Aus der Radiologischen Klinik der Universität Heidelberg
(Geschäftsführender Direktor: Prof. Dr.med. Dr.rer.nat. Jürgen Debus)

Abteilung für RadioOnkologie und Strahlentherapie
(Ärztlicher Direktor: Prof. Dr.med. Dr.rer.nat. Jürgen Debus)

In Zusammenarbeit mit der Klinischen Kooperationseinheit
Translationale Radioonkologie am Deutsches Krebsforschungszentrum
(DKFZ)
(Leiter: Prof. Dr.med. Dr.rer.nat. Amir Abdollahi)

DOSIOMICS-ENHANCED PREDICTION MODELING: AN ARTIFICIAL INTELLIGENCE-BASED WORKFLOW IN RADIATION ONCOLOGY

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Patrick Salome

aus dem
Libanon

2022

Dekan:	Herr Prof. Dr. med. Hans-Georg Kräusslich
Doktorvater:	Herr Prof. Dr. Dr. Jürgen Debus

“What we know is a drop, what we don’t know is an ocean.”

— Isaac Newton

Dedicated to Maha and George Salome

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ADC: apparent diffusion coefficient

AIC: Akaike Information Criterion

BStrap: Bootstrapping

C-I: Concordance index

CIRT: carbon ions radiotherapy

COM: center of mass

CNN: Convolutional Neural Network

CONTRA: contra-lateral

CPH: Cox Proportional Hazard models

CRT: conformational radiotherapy

CS: clinical signature

CSF: cerebrospinal fluid

CT: computer tomography

CTV: Clinical Tumor Volume

CV: cross-validation

DD: dose distribution

DE: dependence entropy

DICOM: Digital Imaging COmmunication in Medicine

DWI: diffusion weighted imaging

DVH: dose volume histogram

DS: dosiomics signature

FLAIR: fluid attenuated inversion recovery

GBM: glioblastoma

GLCM: Gray Level Co-occurence Matrix

GLDM: Gray Level Dependence Matrix

GLRLM: Gray Level Run Length Matrix

GLSZM: Gray Level Size Zone Matrix

GM: gray matter

GMM: Gaussian mixture models

GTV: Gross Tumor Volume

H: high

HIT: Heidelberg Ion-Beam Therapy Center

HM: Nyul-Udupa histrogram matching

HNC: head and neck

HR: Hazard ratio

IBSI: Imaging Biomarker Standardization Initiative

IMRT: intensity modulated radiation therapy

IPSI: ipsi-lateral

ITV: Internal Tumor Volume

KDE: kernel density estimation

KM: Kaplan-Meier

L: low

LGLRE: Long Run Low Gray Level Emphasis

LRHGLE: Long Run High Gray Level Emphasis

LALGLE: Large Area Low Gray Level Emphasis

LoG: Laplacian of Gaussian

LoG3: Laplacian of Gaussian with sigma 3mm

MCCV: Monte Carlo cross-validation

MD: mode

MRI: magnetic resonance imaging

MSE: mean squared error

NN: no normalization

NSCLC: non-small cell lung cancer

NTCP: Normal Tissue Complication Probability

OAR: organs at risk

OS: overall survival

OR: original image

PACS: Picture Archiving and Communication System

POI: Poisson regression models

PTV: Planning Target Volume

RBE: radiobiological effective dose

RDCS: radiomics, dosiomics and clinical signature

RHGG: recurrent high-grade glioma

RMS: Root Mean Squared

ROI: region of interest

RRRS: Re-irradiation risk score

rs: Spearman correlation coefficient

RSF: Random survival forest

RT: radiotherapy

SAE: Small Area Emphasis

SALGLE: Small Area Low Gray Level Emphasis

SE: sensitivity

SD: series description

SOBP: spread-out Bragg peak

SH-CT: sharp kernel reconstructed CT

SM-CT: smooth kernel reconstructed CT

SP: specificity

SRHGLE: Short Run High Gray Level Emphasis

SRLGLE: Short Run Low Gray Level Emphasis

SS: structure set

SWI: susceptibility weighted imaging

SZNUN: Size Zone Non Uniformity

T1w: T1-weighted

T1wce: T1-weighted contrast-enhanced

T2w: T2-weighted

TCP: Tumor Control Probability

TP: Treatment Plan

TPS: Treatment Planning System

TV: target volume

UID: Unique Identifier

UKHD: Heidelberg University Hospital

VMAT: volumetric modulated arc therapy

WHO: World Health Organisation

WM: white matter

WS: white stripe

WV: wavelet filter transformation

1 INTRODUCTION

Accounting for approximately 10 million deaths in the year 2020, cancer - the multi-stage process that alters normal cells into tumorous cells based on the interaction of genetic factors and physical, chemical, and biological carcinogens - is nowadays the second leading cause of death globally (WHO, 2022). Typically, treatment follows three main approaches: surgery, radiotherapy (RT), systemic therapy, e.g., chemotherapy, or immunotherapy, or a combination thereof. In this context, more than half of cancer patients receive RT for curative or palliative purposes (Atun et al., 2015). While ionising radiation targets tumour tissue, healthy tissue is inevitably affected due to dose deposition along the irradiated path. A precise RT treatment plan is thus necessary to spare injuries to the organs at risk around the tumour and destroy malignant cells simultaneously (Joiner and Kogel, 2018). Even if modern RT techniques can minimise or prevent toxicity and improve tumour coverage, several patients still face secondary effects of irradiation or tumour recurrence. Therefore, the determination of factors affecting normal tissue complication probability (NTCP) and tumour control probability (TCP) is an active area of research, especially for tumours with a high risk of recurrence (Reda et al., 2020). Traditional TCP/NTCP modelling focused exclusively on dosimetric predictors (Burman et al., 1991; Seppenwoolde et al., 2003). With the increased medical data availability, such as imaging data in RT planning and follow-up, molecular data for tumour classification or derived RT physical information for treatment delivery, the need to process it to render it useful for patient benefit has also grown (Kang et al., 2015). To this end, machine learning methods have been increasingly adapted for TCP/NTCP modelling, with recent studies focusing on the use of either clinical, dosimetric, radiomics data or combinations of these (Gulliford, 2015; Lambin et al., 2012). Still, there has been little focus on spatial analysis of dose distribution (DD) - dosiomics - and its impact on prediction modelling (Placidi et al., 2021b).

This work aims to incorporate multiple layers of information from medical data, emphasising features extracted from the 3D DD that could potentially improve patient stratification and prognostication. Furthermore, the added benefit for TCP/NTCP modelling of a combined modelling approach, i.e., the use of radiomics, dosiomics, and clinical features, is evaluated in three retrospectively collected patient cohorts treated with either carbon ions or standard RT, each corresponding to a different entity. The curation, preprocessing, and analysis pipeline built with this purpose demonstrated its suitability in different modalities, specifically DD, computer tomography (CT), and multi-parametric magnetic resonance imaging (MRI) for three entities -

brain, head and neck, and lung and different aspects - NTCP and TCP estimations.

1.1 RADIOTHERAPY

Radiation therapy plays a significant role in treating and potentially curing most tumour entities. It is prescribed in more than half of the patients with cancer, either alone or in combination with systemic treatment or surgery (Atun et al., 2015). Over the last decades, how RT is delivered has changed to reduce the dose reaching the organs at risk (OAR) and improve the dose to the tumour tissue (Joiner and Kogel, 2018). A high-quality RT treatment plan is of utmost importance to optimise the delivery. In recent times, external beam RT mostly uses high-energy photons or particles like protons or carbon ions, allowing for good target coverage (Laskar and Kakoti, 2022).

1.1.1 *High-energy photons radiotherapy*

Electron linear accelerators are used to generate radiation for high-energy photon RT. A beam of MeV photons delivers the dose to the tissue at an exponential absorption rate after the initial increase. This renders the maximum delivered dose at around 2-3cm deep in soft tissue. Photons do not deposit significant energy themselves but rather through the ionisation of atoms. Photons transfer their energy to positrons and electrons, which ionise atoms along particle tracks until their energy is lost. Since photons are massless and chargeless, they penetrate deeply into the body while sparing the skin (Bhide and Nutting, 2010). Modern techniques include intensity modulation radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT), which, in comparison to conformational radiotherapy (CRT), improve target volume conformity and reduce radiation-induced toxicities (Nutting, Dearnaley, and Webb, 2000; Teoh et al., 2011; Webb, 1993).

1.1.2 *Hadron therapy*

Hadron therapy includes carbon and proton ions radiation. Producing these ion beams and targeting the tumour requires a dedicated centre, where the therapy process starts from the accelerator complex, i.e., a synchrotron or cyclotron. Due to their increased penetration depth and distinct dose fall-off, carbon ions have been used in clinical settings since 1946 (Rackwitz and Debus, 2019). The depth profiles of carbon ions show a significant but narrow increase in dose at the end of the range - the Bragg peak. Additionally, a monoenergetic proton beam has a defined range in a specific medium. Therefore the Bragg peak has to be extended to cover the extended target volume (TV),

introducing the spread-out Bragg peak (SOBP), which can be achieved by superimposing monoenergetic Bragg curves in the desired range (Bortfeld and Schlegel, 1996). The radiobiological effective dose (RBE) delivered by carbon ions radiotherapy (CIRT) proves to be favourable, with low RBE values at the beginning of the path and at higher depths, with the RBE range depending on the tumour (Amaldi and Kraft, 2005). Currently, CIRT doses are prescribed as RBE-weighted quantities according to Gy(RBE) scales, which until now are not categorically resolved (Molinelli et al., 2016).

1.1.3 *Personalised radiotherapy*

With different RT treatment options available, the selection criteria currently depend mainly on clinical and pathological features, e.g., tumour stage, primary site, or histology (Glatzer et al., 2020; Panje et al., 2018). In recent years, studies started investigating the integration of patient-specific information such as disease subtype or molecular properties of tumours into the RT treatment decision, i.e., going towards personalised treatment planning explicitly tailored to the patient (Fröhlich et al., 2018; Schork, 2015). This is important to improve therapy outcomes and reduce the possibility of RT mistreatment, which can lead to significant side effects impacting the quality and quantity of life (Ford and Terezakis, 2010). In this context, predicting the effectiveness of a treatment for a specific patient would be highly desirable.

1.1.4 *Treatment planning*

RT treatment planning is a time and effort-intensive process that can take days to complete. By starting with a list of target coverage and OAR constraints, a medical physicist has to decide about beam energy, number, angles, etc. After the initial planning and iterative optimisation, the RT treatment plan is available as a Digital Imaging Communication in Medicine (DICOM) object (Mildenberger, Eichelberg, and Martin, 2002). DICOM was developed as a generalised medical image object format to manage medical image acquisition and storage. It enables different imaging devices to communicate with each other and facilitates the export of data from PACS, the Picture Archiving and Communication System (Gudivada and Raghavan, 1995). The treatment plan typically contains the planning imaging data, DD, RT plan, and RT Structure Set (SS). RT plan contains geometric and dosimetric data that summarises the RT treatment delivered. This includes all information about the irradiation based on the beam setup parameters (gantry angle, couch angle, beam modifiers, etc.). RT DD contains the DDs generated by a Treatment Planning System (TPS). Typically, several DD files would be present, each representing a specific format, i.e., the physical or RBE DD, each per beam, per RT fraction, or the entire RT plan. An RT SS includes contoured

structures of the patient's anatomy. The different structures can be split into OAR and the TV, i.e., the gross tumour volume (GTV), clinical target volume (CTV), internal target volume (ITV), and planning target volume (PTV). The different entities are usually identified on TPS or simulation workstations. The contouring and planning are typically performed on CT or MRI-CT co-registered images (Rai et al., 2017). For certain entities, most modern clinics use both MRI and CT for treatment planning to take advantage of the superior soft tissue and tumour contrast of the MRI, mainly for a better GTV delineation, while still using the electron density values provided by the CT, which is necessary for the dose calculation (Rai et al., 2017). The relationship between the different RT-objects is identified by several DICOM Unique Identifiers (UID)s (Newhauser et al., 2014), with each object having a file-specific instance UID and series and study UID unique to the delivered treatment plan. Reference UID tags describe which files are meant to be associated with a particular tag so that they can be retrieved and used by different DICOM files.

1.2 NTCP/TCP MODELLING

The main goal of RT is to deliver an optimal dose to control the tumour tissue while avoiding excessive healthy tissue toxicity. Predictive models have been developed to calculate tumour control and normal tissue complication probabilities. The difference between TCP and NTCP defines the therapeutic window in RT. In this context, current research aims to increase the therapeutic window by improving the efficacy of RT, enhancing the tumour response to irradiation, or decreasing normal tissue toxicity. Radiobiological studies have shown that the dose response follows a sigmoidal curve, suggesting that the dose which induces complication is based on a probability distribution. In general, NTCP can be defined as

$$\mathrm{NTCP}(d) = \int_{-\infty}^{d} p(x)\,dx \tag{1}$$

where d is the irradiation dose, and $p(x)$ is the probability of the dose-inducing complication to be x. The shape of the NTCP function has been modelled over time with various representations such as probit model (Holthusen, 1936) or logistic or log-logistic distribution (Suit, 1965). Lyman (Lyman, 1985) developed a model to consider situations when only part of the organ is irradiated. Later, Kutcher and Burman (Burman et al., 1991) extended the Lyman model into a dose-volume histogram (DVH) reduction scheme, followed by Mohan, who allowed the bypassing of the DVHs (Mohan et al., 1992). As the Lyman model does not include risk factors as comorbidities and does not consider that patients reported without toxicity might have developed it after the follow-up period, Tucker et al. (Tucker et al., 2008) developed the gener-

alised Lyman model which attempts to accounts for these issues. While in the standard Lyman model, the probability of observing a complication after irradiation with dose D of subvolume V is defined as

$$NTCP(D, V) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{u^2/2}\, du \tag{2}$$

where

$$t = \frac{D - TD_{50}/V^n}{m \cdot TD_{50}/V^n} \tag{3}$$

where n is the volume parameter, m is a dimensionless parameter, and TD50 is the subvolume dose resulting in a 50% probability of developing an event after irradiation. If the whole volume is exposed to dose, so that.

$$D_{eff} = (\sum_i v_i \cdot D_i^{1/n})^n \tag{4}$$

To incorporate covariates, the quantity in equation (3) can be replaced as

$$t = \frac{D_{eff} - TD_{50} \cdot exp(\delta_1 \cdot Y_1) \cdot ... \cdot exp(\delta_k \cdot Y_k)}{m \cdot TD_{50} \cdot exp(\delta_1 \cdot Y_1) \cdot ... \cdot exp(\delta_k \cdot Y_k)} \tag{5}$$

where the variables Y_1 to Y_k are the nondosimetric risk factors, and the term $exp(\delta_i Y_i)$ is the dose-modifying factor (DMF) for TD_{50} in the presence of risk factor Y_i. To model the times to toxicity at time τ, a log-normal distribution is commonly assumed and is defined as

$$f(\tau) = \frac{1}{\sigma\tau\sqrt{2\pi}} \cdot e^{-(\ln\tau - \mu)^2/2\sigma^2} \tag{6}$$

with μ and σ as latency parameters. The generalised Lyman model is a mixture of the incidence component, NTCP, and the latency component, $f(\tau)$. Therefore, the contribution to the likelihood of a patient experiencing toxicity at time τ is

$$NTCP(D_{eff}, Y_1, ..., Y_k) \cdot f(\tau) \tag{7}$$

and for a patient without experiencing toxicity is

$$1 - NTCP(D_{eff}, Y_1, ..., Y_k) \cdot F(\tau) \tag{8}$$

with $F(\tau)$ being the cumulative distribution function corresponding to $f(\tau)$.

Another category of modelling is the use of tissue architecture models, which propose that assumptions about the structure of a tissue and its response to radiation can successfully model NTCP and TCP. The most common tissue-architecture models are the Poisson model, proposed by Munro and Gilbert, which assumes that the cell survival curve takes a log-linear form (Munro and Gilbert, 1961). Another tissue-architecture model is the relative seriality model introduced by Kallman, where the organ at risk is divided

into functional subunits (Källman, Ågren, and Brahme, 1992). Data-driven approaches take a new direction and aim to find patterns in high-dimensional data sets to make predictions. In place of hand-crafted features, the models extract informative features from data to predict probabilities of events. The process of extracting insights from data is a topic belonging to machine learning.

1.3 MACHINE LEARNING AND DEEP LEARNING

Briefly, machine learning can be described as the ability of computers to make predictions based on previous experiences (Baştanlar and Özuysal, 2014). Due to the increased storage capacity and the development of more powerful computers, machine learning has been employed in many disciplines, including bioinformatics. Machine learning works by processing the available data and building a computational model of its intrinsic complex relationships, which are usually hard to be observed by humans. The process of building the model is named training, and a trained model can predict outputs for previously unseen input values.

To accurately train a machine learning model, the data must be completely relevant to the problem, since the algorithm maps the extracted features only from the input data to the output values. If the output values exist, the technique is called supervised machine learning. In the absence of output values - and hence the algorithm is to find patterns automatically - it is called unsupervised machine learning (Baştanlar and Özuysal, 2014). A drawback of machine learning models is the need for careful feature engineering and field experts to transform the raw data into feature vectors suitable for the algorithm.

On the other hand, deep learning - a branch of machine learning - is a technique that can automatically discover the optimal data representations for the algorithm. It uses simple (usually non-linear) modules combined into multiple representation layers, which, in a high enough number, can describe very complex problems (LeCun, Bengio, and Hinton, 2015). Commonly, deep learning algorithms are employed for segmentation and classification tasks. A widely used architecture in classification tasks is the residual network (ResNet), introduced in 2015 to deal with the degradation problem, i.e., the degradation of the network accuracy as the depth of the network increases (He et al., 2016). Besides the usual deep convolution neural network architecture for classification purposes, ResNet introduces skip-connections that skip one or more layers, an important concept in medical image segmentation for increasing the speed of convergence and allowing the training of very deep networks (Drozdzal et al., 2016).

1.4 RADIOMICS

The extraction of features from medical images using data characterisation algorithms i.e. radiomics is an upcoming field of research expected to yield (non-invasive) surrogates for important (molecular) characteristics, e.g., in tumours (similar to biomarkers) to predict recurrence patterns or to gain insight about potential adverse effects such as normal tissue toxicity after irradiation (Lambin et al., 2012). Those approaches have the potential to yield detailed information about longitudinal disease development as well as information about tissue or tumour heterogeneity, which is not routinely assessed by biopsies (Sforazzini et al., 2021). Correlation studies quantify features from medical imaging data (e.g., shape or texture features) and compare them to important clinical and molecular covariates, to develop prognostic and predictive models. The potential causal inference of these models can be addressed by evaluating longitudinal alterations (e.g., for normal tissue toxicity) (Scapicchio et al., 2021). Together with RT data, radiomics can build so-called RT predictive models. To this date, radiomics has been successfully applied in various cancer studies for TCP and NTCP modelling after RT (Ding et al., 2021).

1.5 DOSIOMICS

Dosiomics can be regarded as an extension of radiomics - the same concept applied to three-dimensional DD rather than imaging data - to obtain spatial and statistical information. Through dosiomics analysis, the DD can be parametrised into regions of interest (ROIs) by extracting, e.g., textural or shape-based features. The DD can thus be described at a higher complexity level. Currently, dose information is obtained from dose-volume histograms (DVHs) since optimisation and evaluation are based on DVH endpoints, metrics based on DVH, and visual inspection (Placidi et al., 2021b). A disadvantage of solely using DVH is the loss of valuable information on spatial and statistical distribution. Integrating DVH with dosiomics analysis has the potential to reveal new metrics suitable for the evaluation of radiotherapy treatment plans. Furthermore, low or high dose-level value areas in target volumes or organs at risk can be quantified, deriving more information about infiltrative zones, which can be integrated into the RT plan optimisation. Initial studies using dosiomics in survival and radiation-induced toxicity prediction modelling have already been published (a summary is presented in Table 1.1). Interest in evaluating dosiomics features reproducibility and stability has also been seen recently (Adachi et al., 2022; Placidi et al., 2021a; Placidi et al., 2021b; Puttanawarut et al., 2022).

Since spatial information is contained in the DVH plots only in a summarised manner, different dose distributions could have the same DVH curve.

Table 1.1: Summary of published dosiomics studies

Studies	**Aim**
(Buizza et al., 2021)	local control prediction in skull-base chordoma patients (TCP)
(Gabrys, 2020)	prediction of radiation-induced Xerostomia in head and neck cancer patients (NTCP)
(Wu et al., 2020)	predicting the locoregional recurrences in head and neck cancer patients (TCP)
(Adachi et al., 2021; Liang et al., 2019)	prediction of radiation pneumonitis in lung cancer patients (NTCP)
(Lee et al., 2020)	predicting acute-phase weight loss in lung cancer patients (TCP/NTCP)
(Murakami et al., 2022)	correlation between planned dose distribution and biochemical failure in prostate cancer patients (TCP)
(Rossi et al., 2018)	predicting gastrointestinal and genitourinary toxicities in prostate cancer (NTCP)

Extracting features from the DD regarded as an image can thus reveal patterns of variations in the DD and contribute to the prediction of therapy response, as well as provide additional insights regarding uncertain infiltrative zones or organs at risk. Murakami et al., 2022 found that the dosiomics features were significantly correlated to biochemical recurrence in prostate cancer patients, suggesting a need for new ways of evaluating TP quality. Adachi et al., 2021 were able to improve the prediction of radiation pneumonitis incidence by using dosiomics. Similarly, Liang et al., 2019 found that dosiomics features can predict response, but the explainability of the features is not straightforward and needs further targeted studies.

1.6 TYPICAL RADIOMICS AND DOSIOMICS WORKFLOW

A radiomics workflow consists of data preparation (export and curation), ROI delineation, data preprocessing, feature extraction and selection, and modelling. An overview is presented in Fig. 1.1

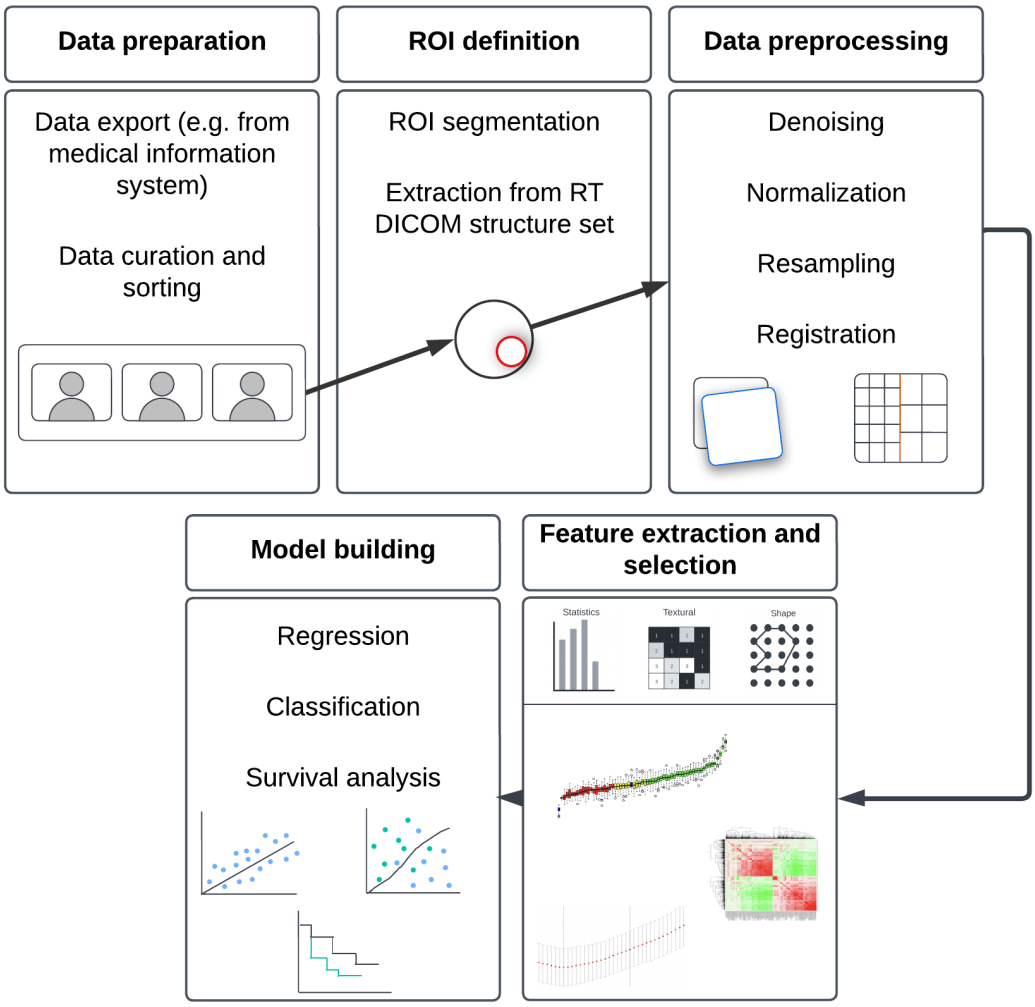


Figure 1.1: Overview of the radiomics and dosiomics workflow

1.6.1 *Cohort preparation*

In a radiomics and dosiomics workflow, since the analysis uses imaging and RT data, it is of utmost importance to have clean and organised data to be processed correctly and meaningfully. In medical studies, data is usually collected from one or more clinical institutions. With different clinics using different imaging devices, i.e., from different vendors, with different protocols, etc., the collected data is not structured similarly across the incoming sources and is thus difficult to use by a machine. The different abbreviations can result in being confounding variables, i.e. variables that would influence both the dependent and independent variables, resulting in spurious correlations. This renders the curation and sorting step extremely important in a radiomics and dosiomics workflow (Lambin et al., 2017). In the curation step, multiple elements of interest are extracted from the DICOM header to be later used in

the analysis. Sorting the data, on the other hand, means structuring the data according to the desired application to be subsequently fed into a computation model.

