by the presence of text that matched grading criteria from earlier grading systems such as RTOG.

Epstein-Barr virus (EBV) and HPV can influence the development of NPC. Future studies may be able to investigate whether they can influence the risk of severe acute OM or dysphagia. The role of HPV status as a predictor of late dysphagia was already reported [139]. The policy for pre-treatment testing for EBV or HPV varies across institutions and is not routinely conducted for all NPC patients. This precludes the inclusion of these factors in model development. Hospitals should consider the cost of additional tests and balance this with the potential benefits. Similarly, the availability of pre-RT blood tests in the PWH dataset varied across patients, depending on their chemotherapy regimen, limiting the availability of blood test data for predictive model development. In general, pre-radiotherapy blood test results were less frequently recorded in the clinical notes unless the patient was receiving neoadjuvant chemotherapy.

Social factors, including smoking status, alcohol consumption, marital status, living conditions, financial status, education level, and family information were generally recorded during the initial consultation with clinicians. However, there was no standardized form for collecting this information, except for a nursing consultation form that was not used for all patients. Consequently, much of this information was missing, and where it was present, there

were significant variations in reporting. The lack of data validation complicated digitization of this data for model development.

### *Variability and ambiguity in clinical data from consultation notes*

Another challenge for model development is the extraction of quantitative feature data from clinical notes. As a result of clinical notes being recorded as raw text / free text, there were substantial variations in abbreviations, use of symbols, and spelling between clinicians. This led to ambiguity in the interpretation of the notes and made extracting quantitative features difficult. For example, toxicity outcomes were sometimes indicated by the name of a condition accompanied by a '+' or '++', or the negative equivalents. This could be interpreted in different ways: the positive sign could indicate the severity or the rate of change of the condition, and reference to a baseline or past consultations made interpretation difficult. Extraction of clinical data from consultation notes generally requires manual inspection and interpretation, checking the record date, considering past records, and the treatment history of the patient. In this study, the median time taken for a researcher to extract the clinical data from a single patient folder was 50 minutes. Therefore, scaling up the sample size requires significant time investment. Any attempt to automate this process would need to ensure accuracy and avoid misinterpretation of the notes. There are common aspects to the structure of each consultation note, however the notes are recorded on different types of forms with inter- and intra-clinician variation in the style of reporting.

### *Determining accurate sample size requirements*

In developing a toxicity prediction model, determination of the required sample size may be based on detection of a statistically significant improvement in the discrimination

performance compared to a reference AUC. Alternatively, the null hypothesis may be that the model discrimination is equivalent to random chance, though this does not ensure sufficient power for statistical comparison between models.

Different methods for sample size calculation for ROC analysis have been proposed in the literature [314, 315]. The equations for the sample size tend to be highly complex and depend on estimation of variance for each AUC. The resulting sample size estimate will depend not only on the expected difference in AUCs, desired significance level, and power, but also on the expected incidence of the outcome, the correlation between models, and the standard deviation within each group. Accurate estimation of the additional parameters is challenging when designing a study, especially if pilot data is not available. The expected difference in AUCs may be set at the minimum clinically significant improvement in the absence of preliminary results from pilot data, though this determination is quite subjective. In practice, the limitations on sample size are more often determined by the number of available patients per institution, and the number of institutions which can feasibly be included. But this approach is useful at the initial stages of study design or protocol development to determine the feasibility of the project.

### *Handling model selection bias*

In CHAPTER 4, CHAPTER 5, and CHAPTER 6, the optimal combination of feature types (clinical, DVH, radiomic, dosiomic, contouromic) and VOIs was investigated by performing model development and evaluation for each combination. The hyperparameters for the model pipeline were optimized for each combination, and the resulting internal and external validation scores were obtained. Higher scores indicated that the selected combination

contained more relevant, generalizable features and fewer irrelevant, non-generalizable features. However, evaluating multiple feature combinations with the same external validation set can result in implicitly tuning the feature set to match the specific properties of the external validation set, rather than obtaining models which would also be generalizable to other institutions. This is a form of selection bias and can result in over-optimistic estimates of external validation performance. The best performing models may happen to be best suited to the PWH dataset but would not be equally generalizable to other hospitals. Therefore, there was a trade-off between identifying the best feature type and VOI combination and the selection bias from multiple comparisons using the same external validation data. However, eliminating this form of bias altogether would have prevented exploration of the role of choice of VOIs and feature types on model generalizability.

### *7.2.3. Proposed solutions for improving performance*

*Multi-centre study design*

Future development of multi-omic models must be based on multi-centre studies. One reason for this is the need for larger sample sizes; there is a limit to the number of recruitable NPC cases treated at a single institution, and combining cohorts from different institutions and countries may be necessary to reach large sample sizes. Such studies will provide more reliable results with greater power and less risk of false positive or false negative errors. The effect of confounding factors can be more easily addressed in studies with larger sample sizes. Another reason for performing multi-centre studies is the need for external validation. Ideally, models would be tested on data from multiple external institutions, to better evaluate the generalizability and robustness to inter-institutional differences. Furthermore, having multi-

centre training data would facilitate the selection of more generalizable models, since performance optimization would be affected by how well the models can perform across different institutions. An example of such a study design is shown in Figure 40. However, use of multi-centre data involves additional challenges. Advanced data harmonization techniques, such as ComBat harmonization, may be necessary for overcoming the effect of differences in scan parameters on radiomic features [316]. Differences in treatment regimen between institutions also pose a significant challenge. If the development dataset has sufficient samples to capture these differences without overfitting, models may be able to learn patterns resulting from these differences in treatment regimen. For example, including three centres for model development and three centres for external validation, each with 200 cases, would provide a significant advantage for developing generalizable models while providing greater statistical power to detect a significant improvement in AUC.

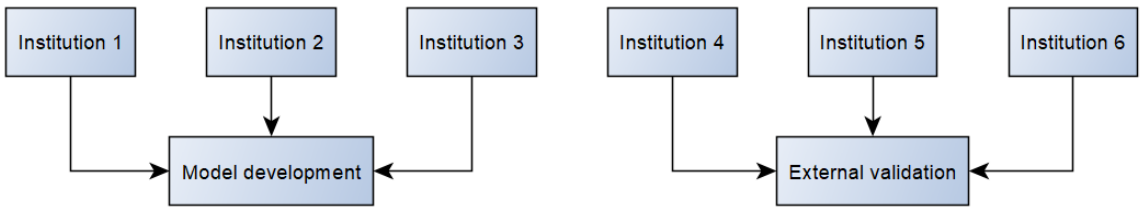



**Figure 40: Example of proposed multi-centre study design**

## *Further development of VOI auto-segmentation*

Deep learning-based auto-segmentation of VOIs is a valuable tool for obtaining a complete and consistent set of contours without requiring additional work from experts. In this work, auto-segmentation models were used to standardize the segmentation of the extended oral cavity, larynx, parotids and pharyngeal constrictor muscles. Prior to auto-segmentation,

there were significant differences in the boundaries of the contours and large differences in the voxel volume, particularly for the larynx and parotid glands. The differences in voxel volume across institutions were not statistically significant after applying auto-segmentation. The auto-segmented VOIs were visually inspected to qualitatively assess their accuracy.

One area of future development for VOI auto-segmentation is quantitative validation of the accuracy of the auto-segmentation contours. Ideally, pre-trained public models with published validation findings should be used. Ideally, if a custom model is trained, quantitative comparison should be made with example contours by clinicians. Such comparisons should include measures of overlap such as the Dice coefficient, and measures of inter- or intra- rater variability using the ICC.

Another area of future development of VOI auto-segmentation is to expand the set of VOIs beyond those included in the guidelines for RT planning for NPC. It may be possible to further optimize the VOIs to be more specific and relevant to discrimination of the risk of severe toxicities. However, because of the limited number of published studies on the prediction of severe OM and dysphagia, there is not a strong consensus on the optimal VOI delineation that should be used. Exploration of a range of VOI variations could be explored. For example, sub-dividing the extended oral cavity into its sub-regions, or including mucosal surface contours of different thicknesses, or converting 2D features from the surface contour. A study on prediction of OM extracted DVH features from a 3mm thick oral mucosa contour [55]. Such a contour may provide more specific information than a VOI including all the oral cavity or extended oral cavity. However, extraction of multi-omic features would involve some additional considerations. Firstly, many shape features would be redundant with the whole-

volume equivalent. Secondly, the extraction of textural features would need to be carefully justified. The thickness of the surface contour would impact on the calculation and interpretation of 3D texture features. Alternatively, the contour could be unfolded into a 2D region-of-interest (ROI) for 2D texture feature extraction. Contouring thin surfaces can involve additional challenges. The surface may be on the boundary of tissue changes, which could increase the impact of inaccurate contouring and reduce the stability of features to inter-observer variation or simulated perturbations to the contours. Resegmentation based on the HU range of the tissues of interest could be applied to address this, but this might lead to a fragmented VOI in the case of inaccurate contours. The use of surface VOIs may hold potential, but careful consideration, clinical justification and review by experts are recommended.

### *Improvements to model development methodology*

While this project aimed to conduct model development carefully and avoid information leakage to the validation set, there remain further improvements which could be implemented in future. One of these pertains to the model selection process, where the machine learning algorithm with the highest discrimination score across training, internal validation, and external validation was selected. Bias could be further minimized by ensuring that model selection is conducted within the model optimization stage and without reference to the external validation score. The external validation score should ideally be evaluated only once after the final model has been selected based on its internal validation performance. This approach ensures that the reported generalization performance reflects the model's true predictive ability, rather than a model that appears to perform better on the external validation set due to random fluctuations or because it has been inadvertently tailored to the specific institution providing

the validation data. However, selecting the model based on the test set is still frequently performed in the literature [317]. Selecting the final model based on the internal validation performance also would require further minimizing the bias in the internal validation score. This results from cross-validation, where the same data is used to identify the best set of hyperparameters and estimate the generalizability of the model on unseen data. Nested cross-validation can reduce this type of bias, albeit at a substantially higher computational cost. With a sufficiently large dataset, the number of folds might be restricted, allowing for a less drastic increase in the computational cost. However, with a limited sample size, 3-fold or 5-fold nested cross validation would leave insufficient data for training, thereby limiting the model's ability to identify patterns. This project only consisted of a single development dataset and a single validation dataset. Exploring different combinations of preprocessing settings, such as the threshold for the hierarchical clustering and feature stability ICC, might lead to some bias in the results. The chosen settings might be optimized for those particular datasets, or yield better results due to statistical fluctuations, but they may not be generalizable. This issue should be less impactful if the datasets included training and testing data each from multiple centres.

### *Further refinement of image perturbation feature stability assessment*

Feature stability is essential for reproducible, reliable models. The simulated perturbations approach has been proposed as a method to assess feature stability without the need for employing multiple different experts to assess inter-observer variability or taking additional scans to assess test-retest reproducibility [266]. However, as seen in **Section 3.7**, the feature stability varied across VOIs. This may be inherent to the challenges of contouring the VOI, but conversely there is the possibility that the perturbation parameters may not be optimal

and require further fine-tuning to fit each VOI. Future studies would benefit from a detailed comparison between features extracted from manual contours by different experts and features extracted from a given set of perturbation settings. This would ensure that only unstable features were excluded.

The study that introduced the image perturbation approach for radiomic feature robustness assessment also proposed a second method of harnessing the perturbations [266]. In this approach, the mean values of the robust features across all perturbations are used for modelling, rather than the features extracted from the original image and VOI. The authors acknowledged that this approach requires calculation of perturbations for all images in the model development cohort and is therefore more computationally expensive. This approach could, however, improve the robustness of the features included in model development.

In the study by Teng et al., the features extracted from the perturbations were treated as separate internal validation cohorts and were used to validate the reliability of the model [318]. The ICC was also calculated across model prediction outcomes to assess the consistency of models across perturbations. This approach could be beneficial in future studies to further assess the model reliability, though comparisons should be made to similar approaches conducted using bootstrapping to determine whether the additional computational cost is merited.

Finally, image perturbations could be used for data augmentation for model development. Since sample size is frequently reported as a limitation in radiomic, dosiomic or contouromic studies, this approach could be used to increase the number of training samples.

Each perturbation could be treated as a separate patient with different multi-omic feature values. Clinical features could be handled by repeating values for binary features and by adding noise for continuous features. This approach would likely result in reduced variability in the development data because of the lower variability between perturbations compared to that between patients. Performance estimates based on the augmented data would therefore be less reliable. However, if sufficient external validation data were available, this approach might help to mitigate the challenges of limited sample sizes, allowing for more subtle patterns to be detected and more complex relationships to be modelled.

### *Comprehensive reporting and data sharing*

In addition to having individual studies with high quality and high levels of evidence, the development of multi-omics for prediction of severe OM and dysphagia would benefit from having consensus between independently conducted studies. Such studies could perform validation of existing models or produce independent findings. In either case, comprehensive reporting is of critical importance to achieve reproducible and transparent research. Insufficient detail in the methodology or results evaluation prevents adequate critique and comparison across the literature. Sharing of imaging, clinical, multi-omic or model data, is often recommended in guidelines such as the CheckList for EvaluAtion of Radiomics research (CLEAR) [227]. It would facilitate independent validation and increase the availability of data for model development. However, sharing of imaging data may not be possible for patient privacy reasons, unless sufficient removal of identifying features is conducted. Feature data should be less affected by privacy concerns; however, authors may be reluctant to make the first generous step, over concerns that other authors may not follow. Research groups may also

be inclined to retain control over their datasets to avoid facilitating competition with other groups on the same research topics. Public datasets remain a useful tool for researchers. Such datasets may be published by government organizations who wish to promote greater cooperation and research activity using local data.

### *Ensuring accurate and comprehensive toxicity outcome data*

Ensuring accurate and comprehensive toxicity outcome data is essential. Toxicity grading must be consistent and adhere to a specific version of a grading system, as criteria can vary over time with different emphases on visual presentation or functional impact. Additionally, it is crucial to ensure that there is sufficient follow-up, covering the relevant period. There is a risk of under-estimating toxicity outcomes when consultation records are insufficient. This issue may stem from data collection practices, limitations of hospital record keeping, or patients repeatedly missing scheduled consultations. Identifying patients with such missing data is advisable to ensure a representative dataset and avoid diluting the incidence of severe toxicity outcomes. Dean et al. discussed a method to mitigate this under-reporting [131]. Prospective studies may offer greater potential to standardize assessment of toxicities within and across institutions. A more standardized assessment schedule could facilitate more in-depth analysis of the time evolution of OM and dysphagia, providing an additional dimension to capture the complex interaction between these toxicities.

### *Inclusion of genetic and biological information*

Increasing the scope of data collection to include genetic information is advisable, given the evidence for the role of particular SNPs in connection with severe OM and dysphagia [67, 85, 104, 115, 126, 127, 217]. This may facilitate radiogenomic analysis, identifying

correlations between radiomic and genomic features to give a more holistic understanding of toxicity and provide a biological explanation for the characterization of tissue provided by radiomics. This would require a prospective study unless a cohort of retrospectively recruited patients had already been enrolled in another trial where genomic data had been collected. In this case, there would be a risk of selection bias from the inclusion criteria of the prior study, if it differed from the intended inclusion criteria. Likewise, comprehensive collection of blood test results or saliva test results, which have also been associated with severe OM and dysphagia [79, 81, 123, 203], would be best achieved in a prospective study design, because in standard clinical practice, the clinical laboratory tests were performed on patients based on their specific needs and conditions. Similarly, the investigation of the role of EBV and HPV in the development of severe toxicity could be further explored in prospective studies, where all included patients could be tested before treatment. In standard clinical practice, this was not the case.

### *More comprehensive collection of patient-related factors*

A more in-depth collection of patient factors would be desirable. Data on pre-treatment performance status was sparse, but this indication of the general physical condition of the patient was reported as a predictive factor for severe OM and dysphagia by several studies [128, 130]. Prospective studies could collect this data as well as data on the pre-treatment dental condition and oral health, such as number of teeth, which have been reported in connection with OM. Quantification of pre-treatment appetite, eating habits, and difficulty swallowing is challenging, but this would be advisable. It would represent a set of risk factors that could be associated with higher chance of severe dysphagia. Subjective factors such as pain perception

could also be measured, providing a link between psychology and development of severe toxicity.

### *Greater characterization of chemotherapy regimen*

A limitation of this work was the lack of in-depth characterization of the chemotherapy regimen received by patients. Specifically, analysis of the differing impact of neoadjuvant and adjuvant chemotherapy, and the role of different chemotherapy drugs, doses, and numbers of cycles. While models did account for the use of chemotherapy versus radiotherapy alone, the significant differences in the usage of neoadjuvant and adjuvant chemotherapy in each centre precluded the inclusion of these features, due to the imbalance between centres and the resulting lack of generalizability of these features. With a multi-centre development cohort, more generalizable models incorporating further chemotherapy characterization could be developed. Information on drug type, dosage, and numbers of cycles was stored in the form of handwritten text that was often very difficult to interpret and had a high level of complexity, precluding straightforward conversion into categorical or continuous feature data. Standardization of chemotherapy regimen is infeasible, since the differences have important clinical justifications, such as using carboplatin instead of cisplatin for patients with reduced renal function. Larger sample sizes would enable development of more complex models that could incorporate a wider range of features characterizing chemotherapy regimen.

### *Inclusion of radiomic features from MRI*

Extraction of radiomic features from magnetic resonance imaging (MRI) may offer additional predictive value. MRI are often acquired prior to RT and are inspected when determining grading and performing VOI segmentation. However, for patients included in this

study, the MRI were typically acquired several weeks before the planning CT, and patients generally received their MRI before neoadjuvant chemotherapy, if applicable. There may therefore be additional morphological changes in the tumour due to this time gap and from the impact of neoadjuvant chemotherapy on weight loss and tumour size. Additionally, MRI were acquired in a different patient position, without immobilization, and there were significant geometric differences between MRI and planning CT. Furthermore, geometric distortion is inherent to MRI. Therefore, extraction of MRI radiomic features necessitates careful geometric registration and checking of each VOI, which would require clinical expertise or a specially trained and validated AI model. Additionally, there may be variation across patients and across institutions in the specific MRI sequences and parameters used for acquisition, resulting in greater challenges in obtaining stable and robust radiomic features. The enhanced soft tissue contrast offered by MRI might offer advantages for model development, though there would also be a trade-off with the resulting increase in dimensionality from having a larger number of extracted features. Exploration of the role of MRI is desirable, provided that these challenges can be overcome.

### *Development of models with prospective data*

Prospective studies could facilitate more structured data collection and enable easier data validation and conversion into categorical or continuous features. Clinicians could be provided with specialized forms during consultations, with separate multi-point scales for assessing toxicities, diet, performance status score and other factors, as well as space to record continuous valued features like weight. If the form were provided in a digital format, then data validation could be provided to prevent variations in spelling and minimize the impact of typos.

Implementation of such a form would need to consider the clinical resources required, such as the clinician's time. Additionally, questionnaires could be provided to obtain information on patient reported toxicity and quality-of-life outcomes. If clinical resources permit, more comprehensive assessment of dysphagia could be conducted using video fluoroscopy swallowing studies or endoscopic evaluation. These would provide more direct assessment of the mechanism of swallowing, in greater isolation from the impact of pain, discomfort or appetite. Prospective studies would also allow the scope of data collection to be expanded include genetic and biological information, as discussed previously.

### *Use of centralized medical data services*

Increasingly, countries are investing in centralized medical data services where 'big data' from hospitals can be stored and analysed under one system. This follows the move towards digitalization of healthcare records. Such services offer a huge potential for the development of AI models, including machine learning-based multi-omic models. Such services offer access to imaging and clinical data from multiple centres, addressing many of the issues raised in this chapter. In Hong Kong, this service is provided by the Hospital Authority Data Collaboration Lab [319]. However, patient privacy and data security is paramount, and there are consequently several measures in place to safeguard the data. Such measures include only allowing data access at specified secure locations. This means that all exploratory analysis, data collection, preprocessing, modelling, and evaluation would typically have to be performed on site, with only aggregate data such as mean performance scores or mean clinical feature values being able to be exported out of the secure site. In practice, this would necessitate many visits to the site by research personnel with strong programming and

database management skills, or extensive preparation of research tools that could facilitate data cleaning and analysis. Such an endeavour would be best achieved by a medium to large research group over a longer time frame. The project would need a highly thorough and detailed protocol, extensive ethics approval steps, and enough time to meet the administrative requirements of the Hospital Authority. These limiting factors have been discussed by local researchers and there may be some future developments to further facilitate and streamline the process of conducting such projects, to better utilize the vast amounts of available data for increased research and development output.

## 7.3. How can multi-omic models be implemented in clinical practice?

### *7.3.1. Requirements for published literature*

Strong evidence of the performance of multi-omic models is critical for any move towards clinical implementation. Such evidence would come from prospective studies, or even better, randomized controlled trials. These types of studies would reduce the risk of selection bias and information bias, and demonstrate a higher level of evidence, while also allowing for greater control of confounders. Establishing this level of evidence would provide confidence that the proposed models can accurately predict severe OM and dysphagia.

The aim of developing predictive models for severe acute OM and dysphagia is to facilitate targeted interventions for prevention and management. In addition to improving model discrimination, the benefit of the intervention must also be demonstrated. This depends on the type of intervention proposed. Photobiomodulation therapy (PBMT), also known as Low

level laser therapy (LLLT), is an intervention proposed for the prevention and treatment of OM. A systematic review and meta-analysis found a beneficial effect of PBMT on OM in HNC, though the authors recommended that future studies should investigate the most effective parameters for PBMT in the management of OM [320]. While PBMT is thought to be safe, there would be costs associated with equipment and hospital personnel to deliver the treatment, as well as the burden on patient in terms of time spent during treatment. Clinicians would need to assess the efficacy, cost, and benefits of a proposed treatment once a set of optimal parameters is determined. PBMT may also have a secondary benefit in reducing the severity of acute dysphagia by tackling the OM symptoms. Additionally, speech and language therapy has been proposed for managing dysphagia [321]. The costs and benefits of providing these services to patients would need to be assessed, considering the expected performance of the prediction model and comparing it to the treat-all and treat-none approaches.

A further consideration is the explainability and interpretability of the model. An advantage of the multi-omics approaches used in this project is that, unlike deep learning 'black box' models, each feature has a pre-defined mathematical definition that corresponds to a known property. However, the interpretation of radiomic, dosiomic or contouromic features can still be difficult, particularly in the case of second-order textural features. Clinicians may be reluctant to employ a model which does not have an intuitive explanation, or which does not use omics which have an established close connection to biology, such as genomics or proteomics. Future development of predictive models for toxicity may involve a broader set of omics, where connections between different omics may provide a more holistic understanding.

For example, radiogenomics promises to uncover links between macro-scale imaging features and micro-scale genetic properties.

While this project focused on acute OM and dysphagia, it is important to acknowledge that late and chronic toxicity are also important, for their impact on quality-of-life post-treatment. For a model to be considered for clinical implementation, it would also be helpful for it to include some investigation of late toxicity. The link between acute and late toxicity has been reported for dysphagia [213], and future studies may be able to further characterize the link between the two.

### *7.3.2. Aspects of implementation*

Clinical implementation must allow for the prediction of the risk of severe OM and dysphagia without incurring significant clinical time or resources beyond the standard of care. As such, the implementation must be in the form of integrated software requiring minimal input from the clinician. The software must be able to access the medical imaging data (CT or MRI) as well as the dose distribution from the treatment planning system. Any required VOI contours from RT planning should also be accessible. Next, any automatic segmentation would be performed, features would be extracted, and models would be evaluated. The predicted probability of severe OM and dysphagia could then be displayed for the clinician, along with the recommended intervention based on previous risk and cost analysis. It is unlikely that these model predictions could be used to further refine RT plans to improve dose sparing, because well-established dose sparing guidelines already exist, and any change would need to be well

justified. The priority of RT planning is to meet the required dose to the tumour along with the dose constraints, to prioritise the patient's survival.

A clinically implemented model should continue to be updated based on new data. The model may be further calibrated and fine-tuned for that specific institution. Alternatively, a federated learning approach could be employed, where models implemented at different institutions each send back model weight updates to further improve the master model. Any such approach must ensure that the security of hospital networks is maintained, and patient privacy is respected by avoiding any identifying information being at risk of interception.

## 7.4. Predicting other treatment-induced toxicities

Multi-omic models have been published for some other treatment-induced toxicities in HNC, as identified in **Section 1.3.3**. These include models for predicting late xerostomia, acute xerostomia, and late hypothyroidism. It would be particularly valuable to incorporate xerostomia into a multi-label toxicity prediction model for severe acute OM and dysphagia, because of the impact of saliva production on these conditions. Multi-omics might also offer potential for the prediction of dysgeusia, radiation dermatitis, hearing loss, and osteoradionecrosis. Inclusion of other toxicities was not possible in this project because of the limitations in the retrospectively collected consultation records. For many of these toxicities, severity gradings were not regularly recorded. Instead, there was reference to the condition accompanied by an indication of whether it was worsening or improving. This was insufficient for determining a severity cutoff. Prospective studies could ensure a standardized assessment during follow-up, and could include additional measurements such as hearing tests,

photographs of radiation dermatitis symptoms, and patient-reported questionnaires on dysgeusia.

# CHAPTER 8 CONCLUSIONS

Severe acute OM and dysphagia, as two of the most common and devastating treatment-induced toxicities for NPC patients, pose a significant adverse impact on patient quality of life, as well as threatening treatment outcome because of pain, weight loss and treatment interruption. With the improvement in survival rates for NPC, it is increasingly important to address the burden of toxicity on patients. Accurate prediction of patients at high risk of severe toxicity should enable more personalised and targeted prevention and management strategies.

In Chapter 1, a detailed literature review was conducted to identify risk factors for severe acute OM and dysphagia. Additionally, published prediction models for each toxicity were identified. Most models lacked external validation, and multi-omic features were under-explored. Furthermore, many studies were developed on mixed cohorts of multiple HNCs, with very few highlighting the unique challenges within NPC. A literature review on the use of multi-omics for toxicity prediction in head and neck cancer confirmed the research gap. This informed the aim and objectives, as outlined in Chapter 2.

In Chapter 3, a comprehensive methodology for model development and assessment was devised, including important steps such as assessment of feature stability, as recommended by the CLEAR guidelines [227].

Chapter 4 reported the development of multi-omic models for pre-treatment prediction of severe acute OM in NPC patients undergoing RT. To our best knowledge, these represent the only externally validated prediction models for severe acute OM to utilize radiomic, dosiomic or contouromic features. Multi-omic models outperformed models developed using

conventional clinical and dosimetric features on the same dataset, and also outperformed the only externally validated model for severe acute OM in the literature. The results suggested that multi-omic features hold predictive value for severe acute OM independently of clinical and dosimetric features and can contribute to improved discrimination.

Chapter 5 reported the development of multi-omic models for pre-treatment prediction of severe acute dysphagia in NPC patients undergoing RT. To our best knowledge, no full-length articles reporting radiomic, dosiomic or contouromic models for severe acute dysphagia have been published. Multi-omic models outperformed models developed using conventional clinical and dosimetric features on the same dataset but did not surpass the performance of the models reported in the literature. Nevertheless, the results suggested that multi-omic features hold predictive value for severe acute dysphagia independently of clinical and dosimetric features and can contribute to improved discrimination, though to a lesser extent than for OM.

Severity of OM can contribute to higher risk of severe dysphagia through its impact on pain during swallowing. Chapter 6 reported the development of multi-label prediction models for severe acute OM and dysphagia with the objective of further improving discrimination by harnessing information about the relationship between the two toxicities. To the extent of our understanding, this work was the first to report multi-label models for predicting multiple toxicities resulting from radiotherapy. While the multi-label models did not outperform the top-scoring models from Chapter 4 and Chapter 5, moderate discrimination scores were achieved in a proof-of-concept. Furthermore, evidence for the role of contouromics was highlighted by the results in accordance with findings from earlier chapters, providing further justification for future research into the role of features that quantify patient geometry.

Limitations identified in this study and related work highlight significant challenges in developing prediction models for severe acute OM and dysphagia. Chapter 7 provided a comprehensive discussion of practical considerations for future research, identifying the key challenges in improving model discrimination and proposing solutions based on the experiences gained during the development of this thesis. Additionally, special consideration was given to the key challenges involved in moving towards clinical implementation.

Further research is required to improve the performance and level of evidence of multi-omic prediction models for severe acute OM and dysphagia prior to their clinical implementation. This thesis serves as a critical foundation for future research towards achieving the aim of targeting preventative interventions and personalized management to patients at high risk of severe toxicities. The comprehensive assessment of the related literature, risk factors, published prediction models, and different approaches to model development, as well as the discussion of the limitations, proposed solutions and recommendations for future research, provide invaluable insights for the design of future studies.

# APPENDIX

# Search strategy for literature review

**Table 58: Search strategy for Section 1.2**

| Database | Search String | N results | Date |
|---|---|---|---|
| Embase | **TITLE** = (toxicit* OR morbidit* OR "side effect*" OR mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR aspiration OR stricture* OR gastronom* OR "oral intake") AND (predict* OR model* OR correlat* OR corresp* OR depend* OR assoc* OR relation* OR interact* OR link* OR "risk factors")<br>**TI/AB/KW** = (mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR gastronom* OR "oral intake") AND (radiation OR chemotherap* OR radiotherap* OR chemoradiation OR chemoradiotherap* OR radiochemotherap* OR pharmacotherap* OR "IMRT" OR "VMAT" OR "3DCRT" OR "CRT")<br>Publication Year >= 2000<br>Abstract OR Article | 1093 | 9/2023 |
| PubMed | ((("Stomatitis"[Mesh] OR "Deglutition Disorders"[Mesh]) AND ("Radiotherapy"[Mesh] OR "Drug Therapy"[Mesh]))) AND ((predict*[Title] OR model*[Title] OR correlat*[Title] OR corresp*[Title] OR depend*[Title] OR assoc*[Title] OR relat*[Title] OR interact*[Title] OR link*[Title] OR "risk*"[Title]))<br>FILTERS<br>Publication Year >= 2000 | 553 | 9/2023 |
| Scopus | **TITLE** = (toxicit* OR morbidit* OR "side effect*" OR mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR aspiration OR stricture* OR gastronom* OR "oral intake") AND (predict* OR model* OR correlat* OR corresp* OR depend* OR assoc* OR relation* OR interact* OR link* OR "risk factors")<br>**TI/AB/KW** = (mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR gastronom* OR "oral intake") AND (radiation OR chemotherap* OR radiotherap* OR chemoradiation OR chemoradiotherap* OR radiochemotherap* OR pharmacotherap* OR "IMRT" OR "VMAT" OR "3DCRT" OR "CRT")<br>Publication Year >= 2000<br>Abstract OR Article | 701 | 9/2023 |
| Web of Science | **TITLE** = (toxicit* OR morbidit* OR "side effect*" OR mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR aspiration OR stricture* OR gastronom* OR "oral intake") AND (predict* OR model* OR correlat* OR corresp* OR depend* OR assoc* OR relation* OR interact* OR link* OR "risk factors")<br>**TOPIC** = (mucositis OR dysphagia OR deglutition OR swallow* OR "tube feed*" OR ryle* OR enteral OR nasogastric OR intubation OR gastronom* OR "oral intake") AND (radiation OR chemotherap* OR radiotherap* OR chemoradiation OR chemoradiotherap* OR radiochemotherap* OR pharmacotherap* OR "IMRT" OR "VMAT" OR "3DCRT" OR "CRT")<br>Publication Year >= 2000<br>Abstract OR Article | 797 | 9/2023 |

# Pearson correlation heatmaps for most frequently selected features for severe OM and dysphagia prediction

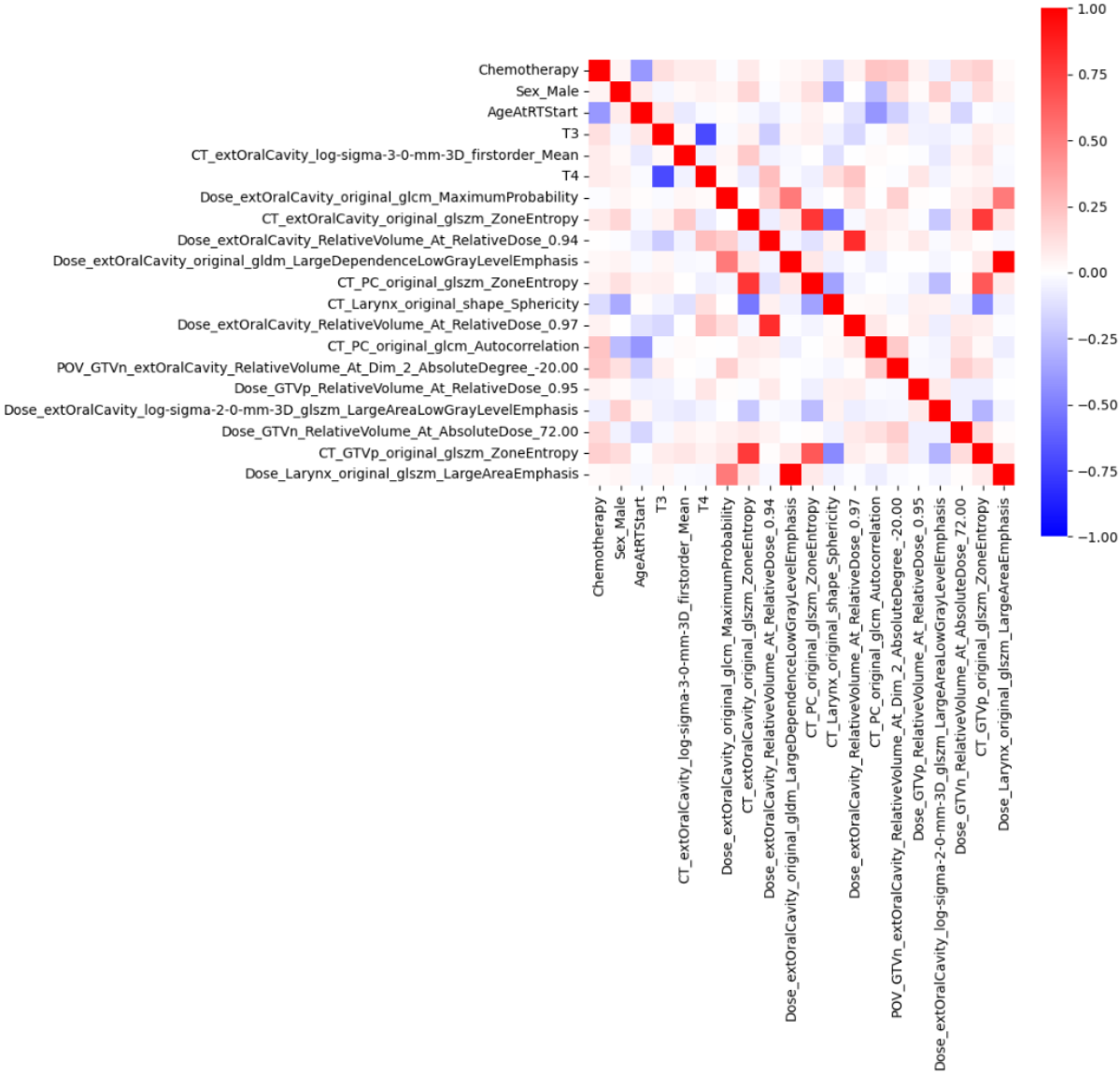



**Figure 41: Pearson correlation coefficients for the most frequently selected features in the top 5% of models for severe acute OM in the development dataset**