1.6.2 *ROI delineation and data preprocessing*

After structuring the datasets, the goal is to limit the image variations so that the images can be directly and correctly comparable inside the computational model. To this end, multiple steps are usually employed, depending on the imaging modality and anatomical region. A common step is segmenting the desired structure out of the original image - this reduces the overall data size and the irrelevant information that could be useless to analyse and would increase computation time for no added value. Registration of multiple modalities, e.g., MRI, and CT data, can also be necessary to use the TV already segmented following institutional guidelines included in the RT SS. Automatic segmentation solutions can be employed in the absence of RT SS. In MRI studies, another crucial step is intensity normalisation. Since MR intensities are acquired in arbitrary units, they are not directly comparable, especially when coming from various scanners (Alam and Rahman, 2018; Collewet, Strzelecki, and Mariette, 2004), influencing the analysis and thus have to be normalised. Grey-level discretisation, i.e. the clustering of pixels based on intensity values to reduce feature calculation time and noise, is also an essential step before feature extraction. Many other imaging preprocessing methods exist; their use depends on the application, and the clinical question addressed.

1.6.3 *Feature extraction and selection*

The process of obtaining meaningful information from data is called feature extraction. Typical features extracted in radiomics and dosiomics studies are presented in Table 1.2. A derived image is an image on which a filter, e.g., wavelet, square root, or Laplacian of Gaussian (LoG), was applied.

Since multiple definitions for calculating features can be used, the Imaging Biomarker Standardization Initiative (IBSI) aims to standardise the process (Zwanenburg et al., 2020). Extracting features using the mentioned methods can yield a rather large number of features, out of which not all will prove relevant for further analysis. To remove redundant features non-informative to the model, feature selection methods such as filter, wrapper, or embedded methods are usually employed (Jović, Brkić, and Bogunović, 2015). The choice of method depends on whether the number of observations is bigger or smaller than the number of variables.

Table 1.2: Typical features extracted in radiomics and dosiomics studies

Feature type	**Description**	**Image type on which features are applied**
first order	distribution of voxel intensities within ROI	original, derived
shape	3D size and shape of ROI	original
texture	variation of intensities inside ROI	original, derived

Filter methods

Filter methods are chosen depending on the task, e.g., regression, classification, or clustering. These methods use a performance measure to select features. Univariate - methods which evaluate a single feature - as well as multivariate - methods which evaluate a whole feature subset - filter methods are common approaches whose choice depends on the number of observations and predictors. Univariate analysis is commonly implemented in the medical field.

Wrapper methods

Wrapper methods evaluate the performance of feature subsets on a modelling algorithm chosen according to the task, e.g., classification or clustering algorithms. These methods are typically slower than filter methods but return feature subsets with improved performance because of using real evaluation models.

Embedded methods

Embedded methods - as the name suggests - are embedded in the algorithm and thus select features during the execution of the model. They typically introduce penalties to the features that do not bring a contribution to the model.

Common feature selection methods for each type are presented in Table 1.3.

1.6.4 *Modelling*

Variables are entities that vary in value and are essential components of modelling. On the one hand, they can be classified into quantitative and qualita-

Table 1.3: Typical feature selection methods

Type	**Methods**
filter	information gain, correlation, chi-square, fisher score, spectral feature selection
wrapper	based on modelling algorithm, e.g., K-means, support vector machines
embedded	Lasso, Boruta

tive variables. That is, quantitative for variables that differ in quantity, e.g., weight and qualitative for variables that vary in quality, e.g., skin type. Quantitative variables can be further classified into discrete or continuous variables. Discrete variables take no values between two given values, while continuous variables can take any value between two given values. Qualitative variables contain categorical variables, e.g., gender, and ordinal variables are similar to categorical but can be put in a specific order, e.g., the scale for severity of an effect. On the other hand, variables can be classified into dependent and independent variables - dependent variables are directly connected to the study's outcome. In contrast, independent variables are not affected by the outcome of the study but can affect the dependent variables if manipulated correctly (Kaliyadan and Kulkarni, 2019).

Linear regression

Linear regression is a method used to find linear relationships between one or multiple variables, applied in forecasting and prediction (Maulud, 2020). The types of linear regression are either simple regression, multivariate regression, or polynomial regression.

Classification

Classification methods try to predict the class of a categorical variable by finding the relationships between input and output. They can be used to predict one or multiple classes. Commonly applied methods for classification are random forests, support vector machines, and k-nearest neighbours (Choubey et al., 2020).

Survival analysis

Survival analysis is a broad modelling topic and has been dedicated to an entire section (1.7).

1.7 SURVIVAL ANALYSIS: TIME TO EVENT MODELLING

Usually, in cancer treatment therapy assessment studies, the time to death and the time between response to treatment and recurrence are interesting to monitor. The event and observation period must be clearly defined in these cases. Typically, in the course of cancer studies, a percentage of individuals have experienced the event. In contrast, the rest have not, making the problem of survival analysis difficult as the survival times will be unknown for a subset of patients. This behaviour is called censoring and it happens either by a patient not reaching the studied event by the end of the study, the patient not being able to be followed up during the study period, or the patient experiencing another event that makes follow-up unsuitable anymore (Clark et al., 2003). The existence of censored data means that specific analysis methods are needed.

1.8 BOTTLENECKS IN USING CLINICAL COHORTS

Data curation is a critical aspect of a radiomics and dosiomics workflow. Nonetheless, to this date, the organisation of clinical cohorts is not standardised, rendering proper data curation a current bottleneck when using clinical cohorts (Lambin et al., 2017). Additionally, when analysing multi-modality data, cohorts are often incomplete, missing certain modalities due to technical or practical reasons, thus introducing the problem of data completion (Cai et al., 2018). Deep learning is a tool that can be successfully used in aiding with both data curation and completion.

1.8.1 *Deep learning for data curation*

Deep learning has been employed in various research areas for data curation purposes, such as medical document triage (Lee et al., 2018) or annotation tool (Demirer et al., 2019). Nonetheless, data curation is a continuous problem needing innovative solutions where deep learning can prove extremely helpful (Thirumuruganathan et al., 2020).

1.8.2 *Deep learning for data completion*

In literature, deep learning has been used to complete multi-modality data by employing deep adversarial learning (Cai et al., 2018), generative networks (Chen et al., 2019), and deep multimodal learning (Li et al., 2020). A successful method used for automatic segmentation is nnU-Net (Isensee et al., 2019a).

2 AIMS

This thesis aims to incorporate multiple medical information layers into a combined modelling approach to improve patient stratification and prognostication while building a data curation, preprocessing, and analysis pipeline for faster deployment of artificial intelligence applications. The combined modelling approach includes using radiomics, dosiomics, and clinical features with the survival and radiation-induced toxicity outcomes in three retrospectively collected cohorts treated with either standard radiotherapy or carbon ions radiotherapy, each corresponding to a different entity, i.e. brain, head and neck, and lung. A schematic overview of the thesis aims is shown in Figure 2.1.

The contributions made by this work w.r.t. the aims are:

- For allowing the analysis of large, heterogeneous cohorts:
  1. Data curation tool for MR images
  2. Data completion approach for missing data points
  3. Study of the impact of MR intensity normalization methods on further analysis. This study was applied for:
     a) the development of a methodology for assessing the impact of normalization methods during preprocessing
     b) the evaluation of the need to report the used normalization method
- For model development with the aim of improving NTCP/TCP prediction:
  1. Development of a combined modelling approach, integrating radiomics, dosiomics, and clinical data. This framework was applied for:
     a) the prediction of overall survival, progression-free survival, and radiation-induced toxicity in recurrent high-grade glioma, non-small cell lung cancer, and head and neck cancer patients
     b) the stratification of patients into different risk groups
     c) the visualization of the different models and the impact of the different modalities on the overall prediction

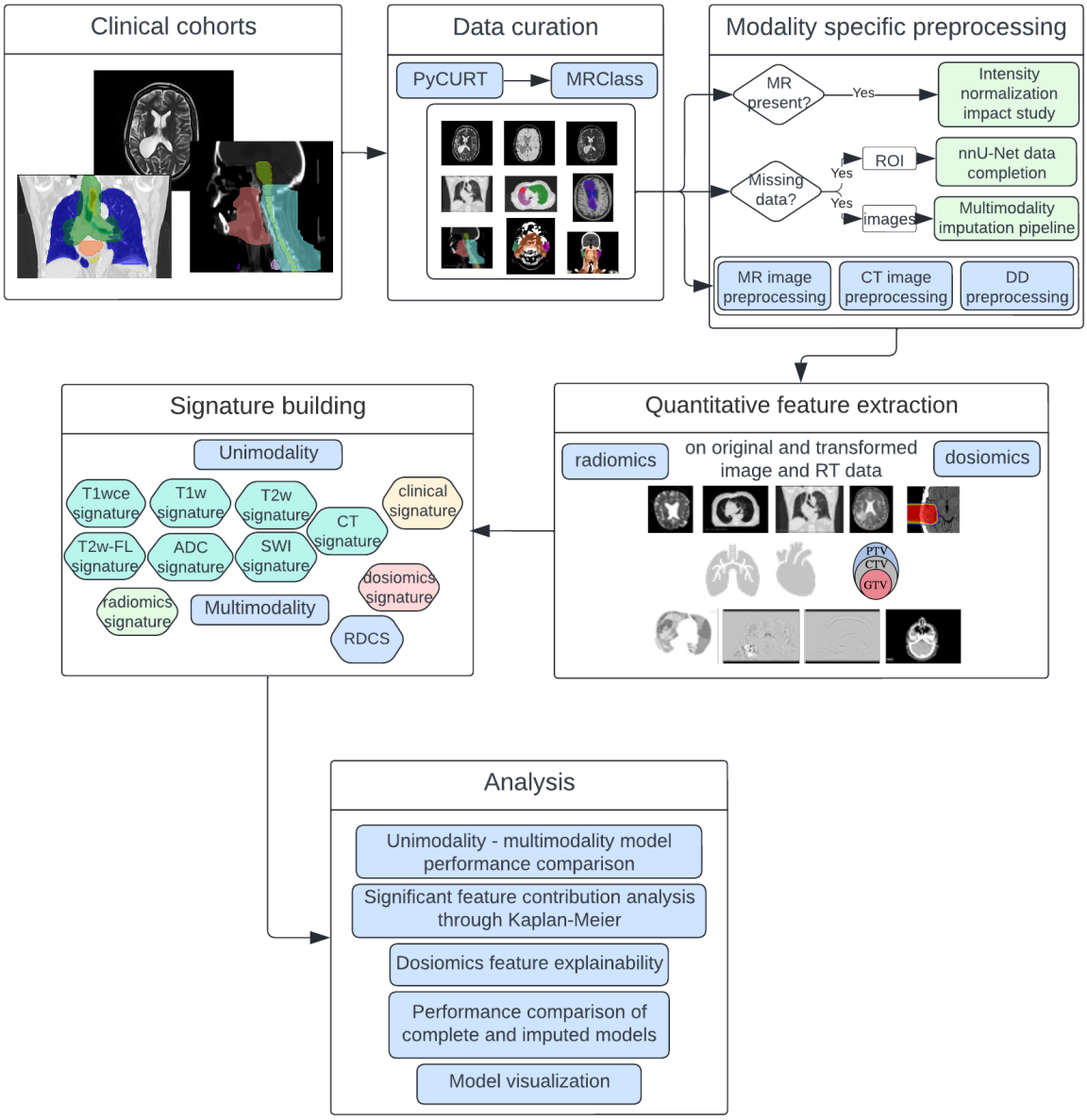


Figure 2.1: Schematic overview of the aims tackled in this work

3 MATERIALS AND METHODS

3.1 COHORTS

Survival and radiation-induced toxicity prediction modelling were performed on three cohorts with three different entities, i.e., recurrent high-grade glioma treated with CIRT, referenced as rHGG, early-stage non-small cell lung cancer treated with external beam stereotactic body radiation therapy (SBRT), referenced as NSCLC and head and neck tumour patients treated with intensity-modulated radiotherapy (IMRT) and helical tomotherapy referenced as HNC. The aim is to assess whether TCP and NTCP predictions can be improved through a multimodality approach incorporating clinical/histological, morphological information through the RT planning medical images, i.e. CT or MR and physical information through the RT DD in the different considered entities. Additionally, it is aimed to understand what are the determinants thereof, e.g. the biology (through radiomics), dose delivered or linear energy transfer (through dosiomics) in the target volumes or organs at risk.

Table 3.1: An overview of the studies conducted on each cohort assessed in this work. pHGG: Primary high-grade glioma, rHGG: Recurrent high-grade glioma, TCGA-GBM: The Cancer Genome Atlas - Glioblastoma, NSCLC: non-small lung cancer, HNC: head and neck

	AI workflow applications		
	Data curation MR-Class (4.1)	Data completion nn-Unet (4.2)	Intensity normalization study (4.3)
pHGG	x	-	x
rHGG	x	-	x
TCGA-GBM	x	-	-
NSCLC	-	x	-
HNC	-	-	-

	TCP/NTCP modeling			
Cohort	Overall survival prediction (4.4)	Progression-free survival prediction (4.4)	Fibrosis prediction (4.4)	Xerostomia prediction (4.4)
pHGG	-	-	-	-
rHGG	x	x	-	-
TCGA-GBM	-	-	-	-
NSCLC	x	x	x	-
HNC	-	-	-	x

The three cohorts, together with two additional cohorts, were used to train and test the different AI-based workflow applications. The additional cohorts are a primary high-grade glioma treated with photon RT referenced as pHGG and a public GBM cohort retrieved from the Cancer Imaging Archive (TCIA) referenced as TCGA-GBM (Scarpace et al., 2016). They were used to train the DCNNs for the data curation of MR images. Furthermore, a subset of the pHGG cohort including patients with at least two MR sequences taken no longer than 30 days before RT was used in the computational experiment that evaluated the impact of MR normalization methods on different MR sequences. An overview of the analyses performed on each cohort can be seen in Table 3.1. In the following sections, summaries of the considered cohorts are given, starting with the discovery, and test sets (80%/20% split) of the three main cohorts where NTCP/TCP modelling was performed, followed by the extra cohorts used to train and validate the built deep learning applications (Summaries of the extra cohorts as well as cohort specifications are shown in Appendix A).

3.1.1 *Recurrent high-grade glioma*

Despite therapy, recurrence of high-grade glioma is extremely common, eventually occurring in most patients (Hervey-Jumper and Berger, 2014). The dataset contained 197 patients with a median of 7 time points (corresponding to a median of 4 months), resulting in 11333 images acquired between 2009 and 2018, retrospectively collected from 15 different scanners at UKHD. The patients were treated with CIRT at the Heidelberg Ion-Beam Therapy Center (HIT), according to the CINDERELLA trial (Combs et al., 2010a). The median dose was 42GyRBE in 14 fractions. The cohort included 71 patients with grade III and 126 with grade IV (GBM) with a median follow-up time point of 34.2 months, median OS of 9 months [3-87], and PFS of 5 months [3-80] at reRT. OS was calculated as the number of days between the start of the reRT and death. PFS was calculated as the number of months between the beginning of the reRT and progression or death. Progression events were derived from the clinical follow-ups' reports. A summary of patient demographics (discovery and test sets with an 80%/20% split) is presented in Table 3.2.

Table 3.2: Recurrent high-grade glioma cohort overview. RRRS: reirradiation risk score developed by (Niyazi et al., 2018). P-value calculated by chi-squared or Fishers' tests

Characteristics	Discovery set (n=157)	Test set (n=40)	P-val
Age at reRT (years)			0.18
Median [range]	51 [16-79]	55 [32-71]	
⩾65	19 (12%)	6 (15%)	
65	138 (88%)	34 (85%)	
Gender			0.28
Male	93 (59%)	28 (70%)	
Female	64 (41%)	12 (30%)	
Karnofsky performance score (KPS)			0.68
⩾ 70	137 (87%)	35 (87.5%)	
<70	20 (13%)	5 (12.5%)	
RRRS*			0.66
Median [range]	-0.01 [-0.62-1.21]	-0.04 [-0.48-1.13]	
Good	47 (29.9%)	14 (35%)	
Intermediate	93 (59.2%)	24 (60%)	
Poor	17 (10.9%)	2 (5%)	
WHO grade, reRT			0.97
III	56 (36%)	15 (37.5%)	
IV	101 (64%)	25 (62.5%)	
MGMT promoter			0.85
Hypermethylated	19 (12%)	10 (25%)	
Not hypermethylated	17 (11%)	8 (20%)	
Unsure	1 (0.6%)	0 (0%)	
Missing	120 (76.4%)	22 (55%)	
IDH1			0.71
Mutant	33 (21%)	8 (24%)	
Wildtype	23 (15%)	14 (31%)	
Missing	101 (64%)	18 (45%)	
1p/19q Codeletion			0.44
Yes	12 (7.6%)	4 (10%)	
No	11 (7.4%)	2 (5%)	
Missing	134 (85%)	34 (85%)	
Reresection			0.84
Yes	21 (13%)	18 (18%)	
No	136 (87%)	32 (82%)	
Temozolomide			0.57
Yes	139 (88.5%)	37 (92.5%)	
Total dose [GyRBE]			0.85
30	8 (5%)	2 (5%)	
33	12 (8%)	3 (7%)	
36	18 (11%)	4 (10%)	
39	17 (11%)	4 (10%)	
42	22 (14%)	6 (15%)	
45	70 (45%)	19 (48%)	
48	10 (6%)	2 (5%)	

3.1.2 *Head and neck cancer*

Head and neck carcinomas form in the linings of the upper respiratory tract and represent the sixth most common cause of cancer. The cohort contained 153 head-and-neck cancer patients treated between 2010– 2015 at UKHD. A summary of the patient demographics (discovery and test sets with an 80%/20% split) is presented in Table 3.3. More details about the dataset are available in (Gabryś et al., 2018).

Table 3.3: Head and neck cancer cohort overview. P-value calculated by chi-squared or Fishers' tests

Characteristics	Discovery set (n=123)	Test set (n=30)	P-val
Patient characteristics			
Age at RT (years)			
<65	83 (67%)	17 (57%)	
65-80	35 (28%)	11 (37%)	
>=80	5 (5%)	2 (6%)	
Gender			0.88
Male	93 (76%)	23 (77%)	
Female	30 (24%)	7 (23%)	
Tumor characteristics			
Tumor site			0.55
hypopharynx	16 (13%)	5 (17%)	
larynx	12 (10%)	4 (13%)	
nasopharynx	8 (2%)	4 (13%)	
oropharynx	83 (67%)	16 (53%)	
other	4 (8%)	1 (4%)	
Treatment characteristics			
Modality			0.96
IMRT	29 (26%)	8 (27%)	
Tomotherapy	94 (74%)	22 (73%)	
Median dose [range]			0.57
ipsilateral	24.12 [0.35, 63.37]	23.41 [0.41, 61.2]	
contralateral	19.87 [0.33, 30.92]	18.9 [0.6, 28.7]	

3.1.3 *Non-small cell lung cancer*

The cohort consisted of 180 non-small cell lung carcinoma (NSCLC) patients, from which a subset of 106 patients had a stage I or II NSCLC. The early-stage NSCLC patients were treated with stereotactic body radiotherapy (SBRT) with a median dose of 60 Gy in 8 fractions between 2008-2019. The staging was derived based on the tumour, node, metastasis (TNM) staging system by the International Association for the Study of Lung Cancer (IASLC) and the American Joint Committee on Cancer (AJCC), lastly revised in 2017 (Detterbeck et al., 2017). The TCP/NTCP modelling was only performed on the 106 early-stage NSCLC patients. No staging was possible for the remaining 74 NSCLC patients due to missing clinical information. They were treated with RT at UKHD between 2008-2019. They were selected solely to increase the accuracy of the segmentation networks (Section 4.2). All patients were added to the training set of the automatic segmentation DCNN to segment the ROI for patients with either a missing or corrupt RT SS in the early-stage NSCLC cohort. A summary of the early-stage NSCLC patient demographics (discovery and test sets with an 80%/20% split) is presented in Table 3.4. Table A.4 summarises the patient demographics of the remaining NSCLC patients.

3.1.4 *Primary high-grade glioma*

High-grade glioma is the primary tumour occurring most often in the central nervous system (Jovčevska, Kočevar, and Komel, 2013; Louis et al., 2007). The cohort consists of 320 patients treated with photons RT with a median of 9 image acquisition time points (corresponding to a median of 4 months), resulting in 20101 MR images acquired between 2006 and 2018. The dataset was collected retrospectively from 23 different scanners at UKHD. The in-plane resolution ranged from 0.45 x 0.45 to 1.40 x 1.40 mm, while the slice thickness ranged from 0.9 to 5 mm in all MR scans. A subset of 141 patients with available RT DICOM data and pre-irradiation planning MR time-points with at least 2 MR sequenced were further selected for the MR intensity normalization impact study (Section 4.3). A summary of the patient demographics can be seen in Table A.2.

3.1.5 *TCGA-GBM*

The TCGA-GBM cohort is a public cohort retrieved from the Cancer Genome Atlas Glioblastoma Multiforme, (Scarpace et al., 2016) including scans from 256 GBM patients with a median of 3 time points (corresponding to a median of 7 months), resulting in 3522 MR images acquired between 1986 and 2019, collected from 17 different scanners. A summary of the patient demographics can be seen in Table A.3.