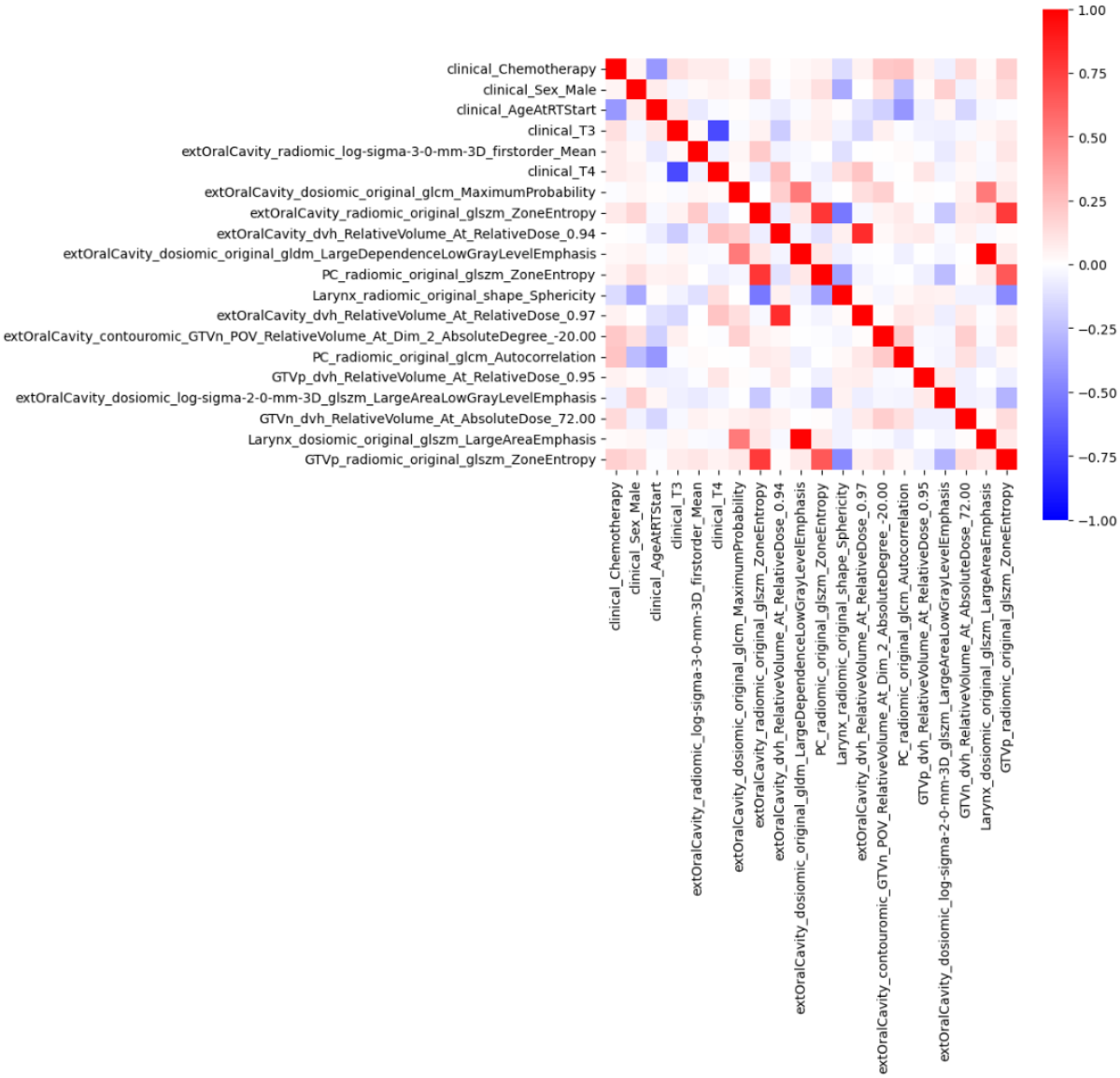



**Figure 42: Pearson correlation coefficients for the most frequently selected features weighted by model AUC in the top 5% of models for severe acute OM in the development dataset**

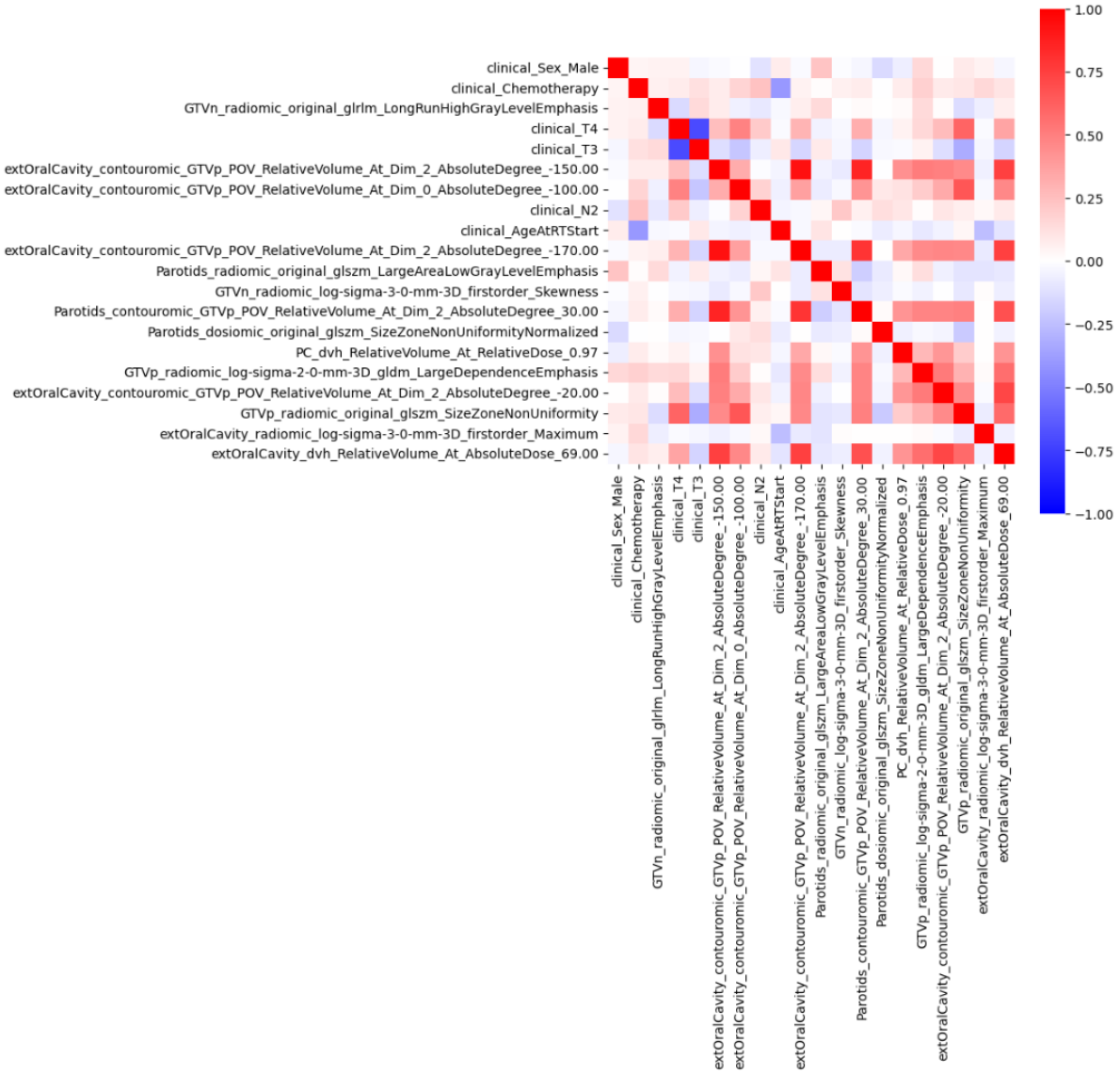



**Figure 43: Pearson correlation coefficients for the most frequently selected features in the top 5% of models for severe acute dysphagia in the development dataset**

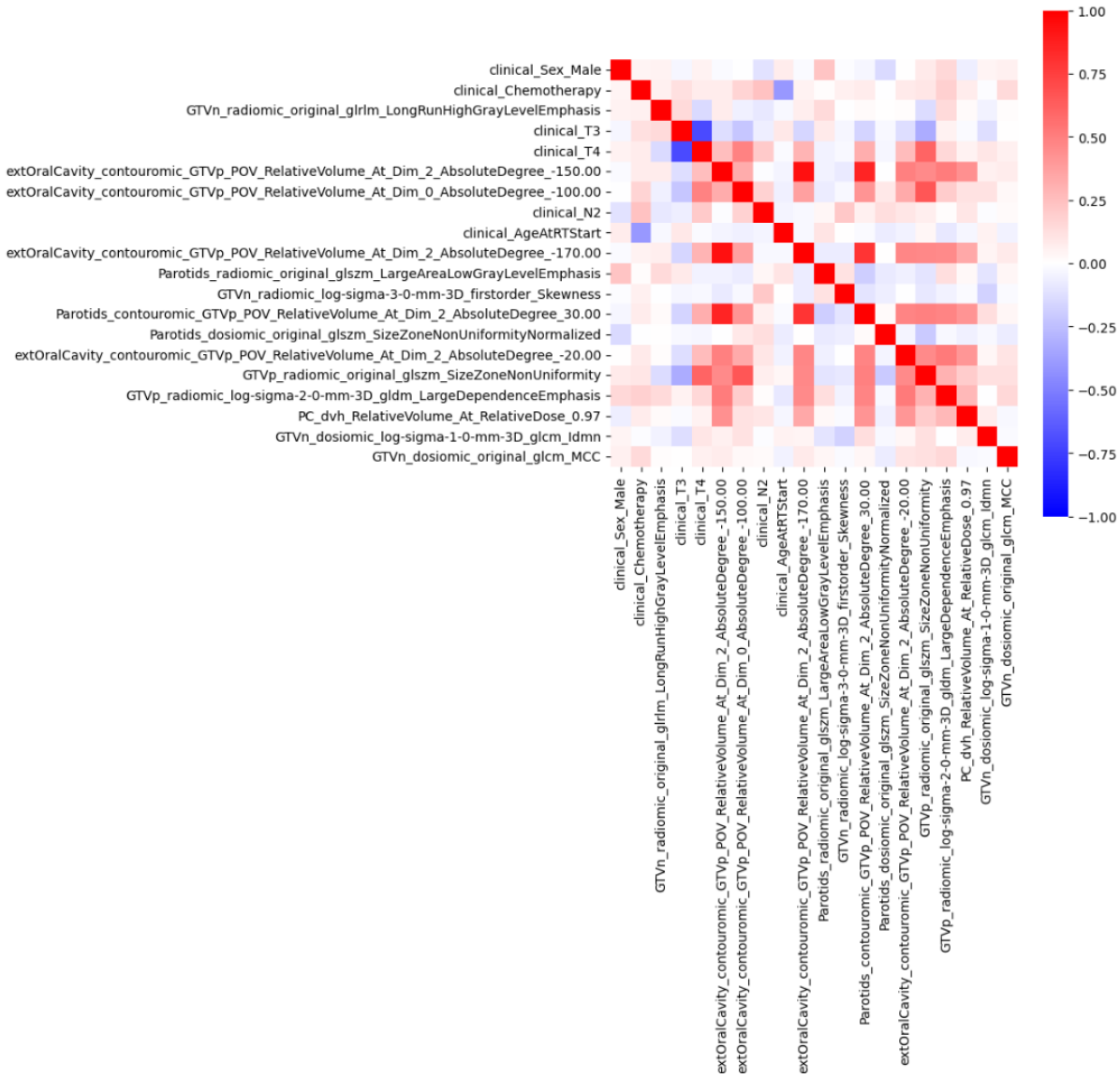



**Figure 44: Pearson correlation coefficients for the most frequently selected features weighted by model AUC in the top 5% of models for severe acute dysphagia in the development dataset**

# Multi-label model settings

**Table 59: Top multi-label model settings for label powerset approach**

| | |
|---|---|
| **Initial feature set** | Clinical, radiomic, contouromic |
| **VOIs** | Extended oral cavity, parotid glands |
| **N features after ICC filter and VIF clustering** | 23 |
| **MRMR K** | 4 |
| **Model** | Gaussian Naïve Bayes<br>Var_smoothing = 1e-9 |

**Table 60: Top multi-label model settings for classifier chain approach with shared feature selection**

| | |
|---|---|
| **Initial feature set** | Clinical, radiomic, contouromic |
| **VOIs** | Extended oral cavity |
| **N features after ICC filter and VIF clustering** | 18 |
| **MRMR K** | 4 |
| **Model** | Gaussian Naïve Bayes<br>Var_smoothing = 1e-9 |

**Table 61: Top multi-label model settings for classifier chain approach with separate feature selection**

| | |
|---|---|
| **Initial feature set** | Clinical, contouromic |
| **VOIs** | Extended oral cavity |
| **N features after ICC filter and VIF clustering** | 23 |
| **MRMR K** | 3 |
| **Model** | Ridge regression<br>Class weights = balanced<br>C = 1 |

# REFERENCES


1. **Nicol, A. J.; Ching, J. C. F.; et al.**, Predictive Factors for Chemoradiation-Induced Oral Mucositis and Dysphagia in Head and Neck Cancer: A Scoping Review. Cancers, **2023**. 15(23): p. 5705.
2. **Sung, H.; Ferlay, J.; 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): p. 209-249.DOI: https://doi.org/10.3322/caac.21660.
3. **Gormley, M.; Creaney, G.; et al.**, Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. British Dental Journal, **2022**. 233(9): p. 780-786.DOI: 10.1038/s41415-022-5166-x.
4. **Johnson, D. E.; Burtness, B.; et al.**, Head and neck squamous cell carcinoma. Nature Reviews Disease Primers, **2020**. 6(1): p. 92.DOI: 10.1038/s41572-020-00224-3.
5. **Amin, M. B.; Edge, S. B.; et al.**, AJCC Cancer Staging Manual. 8 ed. **2017**: Springer Cham.
6. **Osazuwa-Peters, N.; Simpson, M. C.; et al.**, Suicide risk among cancer survivors: Head and neck versus other cancers. Cancer, **2018**. 124(20): p. 4072-4079.DOI: https://doi.org/10.1002/cncr.31675.
7. **Tang, L.-L.; Chen, Y.-P.; et al.**, The Chinese Society of Clinical Oncology (CSCO) clinical guidelines for the diagnosis and treatment of nasopharyngeal carcinoma. Cancer Communications, **2021**. 41(11): p. 1195-1227.DOI: https://doi.org/10.1002/cac2.12218.
8. **Chen, Y.-P.; Chan, A. T. C.; et al.**, Nasopharyngeal carcinoma. The Lancet, **2019**. 394(10192): p. 64-80.DOI: https://doi.org/10.1016/S0140-6736(19)30956-0.
9. **Hong Kong Cancer Registry**, Overview of Hong Kong Cancer Statistics of 2021. **2021**.
10. **Lv, J. W.; Huang, X. D.; et al.**, A National Study of Survival Trends and Conditional Survival in Nasopharyngeal Carcinoma: Analysis of the National Population-Based Surveillance Epidemiology and End Results Registry. Cancer Res Treat, **2018**. 50(2): p. 324-334.DOI: 10.4143/crt.2016.544.
11. **Liu, Q.; Chen, J. O.; et al.**, Trends in the survival of patients with nasopharyngeal carcinoma between 1976 and 2005 in Sihui, China: a population-based study. Chin J Cancer, **2013**. 32(6): p. 325-33.DOI: 10.5732/cjc.012.10189.
12. **Sun, X.-S.; Liu, D.-H.; et al.**, Patterns of Failure and Survival Trends in 3,808 Patients with Stage II Nasopharyngeal Carcinoma Diagnosed from 1990 to 2012: A Large-Scale Retrospective Cohort Study. crt, **2019**. 51(4): p. 1449-1463.DOI: 10.4143/crt.2018.688.
13. **Huang, W.-Y.; Lin, C.-L.; et al.**, Survival outcome of patients with nasopharyngeal carcinoma: a nationwide analysis of 13 407 patients in Taiwan. Clinical Otolaryngology, **2015**. 40(4): p. 327-334.DOI: https://doi.org/10.1111/coa.12371.
14. **Mao, Y. P.; Tang, L. L.; et al.**, Prognostic factors and failure patterns in non-metastatic nasopharyngeal carcinoma after intensity-modulated radiotherapy. Chin J Cancer, **2016**. 35(1): p. 103.DOI: 10.1186/s40880-016-0167-2.
15. **Hong Kong Cancer Registry**. Top Ten Cancers. **2021** 11/2023]; Available from: https://www3.ha.org.hk/cancereg/topten.html.
16. **Census and Statistics Department**. Summary results of 2021 Population Census. **2022** Mar 2024]; Available from: https://www.censtatd.gov.hk/en/press_release_detail.html?id=5156.
17. **Chow, J. C. H.; Tam, A. H. P.; et al.**, Second primary cancer after intensity-modulated radiotherapy for nasopharyngeal carcinoma: A territory-wide study by HKNPCSG. Oral Oncology, **2020**. 111: p. 105012.DOI: https://doi.org/10.1016/j.oraloncology.2020.105012.
18. **Ratko, T. A.; Douglas, G. W.; et al.**, Radiotherapy Treatments for Head and Neck Cancer Update, in Radiotherapy Treatments for Head and Neck Cancer Update. **2014**, Agency for Healthcare Research and Quality (US): Rockville (MD).
19. **Gianfaldoni, S.; Gianfaldoni, R.; et al.**, An Overview on Radiotherapy: From Its History to Its Current Applications in Dermatology. Open Access Maced J Med Sci, **2017**. 5(4): p. 521-525.DOI: 10.3889/oamjms.2017.122.
20. **Wang, K. and Tepper, J. E.**, Radiation therapy-associated toxicity: Etiology, management, and prevention. CA Cancer J Clin, **2021**. 71(5): p. 437-454.DOI: 10.3322/caac.21689.
21. **Pomeroy, A. E.; Schmidt, E. V.; et al.**, Drug independence and the curability of cancer by combination chemotherapy. Trends Cancer, **2022**. 8(11): p. 915-929.DOI: 10.1016/j.trecan.2022.06.009.
22. **Brown, T. J. and Gupta, A.**, Management of Cancer Therapy-Associated Oral Mucositis. JCO Oncol Pract, **2020**. 16(3): p. 103-109.DOI: 10.1200/JOP.19.00652.
23. **Cox, J. D.; Stetz, J.; et al.**, Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European organization for research and treatment of cancer (EORTC). International Journal of Radiation

Oncology*Biology*Physics, **1995**. 31(5): p. 1341-1346.DOI: https://doi.org/10.1016/0360-3016(95)00060-C.
24. **Brook, I.**, Late side effects of radiation treatment for head and neck cancer. Radiat Oncol J, **2020**. 38(2): p. 84-92.DOI: 10.3857/roj.2020.00213.
25. **Pulito, C.; Cristaudo, A.; et al.**, Oral mucositis: the hidden side of cancer therapy. Journal of Experimental & Clinical Cancer Research, **2020**. 39(1): p. 210.DOI: 10.1186/s13046-020-01715-7.
26. **Li, J.; Zhu, C.; et al.**, Incidence and Risk Factors for Radiotherapy-Induced Oral Mucositis Among Patients With Nasopharyngeal Carcinoma: A Meta-Analysis. Asian Nurs Res (Korean Soc Nurs Sci), **2023**. 17(2): p. 70-82.DOI: 10.1016/j.anr.2023.04.002.
27. **Lalla, R. V.; Sonis, S. T.; et al.**, Management of oral mucositis in patients who have cancer. Dent Clin North Am, **2008**. 52(1): p. 61-77, viii.DOI: 10.1016/j.cden.2007.10.002.
28. **Pixabay**. [Image] A close up of a person with their mouth open. Published under Creative Commons license. **2016** [cited 2024; Available from: https://picryl.com/media/dentist-mouth-open-mouth-726fd3.
29. **Elad, S.; Cheng, K. K. F.; et al.**, MASCC/ISOO clinical practice guidelines for the management of mucositis secondary to cancer therapy. Cancer, **2020**. 126(19): p. 4423-4431.DOI: 10.1002/cncr.33100.
30. **Matsuo, K. and Palmer, J. B.**, Anatomy and Physiology of Feeding and Swallowing: Normal and Abnormal. Physical Medicine and Rehabilitation Clinics of North America, **2008**. 19(4): p. 691-707.DOI: https://doi.org/10.1016/j.pmr.2008.06.001.
31. **State Board Colorado Community College System**. [Image] Digestive Structures and Functions. (Cenveo - Creative Commons license). Anatomy and Physiology [cited 2024; Available from: https://pressbooks.ccconline.org/bio106/chapter/digestive-structures-and-functions/.
32. **Schindler, A.; Denaro, N.; et al.**, Dysphagia in head and neck cancer patients treated with radiotherapy and systemic therapies: Literature review and consensus. Critical Reviews in Oncology/Hematology, **2015**. 96(2): p. 372-384.DOI: https://doi.org/10.1016/j.critrevonc.2015.06.005.
33. **Al-Othman, M. O.; Amdur, R. J.; et al.**, Does feeding tube placement predict for long-term swallowing disability after radiotherapy for head and neck cancer? Head Neck, **2003**. 25(9): p. 741-7.DOI: 10.1002/hed.10279.
34. **Nguyen, N. P.; Moltz, C. C.; et al.**, Dysphagia following chemoradiation for locally advanced head and neck cancer. Annals of Oncology, **2004**. 15(3): p. 383-388.DOI: https://doi.org/10.1093/annonc/mdh101.
35. **List, M. A.; Siston, A.; et al.**, Quality of life and performance in advanced head and neck cancer patients on concomitant chemoradiotherapy: A prospective examination. Journal of Clinical Oncology, **1999**. 17(3): p. 1020-1028.DOI: 10.1200/jco.1999.17.3.1020.
36. **Institute of Medicine**, Evolution of Translational Omics: Lessons Learned and the Path Forward, ed. C.M. Micheel, S.J. Nass, and G.S. Omenn. **2012**, Washington, DC: The National Academies Press. 354.
37. **Dai, X. and Shen, L.**, Advances and Trends in Omics Technology Development. Frontiers in Medicine, **2022**. 9.DOI: 10.3389/fmed.2022.911861.
38. **Lambin, P.; Rios-Velazquez, E.; et al.**, Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer, **2012**. 48(4): p. 441-6.DOI: 10.1016/j.ejca.2011.11.036.
39. **van Timmeren, J. E.; Cester, D.; et al.**, Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging, **2020**. 11(1): p. 91.DOI: 10.1186/s13244-020-00887-2.
40. **Gabrys, H. S.; Buettner, F.; et al.**, Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. Front Oncol, **2018**. 8: p. 35.DOI: 10.3389/fonc.2018.00035.
41. **Lam, S. K.; Zhang, Y.; et al.**, Multi-Organ Omics-Based Prediction for Adaptive Radiation Therapy Eligibility in Nasopharyngeal Carcinoma Patients Undergoing Concurrent Chemoradiotherapy. Front Oncol, **2021**. 11: p. 792024.DOI: 10.3389/fonc.2021.792024.
42. **Zwanenburg, A.; Vallières, M.; et al.**, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology, **2020**. 295(2): p. 328-338.DOI: 10.1148/radiol.2020191145.
43. **Gul, M.; Bonjoc, K.-J. C.; et al.**, Diagnostic Utility of Radiomics in Thyroid and Head and Neck Cancers. Frontiers in Oncology, **2021**. 11.DOI: 10.3389/fonc.2021.639326.
44. **High Level Expert Group on Artificial Intelligence**, A definition of AI: Main capabilities and disciplines. **2019**, European Commission. p. 6.
45. **Mitchell, T. M.**, Machine learning. **1997**, McGraw-hill.
46. **IBM**. What is Deep Learning? Mar 2024]; Available from: https://www.ibm.com/topics/deep-learning.

47. **Moroney, L. B.; Helios, J.; et al.**, Patterns of dysphagia and acute toxicities in patients with head and neck cancer undergoing helical IMRT±concurrent chemotherapy. Oral Oncology, **2017**. 64: p. 1-8.DOI: https://doi.org/10.1016/j.oraloncology.2016.11.009.
48. **O'Neill, C. B.; Baxi, S. S.; et al.**, Treatment-related toxicities in older adults with head and neck cancer: A population-based analysis. Cancer, **2015**. 121(12): p. 2083-2089.DOI: https://doi.org/10.1002/cncr.29262.
49. **Tricco, A. C.; Lillie, E.; et al.**, PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine, **2018**. 169(7): p. 467-473.DOI: 10.7326/m18-0850 %m 30178033.
50. **Zhu, X. X.; Yang, X. J.; et al.**, The Potential Effect of Oral Microbiota in the Prediction of Mucositis During Radiotherapy for Nasopharyngeal Carcinoma. EBioMedicine, **2017**. 18: p. 23-31.DOI: 10.1016/j.ebiom.2017.02.002.
51. **Orlandi, E.; Iacovelli, N. A.; et al.**, Multivariable model for predicting acute oral mucositis during combined IMRT and chemotherapy for locally advanced nasopharyngeal cancer patients. Oral Oncol, **2018**. 86: p. 266-272.DOI: 10.1016/j.oraloncology.2018.10.006.
52. **Sharabiani, M.; Clementel, E.; et al.**, Independent external validation using the EORTC HNCG-ROG 1219 DAHANCA trial data of NTCP models for acute oral mucositis. Radiother Oncol, **2021**. 161: p. 35-39.DOI: 10.1016/j.radonc.2021.04.006.
53. **Dean, J. A.; Wong, K. H.; et al.**, Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. Radiother Oncol, **2016**. 120(1): p. 21-7.DOI: 10.1016/j.radonc.2016.05.015.
54. **Liu, Z.; Huang, L.; et al.**, Predicting Nomogram for Severe Oral Mucositis in Patients with Nasopharyngeal Carcinoma during Intensity-Modulated Radiation Therapy: A Retrospective Cohort Study. Current Oncology, **2023**. 30(1): p. 219-232.
55. **Li, P. J.; Li, K. X.; et al.**, Predictive Model and Precaution for Oral Mucositis During Chemo-Radiotherapy in Nasopharyngeal Carcinoma Patients. Front Oncol, **2020**. 10: p. 596822.DOI: 10.3389/fonc.2020.596822.
56. **Hansen, C. R.; Bertelsen, A.; et al.**, Prediction of radiation-induced mucositis of H&N cancer patients based on a large patient cohort. Radiother Oncol, **2020**. 147: p. 15-21.DOI: 10.1016/j.radonc.2020.03.013.
57. **Dong, Y.; Zhang, J.; et al.**, Multimodal Data Integration to Predict Severe Acute Oral Mucositis of Nasopharyngeal Carcinoma Patients Following Radiation Therapy. Cancers, **2023**. 15(7): p. 2032.
58. **Al-Qadami, G.; Bowen, J.; et al.**, Baseline gut microbiota composition is associated with oral mucositis and tumour recurrence in patients with head and neck cancer: a pilot study. Supportive Care in Cancer, **2023**. 31(1): p. 98.DOI: 10.1007/s00520-022-07559-5.
59. **Bansal, A.; Bedi, N.; et al.**, Correlation of oral mucosa dose and volume parameters with Grade 3 mucositis, in patients treated with volumetric modulated arc radiotherapy for oropharyngeal cancer? Japanese Journal of Clinical Oncology, **2022**. 53(4): p. 313-320.DOI: 10.1093/jjco/hyac194.
60. **Chung, A.; Chung, Y.-T.; et al.**, Waldeyer ring microbiome in relation to chemoradiation-induced oral mucositis in patients with nasopharyngeal carcinoma. Head & Neck, **2023**. 45(8): p. 2047-2057.DOI: https://doi.org/10.1002/hed.27431.
61. **Lv, J.; Liao, S.; et al.**, Scheduling radiotherapy for patients with nasopharyngeal carcinoma in the corresponding time window can reduce radiation-induced oral mucositis: A randomized, prospective study. Cancer Medicine, **2023**. 12(15): p. 16032-16040.DOI: https://doi.org/10.1002/cam4.6252.
62. **Xu, J.; Yang, G.; et al.**, Correlations between the severity of radiation-induced oral mucositis and salivary epidermal growth factor as well as inflammatory cytokines in patients with head and neck cancer. Head & Neck, **2023**. 45(5): p. 1122-1129.DOI: https://doi.org/10.1002/hed.27313.
63. **Alterovitz, G.; Tuthill, C.; et al.**, Personalized medicine for mucositis: Bayesian networks identify unique gene clusters which predict the response to gamma-D-glutamyl-L-tryptophan (SCV-07) for the attenuation of chemoradiation-induced oral mucositis. Oral Oncol, **2011**. 47(10): p. 951-5.DOI: 10.1016/j.oraloncology.2011.07.006.
64. **Bentzen, S. M.; Saunders, M. I.; et al.**, Radiotherapy-related early morbidity in head and neck cancer: quantitative clinical radiobiology as deduced from the CHART trial. Radiother Oncol, **2001**. 60(2): p. 123-35.DOI: 10.1016/s0167-8140(01)00358-9.
65. **Bjarnason, G. A.; Mackenzie, R. G.; et al.**, Comparison of toxicity associated with early morning versus late afternoon radiotherapy in patients with head-and-neck cancer: a prospective randomized trial of the National Cancer Institute of Canada Clinical Trials Group (HN3). Int J Radiat Oncol Biol Phys, **2009**. 73(1): p. 166-72.DOI: 10.1016/j.ijrobp.2008.07.009.

66. **Brzozowska, A.; Mlak, R.; et al.**, Polymorphism of regulatory region of APEH gene (c.-521G>C, rs4855883) as a relevant predictive factor for radiotherapy induced oral mucositis and overall survival in head neck cancer patients. Oncotarget, **2018**. 9(51): p. 29644-29653.DOI: 10.18632/oncotarget.25662.
67. **Brzozowska, A.; Powrozek, T.; et al.**, Polymorphism of Promoter Region of TNFRSF1A Gene (-610 T > G) as a Novel Predictive Factor for Radiotherapy Induced Oral Mucositis in HNC Patients. Pathol Oncol Res, **2018**. 24(1): p. 135-143.DOI: 10.1007/s12253-017-0227-1.
68. **Brzozowska, A.; Mlak, R.; et al.**, Status of hydration assessed by bioelectrical impedance analysis: a valuable predictive factor for radiation-induced oral mucositis in head and neck cancer patients. Clin Transl Oncol, **2019**. 21(5): p. 615-620.DOI: 10.1007/s12094-018-1963-8.
69. **Chen, S. C.; Lai, Y. H.; et al.**, Changes and predictors of radiation-induced oral mucositis in patients with oral cavity cancer during active treatment. Eur J Oncol Nurs, **2015**. 19(3): p. 214-9.DOI: 10.1016/j.ejon.2014.12.001.
70. **Chen, H.; Wu, M.; et al.**, Association between XRCC1 single-nucleotide polymorphism and acute radiation reaction in patients with nasopharyngeal carcinoma: A cohort study. Medicine (Baltimore), **2017**. 96(44): p. e8202.DOI: 10.1097/MD.0000000000008202.
71. **Chen, G.; Jiang, H.; et al.**, Pretreatment serum vitamin level predicts severity of radiation-induced oral mucositis in patients with nasopharyngeal carcinoma. Head Neck, **2021**. 43(4): p. 1153-1160.DOI: 10.1002/hed.26576.
72. **Correia, A. V.; Coelho, M. R.; et al.**, Seroprevalence of HSV-1/2 and correlation with aggravation of oral mucositis in patients with squamous cell carcinoma of the head and neck region submitted to antineoplastic treatment. Support Care Cancer, **2015**. 23(7): p. 2105-11.DOI: 10.1007/s00520-014-2558-8.
73. **Desilets, A.; McCarvill, W.; et al.**, Upfront DPYD Genotyping and Toxicity Associated with Fluoropyrimidine-Based Concurrent Chemoradiotherapy for Oropharyngeal Carcinomas: A Work in Progress. Curr Oncol, **2022**. 29(2): p. 497-509.DOI: 10.3390/curroncol29020045.
74. **Devaraju, C. J.; Lokanatha, D.; et al.**, Risk scoring for predicting mucositis in Indian patients with esophageal carcinoma receiving concurrent chemoradiotherapy. Gastrointest Cancer Res, **2009**. 3(1): p. 4-6.
75. **Epstein, J. B.; Gorsky, M.; et al.**, The correlation between epidermal growth factor levels in saliva and the severity of oral mucositis during oropharyngeal radiation therapy. Cancer, **2000**. 89(11): p. 2258-65.DOI: 10.1002/1097-0142(20001201)89:11<2258::aid-cncr14>3.0.co;2-z.
76. **Fanetti, G.; Polesel, J.; et al.**, Prognostic Nutritional Index Predicts Toxicity in Head and Neck Cancer Patients Treated with Definitive Radiotherapy in Association with Chemotherapy. Nutrients, **2021**. 13(4): p. 12.DOI: 10.3390/nu13041277.
77. **Gu, F.; Farrugia, M. K.; et al.**, Daily Time of Radiation Treatment Is Associated with Subsequent Oral Mucositis Severity during Radiotherapy in Head and Neck Cancer Patients. Cancer Epidemiol Biomarkers Prev, **2020**. 29(5): p. 949-955.DOI: 10.1158/1055-9965.EPI-19-0961.
78. **Hanin, S. M. A.; Dharman, S.; et al.**, Association of Salivary Microbes with Oral Mucositis Among Patients Undergoing Chemoradiotherapy in Head and Neck Cancer: A Hospital-Based Prospective Study. Journal of International Oral Health, **2022**. 14(1): p. 53-60.DOI: 10.4103/Jioh.Jioh_161_21.
79. **Homa-Mlak, I.; Brzozowska, A.; et al.**, Neutrophil-to-Lymphocyte Ratio as a Factor Predicting Radiotherapy Induced Oral Mucositis in Head Neck Cancer Patients Treated with Radiotherapy. J Clin Med, **2021**. 10(19): p. 15.DOI: 10.3390/jcm10194444.
80. **Jehmlich, N.; Stegmaier, P.; et al.**, Differences in the whole saliva baseline proteome profile associated with development of oral mucositis in head and neck cancer patients undergoing radiotherapy. J Proteomics, **2015**. 125: p. 98-103.DOI: 10.1016/j.jprot.2015.04.030.
81. **Kawashita, Y.; Kitamura, M.; et al.**, Association of neutrophil-to-lymphocyte ratio with severe radiation-induced mucositis in pharyngeal or laryngeal cancer patients: a retrospective study. BMC Cancer, **2021**. 21(1): p. 1064.DOI: 10.1186/s12885-021-08793-6.
82. **Kawashita, Y.; Soutome, S.; et al.**, Predictive Risk Factors Associated with Severe Radiation-Induced Mucositis in Nasopharyngeal or Oropharyngeal Cancer Patients: A Retrospective Study. Biomedicines, **2022**. 10(10): p. 8.DOI: 10.3390/biomedicines10102661.
83. **Kazmierska, J.; Barczak, W.; et al.**, The kinetics of gamma-H2AX during radiotherapy of head and neck cancer potentially allow for prediction of severe mucositis. Radiol Oncol, **2020**. 54(1): p. 96-102.DOI: 10.2478/raon-2020-0005.