Table 3.4: Early stage NSCLC cohort overview. FEV1: Volume exhaled at the end of the first second of forced expiration, FVC: Forced vital capacity. P-value calculated by chi-squared or Fishers' tests

Characteristics	Discovery set (n=70)	Test set (n=36)	P-val
Patient characteristics			
Age at RT (years)			0.10
<60	3 (4%)	6 (17%)	
60-74	40 (57%)	20 (56%)	
⩾75	27 (41%)	10 (27%)	
Gender			0.41
Male	46 (66%)	20 (56%)	
Female	24 (34%)	16 (44%)	
Karnofsky performance score (KPS)			0.87
⩾70	43 (62%)	29 (81%)	
<70	8 (11%)	7 (19%)	
Unknown	19 (27%)	-	
FEV1/FVC(%) at RT			0.07
⩾70	40 (57%)	26 (72%)	
<70	18 (26%)	10 (28%)	
Unknown	12 (17%)	-	
Cigarettes pack/year			0.31
0	17 (24%)	12 (33%)	
<40	14 (20%)	11 (31%)	
>40	21 (30%)	13 (36%)	
Smoker but unknown	18 (26%)	-	
Tumor characteristics			
NSCLC subtype			0.27
adenocarcinoma	27 (39%)	24 (67%)	
Squamous cell carcinoma	21 (30%)	12 (33%)	
Unknown	22 (31%)	-	
Tumor site			
Upper lobe	39 (56%)	20 (56%)	0.73
Middle lobe	7 (10%)	5 (14%)	0.71
Lower lobe	24 (35%)	11 (30%)	0.85
Treatment characteristics			
Total dose [Gy]			0.55
30	6 (8%)	2 (6%)	
45	22 (31%)	11 (30%)	
48	2 (3%)	0 (0%)	
50	4 (6%)	2 (6%)	
54	3 (5%)	3 (8%)	
60	33 (47%)	18 (50%)	

3.2 DEEP LEARNING FOR DATA CURATION

An essential step in the data preparation phase of AI applications and studies is accurately classifying the medical image modalities present in the cohort since each image communicates specific anatomical or physiological information. However, assuring that the right modalities are used for analysis (classification of sequences) might be a tedious and time-consuming task, especially when dealing with a large amount of data from various sources (multiple scanners, multiple treatment centres) due to possible inconsistent naming schemes. In particular, retrospective data collection yields additional challenges as they usually include non-prespecified protocols and sequences. A previous study demonstrated that classifying medical images based on image metadata (i.e., based on information stored in the DICOM header) is often unreliable (Gueld et al., 2002). DICOM tags and the actual examination protocols applied are not always consistently matched. This is mainly done to improve imaging quality, for example, the implementation of different body region imaging protocols due to variabilities and differences among patients' anatomies (Gueld et al., 2002). Therefore, automatizing medical image retrieval and classification based on the content data would be beneficial in terms of time efficiency, accuracy, and, ultimately, reproducibility. In the context of medical image retrieval and classification using DCNNs, four studies have been identified for the classification of body organs and MR images (Accuracy >90%) (Ayyachamy et al., 2019; Qayyum et al., 2017; Remedios et al., 2018; Voort, Smits, and Klein, 2021). A limitation of these methods is the inability to deal with the open-set recognition problem, i.e., the failure of a classifier trained to classify between a specific number of classes to handle unknown classes (Scheirer et al., 2012). The open-set recognition problem is a common issue when dealing with clinical cohorts since datasets exported from the hospitals' Picture archiving and communication system (PACS) usually include all available medical images and data, resulting in various medical image modalities and sequences. In this section, the open-set recognition problem in automatized medical image classifiers is tackled by training a DCNN-based MR image classifier (MR-Class) using a one-vs-all approach. One-vs-all classification is implemented to deal with the open-set recognition problem and thus would enable the handling of unknown classes. A comparison study of the published DCNNs (mentioned above) for medical image classification was first performed to determine the adopted DCNN model. Then, one-vs-all binary class-specific DCNN classifiers were trained to recognize a particular MR image, thus forming MR-Class. MR-Class consists of multiple one-vs-all binary classifiers rather than a single multiclass classifier, i.e., a classifier trained to classify all classes. The intuition behind training multiple one-vs-all DCNN is the open set recognition problem and that training a DCNN image classifier on every possible MR image is cumber-

some. The training was performed twice using scans from pHGG. The first training included all MR images available in the dataset. The second had only the image volumes of the six considered classes (the same images included in the comparison study during training). The latter was performed to compare the one-vs-all dual-class classifiers (MR-Class) performance against a multi-class DCNN classifier, both trained on the same number of images. Classes for each binary classifier were defined as follows: class 1 included all images corresponding to the targeted class, whereas class 0 contained all remaining images in the dataset. A stratified (by class) 80%-20% dataset split was used for training and validation (Table 3.5).

Table 3.5: Number (%) of MR images from the training cohort pHGG considered for each one-vs-all DCNN classifier. T2w-FL: T2-FLAIR

	Training		**Validation**	
DCNN classifier	Targeted class	Remaining images	Targeted class	Remaining images
T1w vs all	3152 (15.7)	12929 (64.3)	788 (3.9)	3232 (16.1)
T2w vs all	1576 (7.9)	14505 (72.1)	394 (2.0)	3626 (18.0)
T2w-FL vs all	1535 (7.6)	14546 (72.4)	384 (1.9)	3636 (18.1)
ADC vs all	1550 (7.7)	14530 (72.3)	388 (1.9)	3633 (18.1)
SWI vs all	1183 (5.9)	14898 (74.1)	296 (1.5)	3724 (18.5)

3.2.1 *MR scans*

Multiparametric MRIs (mpMRI) were collected from multiple scanners from the pHGG, rHGG, and TCGA-GBM cohorts, resulting in heterogeneous modalities and MR sequence protocols (Appendix A). Conventional multislice (2D) acquired in the axial, sagittal, or coronal plane, as well as 3D scans, are present. The MR sequences found in the cohorts are the widely used sequences for brain tumour imaging (Ellingson et al., 2015) in clinical routines and trials (Combs et al., 2010b; Niyazi et al., 2018). All MR images found in the training cohort were included in the training. However, one-vs-all DCNN classifiers were only trained for T1w, contrast-enhanced T1w (T1wce), T2w, T2w fluid-attenuated inversion recovery (FLAIR), apparent diffusion coefficient (ADC), and susceptibility-weighted imaging (SWI). No SWI scans were found in TCGA-GBM. The in-plane resolution ranged from 0.33 x 0.33 to 2 x 2 mm for pHGG, 0.45 x 0.45 to 1.40 x 1.40 mm for rHGG, and 0.45 x 0.45 to 1.14 x 1.14 mm for TCGA-GBM. Slice thickness ranged from 0.9 to 7.5 mm in all MR scans. Human experts manually labelled each MR image through

an in-house interactive labelling tool. The DICOM attributes "Series Description" (SD) and "Contrast/Bolus Agent" DICOM attribute were then extracted and compared to the derived labels to evaluate the metadata's consistency. Sample images of the classifiable sequences are shown in Figure 3.1.

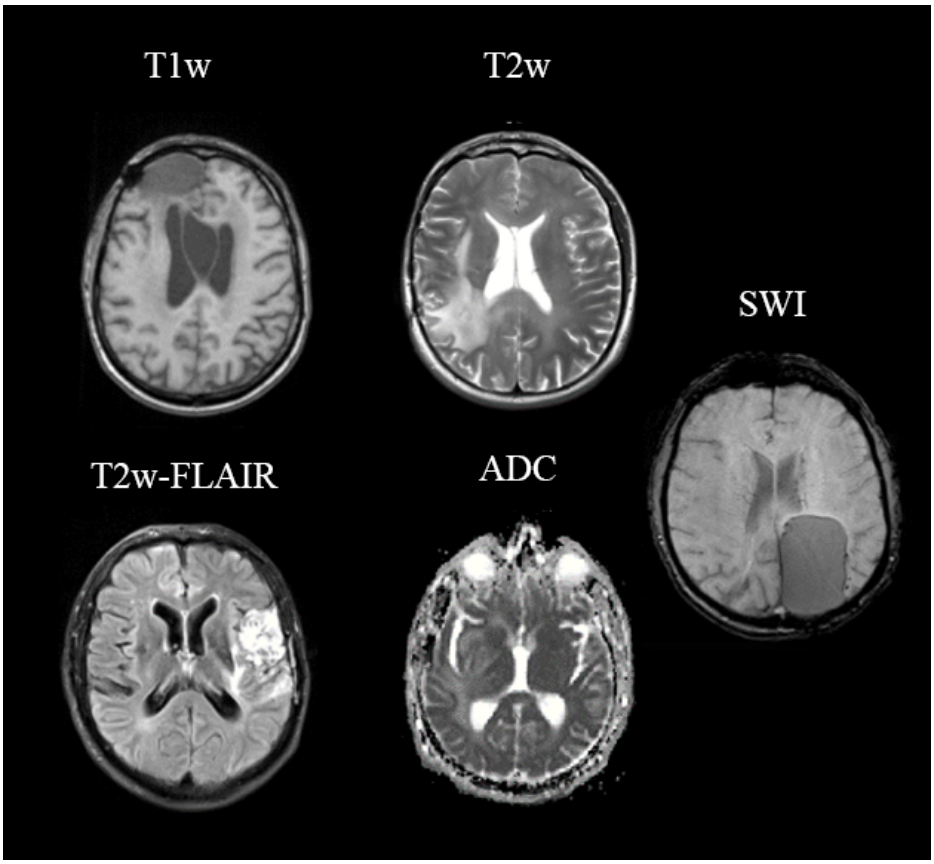


Figure 3.1: Sample images of the different classifiable MR sequences

3.2.2 *MR-Class: Training and preprocessing*

The DCNN architecture chosen for MR-Class is the ResNet-18. Residual Networks (ResNet) were introduced in 2015 to deal with the degradation problem, i.e., the degradation of the network accuracy as the depth of the network increases (He et al., 2016). Besides the usual DCNN architecture for classification purposes (alternating stack of convolutional, activations, and pooling layers), ResNet introduces skip-connections that skip one or more layers. These skip connections fit the unmodified input from the previous layer to the next layer, preserving the original image signal by performing identity mapping. This results in maintaining the norm of the gradient and solving the degradation problem. A softmax layer is appended to the end layer to produce probabilistic predictions of the classes. Schematics of the ResNet architecture is shown in Figure 3.2.

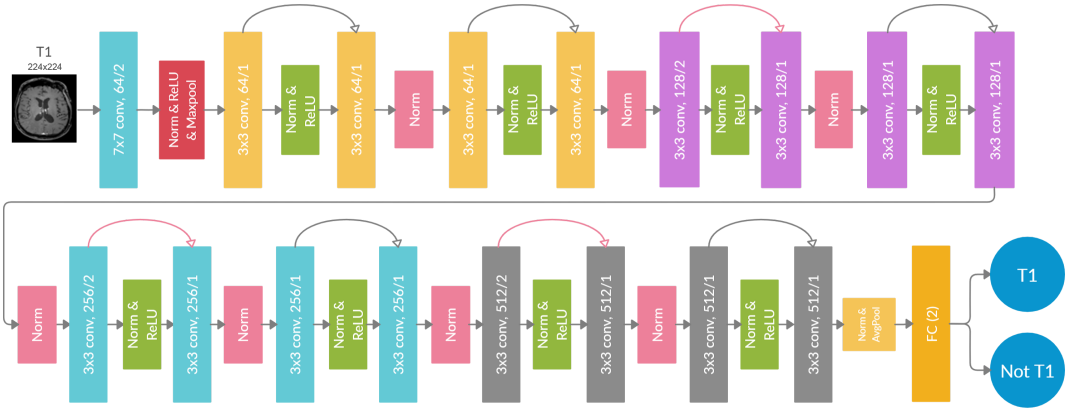


Figure 3.2: The one-vs-all ResNet-18 architecture - An alternating stack of convolutional activations and pooling layers. The skip connections (arrows) fit the unmodified input from the previous to the next layer, preserving the original image signal. FC (2) is a fully connected layer with two neurons as output, representing the sequence and the other possible sequences.

The preprocessing and training approach implemented for the 2D ResNet-18 are likewise applied for the one-vs-all DCNNs; however, further steps were taken to address the imbalanced classes arising from the one-vs-all classification design. First, data augmentation was implemented using the TorchIO python library (Pérez-García, Sparks, and Ourselin, 2021). Specifically, the transformations implemented included adding random Gaussian noise, blurring, performing random affine or elastic deformations, and adding random MR motion artefacts like motion, ghosting, or spikes. Second a weighted binary categorical cross-entropy loss was used, where the weights of a class were equal to the size of the largest class divided by the size of that specific class. For example, for the T2w-vs-all DCNN, if class T2w has 1970 and class all has 18131 MR images, the weights would be 9.2 and 1.0, respectively. Finally, the learning rate scheduler was adjusted to decay based on the targeted class training loss instead of the loss of both classes. A summary of the training workflow can be seen in Figure 3.3.

3.2.3 *MR-Class: Inference and testing*

The cohorts rHGG and TCGA-GBM were used to perform independent testing of MR-Class. In inference mode, the MR images were preprocessed (same as in training) and fed to each DCNN classifier to infer the corresponding class. A classification probability threshold of 0.95 was used. The cutoff threshold value was determined based on the distribution of the probabilities of correct and wrong labelled images when pHGG was inferred back to MR-Class (Figure 4.2). If an image is labelled by more than one classifier, the classifier with the highest probability determines the class. If none of the clas-

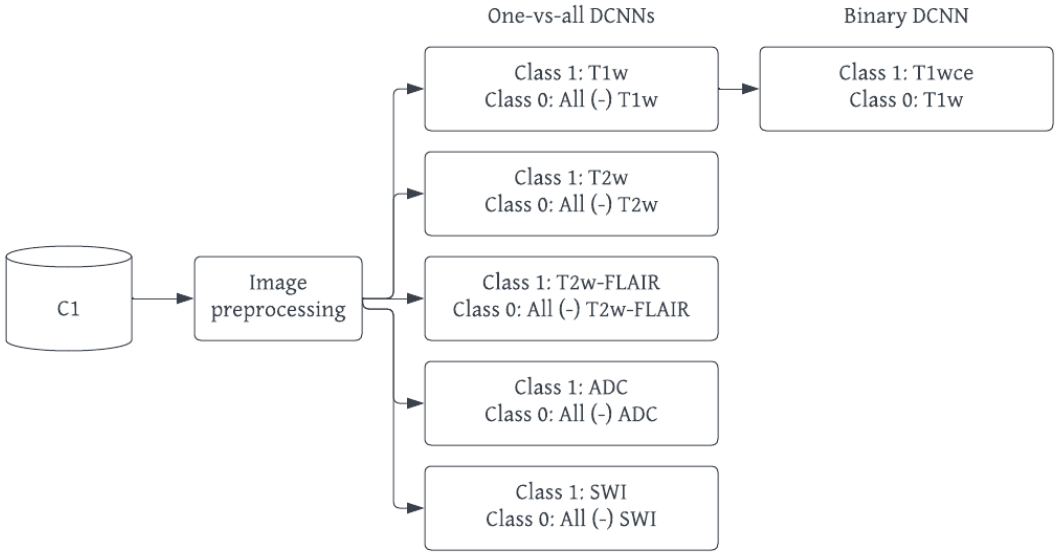


Figure 3.3: MR-Class training workflow - MR-Class comprises five one-vs-all DCNNs, one for each class, and the T1w-vs-T1wce binary DCNN. C1: pHGG

sifiers label an image (i.e., assigned to class 0 by each classifier), it is unclassifiable. The 2D DCNNs classify an MR scan as a class using a majority vote of 10 inferred slices extracted around the middle slice of the corresponding MR acquisition plane. Figure 3.4 shows a summary of the inference workflow.

Classifications were compared to ground truth labels, where the number of correct predictions divided by the total number of images derived the accuracy. The 95% confidence interval (CI) was calculated as the Wilson interval (WS) (Wallis, 2013). Classification sensitivity and specificity were calculated to evaluate the performance of each classifier. Lastly, the misclassified images were analyzed to identify the causes of misclassifications.

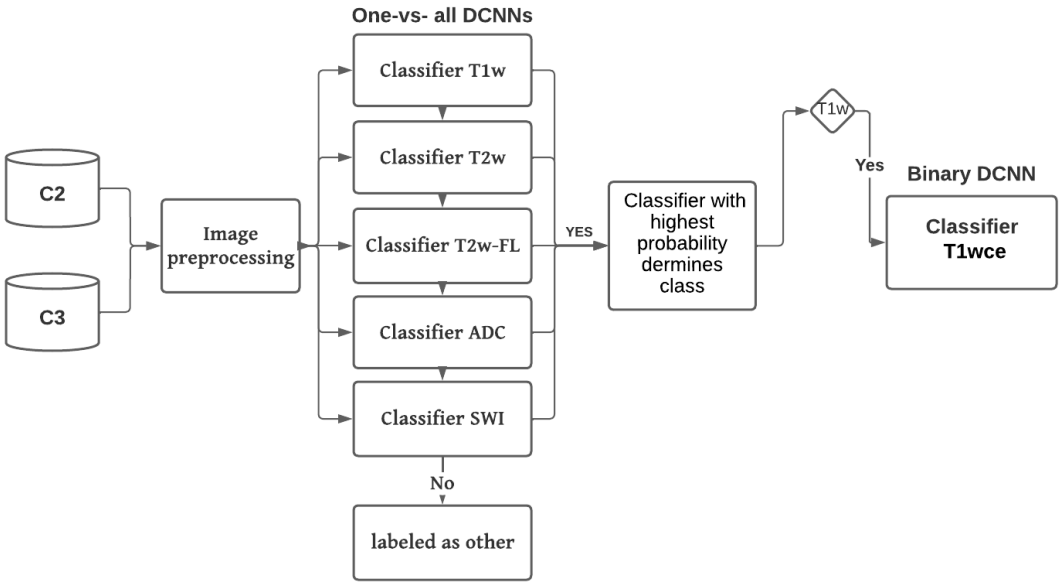


Figure 3.4: MR-Class inference workflow - rHGG and TCGA-GBM were used to test MR-Class. After preprocessing, MR images are passed to the five one-vs-all DCNN classifiers. A classification probability threshold of 0.95 was used. If none of the classifiers label an image, it is rendered as other. If more than one classifier labels a specific image, then the image is labelled by the classifier with the highest probability. C2: rHGG, C3: TCGA-GBM

3.3 DEEP LEARNING FOR DATA COMPLETION

During analysis, 16% of RT SS were observed as missing from cohorts NSCLC and Un-NSCLC, which might affect the overall analysis w.r.t NTCP/TCP prediction performance. To complete the missing data, a deep learning framework - nnU-Net (Isensee et al., 2019a) - was employed. nnU-Net is based on a U-Net structure, allowing for accurate, fast biomedical image segmentation. The preprocessing steps are automated, as well as the tuning of network parameters. 136 images from cohorts NSCLC and Un-NSCLC were used for training, while 15 images were kept for testing. Both 2D and 3D segmentation networks were trained for the automatic segmentation of GTV and the heart.

3.4 PREPROCESSING WORKFLOW

After DICOM dataset curation and MR image classification using pyCURT and MR-Class (Sforazzini et al., 2020), a sequence of different preprocessing steps was applied to the different modalities, i.e., MR, CT, and DD. The image processing diagram is shown in Figure 3.5. After reorienting all images to a common orientation, signal inhomogeneities in T1w images were cor-

rected using the N4 bias field correction algorithm (Tustison et al., 2010). The HD-BET brain extraction tool was next applied to eliminate the skull and background and generate the brain mask (Isensee et al., 2019b). Noting that not all patients had 3D MR sequences available, the 2D transversal, sagittal, and coronal MR scans were corrected for motion and 3D reconstructed to a high-resolution 3D MR scan through NiftyMic (Ebner et al., 2020).

Next, cross-sectional linear co-registrations with 6 degrees of freedom (DOF) of the present MR sequences were performed on the T1wce using advanced normalization tools (ANTs) (Avants, Tustison, Song, et al., 2009). An additional cross-sectional linear co-registration of the T1wce with 6 DOF was performed on the RT planning CT to transform the target volume (TV) segmentations - extracted from the DICOM structure set (SS) objects - to the MR space. Large MR intensity inter-patient variations are present in this cohort since data were collected from multiple scanners (data not shown). The methods applied to the different sequences were derived based on the results obtained by the MR intensity normalization impact study presented in the subsection 3.4.1. With the help of the intensity normalization package (Reinhold et al., 2019) and FAST (FMRIB's Automated Segmentation Tool) (Zhang, Brady, and Smith, 2001), multiple different intensity normalization methods were applied to the cohort, each resulting in a specific intensity normalized dataset, to check and eliminate radiomics features impacted by the choice of the normalization algorithm. The signature and model building were performed on the top-performing identified methods' corresponding normalized dataset - however, after the elimination of the radiomics feature, that showed a low correlation between the different normalization methods (see section 4.3). A Spearman rank-order correlation coefficient rs cutoff of 0.80 was used to eensure a strong correlation.

CT image intensities were clipped to the 5th and 95th percentiles. When single fields or fraction (fx) DD were only present, individual beam dose accumulation and fx-weighted dose conversion were performed to derive the plan RT DD. All DD were next re-calculated using FRoG (Fast Recalculation on GPU), a dose engine benchmarked against the gold standard Monte Carlo (MC) simulation (Mein et al., 2018). All voxel intensities were transformed to the equivalent dose in 2 Gy per fraction (EQD2) using an α/β of 10 Gy. Resampling of the images and TVs to a matrix size of 2x2 mm and a slice thickness of 2 mm were next performed using a cubic spline and linear interpolation, respectively. Finally, image discretization was performed. Five bin counts (16-32-48-64-128) and five bin widths around the interquartile range (IQR) of the intensity range of the cohort in the study were applied to each of the 14 MR intensity normalized discovery datasets. A fixed bin count of 32 was used on the test set for all MR sequences. A fixed bin width of 25 and 0.1 was used for CT and DD, respectively.

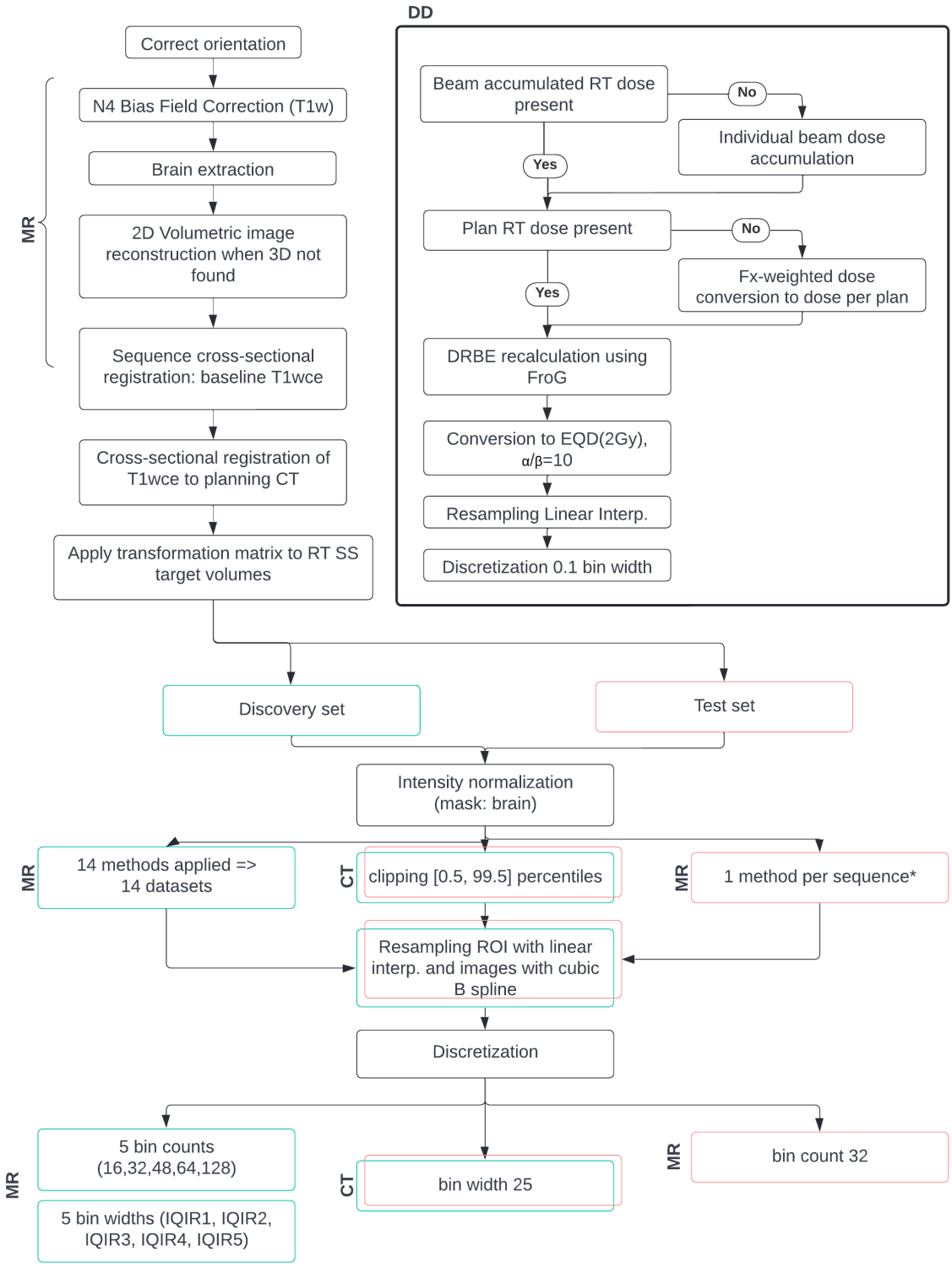


Figure 3.5: MR, CT, and DD preprocessing diagram used on the discovery and test sets. RT SS target volumes represent the target volume segmentation extracted from the DICOM RT structure set. T2w-FL: T2w-FLAIR. ROI: region of interest; black: both sets, green: discovery set, red: test set; DD: dose distribution

3.4.1 *Impact of intensity normalization methods*

Intensity normalization was performed with the help of the intensity normalization package (Reinhold et al., 2019) and the FMRIB's Automated Segmentation Tool (FAST) (Zhang, Brady, and Smith, 2001). The intensity-normalization methods considered are Fuzzy C-Means (FCM) (Bezdek, Ehrlich, and Full, 1984) (9 different masks combinations), kernel density estimation (KDE), Gaussian mixture models (GMM) (Reynolds, 2009), the Nyul's and Udupa's histogram matching-based abbreviated in this study as HM (Nyúl and Udupa, 1999), white-strips (WS) (Shinohara et al., 2014), z-score normalization, and the feature-based batch adjustment method, i.e., Combat (Johnson, Li, and Rabinovic, 2007), resulting in 15 different MRI normalized datasets. A brief description of the methods is given in this section. For a broader definition, the original normalization method papers and the manuscript are referred (Reinhold et al., 2019).

3.4.1.1 *Standard score*

The standard score, also known as the z-score, represents the distance of a raw score from the mean measured in standard deviations. In the context of MR brain image normalization, given that B is the brain mask in the image I, the z-score calculates the mean and standard deviation of the intensities inside the brain image (excluding the background) as follows:

$\mu = \frac{1}{|B|} * \sum_{b \in B} I(b), \quad \sigma = \sqrt{\frac{\sum_{b \in B} (I(b) - \mu)^2}{|B| - 1}}$ with the normalized image being $I_{norm}(x) = \frac{I(x) - \mu}{\sigma}$

This method's downfall is that the images' high intensities will be wrongly diminished.

3.4.1.2 *Fuzzy clustering*

Clustering is a method for analyzing data that aims to discover structures or groups in a data set. Fuzzy clustering (Bezdek, Ehrlich, and Full, 1984) allows a piece of data to be part of more than one cluster. In a fuzzy c-means algorithm, a data point is assigned a membership function, with o being the farthest from a cluster's centre and one being the closest to a cluster's centre, with the data point theoretically being able to belong to all clusters. Used as a normalization technique, the fuzzy c-means algorithm uses a specific tissue mask to normalize the image to the mean intensity of that mask. In brain MRI, the main different tissue types are white matter (wm), grey matter (gm), and cerebrospinal fluid (csf) (Figure 3.6).

If the mean of the tissue is: $\mu = \frac{1}{|T|} * \sum_{t \in T} I(t)$, then the normalized image would be $I_{norm}(x) = \frac{I(x)}{\mu}$, with T as the tissue mask.

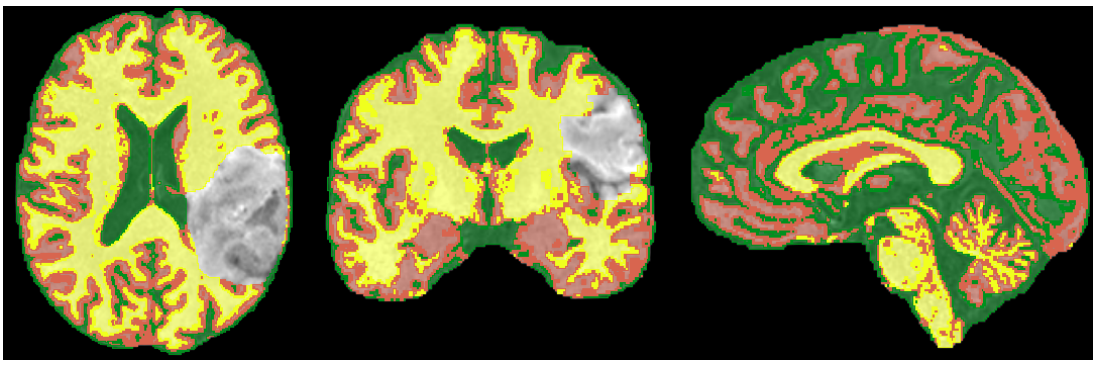


Figure 3.6: Segmentation of the different brain tissue using the segmentation tool FAST. White matter is in yellow, grey matter in red, and the cerebrospinal fluid in green

The brain tissue masks, i.e., white matter (wm), grey matter (gm), and cerebrospinal fluid (csf), segmentations were performed using FSL's FAST (Zhang, Brady, and Smith, 2001). In conjunction with the most common intensity value (mode) in a particular image, nine different mask combinations were implemented to generate nine fuzzy c-means normalized datasets. The masks are csf, gm, wm, csf-gm, wm-csf, wm-gm, csf-mode, wm-mode, and gm-mode.
The normalization with two brain tissue masks is performed as follows: with $\mu_1 = \frac{1}{|T_1|} * \sum_{t \in T} I(t)$ and $\mu_2 = \frac{1}{|T_2|} * \sum_{t \in T} I(t)$, the normalized image would be $I_{norm}(x) = \frac{I(x)-a}{b-a}$ with $a = \min(\mu_1, \mu_2)$ and $b = \max(\mu_1, \mu_2)$.
The normalization with a brain tissue mask and the mode is performed as: as $I_{norm}(x) = \frac{I(x)}{diff}$ with $diff = \mu_T - mode(B)$ with T as the tissue mask and B as the brain mask.

3.4.1.3 *Kernel Density Estimation*

A density estimator aims to find a function for the probability distribution that a dataset was generated from. The kernel density estimation (KDE) is an empirical calculation in a parametrized form. The formula for calculating the KDE for the probability distribution function is:

$$p(x) = \frac{1}{N * M * L * h} * \sum_{i=1}^{N*M*L} K\left(\frac{x - x_i}{h}\right)$$

where N, M, and L are the sizes of the images, K is the kernel (normalized to one), and h is the bandwidth parameter that scales the kernel. This method provides a smoother version of the histogram, making it easier to find the maxima. This is then used to normalize the entire image as $I_{norm}(x) = c * \frac{I(x)}{\pi}$, where c is a positive, real constant. In this study, the KED finds the peak for the white matter histogram and translates it to a standard value.

3.4.1.4 *Gaussian Mixture Models*

A mixture model assumes that a data set comprises subsets whose individual distributions are the respective probability distributions in the overall data set. A specific mixture model is the Gaussian mixture model, where the subsets are considered to be generated from a finite number of Gaussian distributions with undefined parameters. The method used here fits three Gaussian distributions to the histogram of the brain without a skull and normalizes the white matter mean to an expected value (Reinhold et al., 2019).

3.4.1.5 *Landmark-based histogram matching*

The landmark-based histogram matching method (Nyúl and Udupa, 1999) tries to deform the input image intensity histogram to match a reference histogram. Most commonly, the reference histogram is obtained by averaging histograms in a data set and setting the landmarks of interest. Each input image histogram is then matched to the reference one through linear interpolation based on the defined landmarks, usually quantiles.

3.4.1.6 *White Stripe normalization*

The WhiteStripe normalization approach is presented by (Shah et al., 2011). Its main idea is to choose a reference brain tissue, in this case, normal-appearing white matter (NAWM). The NAWM values are obtained by smoothening the image histogram and selecting the largest peak. The so-called white stripe contains intensity values up to 10% around. The white stripe can be defined as

$$\Omega_T = \left\{ I(x) \mid F^{-1}(F(\mu) - \tau) < I(x) < F^{-1}(F(\mu) + \tau) \right\}$$

where F(x) is the cumulative distribution function of the image I and $\tau = 5\%$ is the standard deviation in the white stripe. The normalized image is $I_{norm}(x) = \frac{I(x)-\mu}{\sigma}$.

3.5 RADIOMICS AND DOSIOMICS FEATURE EXTRACTION

After all considered image modalities have been curated and prepossessed, the next step is to extract the radiomics and dosiomics features on which the NTCP/TCP modelling was performed. Radiomics and dosiomics features were extracted from the original image and transformed images using the PyRadiomics (v 3.0) library in Python and the DicomToolboxMatlab (Gabryś et al., 2018; Van Griethuysen et al., 2017). The regions of interest (ROIs) considered for the rHGG cohort are the gross (GTV), clinical (CTV), and planning (PTV) target volumes. For the NSCLC cohort, the ROIs are the GTV, PTV, heart, ipsilateral lung, and contralateral lung. As for the HNC cohort,

the ROIs are the PTV, ipsilateral, and contralateral parotid glands. The derived radiomics images were retrieved from Wavelet filtering, which yielded eight decompositions per level, each representing a combination of either a high or a low pass filter in each of the three dimensions, a Laplacian of Gaussian filter with spatial scaling factors (SSFs) of 2, 3 and 4 mm, and linear binary patterns. Additional transformations were applied on DD, i.e., logarithm, square, gradient, square root, and exponential. The number of shape, first and second-order statistics derived per modality and calculated on both the original and derived images can be seen in Table 3.6. The total yielded 1200 radiomics feature per image modality per ROI, and 1500 dosiomics feature per ROI. Calculations were performed on a Linux workstation (Intel Xeon W-2145 CPU, 16 GB memory).

Table 3.6: Number of shape, first and second-order statistics derived per sequence and calculated on both the original and derived images.

Class	No. features
First-order statistics	19
Shape-based (3D)	16
Second-order statistics	
Grey Level Co-occurrence Matrix	24
Grey Level Run Length Matrix	16
Grey Level Size Zone Matrix	16
Neighbouring grey Tone Difference Matrix	5
Grey Level Dependence Matrix	14

3.6 STATISTICAL ANALYSIS

After all considered image modalities have been curated, prepossessed, and the radiomics and dosiomics features have been extracted, the statistical analysis was performed. Statistical modelling was performed in R (version 3.3.1,2016 28) with the “rms”, “mlr” (Lang et al., 2019), "Hmisc", "PerformanceAnalytics", "Hmisc", "data.table", "doParallel", "foreach", "glmnet", "boruta, "lsmeans", "pheatmap", "plyr" and "dataAnalysisMisc" (Knoll et al., 2020) packages. Plots were generated using the "ggplot2" package. Continuous variables were compared using the Mann-Whitney U test. Testing of differences between two groups was performed using Chi-squared or Fisher’s exact tests. Significance level (p-value) was set at 5% (two-sided). The three datasets used in the NTCP/TCP modelling, i.e. rHGG (Table ,3.2, NSCLC (Table 3.4 and HNC (Table 3.3) were split into discovery and test sets with an 80%/20% split. In

the rHGG cohort, patients with at least a missing or corrupted T1w, T2w, T2w-FLAIR, SWI, or ADC were all assigned to the discovery set. In the HNC cohort, 12 patients had a corrupted CT DICOM image, which were assigned to the discovery set. The same was performed in the NSCLC cohort, where patients with missing RT SS were assigned to the discovery set. All patients in the test sets had all modalities considered in this study, as well as the ROIs present in the RT SS.

3.6.1 *The unimodality and multimodality signature derivation*

After feature extraction, different prepossessing steps were applied to the radiomics and dosiomics feature sets, as well as different tests to check which combination of features is most significantly correlated to the different outcomes modelled in this work. This was performed on the individual considered modalities feature sets to derive the unimodality signatures as well as on the multimodality combined feature sets to derive the multimodality signatures.

The different prepossessing steps and tests were applied as follows. First, z-score normalization was performed, where each feature was normalized as $z = \frac{x-y}{s}$, where x is the feature, y is the mean, and s is the standard deviation. Pairwise correlational tests were next applied between the different intensity normalized and discretized datasets, where features that showed a low correlation in at least 20% of the preprocessing methods' respective processed datasets were eliminated. Low variance and high correlation filters were further implemented on the remaining feature set to drop features with low variance and high correlation. Correlation-based feature elimination was performed using the Spearman correlation formula with a rank-order correlation coefficient (rs) threshold of 0.80 (Wissler, 1905).

OS and PFS correlated features were then derived separately from the clinical (demographics) features and the remaining radiomics and dosiomics features. A combination of three feature selection methods was then used, including a univariate analysis under Cox proportional hazard (CPH) models (Lin and Wei, 1989), a random forest (RF) -based method, i.e., Boruta (Kursa and Rudnicki, 2010), and lasso regression (Muthukrishnan and Rohini, 2016) were applied on 1000 random subsamples (10% left out) of the discovery dataset. Modality-specific significant features identified at least 950 times were further selected. A multi-modality imputation pipeline was then applied using the MICE R package (Van Buuren and Groothuis-Oudshoorn, 2011). The missing significant features observations were imputed using all significant features derived from the other modalities. Finally, a forward-backwards stepwise variable selection procedure using CPH models was applied to the significant features to obtain the unimodality-specific imputed

signatures. Complete signatures, i.e., excluding imputed data, were also derived for performance comparison.

To derive the radiomics signature (RS), all MR modalities and CT significant features were combined into one feature set on which a forward-backwards stepwise variable selection procedure using CPH models (P-val< 0.05) was applied. Similarly, the combined signature (RDCS) was derived, however, with the inclusion of the clinical and dosiomics significant features. Complete signatures with non-imputed data were also derived for performance comparison.

An overview of the unimodality and multimodality signature building is presented in Figure 3.7.

3.6.2 *Model building, evaluation, and comparison*

Time to event information for OS, PFS (i.e. survival in the absence of tumour progression or metastasis) as well as for the radiation-induced toxicities in the NSCLC, i.e. radiation-induced lung fibrosis, and the HNC, i.e. radiation-induced xerostomia were available. For those, multivariate CPH and random survival forest (RSF) models were trained to derive the NTCP and TCP models with the derived signatures. The performance of the unimodality and multimodality models were then evaluated and compared. Model performance comparison was performed in two folds. First, using the CPH models and the discovery set, the resampled concordance index (C-I) was computed for each model and used for performance evaluation. C-I is an established means that quantifies the quality of rankings and evaluate the predictions made by a model and is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs (Steck et al., 2007). Three different resampling approaches, i.e., a 5-fold cross-validation, bootstrapping with 1000 iterations, and Monte-Carlo cross-validation with 1000 iterations. Second, all CPH and RSF models were applied to the test set, and the C-I was computed. Confidence intervals were derived through bootstrapping (n=1000). X-means clustering, an extension of k-means with efficient estimation of the number of clusters through Bayesian Information Criterion (BIC) to approximate the correct number of clusters (Pelleg, Moore, et al., 2000), was next performed separately on the corresponding radiomics and dosiomics features and the multimodality combined signatures and patients were assigned to the derived clusters. A low-dimensional representation of the data was then derived using a t-distributed stochastic neighbour embedding method (t-sne) to visualize the clusters (Maaten and Hinton, 2008). X-means was not performed when single radiomics or dosiomics features formed the corresponding signature, and patients were stratified into two groups based on the median value of the identified feature. Survival and freedom from toxicity curves were then described using Kaplan–Meier (KM) analysis (log-rank test). Effects of the dif-

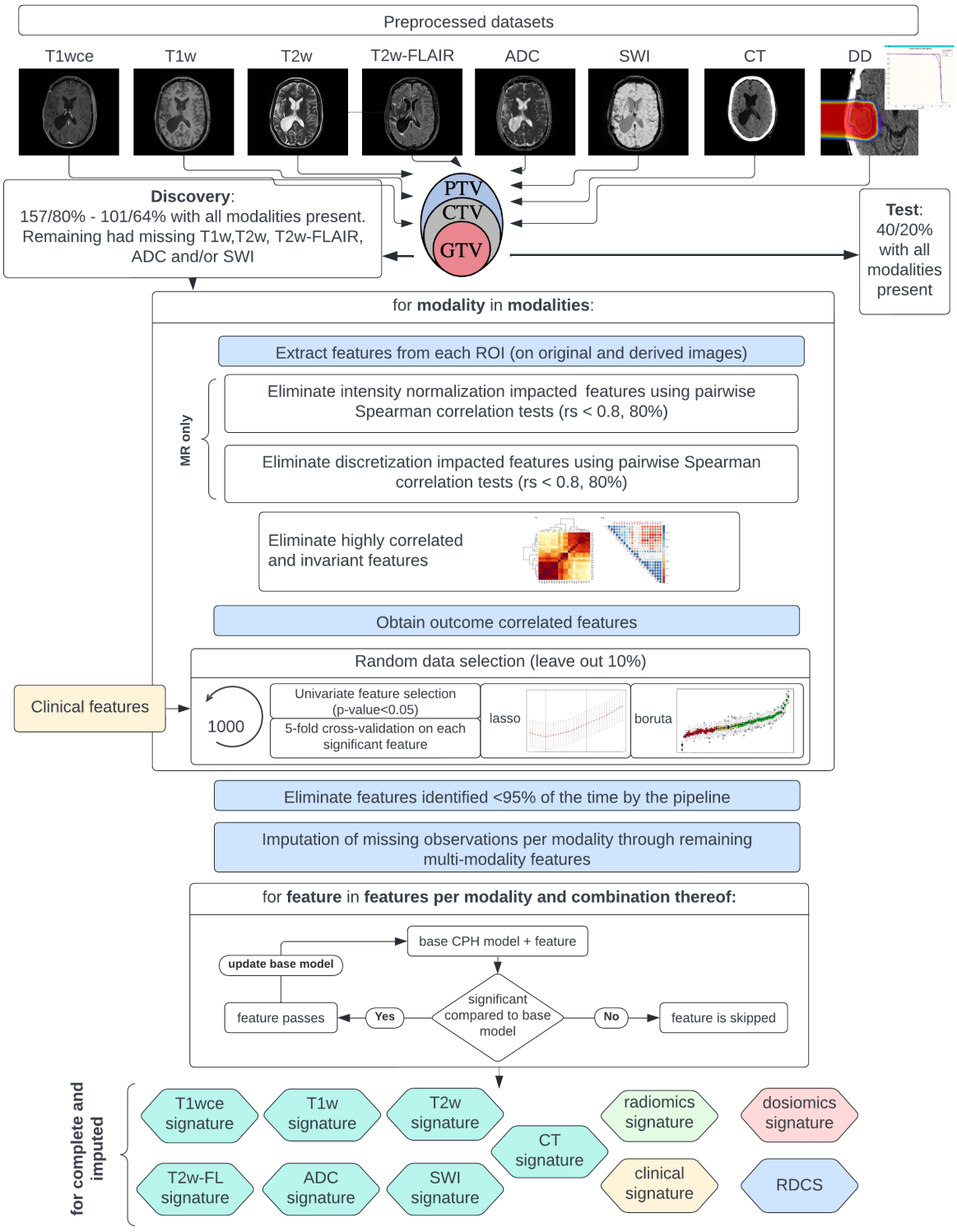


Figure 3.7: Unimodality and multimodality signature building - Following image preprocessing on both discovery and test sets, features were extracted from each set for each modality, and several feature reduction and selection methods were applied through resampling; after identifying the unimodality significant features, the unimodality, and multimodality signatures were built through a forward-backwards feature selection scheme; GTV: gross tumour volume, CTV: clinical target volume; PTV: planning target volume

ferent modalities (clinical, radiomics, and dosiomics) significant features on the prediction and prognostic separation were also assessed. Furthermore, model performances after the derivation of patient subsets based on the clinical features and known biomarkers such as MGMT methylation, IDH1/2 mutation, and 1p/19q codeletion in the rHGG cohort, or the tumour location and site in the NSCLC cohort were analyzed.

3.6.2.1 *Survival function*

Using the Kaplan-Meier method (Kaplan and Meier, 1958), the survival probability can be estimated from observed survival times - either censored or uncensored. The probability of being alive at $t_{j-1}, S(t_j)$ is calculated using $S(t_{j-1}), n_j$ - the number of alive patients before t_j and d_j - the number of events at t_j, as

$$S(t_j) = S(t_{j-1})(1 - \frac{d_j}{n_j}) \tag{9}$$

with $t_0 = 0$ and $S(0) = 1$. $S(t)$ does not change between times of events, which renders a step function whose value changes only at the time of each event (Clark et al., 2003). By plotting the Kaplan-Meier survival curve - survival probability against time - one can estimate, e.g., median survival time.

3.6.2.2 *Hazard function*

The relationship between $S(t)$ and $h(t)$ is given by the formula

$$h(t) = -\frac{d}{dt}[\log S(t)] \tag{10}$$

therefore if one of the functions is known, the other can be calculated. Nonetheless, $h(t)$ cannot be easily estimated. The area under the hazard function - $H(t)$ - between times 0 and t is commonly used to calculate it. $H(t)$ is an intermediate step in estimating $h(t)$ and serves as a tool for checking model validity. Over time, many methods of estimating $H(t)$ have been developed; because of its applicability to a wide array of clinical applications, the Cox proportional hazards model (Cox, 1979), based on order statistics, is the most commonly used one, mainly due to the measure of association, i.e. though the assumption of proportionality of the hazard ratios (Breslow, 1975).

4 RESULTS

This work aimed to incorporate multiple medical information layers into a combined modelling approach to improve patient stratification and prognostication while building a data curation, preprocessing, and analysis pipeline for faster deployment of artificial intelligence applications. This chapter presents:

1. the tools developed to address the issue of analyzing large, heterogeneous cohorts
2. the models developed to improve NTCP/TCP prediction

4.1 DEEP LEARNING FOR DATA CURATION

When working with large, heterogeneous datasets, bottlenecks such as DICOM metadata inconsistencies can slow down the data preprocessing and thus the overall analysis. With the data curation tool, this work aimed to facilitate data preprocessing in projects using large clinical cohorts.

While preparing the different datasets included in this work, several problems were observed. These included DICOM metadata inconsistencies which did not allow for the proper curation of the image modalities. Therefore a one-vs-all DCNN automatic classification pipeline was built to enable fast and accurate image modality classifications based on the image content. The pipeline serviceability is shown in this work through the automatic classification of MR sequences, i.e., MR-Class. Before presenting MR-Class's classification performance, the experiments' results to assess the validity of a one-vs-all DCNN classification pipeline for MR sequences are first reported.

4.1.1 *Metadata consistency*

When analyzing the DICOM series descriptions (SD) of the different MR sequences observed in cohorts pHGG, rHGG, and TCGA-GBM, 2074 SDs were identified. 11.4%, 10.6%, and 10.7% of the SDs for pHGG, rHGG, and TCGA-GBM, respectively, had misleading or inconsistent entries, not allowing for the proper identification of the MR image class (Table 4.1).

Table 4.1: Percentage of DICOM metadata-based labelling errors for each class considered in all three cohorts. T2w-FL: T2w-FLAIR

	pHGG		rHGG		TCGA-GBM	
	n	% error	n	% error	n	% error
T1w	2023	15.1	1189	11.2	433	13.4
T1wce	1917	13.9	4315	13.4	1096	9.9
T2w	1970	9.3	630	11.7	347	10.3
T2w-FL	1919	7.2	811	10.5	389	8.2
ADC	1938	7.6	895	8.4	122	5.5
SWI	1479	6.3	486	6.6	-	-
Other	8855	13.1	3007	7.3	1135	12.1
All	20101	11.4	11333	10.6	3522	10.7

4.1.2 *Multiclass vs one-vs-all classifications*

As for the multiclass vs multiple binary one-vs-all classification experiment, where only the image volumes of the six considered MR sequences were regarded, the validation accuracy was comparable, with 98.6% and 98.1%, respectively. The outputs of an image when inputted into the different convolutional layers (for feature extraction) are the so-called "feature maps". Figure 4.1 illustrates feature maps of different layers for both classification approaches. Only the middle slice of the last map of each convolutional block is shown. They are generated for the positive classification of the six sequences of a single patient to highlight the features learned for an MR sequence classification.

4.1.3 *MR-Class: MR image classification utilizing one-vs-all DCNNs*

In this section, the classification performance of the one-vs-all MR sequence-specific DCNN, as well as the performance of MR-Class, is reported. An evaluation of the misclassified images is also performed.

4.1.4 *Classification performance: One-vs-all DCNNs*

Table 4.2 summarizes the classification accuracies in the validation sets of all six DCNN classifiers on pHGG.