84. **Le, Z.; Niu, X.; et al.**, Predictive single nucleotide polymorphism markers for acute oral mucositis in patients with nasopharyngeal carcinoma treated with radiotherapy. Oncotarget, **2017**. 8(38): p. 63026-63037.DOI: 10.18632/oncotarget.18450.
85. **Li, H.; You, Y.; et al.**, XRCC1 codon 399Gln polymorphism is associated with radiotherapy-induced acute dermatitis and mucositis in nasopharyngeal carcinoma patients. Radiat Oncol, **2013**. 8: p. 31.DOI: 10.1186/1748-717X-8-31.
86. **Li, P.; Du, C. R.; et al.**, Correlation of dynamic changes in gamma-H2AX expression in peripheral blood lymphocytes from head and neck cancer patients with radiation-induced oral mucositis. Radiat Oncol, **2013**. 8: p. 155.DOI: 10.1186/1748-717X-8-155.
87. **Li, K.; Yang, L.; et al.**, Oral Mucosa Dose Parameters Predicting Grade >/=3 Acute Toxicity in Locally Advanced Nasopharyngeal Carcinoma Patients Treated With Concurrent Intensity-Modulated Radiation Therapy and Chemotherapy: An Independent Validation Study Comparing Oral Cavity versus Mucosal Surface Contouring Techniques. Transl Oncol, **2017**. 10(5): p. 752-759.DOI: 10.1016/j.tranon.2017.06.011.
88. **Li, Q.; Liang, Y.; et al.**, Associations of GWAS-Identified Risk Loci with Progression, Efficacy and Toxicity of Radiotherapy of Head and Neck Squamous Cell Carcinoma Treated with Radiotherapy. Pharmgenomics Pers Med, **2021**. 14: p. 1205-1210.DOI: 10.2147/PGPM.S325349.
89. **Manur, J. G. and Vidyasagar, N.**, Correlation of planning target volume with mucositis for head-and-neck cancer patients undergoing chemoradiation. J Cancer Res Ther, **2020**. 16(3): p. 565-568.DOI: 10.4103/jcrt.JCRT_511_19.
90. **Mazzola, R.; Ricchetti, F.; et al.**, Predictors of mucositis in oropharyngeal and oral cavity cancer in patients treated with volumetric modulated radiation treatment: A dose-volume analysis. Head Neck, **2016**. 38 Suppl 1: p. E815-9.DOI: 10.1002/hed.24106.
91. **Mizuno, H.; Miyai, H.; et al.**, Relationship Between Renal Dysfunction and Oral Mucositis in Patients Undergoing Concurrent Chemoradiotherapy for Pharyngeal Cancer: A Retrospective Cohort Study. In Vivo, **2019**. 33(1): p. 183-189.DOI: 10.21873/invivo.11457.
92. **Mlak, R.; Powrozek, T.; et al.**, The relationship between TNF-alpha gene promoter polymorphism (-1211 T > C), the plasma concentration of TNF-alpha, and risk of oral mucositis and shortening of overall survival in patients subjected to intensity-modulated radiation therapy due to head and neck cancer. Support Care Cancer, **2020**. 28(2): p. 531-540.DOI: 10.1007/s00520-019-04838-6.
93. **Morais-Faria, K.; Palmier, N. R.; et al.**, Young head and neck cancer patients are at increased risk of developing oral mucositis and trismus. Support Care Cancer, **2020**. 28(9): p. 4345-4352.DOI: 10.1007/s00520-019-05241-x.
94. **Musha, A.; Shimada, H.; et al.**, Prediction of Acute Radiation Mucositis using an Oral Mucosal Dose Surface Model in Carbon Ion Radiotherapy for Head and Neck Tumors. PLoS One, **2015**. 10(10): p. e0141734.DOI: 10.1371/journal.pone.0141734.
95. **Musha, A.; Fukata, K.; et al.**, Tongue surface model can predict radiation tongue mucositis due to intensity-modulated radiation therapy for head and neck cancer. Int J Oral Maxillofac Surg, **2020**. 49(1): p. 44-50.DOI: 10.1016/j.ijom.2019.06.012.
96. **Nejatinamini, S.; Debenham, B. J.; et al.**, Poor Vitamin Status is Associated with Skeletal Muscle Loss and Mucositis in Head and Neck Cancer Patients. Nutrients, **2018**. 10(9): p. 11.DOI: 10.3390/nu10091236.
97. **Nguyen, H. G.; Avanesov, A.; et al.**, Microcirculatory alterations in patients with oropharyngeal cancer after radiation therapy: A possible correlation with mucositis? Archiv Euromedica, **2020**. 10(4): p. 128-133.DOI: 10.35630/2199-885x/2020/10/4.30.
98. **Nishii, M.; Soutome, S.; et al.**, Factors associated with severe oral mucositis and candidiasis in patients undergoing radiotherapy for oral and oropharyngeal carcinomas: a retrospective multicenter study of 326 patients. Support Care Cancer, **2020**. 28(3): p. 1069-1075.DOI: 10.1007/s00520-019-04885-z.
99. **Porock, D.; Nikoletti, S.; et al.**, The relationship between factors that impair wound healing and the severity of acute radiation skin and mucosal toxicities in head and neck cancer. Cancer Nurs, **2004**. 27(1): p. 71-8.DOI: 10.1097/00002820-200401000-00009.
100. **Rupe, C.; Gioco, G.; et al.**, Oral Candida spp. Colonisation Is a Risk Factor for Severe Oral Mucositis in Patients Undergoing Radiotherapy for Head & Neck Cancer: Results from a Multidisciplinary Mono-Institutional Prospective Observational Study. Cancers (Basel), **2022**. 14(19).DOI: 10.3390/cancers14194746.

101. **Saedi, H. S.; Gerami, H.; et al.**, Frequency of chemoradiotherapy-induced mucositis and related risk factors in patients with the head-and-neck cancers: A survey in the North of Iran. Dent Res J (Isfahan), **2019**. 16(5): p. 354-359.DOI: 10.4103/1735-3327.266088.
102. **Saito, N.; Imai, Y.; et al.**, Low body mass index as a risk factor of moderate to severe oral mucositis in oral cancer patients with radiotherapy. Support Care Cancer, **2012**. 20(12): p. 3373-7.DOI: 10.1007/s00520-012-1620-7.
103. **Saito, N.; Truong, M. T.; et al.**, Correlation of mucositis during head and neck radiotherapy with computed tomography perfusion imaging of the oropharyngeal mucosa. J Comput Assist Tomogr, **2013**. 37(4): p. 499-504.DOI: 10.1097/RCT.0b013e31828aed3f.
104. **Sakamoto, K.; Takeda, S.; et al.**, Association of tumor necrosis factor-alpha polymorphism with chemotherapy-induced oral mucositis in patients with esophageal cancer. Mol Clin Oncol, **2017**. 6(1): p. 125-129.DOI: 10.3892/mco.2016.1081.
105. **Sakashita, T.; Homma, A.; et al.**, Comparison of acute toxicities associated with cetuximab-based bioradiotherapy and platinum-based chemoradiotherapy for head and neck squamous cell carcinomas: A single-institution retrospective study in Japan. Acta Otolaryngol, **2015**. 135(8): p. 853-8.DOI: 10.3109/00016489.2015.1030772.
106. **Sanches, G. L. G.; da Silva Menezes, A. S.; et al.**, Local tissue electrical parameters predict oral mucositis in HNSCC patients: A diagnostic accuracy double-blind, randomized controlled trial. Sci Rep, **2020**. 10(1): p. 9530.DOI: 10.1038/s41598-020-66351-9.
107. **Sanguineti, G.; Sormani, M. P.; et al.**, Effect of radiotherapy and chemotherapy on the risk of mucositis during intensity-modulated radiation therapy for oropharyngeal cancer. Int J Radiat Oncol Biol Phys, **2012**. 83(1): p. 235-42.DOI: 10.1016/j.ijrobp.2011.06.2000.
108. **Schack, L. M. H.; Naderi, E.; et al.**, A genome-wide association study of radiotherapy induced toxicity in head and neck cancer patients identifies a susceptibility locus associated with mucositis. Br J Cancer, **2022**. 126(7): p. 1082-1090.DOI: 10.1038/s41416-021-01670-w.
109. **Schauer, M. C.; Holzmann, B.; et al.**, Interleukin-10 and -12 predict chemotherapy-associated toxicity in esophageal adenocarcinoma. J Thorac Oncol, **2010**. 5(11): p. 1849-54.DOI: 10.1097/JTO.0b013e3181f19028.
110. **Soutome, S.; Yanamoto, S.; et al.**, Risk factors for severe radiation-induced oral mucositis in patients with oral cancer. J Dent Sci, **2021**. 16(4): p. 1241-1246.DOI: 10.1016/j.jds.2021.01.009.
111. **Sunaga, T.; Nagatani, A.; et al.**, The association between cumulative radiation dose and the incidence of severe oral mucositis in head and neck cancers during radiotherapy. Cancer Rep (Hoboken), **2021**. 4(2): p. e1317.DOI: 10.1002/cnr2.1317.
112. **Suresh, A. V.; Varma, P. P.; et al.**, Risk-scoring system for predicting mucositis in patients of head and neck cancer receiving concurrent chemoradiotherapy [rssm-hn]. J Cancer Res Ther, **2010**. 6(4): p. 448-51.DOI: 10.4103/0973-1482.77100.
113. **Tao, Z.; Gao, J.; et al.**, Factors associated with acute oral mucosal reaction induced by radiotherapy in head and neck squamous cell carcinoma: A retrospective single-center experience. Medicine (Baltimore), **2017**. 96(50): p. e8446.DOI: 10.1097/MD.0000000000008446.
114. **van den Broek, G. B.; Balm, A. J.; et al.**, Relationship between clinical factors and the incidence of toxicity after intra-arterial chemoradiation for head and neck cancer. Radiother Oncol, **2006**. 81(2): p. 143-50.DOI: 10.1016/j.radonc.2006.09.002.
115. **Venkatesh, G. H.; Manjunath, V. B.; et al.**, Polymorphisms in radio-responsive genes and its association with acute toxicity among head and neck cancer patients. PLoS One, **2014**. 9(3): p. e89079.DOI: 10.1371/journal.pone.0089079.
116. **Vera-Llonch, M.; Oster, G.; et al.**, Oral mucositis in patients undergoing radiation treatment for head and neck carcinoma. Cancer, **2006**. 106(2): p. 329-36.DOI: 10.1002/cncr.21622.
117. **Wu, C.; Liu, Y.; et al.**, The relationship of serum gastrin-17 and oral mucositis in head and neck carcinoma patients receiving radiotherapy. Discov Oncol, **2022**. 13(1): p. 110.DOI: 10.1007/s12672-022-00570-6.
118. **Yahya, S.; Benghiat, H.; et al.**, Does Dose to an Oral Mucosa Organ at Risk Predict the Duration of Grade 3 Mucositis after Intensity-modulated Radiotherapy for Oropharyngeal Cancer? Clin Oncol (R Coll Radiol), **2016**. 28(12): p. e216-e219.DOI: 10.1016/j.clon.2016.08.009.
119. **Yang, Z. and Liu, Z.**, Potentially functional variants of autophagy-related genes are associated with the efficacy and toxicity of radiotherapy in patients with nasopharyngeal carcinoma. Mol Genet Genomic Med, **2019**. 7(12): p. e1030.DOI: 10.1002/mgg3.1030.

120. **Yang, D. W.; Wang, T. M.; et al.**, Genome-wide association study identifies genetic susceptibility loci and pathways of radiation-induced acute oral mucositis. J Transl Med, **2020**. 18(1): p. 224.DOI: 10.1186/s12967-020-02390-0.
121. **Yu, J.; Huang, Y.; et al.**, Genetic polymorphisms of Wnt/beta-catenin pathway genes are associated with the efficacy and toxicities of radiotherapy in patients with nasopharyngeal carcinoma. Oncotarget, **2016**. 7(50): p. 82528-82537.DOI: 10.18632/oncotarget.12754.
122. **Zahn, K. L.; Wong, G.; et al.**, Relationship of protein and calorie intake to the severity of oral mucositis in patients with head and neck cancer receiving radiation therapy. Head Neck, **2012**. 34(5): p. 655-62.DOI: 10.1002/hed.21795.
123. **Deneuve, S.; Bastogne, T.; et al.**, Predicting acute severe toxicity for head and neck squamous cell carcinomas by combining dosimetry with a radiosensitivity biomarker: a pilot study. Tumori, **2022**: p. 3008916221078061.DOI: 10.1177/03008916221078061.
124. **Otter, S.; Schick, U.; et al.**, Evaluation of the Risk of Grade 3 Oral and Pharyngeal Dysphagia Using Atlas-Based Method and Multivariate Analyses of Individual Patient Dose Distributions. Int J Radiat Oncol Biol Phys, **2015**. 93(3): p. 507-15.DOI: 10.1016/j.ijrobp.2015.07.2263.
125. **Srivastava, S.; Rastogi, M.; et al.**, Correlation of PD-L1 expression with toxicities and response in oropharyngeal cancers treated with definitive chemoradiotherapy. Contemp Oncol (Pozn), **2022**. 26(3): p. 180-186.DOI: 10.5114/wo.2022.118227.
126. **Werbrouck, J.; De Ruyck, K.; et al.**, Acute normal tissue reactions in head-and-neck cancer patients treated with IMRT: influence of dose and association with genetic polymorphisms in DNA DSB repair genes. Int J Radiat Oncol Biol Phys, **2009**. 73(4): p. 1187-95.DOI: 10.1016/j.ijrobp.2008.08.073.
127. **De Ruyck, K.; Duprez, F.; et al.**, A predictive model for dysphagia following IMRT for head and neck cancer: introduction of the EMLasso technique. Radiother Oncol, **2013**. 107(3): p. 295-9.DOI: 10.1016/j.radonc.2013.03.021.
128. **Willemsen, A. C. H.; Kok, A.; et al.**, Development and external validation of a prediction model for tube feeding dependency for at least four weeks during chemoradiotherapy for head and neck cancer. Clin Nutr, **2022**. 41(1): p. 177-185.DOI: 10.1016/j.clnu.2021.11.019.
129. **Gaito, S.; France, A.; et al.**, A Predictive Model for Reactive Tube Feeding in Head and Neck Cancer Patients Undergoing Definitive (Chemo) Radiotherapy. Clin Oncol (R Coll Radiol), **2021**. 33(10): p. e433-e441.DOI: 10.1016/j.clon.2021.05.002.
130. **Willemsen, A. C. H.; Kok, A.; et al.**, Prediction model for tube feeding dependency during chemoradiotherapy for at least four weeks in head and neck cancer patients: A tool for prophylactic gastrostomy decision making. Clin Nutr, **2020**. 39(8): p. 2600-2608.DOI: 10.1016/j.clnu.2019.11.033.
131. **Dean, J.; Wong, K.; 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. Clin Transl Radiat Oncol, **2018**. 8: p. 27-39.DOI: 10.1016/j.ctro.2017.11.009.
132. **Yahya, N.; Linge, A.; et al.**, Assessment of gene expressions from squamous cell carcinoma of the head and neck to predict radiochemotherapy-related xerostomia and dysphagia. Acta Oncol, **2022**. 61(7): p. 856-863.DOI: 10.1080/0284186X.2022.2081931.
133. **Soderstrom, K.; Nilsson, P.; et al.**, Dysphagia - Results from multivariable predictive modelling on aspiration from a subset of the ARTSCAN trial. Radiother Oncol, **2017**. 122(2): p. 192-199.DOI: 10.1016/j.radonc.2016.09.001.
134. **Christianen, M. E.; van der Schaaf, A.; et al.**, Swallowing sparing intensity modulated radiotherapy (SW-IMRT) in head and neck cancer: Clinical validation according to the model-based approach. Radiother Oncol, **2016**. 118(2): p. 298-303.DOI: 10.1016/j.radonc.2015.11.009.
135. **Kalendralis, P.; Sloep, M.; et al.**, Independent validation of a dysphagia dose response model for the selection of head and neck cancer patients to proton therapy. Phys Imaging Radiat Oncol, **2022**. 24: p. 47-52.DOI: 10.1016/j.phro.2022.09.005.
136. **Christianen, M. E.; Schilstra, C.; et al.**, Predictive modelling for swallowing dysfunction after primary (chemo)radiation: results of a prospective observational study. Radiother Oncol, **2012**. 105(1): p. 107-14.DOI: 10.1016/j.radonc.2011.08.009.
137. **Wopken, K.; Bijl, H. P.; et al.**, Development and validation of a prediction model for tube feeding dependence after curative (chemo-) radiation in head and neck cancer. PLoS One, **2014**. 9(4): p. e94879.DOI: 10.1371/journal.pone.0094879.
138. **MD Anderson Head and Neck Cancer Symptom Working Group**, Beyond mean pharyngeal constrictor dose for beam path toxicity in non-target swallowing muscles: Dose-volume correlates of