All six classifiers have high validation accuracies, with the lowest at 97.7% for the T1w-vs-T1wce and the highest at 99.7% for the SWI-vs-all and 99.6% for the ADC-vs-all tasks. Passing back the training set pHGG to MR-Class in

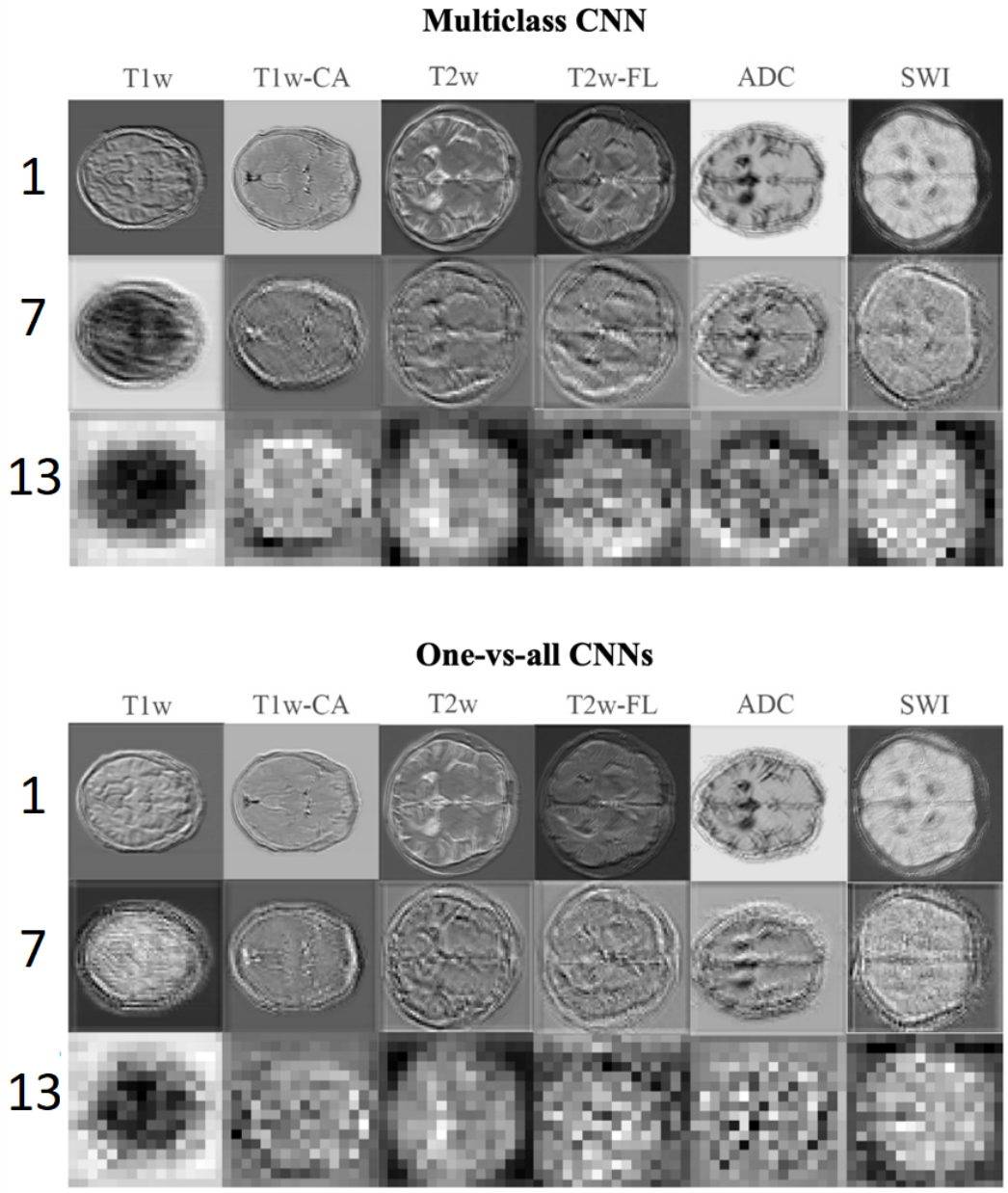


Figure 4.1: Visual inspection of feature maps for multiclass CNN and one-vs-all CNNs. Upper panel: Multiclass CNN, lower panel: multiple binary one-vs-all CNNs - Visual inspection of feature maps for different layers of the true positive sequence classification of a single patient with all 6 MR sequences. Layers 1,7, and 13 are shown in succession in each row (Figure 3.2). The last layer of each convolutional block is shown. Similar feature maps evolution behaviour is observed between the different classification approaches and the different CNNs, as an image propagates throughout the layers

Table 4.2: Validation classification accuracies of all six binary DCNN classifiers on pHGG. T2w-FL: T2w-FLAIR

Classifier	Val Acc (%)	Classifier	Val Acc (%)
T1w-vs-all	99.1	T2w-FL-vs-all	99.4
T1w-vs-T1wce	97.7	ADC-vs-all	99.6
T2w-vs-all	99.3	SWI-vs-all	99.7

inference mode, an accuracy of 97.4% [95% WS CI: 96.2, 98.4] is obtained, i.e., out of 20101 MR scans, MR-Class could not learn 519. As for the multiclass vs multiple binary one-vs-all classification experiment, where only the image volumes of the six considered MR sequences were regarded, the validation accuracy was comparable with 98.6% and 98.1%, respectively. The distribution of the classification probabilities derived by MR-Class for all three cohorts is shown in Figure 4.2. Based on pHGG, a probability cutoff threshold of 0.95 was set for testing MR-Class on rHGG and TCGA-GBM.

4.1.5 *Classification performance: MR-Class*

MR-Class's accuracy against the independent cohort rHGG was 96.7% [95% CI: 95.8, 97.3], i.e., 424 out of 11333 images were misclassified. All DCNNs had a specificity ranging between 93.5% (T2w-vs-all) and 99.6% (SWI-vs-all). The T1w-vs-T1wce and T1w-vs-all had the lowest sensitivity with 91.9% and 96.6%, while all remaining DCNNs had a high sensitivity (>99%) (Figure 4.3-upper, rHGG). In the multiclass normalized confusion matrix (Figure 4.3-lower, rHGG), it is seen that the classification of T1w is the least reliable, with an accuracy of 91.17%. Against the independent TCGA-GBM, MR-Class achieved an accuracy of 94.4% [95% CI: 93.6, 96.1] with 196 misclassified scans out of 3522. The T1w-vs-T1wce had the lowest sensitivity with 97.4%, while all remaining DCNNs had a sensitivity larger than 98%. Specificity ranged between 91.3% (T2w-vs-all) and 98.8% (T1w-vs-T1wce) (Figure 4.3-upper, TCGA-GBM). In the multi-class confusion matrix (Figure 4.3-lower, TCGA-GBM), it is seen that the classification of T2w is the least reliable, with an accuracy of 91.35%, with 8.65% classified as "other". Investigations on the misclassified images were performed in the next section.

4.1.5.1 *Analyses of misclassified images.*

In this section, misclassified images were analyzed to identify the causes of misclassifications. Out of the 14855 inferred images from rHGG and TCGA-GBM, MR-Class classified 620 images incorrectly. The misclassifications can be sorted into different categories: MR artefact-middle slice blurring, MR

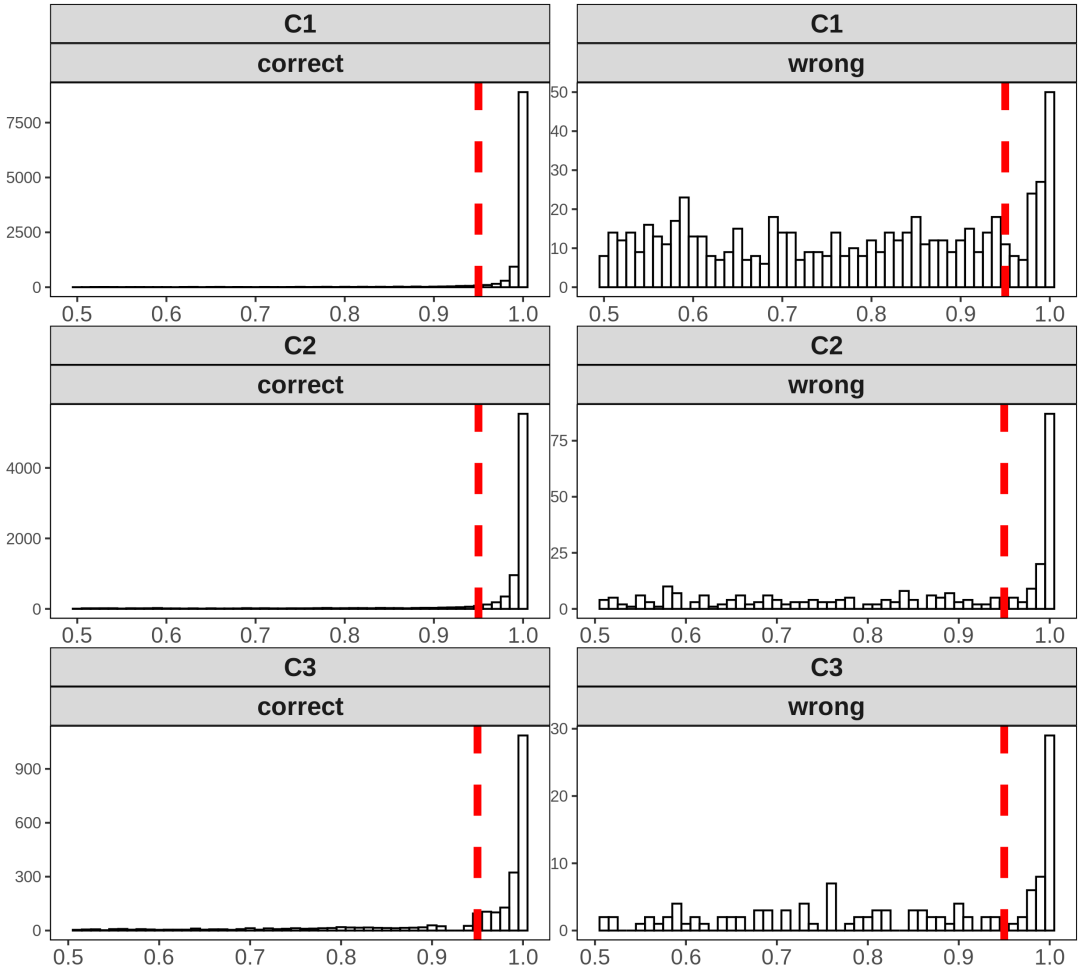


Figure 4.2: Distribution of the probabilities of correct and wrong labelled images for all three cohorts when inferred to MR-Class. Based on the distributions of pHGG (C1), a cutoff classification threshold probability of 0.95 was used. Histogram bin width = 0.01. C2: rhGG, C3: TCGA-GBM

artifacts-other, similar image content for different MR sequences (e.g., a T1w-FLAIR sequence instead of T2w), misclassified diffusion-weighted imaging (DWI) as T2w, and DICOM corrupted scans (sample images shown in Figure 4.4). A manual evaluation revealed frequent misclassification (n=122, 19.68 %) if the architecture of the ventricles was altered, e.g., displaced by large tumours. This was assessed in detail: 122 random, correctly labelled images were used as a reference group. After manual segmentation of the tumour volume and brain, the Euclidean distance between the brain's centre of mass (CoM) and the CoM of the tumour volume was calculated. A t-test was then performed between the reference and misclassified CoM distributions. The t-test returned a p-value of 0.04, with a median CoMs distance of 46.15 voxels for the correctly labelled images and 66.31 for the misclassified images. This result shows a statistical difference between the groups, i.e., the further the

A

		T1w vs all		T1w vs T1wce		T2w vs all		T2w-FL vs all		ADC vs all		SWI vs all	
		0	1	0	1	0	1	0	1	0	1	0	1
	0	5633	196	1084	96	10653	50	10502	20	10431	7	10836	11
C2	1	75	5429	43	4206	41	589	46	765	5	890	2	484
	SE	96.6%		91.9%		99.5%		99.8%		99.9%		99.9%	
	SP	98.6%		99.0%		93.5%		94.3%		99.4%		99.6%	

C2	T1w	T1wce	T2w	T2w-FL	ADC	SWI	Other
T1w	91.17	8.07	0.08	0.17	0.00	0.00	0.50
T1wce	1.00	97.47	0.19	0.28	0.00	0.02	1.04
T2w	0.00	0.32	93.49	0.63	0.00	0.95	4.60
T2w-FLAIR	0.00	0.37	0.37	94.33	0.00	0.00	4.93
ADC	0.00	0.00	0.00	0.00	99.44	0.00	0.56
SWI	0.00	0.00	0.00	0.00	0.00	99.59	0.41
Other	0.07	2.10	1.26	0.07	0.23	0.13	96.14
n	1189	4315	630	811	895	486	3007

0-5 5-10 90-95 95-100

B

		T1w vs all		T1w vs T1wce		T2w vs all		T2w-FL vs all		ADC vs all		SWI vs all	
		0	1	0	1	0	1	0	1	0	1	0	1
	0	1963	30	411	9	3156	19	3111	22	3387	13	3505	17
C3	1	33	1496	13	1063	30	317	16	373	6	116	0	0
	SE	98.5%		97.8%		99.4%		99.3%		99.6%		99.5%	
	SP	97.8%		98.8%		91.3%		95.9%		95.4%		-	

C3	T1w	T1wce	T2w	T2w-FL	ADC	SWI	Other
T1w	94.92	2.08	0.00	0.23	1.15	0.00	1.62
T1w-Ca	1.19	96.99	0.00	0.18	0.18	0.00	1.46
T2w	0.00	0.00	91.35	0.00	0.00	0.00	8.65
T2w-FLAIR	0.00	0.26	0.26	95.89	0.00	0.00	3.60
ADC	0.00	0.00	0.00	0.00	95.08	0.00	4.92
SWI	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.70	1.85	1.59	1.67	0.53	1.50	92.16
n	433	1096	347	389	122	-	1135

Figure 4.3: Confusion matrices for rHGG and TCGA-GBM. Upper panel: Confusion matrices of the 6 DCNNs. Lower panel: MR-Class normalized confusion matrix(%). 'Other': image not labelled by the DCNNs; n: number of scans per class. C2: rHGG cohort, C3: TCGA-GBM cohort

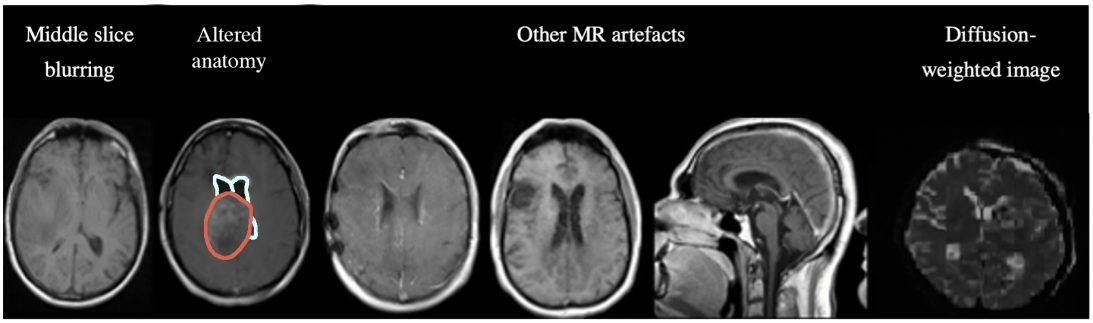


Figure 4.4: Examples of misclassified images. The first image is an example of a misclassified MR due to blurriness. The following two misclassifications are due to alterations in expected anatomy, e.g., displaced ventricles (in light blue) from overlapping tumours (red). The next two MR images show incorrect predictions due to different MR artefacts (Shading, motion, aliasing). All of these images are falsely classified as "other". The last image is a DWI, specifically, a Trace-DWI misclassified as T2w

.

tumour is from the ventricles, the less likely the image is misclassified. The frequencies of the misclassification categories are shown in Table 4.3.

Table 4.3: Frequency of the misclassified images. n represents the number of images per category, % is the percentage of images out of the total misclassified images

Category	n	%
MR artifact-other	146	26.84
MR artifact-middle slice blurring	127	23.35
Tumor displacing ventricles	122	22.43
Similar content- different sequence	80	14.71
DWI as T2w	76	13.97
DICOM corrupted images	69	12.68

4.2 DEEP LEARNING FOR DATA COMPLETION

When analyzing multimodality data, cohorts are often missing certain modalities due to technical or practical reasons, thus introducing the problem of data completion. This work employed deep learning to complete missing data points and thus enabled accurate modelling.

16% of the RT SS in cohort NSCLC are missing or corrupted. No missing RT SS were identified in the rHGG and HNC cohorts; thus, this section only

focuses on the NSCLC cohort Parameters and test results of the 2D and 3D nn-Unet automatic ROI segmentation used to deal with the missing structures from the DICOM RT SS are reported in Table4.4. The 2D network yielded the best performance for heart segmentation, while the 3D network for the GTV segmentation. ROI segmentation of the patient with the worst and best dice received in the test set is shown in Figure 4.5.

Table 4.4: Summary statistics and performance of the 2D and 3D nn-unet segmentation for GTV and heart segmentation. The best-performing network for each ROI is highlighted in grey

Type	Nr. images training/validation		Training time	Epochs	Validation dice score
GTV	2D	108/33	3.5 days	544	0.78 [0.72 -0.83]
	3D	108/33	5 days	641	0.83 [0.77-0.86]
Heart	2D	103/31	1.5 days	367	0.94 [0.90 -0.96]
	3D	103/31	3 days	272	0.92 [0.88 -0.94]

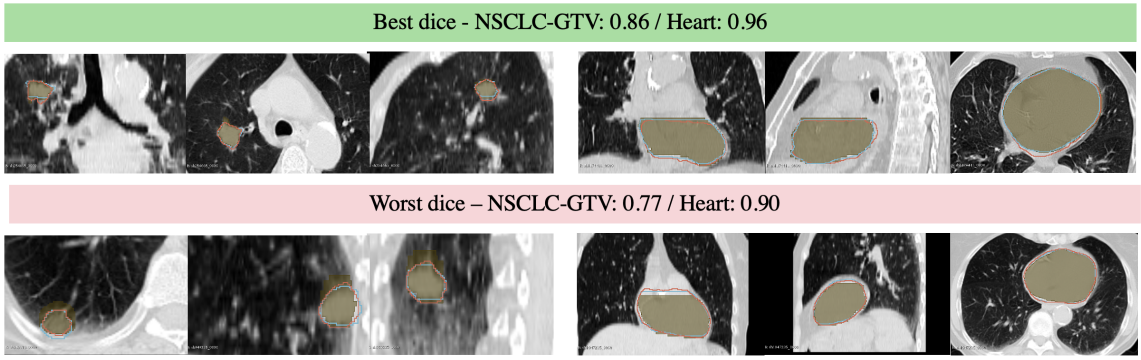


Figure 4.5: GTV and heart segmentation results on the test set. Reference ROI in black, segmentation from the 2D network in red, and 3D network in green

4.3 IMPACT OF MR INTENSITY NORMALIZATION METHODS

In MR studies, a crucial step is the normalization of the image intensities since MR intensities are acquired in arbitrary units and are thus not directly comparable. This work aimed to develop a methodology for assessing the impact of the used normalization method on the performance of radiomics models and choosing the most appropriate method for a certain entity.

This section reports the results of the intensity normalization computation experiment.

4.3.1 *Performance assessment*

Scatter plots of the CPH averaged (over the five bin counts investigated) C-I and POI averaged mse, plotted against the respective AIC, for the 15 different intensity normalization specific OS models derived from cohort pHGG and rHGG are shown in figure 4.6, and 4.7. The OS model derived from the non-normalized (nn) dataset is also plotted. The hard endpoint OS was used in this study as a possible appropriate surrogate since multiple MR scanners were found in both cohorts, where some have been withdrawn from clinical practice. Therefore, the application of phantoms to assess the actual impact of the IN methods could not be performed.

Table 4.5 summarizes and ranks the performance scores of the intensity normalization methods for each of the four MR sequences considered in cohorts pHGG and rHGG.

The white stripe method is ranked first for T1wce in both cohorts (pHGG/rHGG 10-fold CV C-I: 0.71/0.65, AIC: 1033/547, 10-CV mse: 0.21/0.14, AIC: 410/252). For T1w, the feature-based batch adjustment method, i.e. Combat had the best performance in pHGG (0.68, 964, 0.22), while z-score transformation in rHGG (0.65, 494, 0.15, 239). Nevertheless, the HM method was ranked second for both cohorts (pHGG/rHGG, 0.66/0.64, 970/494, 0.21/0.15, 389/2371). Furthermore, the top two ranked methods for T2w in both cohorts were Combat (pHGG/rHGG 0.62/0.67, 661/417, 0.22/0.13, 292/199) and the HM method (pHGG/rHGG 0.65/0.67, 667/415, 0.22/0.13, 294/200). As for T2w-FLAIR, the Fuzzy C-Means algorithm showed the best performance in pHGG and rHGG, however, with different masks. For pHGG, the mask combination of wm and mode (0.67, 907, 0.21, 366) had the best performance, while the mask combination of wm and csf (0.72, 508, 0.15, 230) showed the best results for rHGG. Nevertheless, the former was ranked second in rHGG (0.72, 517, 0.18, 235). Performance metrics of the remaining models in both cohorts are summarized in Table 4.6.