chronic radiation-associated dysphagia (RAD) after oropharyngeal intensity modulated radiotherapy. Radiother Oncol, **2016**. 118(2): p. 304-14.DOI: 10.1016/j.radonc.2016.01.019.
139. **Wentzel, A.; Hanula, P.; et al.**, Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. Radiother Oncol, **2020**. 148: p. 245-251.DOI: 10.1016/j.radonc.2020.05.023.
140. **Wopken, K.; Bijl, H. P.; et al.**, Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer. Radiother Oncol, **2014**. 113(1): p. 95-101.DOI: 10.1016/j.radonc.2014.09.013.
141. **Alexidis, P.; Bangeas, P.; et al.**, Investigating factors associated to dysphagia and need for percutaneous endoscopic gastrostomy in patients with head and neck cancer receiving radiation therapy. J Cancer, **2022**. 13(5): p. 1523-1529.DOI: 10.7150/jca.69130.
142. **Anderson, N. J.; Jackson, J. E.; et al.**, Pretreatment risk stratification of feeding tube use in patients treated with intensity-modulated radiotherapy for head and neck cancer. Head Neck, **2018**. 40(10): p. 2181-2192.DOI: 10.1002/hed.25316.
143. **Awan, M. J.; Mohamed, A. S.; et al.**, Late radiation-associated dysphagia (late-RAD) with lower cranial neuropathy after oropharyngeal radiotherapy: a preliminary dosimetric comparison. Oral Oncol, **2014**. 50(8): p. 746-52.DOI: 10.1016/j.oraloncology.2014.05.003.
144. **Barnhart, M. K.; Ward, E. C.; et al.**, Pretreatment factors associated with functional oral intake and feeding tube use at 1 and 6 months post-radiotherapy (+/- chemotherapy) for head and neck cancer. Eur Arch Otorhinolaryngol, **2017**. 274(1): p. 507-516.DOI: 10.1007/s00405-016-4241-9.
145. **Best, S. R.; Ha, P. K.; et al.**, Factors associated with pharyngoesophageal stricture in patients treated with concurrent chemotherapy and radiation therapy for oropharyngeal squamous cell carcinoma. Head Neck, **2011**. 33(12): p. 1727-34.DOI: 10.1002/hed.21657.
146. **Bhayani, M. K.; Hutcheson, K. A.; et al.**, Gastrostomy tube placement in patients with oropharyngeal carcinoma treated with radiotherapy or chemoradiotherapy: factors affecting placement and dependence. Head Neck, **2013**. 35(11): p. 1634-40.DOI: 10.1002/hed.23200.
147. **Bhide, S. A.; Gulliford, S.; et al.**, Correlation between dose to the pharyngeal constrictors and patient quality of life and late dysphagia following chemo-IMRT for head and neck cancer. Radiother Oncol, **2009**. 93(3): p. 539-44.DOI: 10.1016/j.radonc.2009.09.017.
148. **Caglar, H. B.; Tishler, R. B.; et al.**, Dose to larynx predicts for swallowing complications after intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys, **2008**. 72(4): p. 1110-8.DOI: 10.1016/j.ijrobp.2008.02.048.
149. **Caudell, J. J.; Schaner, P. E.; et al.**, Factors associated with long-term dysphagia after definitive radiotherapy for locally advanced head-and-neck cancer. Int J Radiat Oncol Biol Phys, **2009**. 73(2): p. 410-5.DOI: 10.1016/j.ijrobp.2008.04.048.
150. **Caudell, J. J.; Schaner, P. E.; et al.**, Dosimetric factors associated with long-term dysphagia after definitive radiotherapy for squamous cell carcinoma of the head and neck. Int J Radiat Oncol Biol Phys, **2010**. 76(2): p. 403-9.DOI: 10.1016/j.ijrobp.2009.02.017.
151. **Charters, E.; Bogaardt, H.; et al.**, Functional swallowing outcomes related to radiation exposure to dysphagia and aspiration-related structures in patients with head and neck cancer undergoing definitive and postoperative intensity-modulated radiotherapy. Head Neck, **2022**. 44(2): p. 399-411.DOI: 10.1002/hed.26936.
152. **Chera, B. S.; Fried, D.; et al.**, Dosimetric Predictors of Patient-Reported Xerostomia and Dysphagia With Deintensified Chemoradiation Therapy for HPV-Associated Oropharyngeal Squamous Cell Carcinoma. Int J Radiat Oncol Biol Phys, **2017**. 98(5): p. 1022-1027.DOI: 10.1016/j.ijrobp.2017.03.034.
153. **Deantonio, L.; Masini, L.; et al.**, Dysphagia after definitive radiotherapy for head and neck cancer. Correlation of dose-volume parameters of the pharyngeal constrictor muscles. Strahlenther Onkol, **2013**. 189(3): p. 230-6.DOI: 10.1007/s00066-012-0288-8.
154. **Dirix, P.; Abbeel, S.; et al.**, Dysphagia after chemoradiotherapy for head-and-neck squamous cell carcinoma: dose-effect relationships for the swallowing structures. Int J Radiat Oncol Biol Phys, **2009**. 75(2): p. 385-92.DOI: 10.1016/j.ijrobp.2008.11.041.
155. **Eisbruch, A.; Levendag, P. C.; et al.**, Can IMRT or brachytherapy reduce dysphagia associated with chemoradiotherapy of head and neck cancer? The Michigan and Rotterdam experiences. Int J Radiat Oncol Biol Phys, **2007**. 69(2 Suppl): p. S40-2.DOI: 10.1016/j.ijrobp.2007.04.083.

156. **Eisbruch, A.; Kim, H. M.; et al.**, Chemo-IMRT of oropharyngeal cancer aiming to reduce dysphagia: swallowing organs late complication probabilities and dosimetric correlates. Int J Radiat Oncol Biol Phys, **2011**. 81(3): p. e93-9.DOI: 10.1016/j.ijrobp.2010.12.067.
157. **Feng, F. Y.; Kim, H. M.; et al.**, Intensity-modulated radiotherapy of head and neck cancer aiming to reduce dysphagia: early dose-effect relationships for the swallowing structures. Int J Radiat Oncol Biol Phys, **2007**. 68(5): p. 1289-98.DOI: 10.1016/j.ijrobp.2007.02.049.
158. **Frowen, J.; Hornby, C.; et al.**, Reducing posttreatment dysphagia: Support for the relationship between radiation dose to the pharyngeal constrictors and swallowing outcomes. Pract Radiat Oncol, **2013**. 3(4): p. e187-94.DOI: 10.1016/j.prro.2012.11.009.
159. **Fua, T. F.; Corry, J.; et al.**, Intensity-modulated radiotherapy for nasopharyngeal carcinoma: clinical correlation of dose to the pharyngo-esophageal axis and dysphagia. Int J Radiat Oncol Biol Phys, **2007**. 67(4): p. 976-81.DOI: 10.1016/j.ijrobp.2006.10.028.
160. **Ghadjar, P.; Simcock, M.; et al.**, Predictors of severe late radiotherapy-related toxicity after hyperfractionated radiotherapy with or without concomitant cisplatin in locally advanced head and neck cancer. Secondary retrospective analysis of a randomized phase III trial (SAKK 10/94). Radiother Oncol, **2012**. 104(2): p. 213-8.DOI: 10.1016/j.radonc.2012.05.004.
161. **Goepfert, R. P.; Lewin, J. S.; et al.**, Predicting two-year longitudinal MD Anderson Dysphagia Inventory outcomes after intensity modulated radiotherapy for locoregionally advanced oropharyngeal carcinoma. Laryngoscope, **2017**. 127(4): p. 842-848.DOI: 10.1002/lary.26153.
162. **Guo, G. Z.; Sutherland, K. R.; et al.**, Prospective swallowing outcomes after IMRT for oropharyngeal cancer: Dosimetric correlations in a population-based cohort. Oral Oncol, **2016**. 61: p. 135-41.DOI: 10.1016/j.oraloncology.2016.08.021.
163. **Haderlein, M.; Semrau, S.; et al.**, Dose-dependent deterioration of swallowing function after induction chemotherapy and definitive chemoradiotherapy for laryngopharyngeal cancer. Strahlenther Onkol, **2014**. 190(2): p. 192-8.DOI: 10.1007/s00066-013-0493-0.
164. **Harms, A.; Kansara, S.; et al.**, Swallowing Function in Survivors of Oropharyngeal Cancer Is Associated With Advanced T Classification. Ann Otol Rhinol Laryngol, **2019**. 128(8): p. 696-703.DOI: 10.1177/0003489419839091.
165. **Jensen, K.; Lambertsen, K.; et al.**, Late swallowing dysfunction and dysphagia after radiotherapy for pharynx cancer: frequency, intensity and correlation with dose and volume parameters. Radiother Oncol, **2007**. 85(1): p. 74-82.DOI: 10.1016/j.radonc.2007.06.004.
166. **Kanayama, N.; Kierkels, R. G. J.; et al.**, External validation of a multifactorial normal tissue complication probability model for tube feeding dependence at 6 months after definitive radiotherapy for head and neck cancer. Radiother Oncol, **2018**. 129(2): p. 403-408.DOI: 10.1016/j.radonc.2018.09.013.
167. **Karsten, R. T.; Stuiver, M. M.; et al.**, From reactive to proactive tube feeding during chemoradiotherapy for head and neck cancer: A clinical prediction model-based approach. Oral Oncol, **2019**. 88: p. 172-179.DOI: 10.1016/j.oraloncology.2018.11.031.
168. **Kierkels, R. G. J.; Wopken, K.; et al.**, Multivariable normal tissue complication probability model-based treatment plan optimization for grade 2-4 dysphagia and tube feeding dependence in head and neck radiotherapy. Radiother Oncol, **2016**. 121(3): p. 374-380.DOI: 10.1016/j.radonc.2016.08.016.
169. **Kimura, H.; Hamauchi, S.; et al.**, Pretreatment predictive factors for feasibility of oral intake in adjuvant concurrent chemoradiotherapy for patients with locally advanced squamous cell carcinoma of the head and neck. Int J Clin Oncol, **2020**. 25(2): p. 258-266.DOI: 10.1007/s10147-019-01560-5.
170. **Koiwai, K.; Shikama, N.; et al.**, Risk factors for severe Dysphagia after concurrent chemoradiotherapy for head and neck cancers. Jpn J Clin Oncol, **2009**. 39(7): p. 413-7.DOI: 10.1093/jjco/hyp033.
171. **Langendijk, J. A.; Doornaert, P.; et al.**, A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer. Radiother Oncol, **2009**. 90(2): p. 189-95.DOI: 10.1016/j.radonc.2008.12.017.
172. **Lango, M. N.; Egleston, B.; et al.**, Impact of neck dissection on long-term feeding tube dependence in patients with head and neck cancer treated with primary radiation or chemoradiation. Head Neck, **2010**. 32(3): p. 341-7.DOI: 10.1002/hed.21188.
173. **Lee, W. T.; Akst, L. M.; et al.**, Risk factors for hypopharyngeal/upper esophageal stricture formation after concurrent chemoradiation. Head Neck, **2006**. 28(9): p. 808-12.DOI: 10.1002/hed.20427.
174. **Levendag, P. C.; Teguh, D. N.; et al.**, Dysphagia disorders in patients with cancer of the oropharynx are significantly affected by the radiation therapy dose to the superior and middle constrictor muscle: a dose-effect relationship. Radiother Oncol, **2007**. 85(1): p. 64-73.DOI: 10.1016/j.radonc.2007.07.009.

175. **Li, B.; Li, D.; et al.**, Clinical-dosimetric analysis of measures of dysphagia including gastrostomy-tube dependence among head and neck cancer patients treated definitively by intensity-modulated radiotherapy with concurrent chemotherapy. Radiat Oncol, **2009**. 4: p. 52.DOI: 10.1186/1748-717X-4-52.
176. **Lim, S. B.; Lee, N.; et al.**, Can the Risk of Dysphagia in Head and Neck Radiation Therapy Be Predicted by an Automated Transit Fluence Monitoring Process During Treatment? A First Comparative Study of Patient Reported Quality of Life and the Fluence-Based Decision Support Metric. Technol Cancer Res Treat, **2021**. 20: p. 15330338211027906.DOI: 10.1177/15330338211027906.
177. **Liu, H. C.; Williamson, C. W.; et al.**, Quantitative prediction of aspiration risk in head and neck cancer patients treated with radiation therapy. Oral Oncol, **2022**. 136: p. 106247.DOI: 10.1016/j.oraloncology.2022.106247.
178. **Logemann, J. A.; Rademaker, A. W.; et al.**, Site of disease and treatment protocol as correlates of swallowing function in patients with head and neck cancer treated with chemoradiation. Head Neck, **2006**. 28(1): p. 64-73.DOI: 10.1002/hed.20299.
179. **Loser, A.; Grohmann, M.; et al.**, Impact of dosimetric factors on long-term percutaneous enteral gastrostomy (PEG) tube dependence in head and neck cancer patients after (chemo)radiotherapy-results from a prospective randomized trial. Strahlenther Onkol, **2022**. 198(11): p. 1016-1024.DOI: 10.1007/s00066-022-01992-5.
180. **Machtay, M.; Moughan, J.; et al.**, Factors associated with severe late toxicity after concurrent chemoradiation for locally advanced head and neck cancer: an RTOG analysis. J Clin Oncol, **2008**. 26(21): p. 3582-9.DOI: 10.1200/JCO.2007.14.8841.
181. **Mangar, S.; Slevin, N.; et al.**, Evaluating predictive factors for determining enteral nutrition in patients receiving radical radiotherapy for head and neck cancer: a retrospective review. Radiother Oncol, **2006**. 78(2): p. 152-8.DOI: 10.1016/j.radonc.2005.12.014.
182. **Mattei, P.; Thamphya, B.; et al.**, Therapeutic strategies, oncologic and swallowing outcomes and their predictive factors in patients with locally advanced hypopharyngeal cancer. Eur Arch Otorhinolaryngol, **2022**. 279(7): p. 3629-3637.DOI: 10.1007/s00405-021-07196-4.
183. **Mazzola, R.; Ricchetti, F.; et al.**, Dose-volume-related dysphagia after constrictor muscles definition in head and neck cancer intensity-modulated radiation treatment. Br J Radiol, **2014**. 87(1044): p. 20140543.DOI: 10.1259/bjr.20140543.
184. **Mierzwa, M. L.; Gharzai, L. A.; et al.**, Early MRI Blood Volume Changes in Constrictor Muscles Correlate With Postradiation Dysphagia. Int J Radiat Oncol Biol Phys, **2021**. 110(2): p. 566-573.DOI: 10.1016/j.ijrobp.2020.12.018.
185. **Monti, S.; Palma, G.; et al.**, Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. Sci Rep, **2017**. 7(1): p. 7220.DOI: 10.1038/s41598-017-07586-x.
186. **Mortensen, H. R.; Overgaard, J.; et al.**, Factors associated with acute and late dysphagia in the DAHANCA 6 & 7 randomized trial with accelerated radiotherapy for head and neck cancer. Acta Oncol, **2013**. 52(7): p. 1535-42.DOI: 10.3109/0284186X.2013.824609.
187. **Mortensen, H. R.; Jensen, K.; et al.**, Late dysphagia after IMRT for head and neck cancer and correlation with dose-volume parameters. Radiother Oncol, **2013**. 107(3): p. 288-94.DOI: 10.1016/j.radonc.2013.06.001.
188. **Mouw, K. W.; Haraf, D. J.; et al.**, Factors associated with long-term speech and swallowing outcomes after chemoradiotherapy for locoregionally advanced head and neck cancer. Arch Otolaryngol Head Neck Surg, **2010**. 136(12): p. 1226-34.DOI: 10.1001/archoto.2010.218.
189. **Murono, S.; Tsuji, A.; et al.**, Factors associated with gastrostomy tube dependence after concurrent chemoradiotherapy for hypopharyngeal cancer. Support Care Cancer, **2015**. 23(2): p. 457-62.DOI: 10.1007/s00520-014-2388-8.
190. **Nevens, D.; Goeleven, A.; et al.**, Does the total dysphagia risk score correlate with swallowing function examined by videofluoroscopy? Br J Radiol, **2018**. 91(1083): p. 20170714.DOI: 10.1259/bjr.20170714.
191. **Nguyen, N. P.; Frank, C.; et al.**, Analysis of factors influencing aspiration risk following chemoradiation for oropharyngeal cancer. Br J Radiol, **2009**. 82(980): p. 675-80.DOI: 10.1259/bjr/72852974.
192. **Orlandi, E.; Miceli, R.; et al.**, Predictors of Patient-Reported Dysphagia Following IMRT Plus Chemotherapy in Oropharyngeal Cancer. Dysphagia, **2019**. 34(1): p. 52-62.DOI: 10.1007/s00455-018-9913-8.