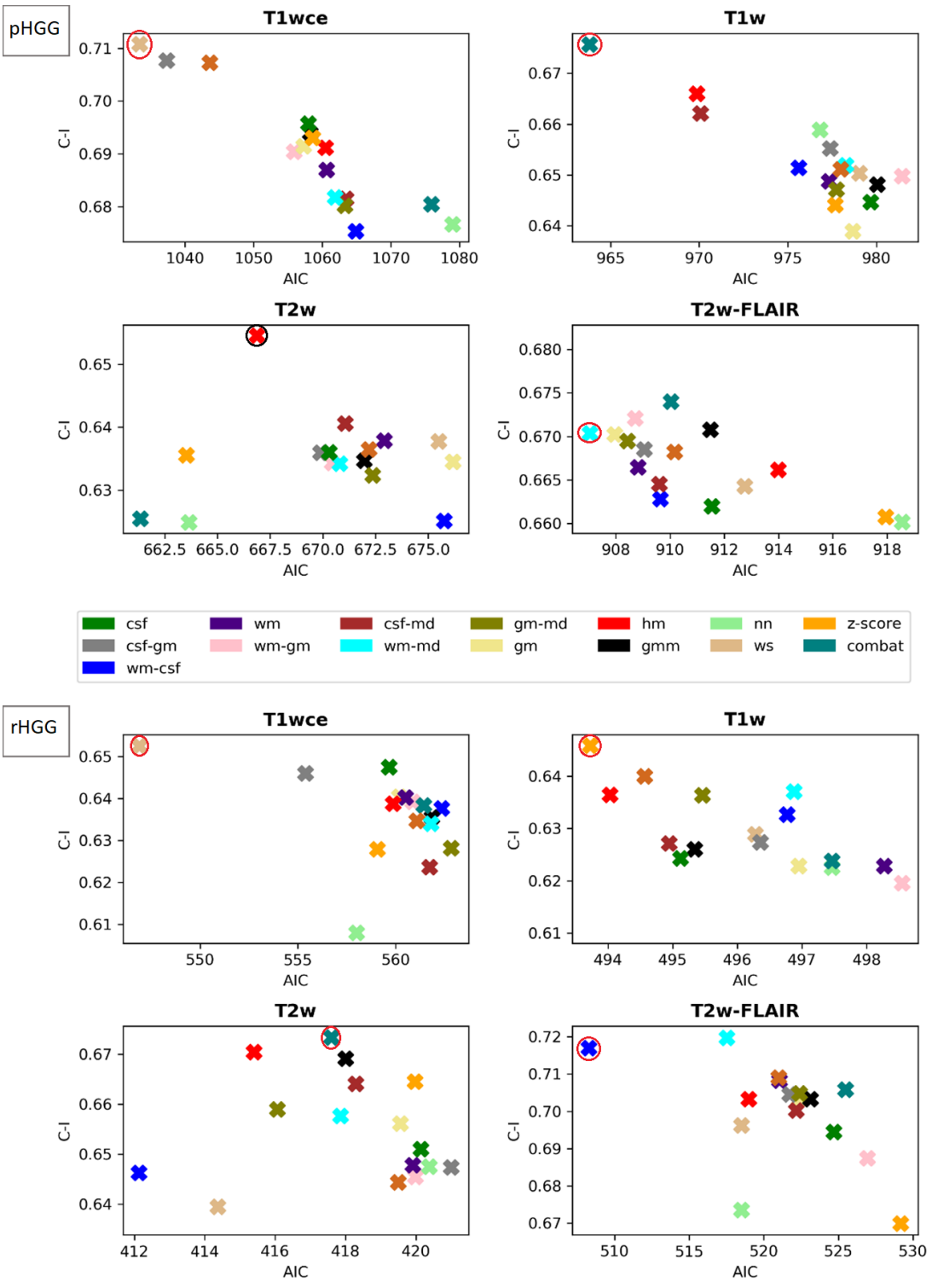


Figure 4.6: Scatter plots of the averaged (over the five bin counts considered) C-I vs AIC were obtained by the CPH models for all four sequences. The best-performing method is circled. Variability in the different MR sequences results was observed, i.e., the intensity normalization algorithm performance is correlated with the MR sequence. Upper panel: cohort pHGG, lower panel: cohort rHGG. C-I: Concordance-index, AIC: Akaike Information Criterion, CPH: Cox Proportional Hazard

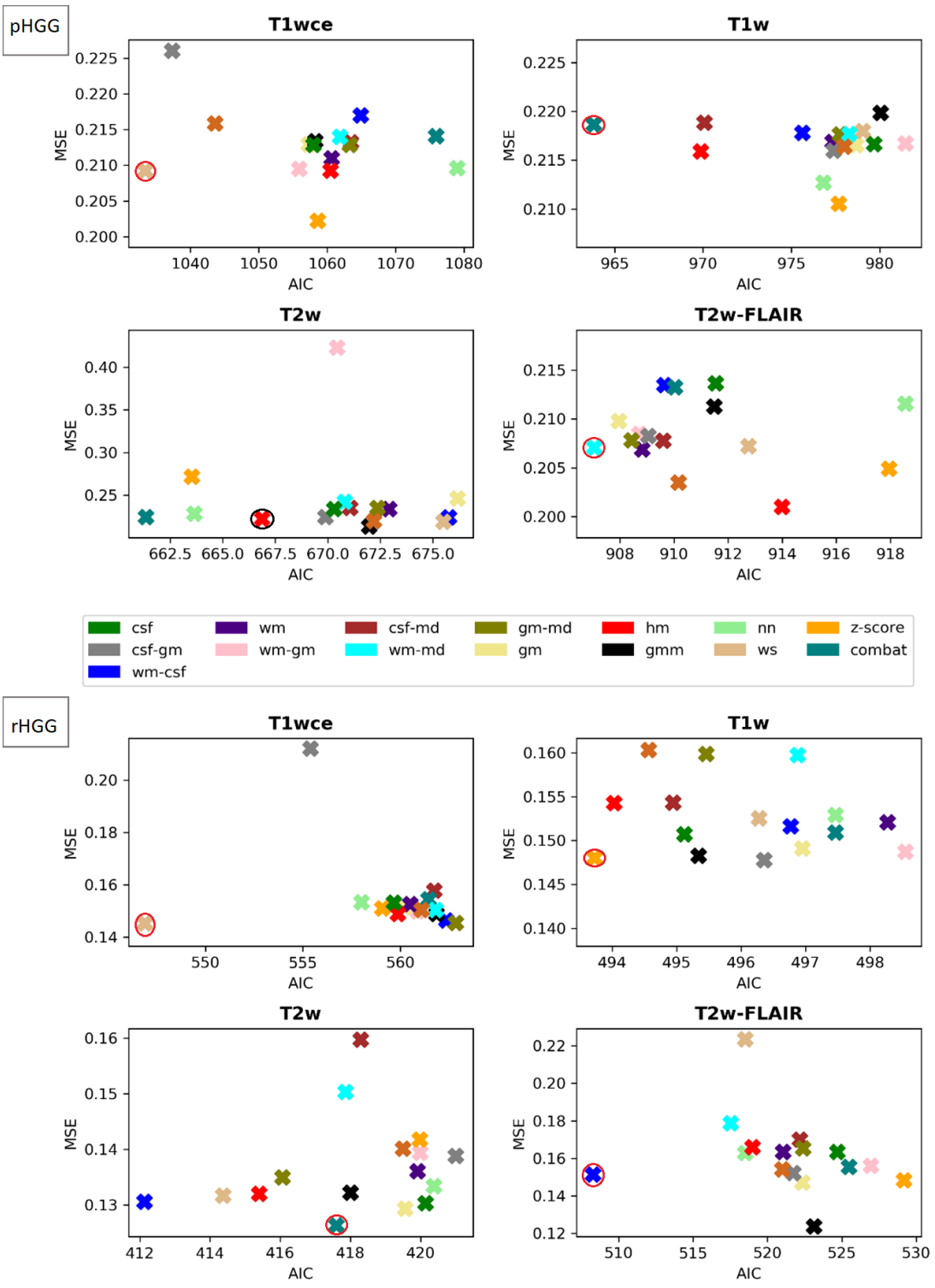


Figure 4.7: Scatter plots of the averaged (over the five bin counts considered) mse vs AIC were obtained by the POI models for all four sequences. The best-performing method is circled. Variability in the different MR sequences results was observed, i.e., the intensity normalization algorithm performance is correlated with the MR sequence. Upper panel: cohort pHGG, Lowe panel: cohort rHGG. POI: Poisson, mse: mean squared error, AIC: Akaike Information Criterion

Table 4.5: Ranking with scores of the intensity normalization method for each MR sequence in cohorts pHGG and rHGG. Methods that showed good performance in both cohorts are highlighted in grey. T2w-FL: T2w-FLAIR. Norm. method: Intensity normalization method

pHGG	T1wce		T1w		T2w		T2w-FL	
	Norm. method	Score	Norm. method	Score	Norm. method	Score	Norm. method	Score
1	ws	0.71	combat	0.13	hm	0.27	wm-md	0.02
2	kde	-0.13	hm	-0.28	combat	-0.03	wm-gm	-0.11
3	csf-gm	-0.20	csf-md	-0.90	z-score	-0.28	kde	-0.13
4	z-score	-0.48	nn	-1.00	gmm	-0.38	gm-md	-0.23
5	wm-gm	-0.85	z-score	-1.14	csf-gm	-0.61	gm	-0.24
6	csf	-0.97	csf-gm	-1.58	kde	-0.71	wm	-0.42
7	hm	-1.04	wm-csf	-1.65	nn	-0.76	csf-gm	-0.46
8	gmm	-1.11	wm	-1.85	csf	-0.78	combat	-0.77
9	gm	-1.13	kde	-1.88	wm-md	-0.80	csf-md	-0.77
10	wm	-1.24	wm-md	-1.95	gm-md	-0.96	hm	-0.80
11	wm-md	-1.67	gm-md	-2.05	csf-md	-1.09	wm-csf	-1.01
12	csf-md	-1.71	ws	-2.15	ws	-1.18	gmm	-1.02
13	gm-md	-1.72	csf	-2.16	wm	-1.22	ws	-1.29
14	wm-csf	-2.16	gm	-2.23	wm-gm	-1.72	csf	-1.75
15	combat	-2.25	wm-gm	-2.37	gm	-1.79	z-score	-2.21
16	nn	-2.27	gmm	-2.48	wm-csf	-2.01	nn	-2.65
rHGG								
1	ws	1.00	z-score	0.64	combat	0.07	wm-csf	0.66
2	csf	-0.54	hm	-0.11	hm	-0.09	wm-md	-0.32
3	hm	-0.73	csf	-0.34	gm-md	-0.21	gmm	-0.56
4	z-score	-0.76	gmm	-0.35	wm-csf	-0.24	kde	-0.63
5	gm	-0.77	csf-md	-0.81	gmm	-0.41	csf-gm	-0.71
6	wm	-0.87	kde	-0.93	wm-md	-0.78	wm	-0.72
7	wm-gm	-0.87	gm-md	-0.97	gm	-1.00	gm	-0.76
8	csf-gm	-0.96	csf-gm	-0.97	csf-md	-1.12	hm	-0.81
9	kde	-0.98	ws	-1.04	ws	-1.13	gm-md	-0.90
10	wm-csf	-1.07	gm	-1.18	z-score	-1.21	csf-md	-1.05
11	gmm	-1.10	combat	-1.20	csf	-1.31	nn	-1.25
12	wm-md	-1.13	nn	-1.41	kde	-1.36	csf	-1.35
13	combat	-1.19	wm-csf	-1.43	wm	-1.52	combat	-1.42
14	gm-md	-1.28	wm-md	-1.64	nn	-1.60	ws	-1.50
15	csf-md	-1.39	wm-gm	-2.01	wm-gm	-1.69	wm-gm	-1.59
16	nn	-1.82	wm	-2.11	csf-gm	-1.81	z-score	-2.11

Table 4.6: Model performance metrics for each MR sequence and normalization method for rHGG and pHGG. Methods that showed good performance in both cohorts are highlighted in grey

rHGG	T1wce					T1w					T2w					T2w-FL				
	No. feat.	AIC CPH	AIC POI	C-I	mse	No. feat.	AIC CPH	AIC POI	C-I	mse	No. feat.	AIC CPH	AIC POI	C-I	mse	No. feat.	AIC CPH	AIC POI	C-I	mse
combat	3.6	561.43	262.01	0.638	0.154	4.4	497.46	238.44	0.624	0.151	3.4	417.59	199.56	0.673	0.126	6	525.47	243.63	0.706	0.156
csf	3.2	559.66	258.3	0.647	0.153	3.2	495.12	236.39	0.624	0.151	3.8	420.13	199.39	0.651	0.13	6.4	524.68	239.19	0.694	0.164
csf-gm	2.8	555.38	255.67	0.646	0.212	4.4	494.94	238.34	0.627	0.154	3.6	421	201.05	0.647	0.139	6.8	521.74	236.78	0.705	0.152
csf-md	2.2	561.73	259.62	0.624	0.158	3.6	496.35	240.74	0.627	0.148	5	418.29	198.62	0.664	0.16	5.2	522.18	237.62	0.7	0.17
gm	5	560.18	259.12	0.64	0.151	3.2	496.95	239.6	0.623	0.149	4.8	419.55	200.26	0.656	0.129	6.6	522.35	237.27	0.703	0.147
gm-md	3.4	562.82	260.87	0.628	0.146	4	495.46	238.15	0.636	0.16	4.8	416.06	198.14	0.659	0.135	6.6	522.38	237.17	0.705	0.165
gmm	4	561.83	260.84	0.635	0.149	3.2	495.34	237.76	0.626	0.148	3.8	418	198.07	0.669	0.132	6.4	523.14	237.25	0.703	0.124
hm	2.6	559.87	258.8	0.639	0.149	2.8	494.03	237.45	0.636	0.154	3.4	415.39	200.42	0.67	0.132	5.8	518.98	237.71	0.703	0.166
kde	3.4	561.07	259.57	0.635	0.15	4.6	494.56	239.62	0.64	0.16	4.6	419.49	199.16	0.644	0.14	5.6	520.99	237.08	0.709	0.154
nn	3	558	264.03	0.608	0.153	3.2	497.46	238.44	0.623	0.153	3.8	420.37	199.49	0.648	0.133	8.4	518.5	236.25	0.674	0.163
wm	2.8	560.5	259.71	0.64	0.153	3	498.27	242.12	0.623	0.152	3.6	419.91	198.23	0.648	0.136	7.6	521.05	236.72	0.708	0.164
wm-csf	3.2	562.35	261.09	0.638	0.146	3.8	496.77	242.35	0.633	0.152	3.4	412.13	195.97	0.646	0.131	6	508.27	230.37	0.717	0.151
wm-gm	2.4	560.87	259.73	0.639	0.15	2.8	498.54	242	0.62	0.149	4	419.99	200.47	0.645	0.139	6.4	526.95	239.97	0.687	0.156
wm-md	2.8	561.8	260.5	0.634	0.151	3.8	496.87	240.59	0.637	0.16	4.6	417.86	198.19	0.658	0.15	5.4	517.53	234.81	0.72	0.179
ws	2.4	546.89	252.45	0.652	0.145	4	496.28	239.31	0.629	0.153	3.6	414.37	198.81	0.639	0.132	5	518.51	237.6	0.696	0.223
z-score	3.6	559.05	256.55	0.628	0.151	4.4	493.72	238.37	0.646	0.148	5	419.98	203.52	0.665	0.142	5.8	529.17	241.86	0.67	0.148
rHGG																				
combat	4.4	1075.92	424.71	0.68	0.214	4	963.81	387.77	0.676	0.219	1.6	661.34	291.7	0.625	0.225	6	910.03	368.51	0.674	0.213
csf	6.6	1057.96	418.54	0.696	0.213	3.8	979.68	397.75	0.645	0.217	2.6	670.3	294.2	0.636	0.234	4.8	911.54	368.16	0.662	0.214
csf-gm	7.4	1037.33	410.22	0.708	0.226	4.2	977.4	396.51	0.655	0.216	3	669.88	295.01	0.636	0.225	5	909.05	366.76	0.669	0.208
csf-md	5.4	1063.44	421.71	0.682	0.213	3.6	970.08	391.51	0.662	0.219	2.4	671.05	295.44	0.641	0.235	5	909.6	366.94	0.665	0.208
gm	6.6	1057.27	419.28	0.691	0.213	4	978.68	397.49	0.639	0.217	2.8	676.17	297.95	0.635	0.246	5.4	907.96	365.91	0.67	0.21
gm-md	5.2	1063.3	421.46	0.68	0.213	4.2	977.74	397.17	0.647	0.218	2.2	672.36	295.78	0.632	0.235	5.4	908.43	366.3	0.669	0.208
gmm	6.4	1058.22	419.21	0.694	0.213	3.8	980.06	398.36	0.648	0.22	2.4	671.96	291.7	0.635	0.214	5.4	911.49	368.84	0.671	0.211
hm	6.6	1060.47	419.17	0.691	0.209	4.4	969.87	389.05	0.666	0.216	2	666.87	293.7	0.655	0.222	5.2	913.99	369.14	0.666	0.201
kde	6.4	1043.62	413.4	0.707	0.216	4.2	977.98	397.6	0.651	0.216	3	672.18	294.37	0.636	0.22	5.4	910.16	366.24	0.668	0.203
nn	4.4	1078.98	425.38	0.677	0.21	4.6	976.81	395.25	0.659	0.213	2.6	663.64	294.21	0.625	0.228	5.2	918.55	370.59	0.66	0.212
wm	6	1060.63	419.24	0.687	0.211	4.4	977.34	396.98	0.649	0.217	2.4	672.91	295.94	0.638	0.234	5	908.82	366.33	0.666	0.207
wm-csf	6	1064.88	423.03	0.675	0.217	4.4	975.62	394.81	0.651	0.218	3	675.76	298.16	0.625	0.224	4.6	909.64	364.39	0.663	0.213
wm-gm	6.4	1055.89	417.18	0.69	0.21	3.6	981.46	400.79	0.65	0.217	2.8	670.43	294.45	0.634	0.423	5	908.71	366.22	0.672	0.208
wm-md	5.4	1061.88	420.72	0.682	0.214	4.4	978.27	397.1	0.652	0.218	2.8	670.8	294.97	0.634	0.242	5.2	907.05	366.15	0.67	0.207
ws	6.8	1033.41	409.77	0.711	0.209	4.2	979.04	398.12	0.65	0.218	3	675.5	295.8	0.638	0.219	4.6	912.75	368.97	0.664	0.207
z-score	5.8	1058.63	416.4	0.693	0.202	3	977.69	394.24	0.644	0.211	2.4	663.53	293.13	0.636	0.272	4	917.95	371.92	0.661	0.205

4.3.2 *Significant feature correlation between the normalized datasets*

To assess which features are affected by the intensity normalization methods so that they can be further eliminated before moving forward to the modelling, pairwise correlation tests were performed. Spearman correlation heatmaps between the different normalization methods of the significant features identified for each of the bin counts considered were plotted. Even though they are significantly correlated to OS, it was observed that certain features had different distributions when different intensity normalization methods were applied. An example of T1wce significant features from pHGG and bin count 64 is shown in Figure 4.8.

The 10-CV C-I and mse of the CPH and POI models with only the stable features with a high correlation (rs >0.8) between at least 12 methods are reported in Table 4.7. Box plots of the difference for each modality in both cohorts are shown in Figure 4.9.

Table 4.7: Performance of the top-ranked image normalization method before and after the elimination of the intensity normalization impacted significant features for cohorts pHGG and rHGG for each MR sequence. The average (across all bin counts) 10-CV C-I/MSE with the 95% confidence intervals is reported. Performance of both models was similarly affected after the elimination of the intensity normalization impacted significant features, with a mean decrease in the 10-CV C-I and 10-CV MSE of 0.05 and 0.03 in all four sequences across both cohorts

.

	pHGG		**rHGG**	
	Before	After	Before	After
T1wce	0.71 [0.69 0.74] / 0.21 [0.19 0.23]	0.65 [0.63 0.69] / 0.23 [0.21 0.25]	0.65 [0.62 0.67] / 0.15 [0.13 0.17]	0.62 [0.60 0.65] / 0.19 [0.17 0.21]
T1w	0.68 [0.64 0.70] / 0.22 [0.20 0.25]	0.63 [0.61 0.67] / 0.24 [0.22 0.26]	0.65 [0.61 0.69] / 0.15 [0.12 0.18]	0.62 [0.58 0.65] / 0.18 [0.15 0.20]
T2w	0.65 [0.62 0.67] / 0.22 [0.19 0.25]	0.63 [0.60 0.67] / 0.25 [0.22 0.28]	0.67 [0.64 0.69] / 0.13 [0.11 0.17]	0.60 [0.58 0.65] / 0.16 [0.14 0.20]
T2w-FL	0.67 [0.64 0.69] / 0.20 [0.18 0.23]	0.62 [0.59 0.67] / 0.23 [0.21 0.25]	0.72 [0.65 0.76] / 0.18 [0.15 0.21]	0.66 [0.64 0.69] / 0.20 [0.17 0.22]

4.3.3 *Performance comparison of the feature-based and top-ranked image-based normalization methods*

Table 4.8 summarizes the performance of the top-ranked image normalization method separate and in combination with the feature-based method Combat

for cohorts pHGG and rHGG. Since Combat ranked first for the T1w models from pHGG and T2w models from rHGG, the second-ranked method, i.e., the HM method, was the image-based intensity normalization method for these two datasets.

Table 4.8: Performance of the top-ranked image normalization method separate and in combination with the feature-based method Combat for cohorts pHGG and rHGG for each MR sequence. The average (across all bin counts) 10-CV C-I/MSE with the 95% confidence intervals is reported

	pHGG		
	Combat	I. norm.	Combined
T1wce	0.68 [0.66 0.70] / 0.21 [0.19 0.23]	0.71 [0.690.74] / 0.21 [0.19 0.23]	0.68 [0.66 0.69] / 0.21 [0.19 0.23]
T1w	0.68 [0.64 0.70] / 0.22 [0.20 0.24]	0.66 [0.64 0.68] / 0.22 [0.19 0.24]	0.62 [0.59 0.64] / 0.23 [0.20 0.26]
T2w	0.62 [0.59 0.64] / 0.23 [0.21 0.23]	0.65 [0.62 0.67] / 0.22 [0.19 0.25]	0.61 [0.58 0.63] / 0.25 [0.23 0.27]
T2w-FL	0.67 [0.64 0.69]/ 0.21 [0.19 0.24]	0.67 [0.64 0.69] / 0.20 [0.18 0.23]	0.64 [0.61 0.66] / 0.24 [0.22 0.26]
	rHGG		
	Combat	I. norm.	Combined
T1wce	0.64 [0.62 0.68] / 0.15 [0.13 0.17]	0.65 [0.62 0.67] / 0.15 [0.13 0.17]	0.63 [0.61 0.66] / 0.17 [0.15 0.19]
T1w	0.62 [0.60 0.66] / 0.15 [0.12 0.17]	0.65 [0.61 0.69] / 0.15 [0.12 0.18]	0.62 [0.59 0.65] / 0.15 [0.11 0.16]
T2w	0.67 [0.64 0.69] / 0.13 [0.11 0.17]	0.67 [0.64 0.69] / 0.13 [0.11 0.15]	0.62 [0.59 0.65] / 0.15 [0.13 0.19]
T2w-FL	0.70 [0.67 0.72] / 0.16 [0.14 0.19]	0.72 [0.65 0.76] / 0.14 [0.12 0.17]	0.68 [0.65 0.70] / 0.17 [0.15 0.21]

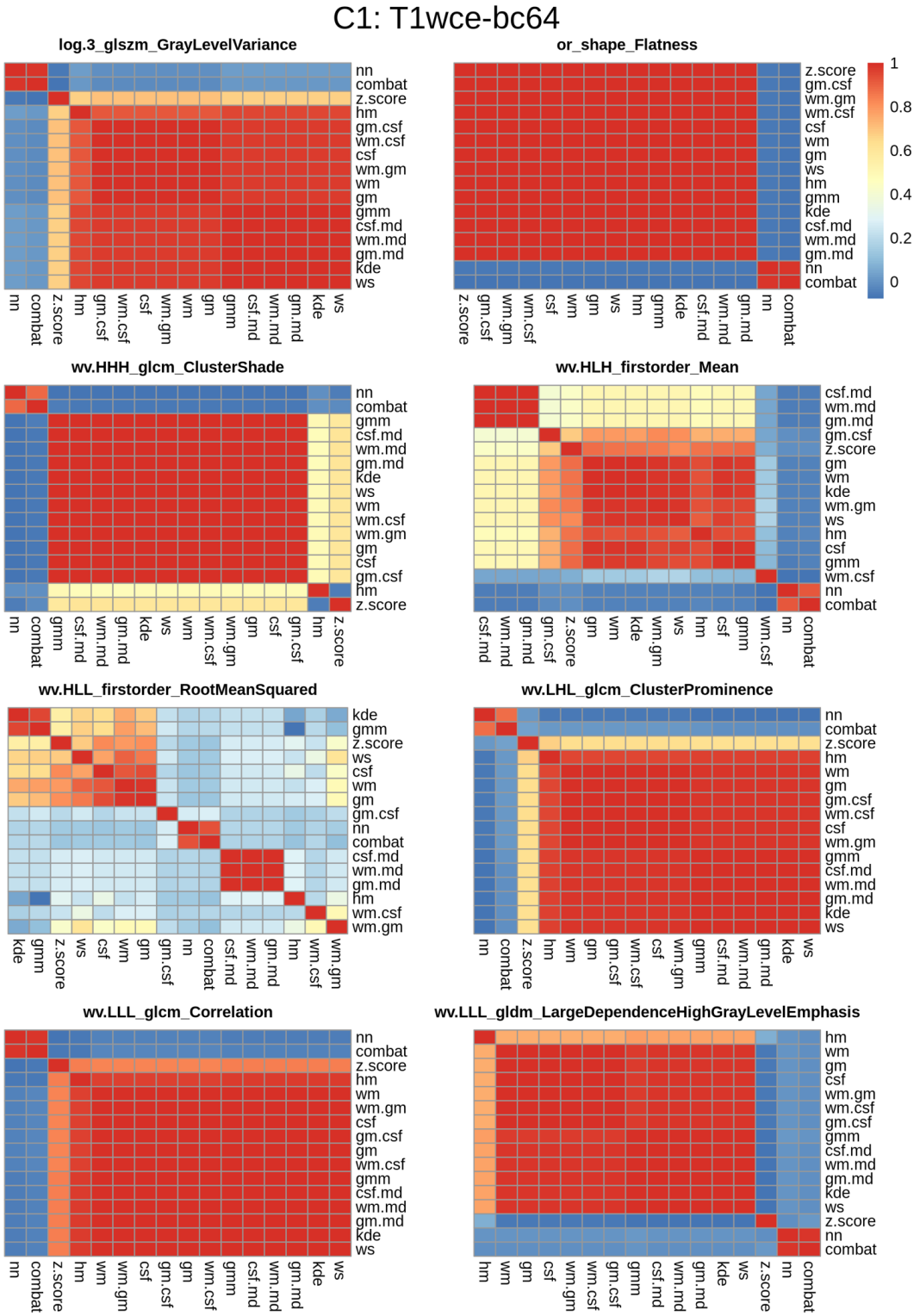


Figure 4.8: Correlation heatmaps between the 15 different normalization methods considered and the reference non-normalized dataset for T1wce images from cohort pHGG discretized with a bin count of 64. Features with a high correlation (rs>0.8) between at least 12 methods were selected as robust features in subsequent feature selection steps

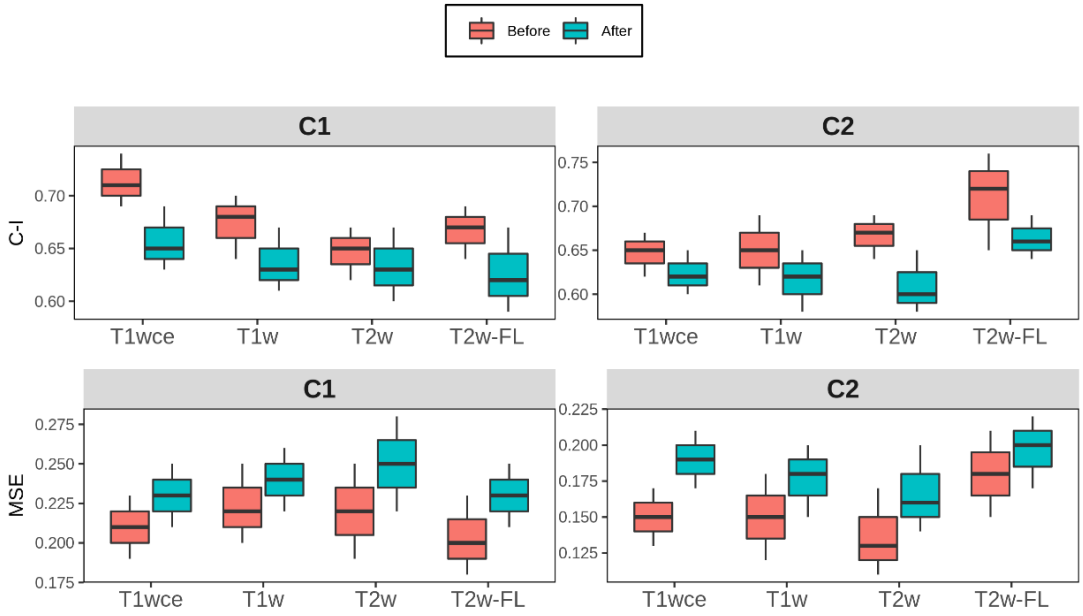


Figure 4.9: Box plots of the top-ranked image normalization method evaluation metrics C-I and MSE before and after the elimination of the intensity normalization impacted significant features for cohorts pHGG and rHGG for each MR sequence. The average (across all bin counts) 10-CV C-I/MSE with the 95% confidence intervals are plotted. Performance of both models was similarly affected after the elimination of the intensity normalization impacted significant features, with a mean decrease in the 10-CV C-I and 10-CV MSE of 0.05 and 0.03 in all four sequences across both cohorts C-I: Concordance-index, mse: mean squared error

4.4 DATA-BASED DRIVEN TCP AND NTCP MODELLING

This work incorporated multiple layers of information from medical data, placing emphasis on features extracted from the 3D dose distributions to evaluate the potential improvement of patient stratification and prognostication and assess the added benefit for TCP/NTCP modelling of a combined modelling approach, i.e., the use of radiomics, dosiomics, and clinical features, and the survival and radiation-induced toxicity predictions in recurrent high-grade glioma, early-stage non-small cell lung cancer and head and neck cancer patients treated with CIRT, SBRT and IMRT or helical tomotherapy respectively.