193. **Ortigara, G. B.; Schulz, R. E.; et al.**, Association between trismus and dysphagia-related quality of life in survivors of head and neck cancer in Brazil. Oral Surg Oral Med Oral Pathol Oral Radiol, **2019**. 128(3): p. 235-242.DOI: 10.1016/j.oooo.2019.05.009.
194. **Ottosson, S.; Lindblom, U.; et al.**, Weight loss and body mass index in relation to aspiration in patients treated for head and neck cancer: a long-term follow-up. Support Care Cancer, **2014**. 22(9): p. 2361-9.DOI: 10.1007/s00520-014-2211-6.
195. **Petersson, K.; Finizia, C.; et al.**, Predictors of severe dysphagia following radiotherapy for head and neck cancer. Laryngoscope Investig Otolaryngol, **2021**. 6(6): p. 1395-1405.DOI: 10.1002/lio2.676.
196. **Petras, K. G.; Rademaker, A. W.; et al.**, Dose-volume relationship for laryngeal substructures and aspiration in patients with locally advanced head-and-neck cancer. Radiat Oncol, **2019**. 14(1): p. 49.DOI: 10.1186/s13014-019-1247-7.
197. **Poulsen, M. G.; Riddle, B.; et al.**, Predictors of acute grade 4 swallowing toxicity in patients with stages III and IV squamous carcinoma of the head and neck treated with radiotherapy alone. Radiother Oncol, **2008**. 87(2): p. 253-9.DOI: 10.1016/j.radonc.2008.03.010.
198. **Prameela, C. G.; Ravind, R.; et al.**, Radiation dose to dysphagia aspiration-related structures and its effect on swallowing: Comparison of three-dimensional conformal radiotherapy and intensity-modulated radiation therapy plans. J Cancer Res Ther, **2016**. 12(2): p. 845-51.DOI: 10.4103/0973-1482.163676.
199. **Pu, D.; Lee, V. H. F.; et al.**, The Relationships Between Radiation Dosage and Long-term Swallowing Kinematics and Timing in Nasopharyngeal Carcinoma Survivors. Dysphagia, **2022**. 37(3): p. 612-621.DOI: 10.1007/s00455-021-10311-6.
200. **Ray, X.; Sumner, W.; et al.**, Evaluating predictive factors for toxicities experienced by head & neck cancer patients undergoing radiotherapy. J Transl Med, **2021**. 19(1): p. 380.DOI: 10.1186/s12967-021-03047-2.
201. **Rwigema, J. M.; Langendijk, J. A.; et al.**, A Model-Based Approach to Predict Short-Term Toxicity Benefits With Proton Therapy for Oropharyngeal Cancer. Int J Radiat Oncol Biol Phys, **2019**. 104(3): p. 553-562.DOI: 10.1016/j.ijrobp.2018.12.055.
202. **Sachdev, S.; Refaat, T.; et al.**, Age most significant predictor of requiring enteral feeding in head-and-neck cancer patients. Radiat Oncol, **2015**. 10: p. 93.DOI: 10.1186/s13014-015-0408-6.
203. **Saito, H.; Shodo, R.; et al.**, The association between oral candidiasis and severity of chemoradiotherapy-induced dysphagia in head and neck cancer patients: A retrospective cohort study. Clin Transl Radiat Oncol, **2020**. 20: p. 13-18.DOI: 10.1016/j.ctro.2019.10.006.
204. **Salama, J. K.; Stenson, K. M.; et al.**, Characteristics associated with swallowing changes after concurrent chemotherapy and radiotherapy in patients with head and neck cancer. Arch Otolaryngol Head Neck Surg, **2008**. 134(10): p. 1060-5.DOI: 10.1001/archotol.134.10.1060.
205. **Sanguineti, G.; Gunn, G. B.; et al.**, Weekly dose-volume parameters of mucosa and constrictor muscles predict the use of percutaneous endoscopic gastrostomy during exclusive intensity-modulated radiotherapy for oropharyngeal cancer. Int J Radiat Oncol Biol Phys, **2011**. 79(1): p. 52-9.DOI: 10.1016/j.ijrobp.2009.10.057.
206. **Schwartz, D. L.; Hutcheson, K.; et al.**, Candidate dosimetric predictors of long-term swallowing dysfunction after oropharyngeal intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys, **2010**. 78(5): p. 1356-65.DOI: 10.1016/j.ijrobp.2009.10.002.
207. **Setton, J.; Lee, N. Y.; et al.**, A multi-institution pooled analysis of gastrostomy tube dependence in patients with oropharyngeal cancer treated with definitive intensity-modulated radiotherapy. Cancer, **2015**. 121(2): p. 294-301.DOI: 10.1002/cncr.29022.
208. **Strom, T.; Trotti, A. M.; et al.**, Risk factors for percutaneous endoscopic gastrostomy tube placement during chemoradiotherapy for oropharyngeal cancer. JAMA Otolaryngol Head Neck Surg, **2013**. 139(11): p. 1242-6.DOI: 10.1001/jamaoto.2013.5193.
209. **Teguh, D. N.; Levendag, P. C.; et al.**, Risk model and nomogram for dysphagia and xerostomia prediction in head and neck cancer patients treated by radiotherapy and/or chemotherapy. Dysphagia, **2013**. 28(3): p. 388-94.DOI: 10.1007/s00455-012-9445-6.
210. **Truong, M. T.; Lee, R.; et al.**, Correlating computed tomography perfusion changes in the pharyngeal constrictor muscles during head-and-neck radiotherapy to dysphagia outcome. Int J Radiat Oncol Biol Phys, **2012**. 82(2): p. e119-27.DOI: 10.1016/j.ijrobp.2011.04.058.
211. **Tsai, C. J.; Jackson, A.; et al.**, Modeling Dose Response for Late Dysphagia in Patients With Head and Neck Cancer in the Modern Era of Definitive Chemoradiation. JCO Clin Cancer Inform, **2017**. 1: p. 1-7.DOI: 10.1200/CCI.17.00070.

212. **Ursino, S.; Giuliano, A.; et al.**, Incorporating dose-volume histogram parameters of swallowing organs at risk in a videofluoroscopy-based predictive model of radiation-induced dysphagia after head and neck cancer intensity-modulated radiation therapy. Strahlenther Onkol, **2021**. 197(3): p. 209-218.DOI: 10.1007/s00066-020-01697-7.
213. **van der Laan, H. P.; Bijl, H. P.; et al.**, Acute symptoms during the course of head and neck radiotherapy or chemoradiation are strong predictors of late dysphagia. Radiother Oncol, **2015**. 115(1): p. 56-62.DOI: 10.1016/j.radonc.2015.01.019.
214. **van der Molen, L.; Heemsbergen, W. D.; et al.**, Dysphagia and trismus after concomitant chemo-Intensity-Modulated Radiation Therapy (chemo-IMRT) in advanced head and neck cancer; dose-effect relationships for swallowing and mastication structures. Radiother Oncol, **2013**. 106(3): p. 364-9.DOI: 10.1016/j.radonc.2013.03.005.
215. **Vangelov, B. and Smee, R. I.**, Clinical predictors for reactive tube feeding in patients with advanced oropharynx cancer receiving radiotherapy +/- chemotherapy. Eur Arch Otorhinolaryngol, **2017**. 274(10): p. 3741-3749.DOI: 10.1007/s00405-017-4681-x.
216. **Vidyasagar, N. and Manur Gururajachar, J.**, Predicting toxicity for head and neck cancer patients undergoing radiation therapy: an independent and external validation of MDASI-HN based nomogram. Rep Pract Oncol Radiother, **2020**. 25(3): p. 355-359.DOI: 10.1016/j.rpor.2020.03.005.
217. **Wang, Y.; Xiao, F.; et al.**, A two-stage genome-wide association study to identify novel genetic loci associated with acute radiotherapy toxicity in nasopharyngeal carcinoma. Mol Cancer, **2022**. 21(1): p. 169.DOI: 10.1186/s12943-022-01631-8.
218. **Wentzel, A.; Luciani, T.; et al.**, Precision association of lymphatic disease spread with radiation-associated toxicity in oropharyngeal squamous carcinomas. Radiother Oncol, **2021**. 161: p. 152-158.DOI: 10.1016/j.radonc.2021.06.016.
219. **Yang, W.; McNutt, T. R.; et al.**, Predictive Factors for Prophylactic Percutaneous Endoscopic Gastrostomy (PEG) Tube Placement and Use in Head and Neck Patients Following Intensity-Modulated Radiation Therapy (IMRT) Treatment: Concordance, Discrepancies, and the Role of Gabapentin. Dysphagia, **2016**. 31(2): p. 206-13.DOI: 10.1007/s00455-015-9679-1.
220. **Alexidis, P.; Kolias, P.; et al.**, Investigating Predictive Factors of Dysphagia and Treatment Prolongation in Patients with Oral Cavity or Oropharyngeal Cancer Receiving Radiation Therapy Concurrently with Chemotherapy. Current Oncology, **2023**. 30(5): p. 5168-5178.
221. **Beddok, A.; Maynadier, X.; et al.**, Predictors of toxicity after curative reirradiation with intensity modulated radiotherapy or proton therapy for recurrent head and neck carcinoma: new dose constraints for pharyngeal constrictors muscles and oral cavity. Strahlentherapie und Onkologie, **2023**.DOI: 10.1007/s00066-023-02080-y.
222. **Vasquez Osorio, E.; Abravan, A.; et al.**, Dysphagia at 1 Year is Associated With Mean Dose to the Inferior Section of the Brain Stem. International Journal of Radiation Oncology, Biology, Physics.DOI: 10.1016/j.ijrobp.2023.06.004.
223. **Alexandra, G.; Alexandru, M.; et al.**, Blood Group Type Association with Head and Neck Cancer. Hematol Rep, **2022**. 14(1): p. 24-30.DOI: 10.3390/hematolrep14010005.
224. **Carbonara, R.; Bonomo, P.; et al.**, Investigation of Radiation-Induced Toxicity in Head and Neck Cancer Patients through Radiomics and Machine Learning: A Systematic Review. Journal of Oncology, **2021**. 2021: p. 5566508.DOI: 10.1155/2021/5566508.
225. **Araújo, A. L. D.; Moraes, M. C.; et al.**, Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis. Oral Oncology, **2023**. 140: p. 106386.DOI: https://doi.org/10.1016/j.oraloncology.2023.106386.
226. **Tan, D.; Mohamad Salleh, S. A.; et al.**, Delta-radiomics-based models for toxicity prediction in radiotherapy: A systematic review and meta-analysis. J Med Imaging Radiat Oncol, **2023**. 67(5): p. 564-579.DOI: 10.1111/1754-9485.13546.
227. **Kocak, B.; Baessler, B.; et al.**, CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights into Imaging, **2023**. 14(1): p. 75.DOI: 10.1186/s13244-023-01415-8.
228. **Patzer, R. E.; Kaji, A. H.; et al.**, TRIPOD Reporting Guidelines for Diagnostic and Prognostic Studies. JAMA Surgery, **2021**. 156(7): p. 675-676.DOI: 10.1001/jamasurg.2021.0537.
229. **Nardone, V.; Tini, P.; et al.**, Texture analysis as a predictor of radiation-induced xerostomia in head and neck patients undergoing IMRT. La radiologia medica, **2018**. 123(6): p. 415-423.DOI: 10.1007/s11547-017-0850-7.

230. **Zhou, L.; Zheng, W.; et al.**, Integrated radiomics, dose-volume histogram criteria and clinical features for early prediction of saliva amount reduction after radiotherapy in nasopharyngeal cancer patients. Discover Oncology, **2022**. 13(1): p. 145.DOI: 10.1007/s12672-022-00606-x.
231. **Pota, M.; Scalco, E.; et al.**, Early prediction of radiotherapy-induced parotid shrinkage and toxicity based on CT radiomics and fuzzy classification. Artificial Intelligence in Medicine, **2017**. 81: p. 41-53.DOI: https://doi.org/10.1016/j.artmed.2017.03.004.
232. **Abdollahi, H.; Mostafaei, S.; et al.**, Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. Physica Medica, **2018**. 45: p. 192-197.DOI: https://doi.org/10.1016/j.ejmp.2017.10.008.
233. **Ren, W.; Liang, B.; et al.**, Dosiomics-based prediction of radiation-induced hypothyroidism in nasopharyngeal carcinoma patients. Physica Medica, **2021**. 89: p. 219-225.DOI: https://doi.org/10.1016/j.ejmp.2021.08.009.
234. **Wu, H.; Chen, X.; et al.**, Early Prediction of Acute Xerostomia During Radiation Therapy for Head and Neck Cancer Based on Texture Analysis of Daily CT. International Journal of Radiation Oncology*Biology*Physics, **2018**. 102(4): p. 1308-1318.DOI: https://doi.org/10.1016/j.ijrobp.2018.04.059.
235. **Thor, M.; Tyagi, N.; et al.**, A magnetic resonance imaging-based approach to quantify radiation-induced normal tissue injuries applied to trismus in head and neck cancer. Physics and Imaging in Radiation Oncology, **2017**. 1: p. 34-40.DOI: https://doi.org/10.1016/j.phro.2017.02.006.
236. **Berger, T.; Noble, D. J.; et al.**, Predicting radiotherapy-induced xerostomia in head and neck cancer patients using day-to-day kinetics of radiomics features. Physics and Imaging in Radiation Oncology, **2022**. 24: p. 95-101.DOI: https://doi.org/10.1016/j.phro.2022.10.004.
237. **Qin, Y.; Chang, C.; et al.**, Predicting late radiation-induced xerostomia in nasopharyngeal carcinoma based on radiomics features extracted from T2WI images of parotids. Radiation Medicine and Protection, **2023**. 4(3): p. 125-129.DOI: https://doi.org/10.1016/j.radmp.2023.06.002.
238. **van Dijk, L. V.; Noordzij, W.; et al.**, 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology, **2018**. 126(1): p. 89-95.DOI: https://doi.org/10.1016/j.radonc.2017.08.024.
239. **van Dijk, L. V.; Thor, M.; et al.**, Parotid gland fat related Magnetic Resonance image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology, **2018**. 128(3): p. 459-466.DOI: https://doi.org/10.1016/j.radonc.2018.06.012.
240. **Li, Y.; Sijtsema, N. M.; et al.**, Validation of the 18F-FDG PET image biomarker model predicting late xerostomia after head and neck cancer radiotherapy. Radiotherapy and Oncology, **2023**. 180: p. 109458.DOI: https://doi.org/10.1016/j.radonc.2022.109458.
241. **Busato, F.; Fiorentin, D.; et al.**, Dosiomic-based prediction of dysgeusia in head & neck cancer patients treated with radiotherapy. Radiotherapy and Oncology, **2023**. 188: p. 109896.DOI: https://doi.org/10.1016/j.radonc.2023.109896.
242. **van Dijk, L. V.; Langendijk, J. A.; et al.**, Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. Scientific Reports, **2019**. 9(1): p. 12483.DOI: 10.1038/s41598-019-48184-3.
243. **Ritlumlert, N.; Wongwattananard, S.; et al.**, Improved prediction of radiation-induced hypothyroidism in nasopharyngeal carcinoma using pre-treatment CT radiomics. Scientific Reports, **2023**. 13(1): p. 17437.DOI: 10.1038/s41598-023-44439-2.
244. **Berger, T.; Noble, D. J.; et al.**, Sub-regional analysis of the parotid glands: model development for predicting late xerostomia with radiomics features in head and neck cancer patients. Acta Oncologica, **2023**. 62(2): p. 166-173.DOI: 10.1080/0284186X.2023.2179895.
245. **Abdollahi, H.; Dehesh, T.; et al.**, Radiomics and dosiomics-based prediction of radiotherapy-induced xerostomia in head and neck cancer patients. International Journal of Radiation Biology, **2023**. 99(11): p. 1669-1683.DOI: 10.1080/09553002.2023.2214206.
246. **Sheikh, K.; Lee, S. H.; 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): p. 131.DOI: 10.1186/s13014-019-1339-4.
247. **Cheng, Z.; Nakatsugawa, M.; et al.**, Utility of a Clinical Decision Support System in Weight Loss Prediction After Head and Neck Cancer Radiotherapy. JCO Clin Cancer Inform, **2019**. 3: p. 1-11.DOI: 10.1200/CCI.18.00058.