4.4.1 *The unimodality and multimodality signatures*

After having curated, prepared, and completed missing data, and the application of the preprocessing pipeline, the rejection of features impacted by preprocessing steps to increase reproducibility, and the identification of significantly outcome-correlated features, the uni- and multimodality signatures were built to predict overall and progression-free survival and radiation-induced toxicity in the rHGG, NSCLC and HNC cohorts. The identifiable radiomics and dosiomics features for all considered cohorts and endpoints are summarized in Table B.1 (Appendix B)

OS and PFS TCP models

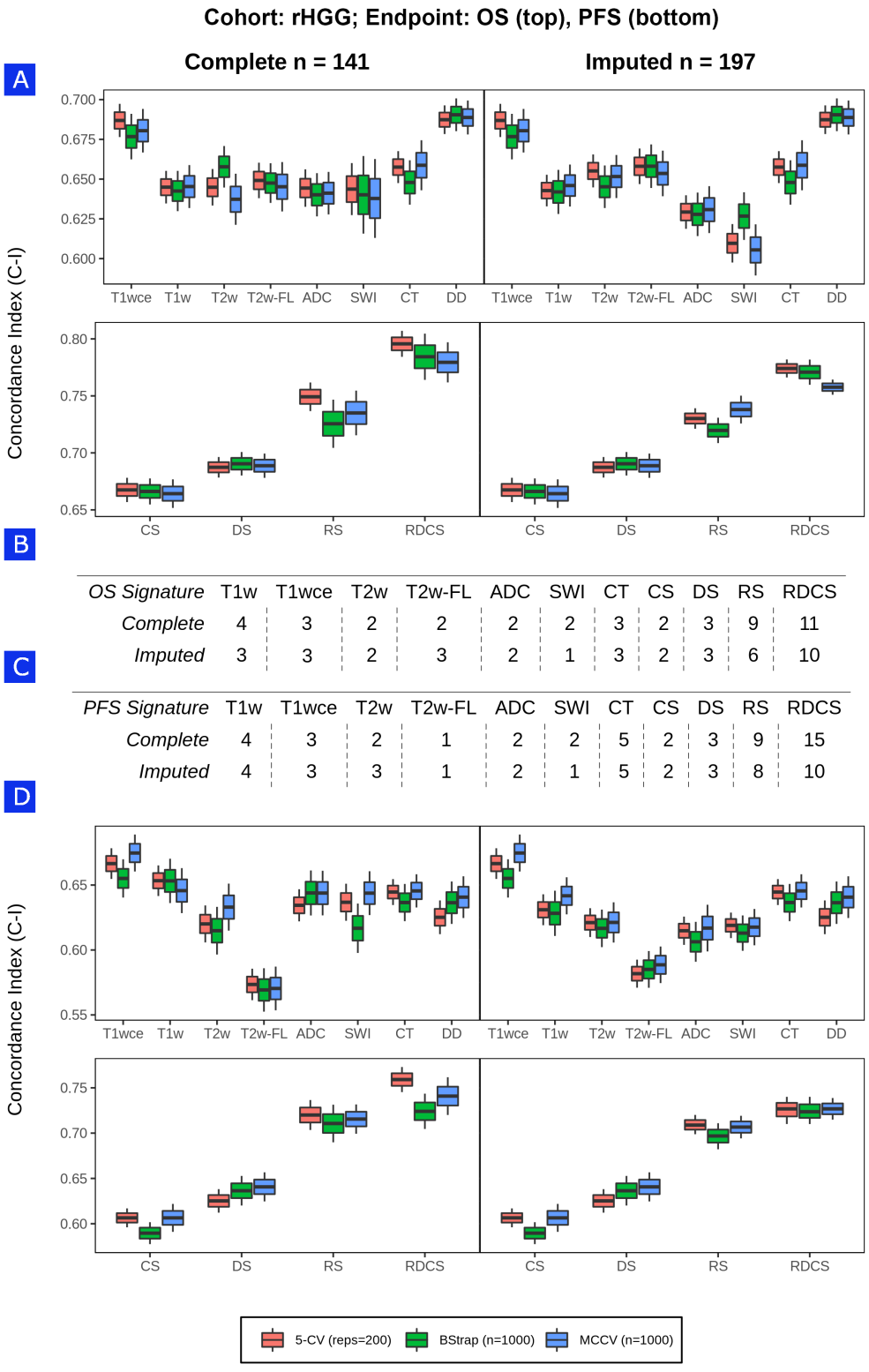
Figures 4.10 and 4.11 show the box plots of the 1st - 99th discovery set C-Is resulting from the three resampling approaches following the fitting of the OS and PFS CPH models by the complete and imputed unimodality and multimodality signatures, i.e., the radiomics signature (RS), and the combined RS, dosiomics (DS) and clinical signature (CS) (RDCS) in the rHGG and NSCLC cohorts. The number of features per modality/signature is also shown in panels B and C. The average CPH discovery set C-Is, with the minimum and maximum C-I achieved, across the three different resampling approaches for all outcomes considered is summarised in Tables C.3, and C.1 (Appendix C).

The top-ranked rHGG OS models achieved on the DS are from the DD (C-I 0.69 [0.66 0.70]) and T1wce (0.68 [0.66 0.70]). The worst ranked model is from the SWI (complete 0.64 [0.61 0.66], imputed 0.61 [0.59 0.64]). The multimodality radiomics, dosiomics and clinical (RDCS) models improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.79 [0.76 0.81]) and imputed (0.77 [0.75 0.78]) datasets. The top-ranked NSCLC PFS models achieved on the DS is from the T1wce (0.68 [0.66 0.70]). The worst ranked model is from the T2w-FLAIR model (Com-

plete 0.57 [0.55 0.59], Imputed 0.59 [0.57 0.60]). The multimodality radiomics, dosiomics and clinical (RDCS) models improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.74 [0.70 0.77]) and imputed (0.72 [0.71 0.74]) datasets. No significant difference was observed between the three different resampling approaches.

As for the NSCLC cohort, the top-ranked OS model is from the SH-CT (complete 0.70 [0.65 0.72]), imputed 0.69 [0.68 0.71] and the CS (0.70 [0.69 0.72]). The worst-ranked model is from the SM-CT (0.68 [0.65 0.71]). The multimodality radiomics, dosiomics and clinical (RDCS) models improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.74 [0.69 0.78]) and imputed (0.78 [0.76 0.80]) datasets. Wider CI (except for the CS and SM-CT models) were observed by the bootstrapping method. No significant difference was observed between the remaining two resampling approaches. The top-ranked NSCLC PFS models achieved on the DS is from the DD (0.69 [0.68 0.71]). The worst-ranked model is from the SM-CT model (0.61 [0.59 0.63]). The radiomics model (complete 0.62 [0.61 0.64]), imputed 0.62 [0.60 0.64]), combining both the SM-CT and SH-CT, had a worst performance than the individual unimodality models. The multimodality radiomics, dosiomics, and clinical (RDCS) model derived from the imputed dataset (0.66 [0.64 0.68]) did not have a better prediction than the SH-CT model. The RDCS model derived from the complete dataset improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.72 [0.69 0.74]). No significant difference was observed between the three different resampling approaches.

Table 4.10, and 4.12 summarise the test set C-Is of the unimodality and multimodality CPH and RSF fitted by the DS. The 95% CI was derived through bootstrapping (n=1000). Similar to the results obtained by the DSs, the RDCS models yielded the best prediction performance in the TSs for all models considered.



OS Signature	T1w	T1wce	T2w	T2w-FL	ADC	SWI	CT	CS	DS	RS	RDCS
Complete	4	3	2	2	2	2	3	2	3	9	11
Imputed	3	3	2	3	2	1	3	2	3	6	10

PFS Signature	T1w	T1wce	T2w	T2w-FL	ADC	SWI	CT	CS	DS	RS	RDCS
Complete	4	3	2	1	2	2	5	2	3	9	15
Imputed	4	3	3	1	2	1	5	2	3	8	10

Figure 4.10: Box plots of the 1st - 99th rHGG C-Is achieved by the unimodality and multimodality models fitted by the imputed and complete signatures after three resampling approaches in the prediction of OS and PFS.A) Overall survival (OS) prediction unimodality and single sequence radiomics features (upper row), and multimodality models, B) number of features per OS signature, C) number of features per PFS signature, and D) Progression-free survival (PFS) prediction models prediction unimodality and single sequence radiomics features (upper row), and multimodality models

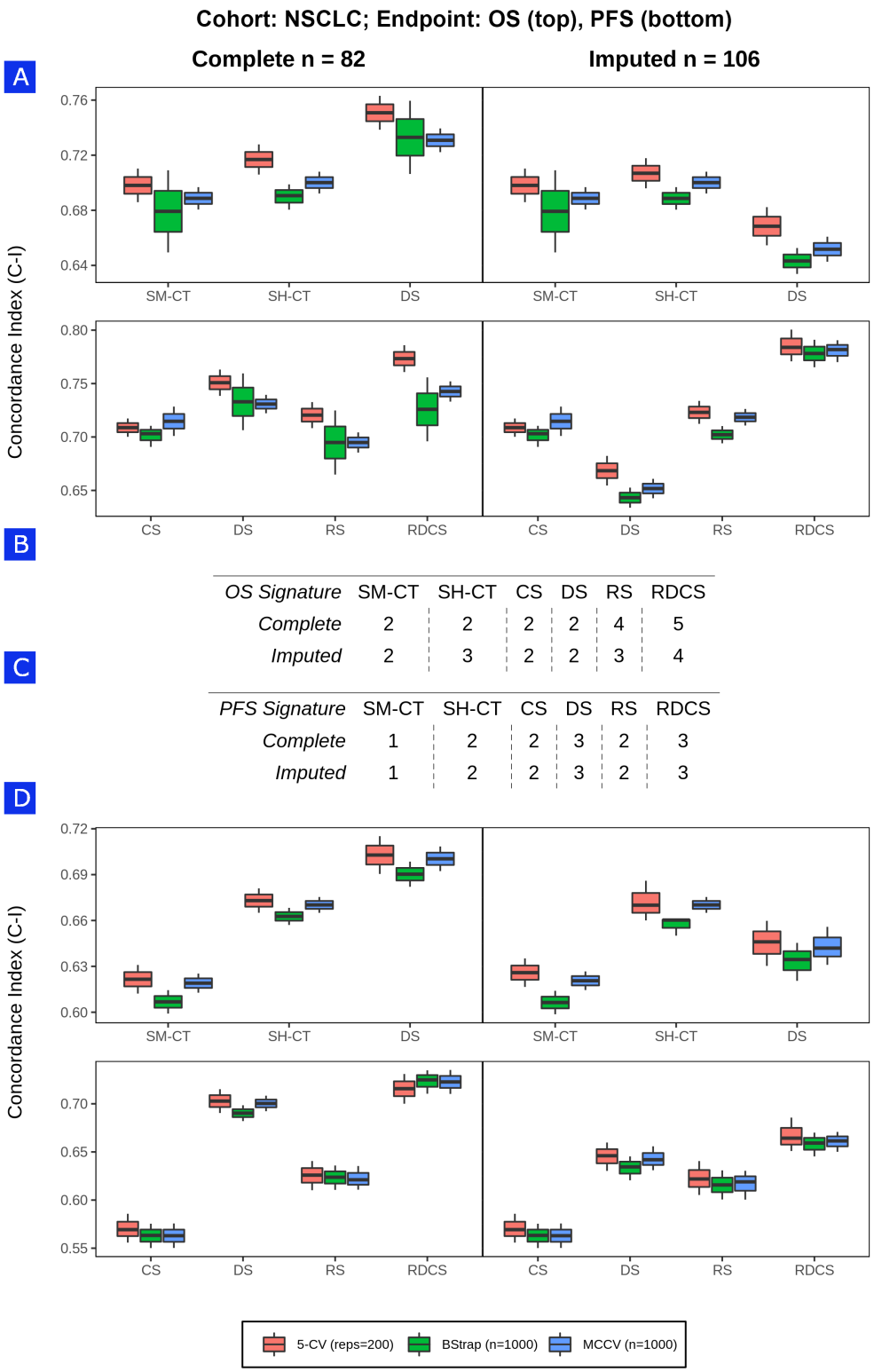


Figure 4.11: Box plots of the 1st - 99th C-Is achieved by the unimodality and multimodality models fitted by the imputed and complete signatures after three resampling approaches for cohort NSCLC. A) Overall survival (OS) prediction unimodality and single sequence radiomics features (upper row), and multimodality models, B) number of features per OS signature, C) number of features per PFS signature, and D) Progression-free survival (PFS) prediction models prediction unimodality and single sequence radiomics features (upper row), and multimodality models

Table 4.9: OS/PFS test set C-Is for the different unimodality and single sequence radiomics features (upper row) and multimodality CPH and RSF models fitted using both the imputed and complete signatures in NSCLC. Range showing the minimum and maximum C-I achieved across the three different resampling approaches. SM-CT: smooth-kernel CT, SH-CT: sharp-kernel CT, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature, CPH: Cox Proportional Hazard Model, RSF: Random Survival Forest

	CPH	
	Complete	Imputed
CS	0.67 [0.66 0.69]/0.63 [0.61 0.66]	
SM-CT	0.69 [0.67 0.72]/0.66 [0.64 0.69]	
SH-CT	0.66 [0.64 0.68]/0.61 [0.58 0.63]	
DD	0.68 [0.67 0.71]/0.67 [0.65 0.71]/	0.68 [0.65 0.70]/0.66 [0.64 0.68]
RS	0.71 [0.68 0.73]/0.69 [0.67 0.72]	0.70 [0.68 0.72]/0.67 [0.65 0.70]
RDCS	0.75 [0.72 0.77]/0.73 [0.71 0.75]	0.73 [0.70 0.75]/0.72 [0.70 0.73]
	RSF	
CS	0.67 [0.66 0.69]/0.63 [0.60 0.65]	
SM-CT	0.67 [0.65 0.70] / 0.62 [0.59 0.64]	
SH-CT	0.69 [0.68 0.70]/0.57 [0.55 0.60]	
DD	0.68 [0.66 0.70]/0.66 [0.64 0.68]	0.66 [0.64 0.69]/0.63 [0.61 0.66]
RS	0.70 [0.69 0.71]/0.68 [0.66 0.70]	0.71 [0.70-0.73]/0.66 [0.64 0.68]
RDCS	0.72 [0.70 0.74]/0.70 [0.68 0.71]	0.73 [0.71-0.77]/0.70 [0.69 0.72]

Table 4.10: OS/PFS test set C-Is for the different unimodality and single sequence radiomics features (upper row) and multimodality CPH and RSF models fitted using both the imputed and complete signatures in rHGG. Range showing the minimum and maximum C-I achieved across the three different resampling approaches. T2w-FL: T2w- FLAIR.

CPH		
	Complete	Imputed
CS	0.66 [0.66 0.68] / 0.61 [0.60 0.62]	
CT	0.67 [0.66 0.67] / 0.67 [0.66 0.68]	
DD	0.72 [0.71 0.72] / 0.66 [0.65 0.66]	
T1wce	0.73 [0.72 0.73] / 0.66 [0.66 0.67]	
T1w	0.74 [0.74 0.75]/0.66 [0.65 0.66]	0.73 [0.73 0.74]/0.70 [0.68 0.70]
T2w	0.66 [0.66 0.67]/0.65 [0.64 0.66]	0.66 [0.65 0.67]/0.62 [0.62 0.63]
T2w-FL	0.63 [0.63 0.64]/0.53 [0.52 0.54]	0.62 [0.62 0.63] /0.53 [0.52 0.54]
ADC	0.63 [0.63 0.64]/0.61 [0.61 0.62]	0.60 [0.56 0.61]/0.60 [0.59 0.60]
SWI	0.68 [0.68 0.69]/0.69 [0.69 0.70]	0.60 [0.59 0.60]/0.64 [0.63 0.64]
RS	0.79 [0.79 0.80]/0.74 [0.72 0.77]	0.72 [0.72 0.73]/0.68 [0.68-0.69]
RDCS	0.81 [0.78 0.84]/0.80 [0.78 0.83]	0.79 [0.78 0.82]/0.77 [0.75 0.79]
RSF		
	Complete	Imputed
CS	0.64 [0.63 0.65] / 0.62 [0.61 0.63]	
CT	0.56 [0.55 0.56] / 0.67 [0.66 0.69]	
DD	0.69 [0.68 0.70] / 0.66 [0.65 0.67]	
T1wce	0.68 [0.68 0.69] / 0.61 [0.59 0.64]	
T1w	0.74 [0.73 0.74]/0.65 [0.65 0.66]	0.71 [0.70-0.72]/0.65 [0.65-0.66]
T2w	0.68 [0.67 0.69]/0.55 [0.54 0.56]	0.67 [0.66-0.68]/0.54 [0.53-0.55]
T2w-FL	0.59 [0.58 0.59]/0.42 [0.41 0.43]	0.60 [0.59-0.60]/0.47 [0.46-0.49]
ADC	0.57 [0.56 0.58]/0.58 [0.57 0.59]	0.58 [0.57-0.58]/0.61 [0.60-0.62]
SWI	0.58 [0.57 0.58]/0.65 [0.65 0.66]	0.59 [0.58-0.60]/0.62 [0.61-0.62]
RS	0.78 [0.77 0.78] /0.74 [0.73 0.75]	0.71 [0.70-0.73]/0.71 [0.69-0.73]
RDCS	0.80 [0.78 0.81]/0.75 [0.73 0.77]	0.77 [0.76-0.78] /0.76 [0.75 0.78]

RILF and XT NTCP models

Figures 4.12 and 4.13 show the box plots of the 1st - 99th discovery set C-Is resulting from the three resampling approaches following the fitting of the XT and RILF CPH models by the complete and imputed unimodality and multimodality signatures, i.e., the radiomics signature (RS), and the combined RS, dosiomics (DS) and clinical signature (CS) (RDCS). The number of features per modality/signature is also shown in panels B and C.

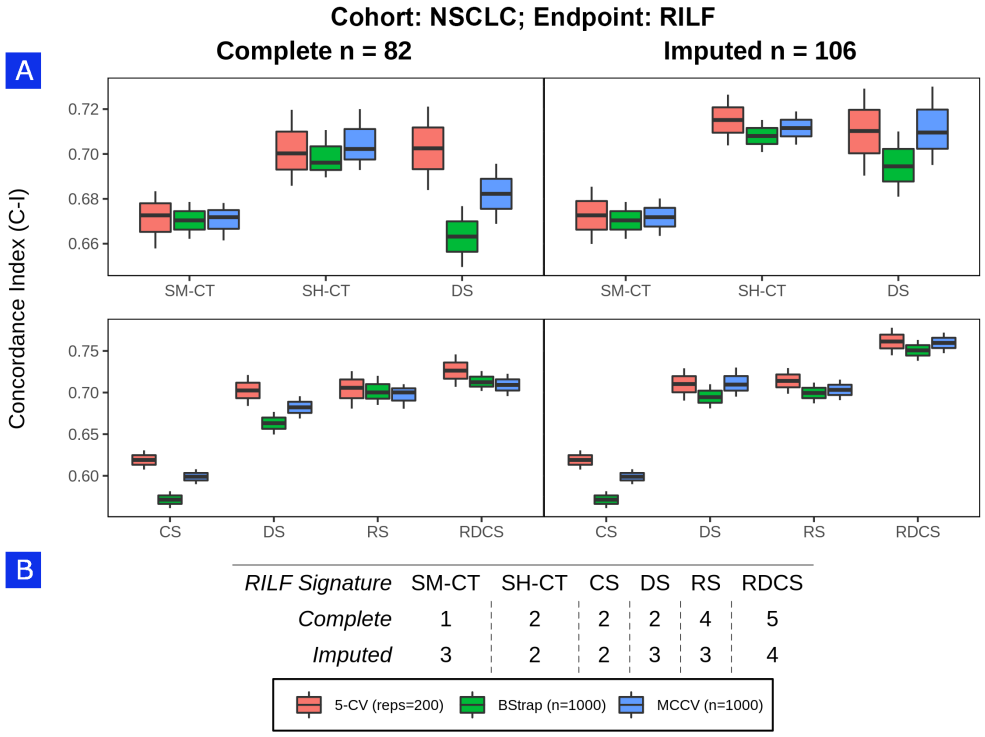


Figure 4.12: Box plots of the 1st - 99th NSCLC C-Is achieved by the unimodality and single sequence radiomics features (upper row) and multimodality models fitted by the imputed and complete signatures after three resampling approaches in the prediction of RILF. A) Radiation-induced lung fibrosis (RILF) prediction unimodality and single sequence radiomics features (upper row), and multimodality models, B) number of features per RILF signature. SM-CT: smooth-kernel CT, SH-CT: sharp-kernel CT, CS: clinical signature, DS: dosiomics signature, RS: radiomics signature, RDCS: combined RS, DS and CS signature, 5-CV: 5-fold cross-validation, BStrap: bootstrap, MCCV: Monte Carlo cross-validation

The top-ranked NSCLC RILF model achieved on the DS are from the DD (complete C-I 0.68 [0.64 0.71, imputed 0.70 [0.68 0.73]) and SH-CT (complete 0.69 [0.67 0.71], imputed 0.69 [0.68 0.70]). The worst-ranked model is from the CS (0.59 [0.56 0.63]). The multimodality radiomics, dosiomics and clinical (RDCS) models improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.72 [0.69 0.72]) and imputed

(0.75 [0.73 0.77]) datasets. No significant difference was observed between the three different resampling approaches.