248. **Liu, Y.; Shi, H.; et al.**, Early prediction of acute xerostomia during radiation therapy for nasopharyngeal cancer based on delta radiomics from CT images. Quantitative Imaging in Medicine and Surgery, **2019**. 9(7): p. 1288-1302.
249. **Wang, J.; Liu, R.; et al.**, A predictive model of radiation-related fibrosis based on the radiomic features of magnetic resonance imaging and computed tomography. Translational Cancer Research, **2020**. 9(8): p. 4726-4738.
250. **Liu, W.; Zeng, C.; et al.**, A combined predicting model for benign esophageal stenosis after simultaneous integrated boost in esophageal squamous cell carcinoma patients (GASTO1072). Frontiers in Oncology, **2022**. 12.DOI: 10.3389/fonc.2022.1026305.
251. **Barua, S.; Elhalawani, H.; et al.**, Computed Tomography Radiomics Kinetics as Early Imaging Correlates of Osteoradionecrosis in Oropharyngeal Cancer Patients. Frontiers in Artificial Intelligence, **2021**. 4.DOI: 10.3389/frai.2021.618469.
252. **Calamandrei, L.; Mariotti, L.; et al.**, Morphological, Functional and Texture Analysis Magnetic Resonance Imaging Features in the Assessment of Radiotherapy-Induced Xerostomia in Oropharyngeal Cancer. Applied Sciences, **2023**. 13(2): p. 810.
253. **Smyczynska, U.; Grabia, S.; et al.**, Prediction of Radiation-Induced Hypothyroidism Using Radiomic Data Analysis Does Not Show Superiority over Standard Normal Tissue Complication Models. Cancers, **2021**. 13(21): p. 5584.
254. **Zhang, J.; Teng, X.; et al.**, Comparing effectiveness of image perturbation and test retest imaging in improving radiomic model reliability. Scientific Reports, **2023**. 13(1): p. 18263.DOI: 10.1038/s41598-023-45477-6.
255. MedCalc Statistical Software. **2020**, MedCalc Software Ltd, Ostend, Belgium.
256. **Brierley, J. D.**, TNM Classification of Malignant Tumours. 8 ed. **2016**: Wiley-Blackwell.
257. **Isensee, F.; Jaeger, P. F.; et al.**, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods, **2021**. 18(2): p. 203-211.DOI: 10.1038/s41592-020-01008-z.
258. **Brouwer, C. L.; Steenbakkers, R. J. H. M.; et al.**, CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiotherapy and Oncology, **2015**. 117(1): p. 83-90.DOI: https://doi.org/10.1016/j.radonc.2015.07.041.
259. **Nachalon, Y.; Nativ-Zeltzer, N.; et al.**, Cervical Fibrosis as a Predictor of Dysphagia. Laryngoscope, **2021**. 131(3): p. 548-552.DOI: 10.1002/lary.28880.
260. **Li, S.; Wan, X.; et al.**, Predicting prognosis of nasopharyngeal carcinoma based on deep learning: peritumoral region should be valued. Cancer Imaging, **2023**. 23(1): p. 14.DOI: 10.1186/s40644-023-00530-5.
261. **Xu, H.; Wang, A.; et al.**, Intra- and peritumoral MRI radiomics assisted in predicting radiochemotherapy response in metastatic cervical lymph nodes of nasopharyngeal cancer. BMC Medical Imaging, **2023**. 23(1): p. 66.DOI: 10.1186/s12880-023-01026-1.
262. **Zhang, J.**, Radiotherapy data analysis and reporting (RADAR) toolkit : an end-to-end artificial intelligence development solution for precision medicine, in Department of Health Technology and Informatics. **2023**, The Hong Kong Polytechnic University.
263. **van Griethuysen, J. J. M.; Fedorov, A.; et al.**, Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, **2017**. 77(21): p. e104-e107.DOI: 10.1158/0008-5472.Can-17-0339.
264. **Puttanawarut, C.; Sirirutbunkajorn, N.; et al.**, Biological dosiomic features for the prediction of radiation pneumonitis in esophageal cancer patients. Radiation Oncology, **2021**. 16(1): p. 220.DOI: 10.1186/s13014-021-01950-y.
265. **Zhang, J.; Teng, X.; et al.**, Quantitative Spatial Characterization of Lymph Node Tumor for N Stage Improvement of Nasopharyngeal Carcinoma Patients. Cancers (Basel), **2022**. 15(1).DOI: 10.3390/cancers15010230.
266. **Zwanenburg, A.; Leger, S.; et al.**, Assessing robustness of radiomic features by image perturbation. Scientific Reports, **2019**. 9(1): p. 614.DOI: 10.1038/s41598-018-36938-4.
267. **Panyura, P.**, Effect of Inter-observer Delineation Variability on Radiomics Features in Nasopharyngeal Cancer. **2022**, Chulalongkorn University.
268. **Vallat, R.**, Pingouin: statistics in Python. Journal of Open Source Software, **2018**. 3(31): p. 1026.DOI: 10.21105/joss.01026.

269. **Pandey, U.; Saini, J.; et al.**, Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images. Journal of Magnetic Resonance Imaging, **2021**. 53(2): p. 394-407.DOI: https://doi.org/10.1002/jmri.27349.
270. **Duan, J.; Qiu, Q.; et al.**, Reproducibility for Hepatocellular Carcinoma CT Radiomic Features: Influence of Delineation Variability Based on 3D-CT, 4D-CT and Multiple-Parameter MR Images. Frontiers in Oncology, **2022**. 12.DOI: 10.3389/fonc.2022.881931.
271. **Denzler, S.; Vuong, D.; et al.**, Impact of CT convolution kernel on robustness of radiomic features for different lung diseases and tissue types. British Journal of Radiology, **2021**. 94(1120).DOI: 10.1259/bjr.20200947.
272. **Project Jupyter**. JupyterLab. [cited 2024 Mar]; Available from: https://github.com/jupyterlab/jupyterlab.
273. **Li, Y.; Mansmann, U.; et al.**, Benchmark study of feature selection strategies for multi-omics data. BMC Bioinformatics, **2022**. 23(1): p. 412.DOI: 10.1186/s12859-022-04962-x.
274. **Xue, C.; Yuan, J.; et al.**, Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review. Quant Imaging Med Surg, **2021**. 11(10): p. 4431-4460.DOI: 10.21037/qims-21-86.
275. **Thompson, C. G.; Kim, R. S.; et al.**, Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. Basic and Applied Social Psychology, **2017**. 39(2): p. 81-90.DOI: 10.1080/01973533.2016.1277529.
276. **Cheng, J.; Sun, J.; et al.**, A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta A Mol Biomol Spectrosc, **2022**. 268: p. 120652.DOI: 10.1016/j.saa.2021.120652.
277. **Pedregosa, F.; Varoquaux, G.; et al.**, Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, **2011**. 12: p. 2825-2830.
278. **Chawla, N. V.; Bowyer, K. W.; et al.**, SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, **2002**. 16: p. 321-357.
279. **Lemaitre , G.; Nogueira, F.; et al.**, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, **2017**. 18: p. 1-5.
280. **Bolón-Canedo, V.; Sánchez-Maroño, N.; et al.**, A review of feature selection methods on synthetic data. Knowledge and Information Systems, **2013**. 34(3): p. 483-519.DOI: 10.1007/s10115-012-0487-8.
281. **Ding, C. and Peng, H.**, Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, **2005**. 03(02): p. 185-205.DOI: 10.1142/s0219720005001004.
282. **Mazzanti, S.** MRMR-selection. [cited 2023; Available from: https://github.com/smazzanti/mrmr.
283. **Iglewicz, B.**, Robust scale estimators and confidence intervals for location. Understanding robust and exploratory data analysis, **1983**: p. 405431.
284. **Hoerl, A. E. and Kennard, R. W.**, Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, **1970**. 12(1): p. 55-67.DOI: 10.2307/1267351.
285. **Alzubi, J.; Nayyar, A.; et al.**, Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series, **2018**. 1142(1): p. 012012.DOI: 10.1088/1742-6596/1142/1/012012.
286. XGBoost. Available from: https://github.com/dmlc/xgboost/.
287. **Geroldinger, A.; Lusa, L.; et al.**, Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study. Diagnostic and Prognostic Research, **2023**. 7(1): p. 9.DOI: 10.1186/s41512-023-00146-0.
288. **Lundberg, S. M. a. L., Su-In**, A Unified Approach to Interpreting Model Predictions, in Advances in Neural Information Processing Systems. **2017**, Curran Associates, Inc.
289. **Breiman, L.**, Random Forests. Machine Learning, **2001**. 45(1): p. 5-32.DOI: 10.1023/A:1010933404324.
290. **Niculescu-Mizil, A. and Caruana, R.**, Predicting good probabilities with supervised learning, in Proceedings of the 22nd international conference on Machine learning. **2005**, Association for Computing Machinery: Bonn, Germany. p. 625–632.
291. **Hajian-Tilaki, K.**, Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med, **2013**. 4(2): p. 627-35.
292. **Piovani, D.; Sokou, R.; et al.**, Optimizing Clinical Decision Making with Decision Curve Analysis: Insights for Clinical Investigators. Healthcare (Basel), **2023**. 11(16).DOI: 10.3390/healthcare11162244.
293. **Van Calster, B.; Wynants, L.; et al.**, Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology, **2018**. 74(6): p. 796-804.DOI: https://doi.org/10.1016/j.eururo.2018.08.038.

294. **Nicol, A. J.; Lam, S.-K.; et al.**, A Multi-Centre, Multi-Organ, Multi-Omic Prediction Model for Treatment-Induced Severe Oral Mucositis in Nasopharyngeal Carcinoma.Radiol med, **2024**. DOI: https://doi.org/10.1007/s11547-024-01901-z
295. **Otter, S.; Schick, U.; et al.**, Evaluation of the Risk of Grade 3 Oral and Pharyngeal Dysphagia Using Atlas-Based Method and Multivariate Analyses of Individual Patient Dose Distributions. International Journal of Radiation Oncology*Biology*Physics, **2015**. 93(3): p. 507-515.DOI: https://doi.org/10.1016/j.ijrobp.2015.07.2263.
296. **Vissink, A.; Jansma, J.; et al.**, Oral sequelae of head and neck radiotherapy. Crit Rev Oral Biol Med, **2003**. 14(3): p. 199-212.DOI: 10.1177/154411130301400305.
297. **Chinn, S.**, A Simple Method for Converting An Odds Ratio to Effect Size for use in Meta-analysis. Statistics in medicine, **2000**. 19: p. 3127-31.DOI: 10.1002/1097-0258(20001130)19:223.3.CO;2-D.
298. **Sanguineti, G.; Sormani, M. P.; et al.**, Effect of Radiotherapy and Chemotherapy on the Risk of Mucositis During Intensity-Modulated Radiation Therapy for Oropharyngeal Cancer. International Journal of Radiation Oncology, Biology, Physics, **2012**. 83(1): p. 235-242.DOI: 10.1016/j.ijrobp.2011.06.2000.
299. **Köstler, W. J.; Hejna, M.; et al.**, Oral Mucositis Complicating Chemotherapy and/or Radiotherapy: Options for Prevention and Treatment. CA: A Cancer Journal for Clinicians, **2001**. 51(5): p. 290-315.DOI: https://doi.org/10.3322/canjclin.51.5.290.
300. **Dodd, M.**, The pathogenesis and characterization of oral mucositis associated with cancer therapy. Oncol Nurs Forum, **2004**. 31(4 Suppl): p. 5-11.DOI: 10.1188/04.ONF.S4.5-11.
301. **Wennmann, M.; Rotkopf, L. T.; et al.**, Reproducible Radiomics Features from Multi-MRI-Scanner Test–Retest-Study: Influence on Performance and Generalizability of Models. Journal of Magnetic Resonance Imaging. n/a(n/a).DOI: https://doi.org/10.1002/jmri.29442.
302. **Wohlschlaeger, A.**, Prevention and Treatment of Mucositis: A Guide for Nurses. Journal of Pediatric Oncology Nursing, **2004**. 21(5): p. 281-287.DOI: 10.1177/1043454204265840.
303. **US Department of Health and Human Services**, Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0. . **2017**.
304. **Chen, A. Y.; Frankowski, R.; et al.**, The Development and Validation of a Dysphagia-Specific Quality-of-Life Questionnaire for Patients With Head and Neck Cancer: The M. D. Anderson Dysphagia Inventory. Archives of Otolaryngology–Head & Neck Surgery, **2001**. 127(7): p. 870-876.DOI: 10-1001/pubs.Arch Otolaryngol. Head Neck Surg.-ISSN-0886-4470-127-7-ooa00162.
305. **Zhang, M. L. and Zhou, Z. H.**, A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, **2014**. 26(8): p. 1819-1837.DOI: 10.1109/TKDE.2013.39.
306. **Xu, Y.; Jiao, L.; et al.**, Multi-label classification for colon cancer using histopathological images. Microsc Res Tech, **2013**. 76(12): p. 1266-77.DOI: 10.1002/jemt.22294.
307. **Jamthikar, A.; Gupta, D.; et al.**, A machine learning framework for risk prediction of multi-label cardiovascular events based on focused carotid plaque B-Mode ultrasound: A Canadian study. Computers in Biology and Medicine, **2022**. 140: p. 105102.DOI: https://doi.org/10.1016/j.compbiomed.2021.105102.
308. **Ceylan, Z. and Pekel, E.**, Comparison of Multi-Label Classification Methods for Prediagnosis of Cervical Cancer. International Journal of Intelligent Systems and Applications in Engineering, **2017**. 5(4): p. 232-236.DOI: 10.18201/ijisae.2017533896.
309. **Azzini, A.; Cortesi, N.; et al.**, A Multi-Label Machine Learning Approach to Support Pathologist's Histological Analysis. ENTRENOVA - ENTerprise REsearch InNOVAtion, **2019**. 5(1): p. 165-176.
310. **Ren, Y.; Chakraborty, T.; et al.**, Multi-label classification for multi-drug resistance prediction of Escherichia coli. Comput Struct Biotechnol J, **2022**. 20: p. 1264-1270.DOI: 10.1016/j.csbj.2022.03.007.
311. **Yap, X. H. and Raymer, M.**, Multi-label classification and label dependence in in silico toxicity prediction. Toxicology in Vitro, **2021**. 74: p. 105157.DOI: https://doi.org/10.1016/j.tiv.2021.105157.
312. **Lu, S.-H.; Cheng, J. C.-H.; et al.**, Volumetric modulated arc therapy for nasopharyngeal carcinoma: A dosimetric comparison with TomoTherapy and step-and-shoot IMRT. Radiotherapy and Oncology, **2012**. 104(3): p. 324-330.DOI: https://doi.org/10.1016/j.radonc.2011.11.017.
313. **Bosse, C.; Narayanasamy, G.; et al.**, Dose Calculation Comparisons between Three Modern Treatment Planning Systems. Journal of Medical Physics, **2020**. 45(3).
314. **Hanley, J. A. and McNeil, B. J.**, The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, **1982**. 143(1): p. 29-36.DOI: 10.1148/radiology.143.1.7063747.
315. **Jung, S.-H.**, Sample size calculation for comparing two ROC curves. Pharmaceutical Statistics, **2024**. n/a(n/a).DOI: https://doi.org/10.1002/pst.2371.

316. **Orlhac, F.; Eertink, J. J.; et al.**, A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. J Nucl Med, **2022**. 63(2): p. 172-179.DOI: 10.2967/jnumed.121.262464.
317. **Gidwani, M.; Chang, K.; et al.**, Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models. Radiology, **2023**. 307(1): p. e220715.DOI: 10.1148/radiol.220715.
318. **Teng, X.; Zhang, J.; et al.**, Building reliable radiomic models using image perturbation. Scientific Reports, **2022**. 12(1): p. 10035.DOI: 10.1038/s41598-022-14178-x.
319. **Hong Kong Hospital Authority**. Hospital Authority Data Sharing Portal. [cited 2024 July]; Available from: https://www3.ha.org.hk/data/DCL/Index/.
320. **Shen, B.; Zhou, Y.; et al.**, Efficacy of photobiomodulation therapy in the management of oral mucositis in patients with head and neck cancer: A systematic review and meta-analysis of randomized controlled trials. Head Neck, **2024**. 46(4): p. 936-950.DOI: 10.1002/hed.27655.
321. **Moroney, L. B.; Ward, E. C.; et al.**, Evaluation of a speech pathology service delivery model for patients at low dysphagia risk during radiotherapy for HNC. Support Care Cancer, **2020**. 28(4): p. 1867-1876.DOI: 10.1007/s00520-019-04992-x.