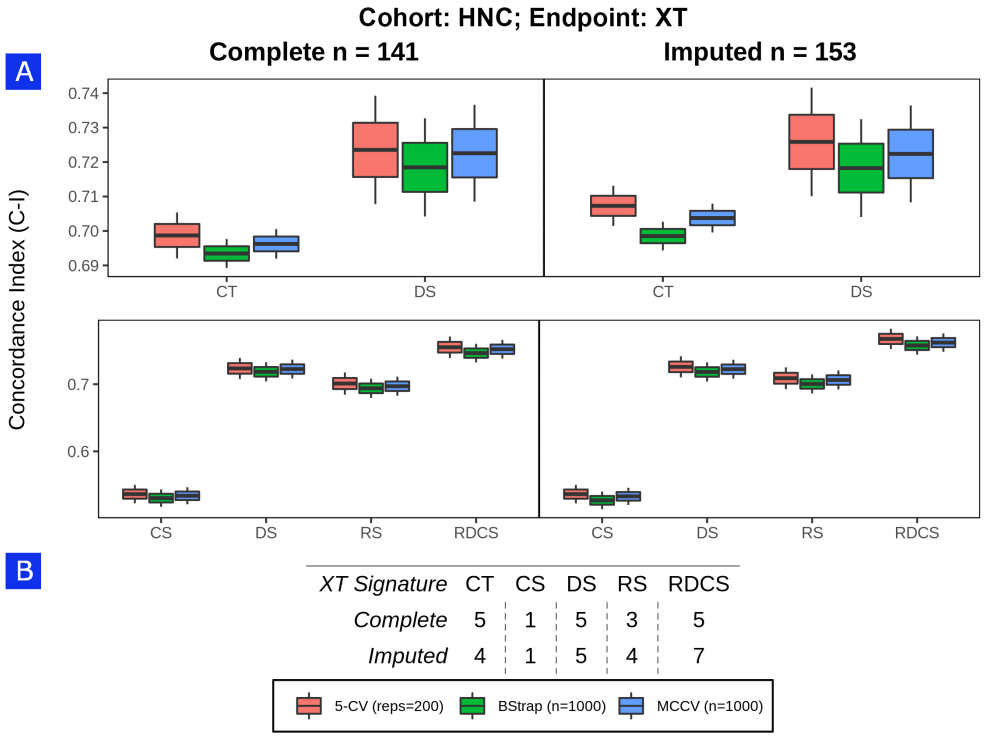


Figure 4.13: Box plots of the 1st - 99th HNC C-Is achieved by the unimodality and multimodality models fitted by the imputed and complete signatures after three resampling approaches in the prediction of XT. A) Xerostomia (XT) prediction unimodality and single sequence radiomics features (upper row), and multimodality models, B) number of features per XT signature

The top-ranked HNC XT model achieved on the DS is from the DD (0.72 [0.70 0.74]). The worst-ranked model is from the CS (0.51 [0.50 0.53]). The multimodality radiomics, dosiomics and clinical (RDCS) models improved prediction compared to the top-ranked unimodality and clinical models in both the complete (0.75 [0.73 0.77]) and imputed (0.76 [0.74 0.78]) datasets. No significant difference was observed between the three different resampling approaches. Table 4.10, 4.12, and 4.11 summarise the test set C-Is of the unimodality and multimodality CPH and RSF fitted by the DS. The 95% CI was derived through bootstrapping (n=1000). Similar to the results obtained by the DSs, the RDCS models yielded the best prediction performance in the TSs for all models considered.

Table 4.11: XT test set C-Is for the different unimodality and multimodality CPH and RSF models fitted using both the imputed and complete signatures in HNC. Range showing the minimum and maximum C-I achieved across the three different resampling approaches. XT: xerostomia, C-I: Concordance index, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature, CPH: Cox Proportional Hazard Model, RSF: Random Survival Forest

	CPH		**RSF**	
	Complete	Imputed	Complete	Imputed
CS	0.55 [0.53 0.58]		0.50 [0.48 0.54]	
DD	0.75 [0.72 0.77]		0.72 [0.70 0.74]	
RS	0.69 [0.67-0.71]	0.68 [0.67 0.69]	0.65 [0.64 0.67]	0.67 [0.65 0.68]
RDCS	0.78 [0.75 0.80]	0.77 [0.74 0.80]	0.73 [0.71-0.75]	0.72 [0.70-0.74]

Table 4.12: RILF test set C-Is for the different unimodality and multimodality CPH and RSF models fitted using both the imputed and complete signatures in NSCLC. Range showing the minimum and maximum C-I achieved across the three different resampling approaches. SM-CT: smooth-kernel CT, SH-CT: sharp-kernel CT, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature, CPH: Cox Proportional Hazard Model, RSF: Random Survival Forest

	CPH		**RSF**	
	Complete	Imputed	Complete	Imputed
CS	0.67 [0.66 0.70]		0.64 [0.62 0.65]	
SM-CT	0.68 [0.65 0.71]		0.65 [0.63 0.68]	
SH-CT	0.65 [0.60 0.66]		0.63 [0.60 0.66]	
DD	0.72 [0.70 0.74]	0.71 [0.68 0.73]	0.70 [0.68 0.72]	0.67 [0.65 0.70]
RS	0.72 [0.70 0.74]	0.71 [0.68 0.73]	0.70 [0.68 0.71]	0.69 [0.67 0.70]
RDCS	0.77 [0.73 0.79]	0.74 [0.72 0.76]	0.71 [0.70-0.73]	0.71 [0.69-0.74]

4.4.2 *The clinical, radiomics and dosiomics feature forming the RDCS signature*

In the following sections, the imputed dataset models are further considered for analysis, where a multivariate Cox regression analysis incorporating the different multimodality factors, i.e., clinical, radiomics and dosiomics, forming the cohort-specific signatures was used to calculate the corrected HR and 95% CI. Forest plots were created with the multivariate regression models to show the correlation of each feature to the modelled outcomes. Furthermore, X-means clustering was performed separately on the corresponding radiomics and dosiomics features and on the multimodality combined signatures, and patients were assigned to the derived clusters. A low-dimensional representation of the data was derived using t-sne (Appendix D). X-means was not performed when single radiomics or dosiomics features formed the corresponding signature and patients were stratified into two groups based on the median value of the identified feature.

OS and PFS TCP models

Figure 4.14, 4.15, 4.16, and 4.17 show the forest plots of the OS and PFS CPH RDCS models and the origin of the different features forming the signatures in the rHGG and the NSCLC cohorts. Furthermore, the rs with the prescribed dose and the ROI volume are also reported.

Feature	Modality	ROI	Filter	Param.	Class	Name	rs (volume)	rs (dose)	P-Value	HR
F1	Clinical					RRRS	0.12	0.09	<0.01	3.87
F2	Clinical					Grade	0.18	0.05	<0.01	3.1
F3	CT	GTV	Wavelet	LHL	Glszm	SAE	0.51	0.03	<0.01	1.5
F4	ADC	PTV	Log	3mm	Glrlm	LGLRE	-0.44	-0.19	<0.01	0.62
F5	DD	PTV	Wavelet	HLH	Firstorder	Maximum	0.31	0.25	<0.01	1.46
F6	CT	PTV	Wavelet	LLL	Glcm	Idmn	0.21	-0.02	<0.01	1.46
F7	T1wce	CTV	Original		Firstorder	RMS	0.11	0.1	<0.01	1.79
F8	DD	PTV	Wavelet	LHL	Glrlm	SRLGLE	-0.68	-0.2	<0.01	0.68
F9	T1w	PTV	Wavelet	LLL	Glcm	Imc1	-0.39	0.01	0.01	0.78
F10	T2w	CTV	Wavelet	LHL	Glszm	SZNU	0.51	0.15	0.26	1.07

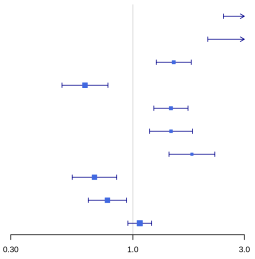


Figure 4.14: Forest plot to show the results of the multivariate CPH OS RDCS model for cohort rHGG with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value is calculated through the likelihood ratio test. RDCS: radiomics dosiomics clinical signature, OS: overall survival

Considering the rHGG cohort, the multimodality rHGG RDCS OS signature includes the two clinical features, i.e., the tumour grade and RRRS, the square-root of the mean (RMS) of all T1wce squared intensity values in the CTV, two PTV texture features from the DD, and seven texture radiomics features, specifically one from the modality-ROI T1w-PTV, one T2w-CTV, one ADC-PTV, one CT-GTV, and one CT-PTV. As for the multimodality rHGG

RDCS PFS signature, it includes the two clinical features, i.e., the tumour grade and RRRS, one PTV texture feature from the DD, and eight texture radiomics features, specifically one from the modality-ROI, one T1wce-GTV, one T1w-PTV, one SWI-CTV, one ADC-CTV, one ADC-PTV, one CT-GTV, and one CT-CTV. Other than SWI in the RDCS OS models, all modalities contributed to the signatures. The RRRS has the highest hazard ratio in the OS model (HR: 3.87), while the tumour grade in the PFS model (HR:2.03). X-means yielded two clusters from the radiomics and dosiomics features in the OS RDCS model and the radiomics features from the PFS RDCS model (Figure D.1). Since a single dosiomics feature is identified in the PFS RDCS model, patient stratification into two subgroups was performed using the dosiomics feature median value.

Feature	Modality	ROI	Filter	Param.	Class	Name	rs (volume)	rs (dose)	P-Value	HR
F1	Clinical					RRRS	0.12	0.09	<0.01	1.42
F2	Clinical					Grade	0.18	0.05	<0.01	2.03
F3	ADC	CTV	Log	2mm	Firstorder	Skewness	-0.18	-0.02	<0.01	0.73
F4	SWI	CTV	Wavelet	LHL	Glszm	ZoneEntropy	0.15	-0.12	<0.01	0.69
F5	T1w	PTV	Wavelet	HLL	Glrlm	LRLGLE	0.42	0.11	<0.01	1.5
F6	DD	PTV	Wavelet	HHL	Firstorder	Median	0.26	0.2	<0.01	1.29
F7	CT	GTV	Wavelet	LHL	Firstorder	Skewness	0.05	0.07	<0.01	1.34
F8	T1wce	GTV	Logarithm		Firstorder	RMS	-0.03	0.13	<0.01	1.38
F9	CT	CTV	Wavelet	HHL	Glszm	SZNUN	-0.12	-0.03	<0.01	0.78
F10	ADC	PTV	Wavelet	LHL	Glrlm	SRLGLE	-0.29	-0.06	<0.01	0.69

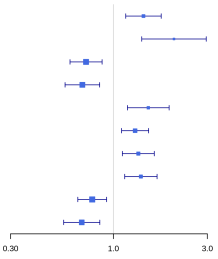


Figure 4.15: Forest plot to show the results of the multivariate CPH PFS RDCS model for cohort rHGG with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value is calculated through the likelihood ratio test. RDCS: radiomics dosiomics clinical signature, PFS: progression-free survival

As for the NSCLC cohort, the RDCS OS signature includes the tumour location (central versus peripheral), one texture feature from the DD in the ipsilateral lung, and two texture radiomics, specifically one from the SH-CT-GTV and one from the SH-CT-heart. X-means yielded two clusters from the radiomics features (Figure D.2). Since a single dosiomics feature is identified, patient stratification into two subgroups was performed using the dosiomics feature median value.

The NSCLC RDCS PFS signature includes the tumour location (central versus peripheral), a texture feature from the DD and the ipsilateral lung, and a radiomics texture feature from the SH-CT-GTV. The tumour location has the highest HR for both the OS (4.37) and PFS (2.51) models. As single radiomics and dosiomics features were derived for all three models, X-means was not performed, and the median feature value was used for clustering.

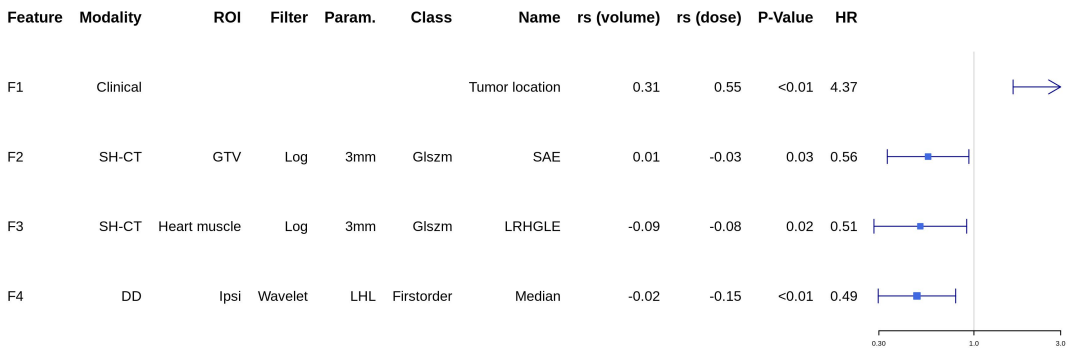


Figure 4.16: Forest plot to show the results of the multivariate CPH OS RDCS model for cohort NSCLC with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value calculated through the likelihood ratio test

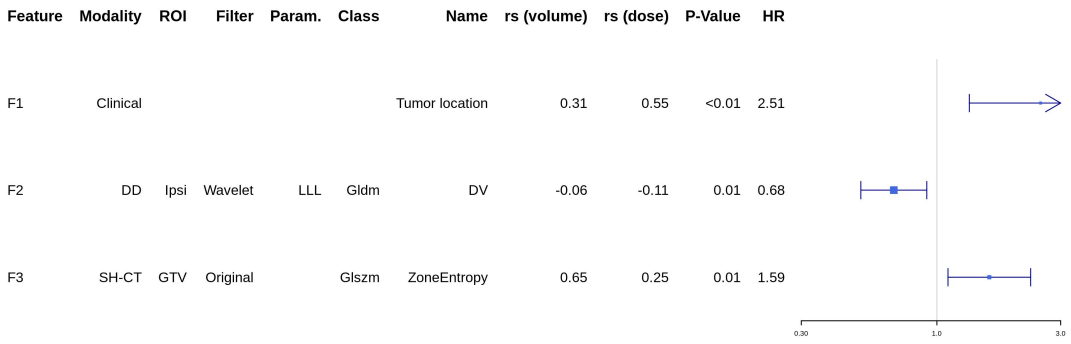


Figure 4.17: Forest plot to show the results of the multivariate CPH PFS RDCS model for cohort NSCLC with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value calculated through the likelihood ratio test

RILF and XT NTCP models

Figure 4.18, and 4.19 show the forest plots of the RILF and XT CPH RDCS models and the origin of the different features forming the signatures. Furthermore, the rs with the prescribed dose and the ROI volume are also reported.

Feature	Modality	ROI	Filter	Param.	Class	Name	rs (volume)	rs (dose)	P-Value	HR
F1	Clinical					ICD-O C34.2	-0.03	-0.01	0.66	1.59
F2	DD	PTV	Wavelet	HLH	Glszm	SAE	0.14	0.08	0.02	0.61
F3	SH-CT	Contra	Log	4mm	Glszm	SAHGLE	-0.13	-0.06	0.02	0.57
F4	SH-CT	GTV	Wavelet	HLH	Glszm	SZNUN	0.13	0.08	0.02	0.61
F5	SM-CT	PTV	Wavelet	LHH	Glszm	SZNUN	0.03	0.18	0.13	1.41

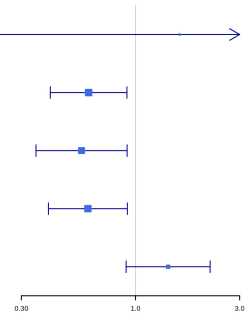


Figure 4.18: Forest plot to show the results of the multivariate CPH RILF RDCS models for cohort NSCLC with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value calculated through the likelihood ratio test

Considering the NSCLC cohort, the RDCS RILF signature includes the tumour site, one texture feature from the DD and the PTV, and three texture radiomics features, specifically one from the SH-CT and the contralateral lung, one SH-CT-GTV, and one SM-CT-PTV. The radiomics texture feature, the GLsZM-DE calculated from the CT wavelet transformation (LHH) and the PTV, had the highest significant HR (1.41). X-means yielded two clusters from the radiomics features (Figure D.2). Since a single dosiomics feature is identified, patient stratification into two subgroups was performed using the dosiomics feature median value.

As for the HNC cohort, the RDCS signature includes three dosiomics texture features (one from the PTV, one from the ipsilateral, and one from the contralateral parotid gland) and four radiomics texture features (one from the PTV, one from the ipsilateral, and two from the contralateral parotid gland). The dosiomics texture feature, specifically the GLDM-DE calculated from the wavelet transformation (LHL) of the DD and the contra-lung, had the highest significant HR (2.16). X-means yielded two clusters from the radiomics and dosiomics features in the XT RDCS models (Figure D.3).

Feature	Modality	ROI	Filter	Param.	Class	Name	rs (volume)	rs (dose)	P-Value	HR
F1	CT	Contra	Wavelet	HHL	Glszm	LRHGLE	-0.03	0.04	<0.01	1.61
F2	DD	PTV	Wavelet	HHH	Firstorder	Mean	0.02	-0.15	<0.01	1.45
F3	CT	Ipsi	Wavelet	LHL	Glszm	ZoneEntropy	0.15	-0.12	0.01	0.68
F4	CT	Contra	Log	3mm	Glrlm	LGLZE	0.01	0.07	0.02	0.58
F5	DD	Contra	Wavelet	LHL	Gldm	DE	0.13	0.17	0.01	2.16
F6	DD	Ipsi	Wavelet	LHH	Glcm	ClusterShade	0.03	0	<0.01	0.76
F7	CT	PTV	Wavelet	LHL	Firstorder	Median	0.05	-0.14	0.02	0.67

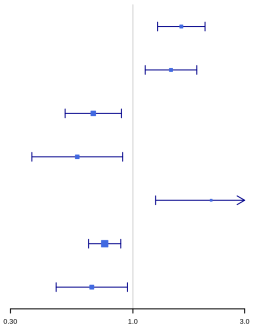


Figure 4.19: Forest plot to show the results of the multivariate CPH XT RDCS model for cohort HNC with the origin summary of the corresponding features. The forest plot of the hazard ratio of each feature in the RDCS models are derived from the discovery set. P-value calculated through the likelihood ratio test

4.5 KAPLAN-MEIER ANALYSIS: RISK STRATIFICATION AND PROGNOSIS COMPARISON

Curves for survival and toxicity were described using Kaplan-Meier analysis, with the aim of assessing and discovering prognostic subgroups using the multimodality signature. Effects of the different modalities' significant features on the prediction and prognostic separation were also evaluated.

In this section, Kaplan-Meier (KM) analysis was used to stratify patients based on the risk of events by associating the survival and radiation-induced toxicity information with the clusters derived by the multimodality signature as well as the individually-identified features from each considered modality. Significant stratification was concluded when a P-value < 0.05 (log-rank test) was obtained by the KM analysis.

OS and PFS TCP models

The KM plots for the rHGG and NSCLC OS models are shown in this section. The remaining plots are summarized in Appendix E. Figure 4.21, E.1, 4.20, and E.6, show that the KM analysis of the RDCS-derived clusters significantly discriminated the patients into high and low risk of survival and progression in both the discovery and test sets (P-value<0.05) in all TCP models considered.

For cohorts rHGG, Figures 4.21-B, C, and E.1-B, C show significant stratification between the radiomics (R0, R1) and dosiomics (D0, D1) clusters (Figure D.1). When all 4 clusters were analyzed, it was observed that the cluster combination of D0-R0 and D1-R0 showed a similar risk for OS (Figure 4.21-D) and thus was combined into one group for further feature effect analysis (Figure 4.21-E). KM analysis of the three clustering combination groups, i.e., D0-R0 & D1-R0, D0-R1, and D1-R1, showed significant stratification (Figure 4.21-E). As for the PFS model, all clustering combinations except for the D1-R1 group showed similar risk (Figure E.1-D), where KM analysis of the newly derived groups showed significant stratification (Figure E.1-E). Figure E.2 (OS) and E.3 (PFS) show the KM curves of the combination clusters when patients were subsetted based on the RRRS (Figures E.2-A, and E.3-A: good; E.2-B, and E.3-B: intermediate; E.2-C, and E.3-C: poor) and the tumour histology (grade III, Figure E.2-D, and E.3-D: grade III; E.2-E, and E.3-E: grade IV, GBM). Finally, KM analysis on patient subset based on validated biomarkers, i.e., MGMT methylation, IDH1/2 mutation, and p/19q codeletion, were analyzed (Figure E.4, and E.5).

As for the NSCLC cohort, Figures 4.20-B, C, and E.6-B, C show that non-significant patient stratification was observed between the radiomics (R0, R1)

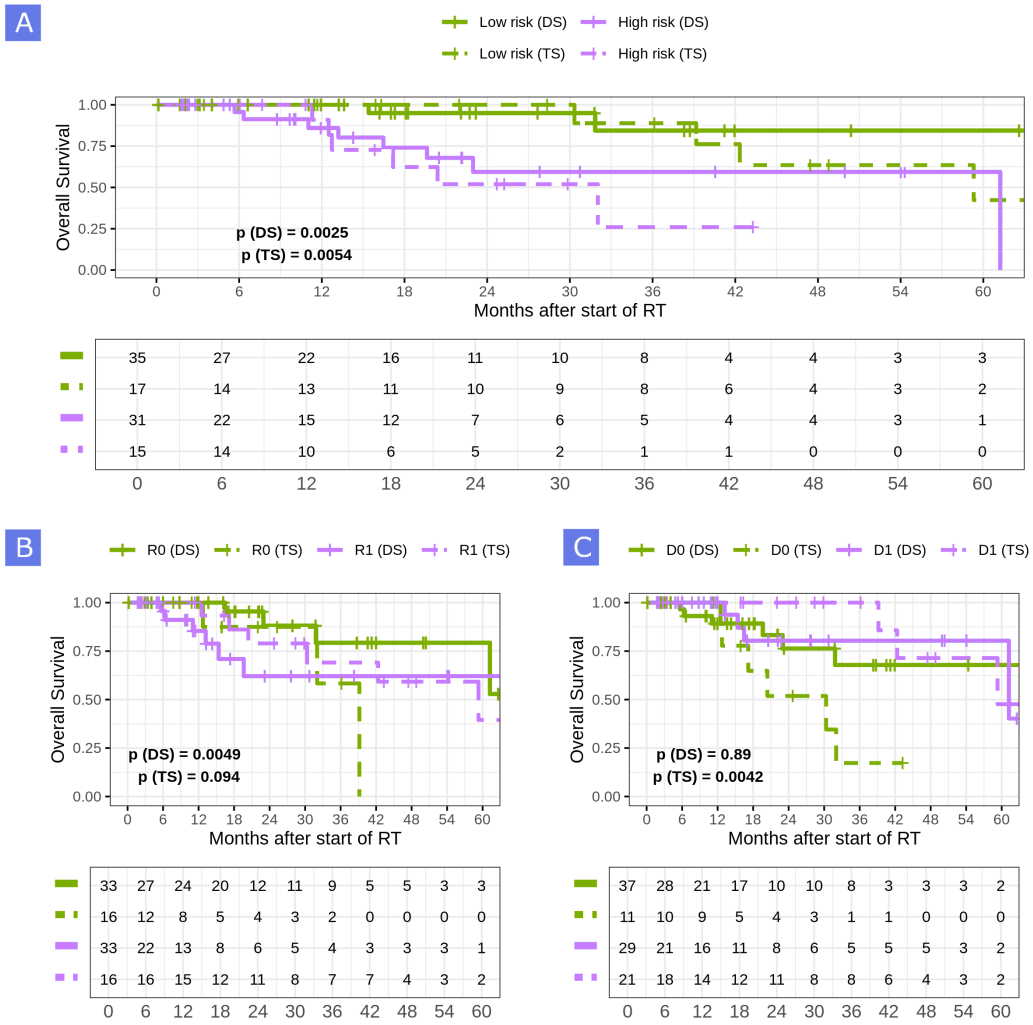


Figure 4.20: Kaplan-Meier OS curves for the NSCLC cohort. A) RDCS clusters, B) RS clusters, and C) DS clusters. Non-significant patient stratification was observed between the unimodality clusters, i.e. only the multimodality RDCS-derived clusters showed different prognoses

and dosiomics (Do, D1) clusters (Figure D.2), i.e. only the multimodality RDCS-derived clusters showed different prognosis.

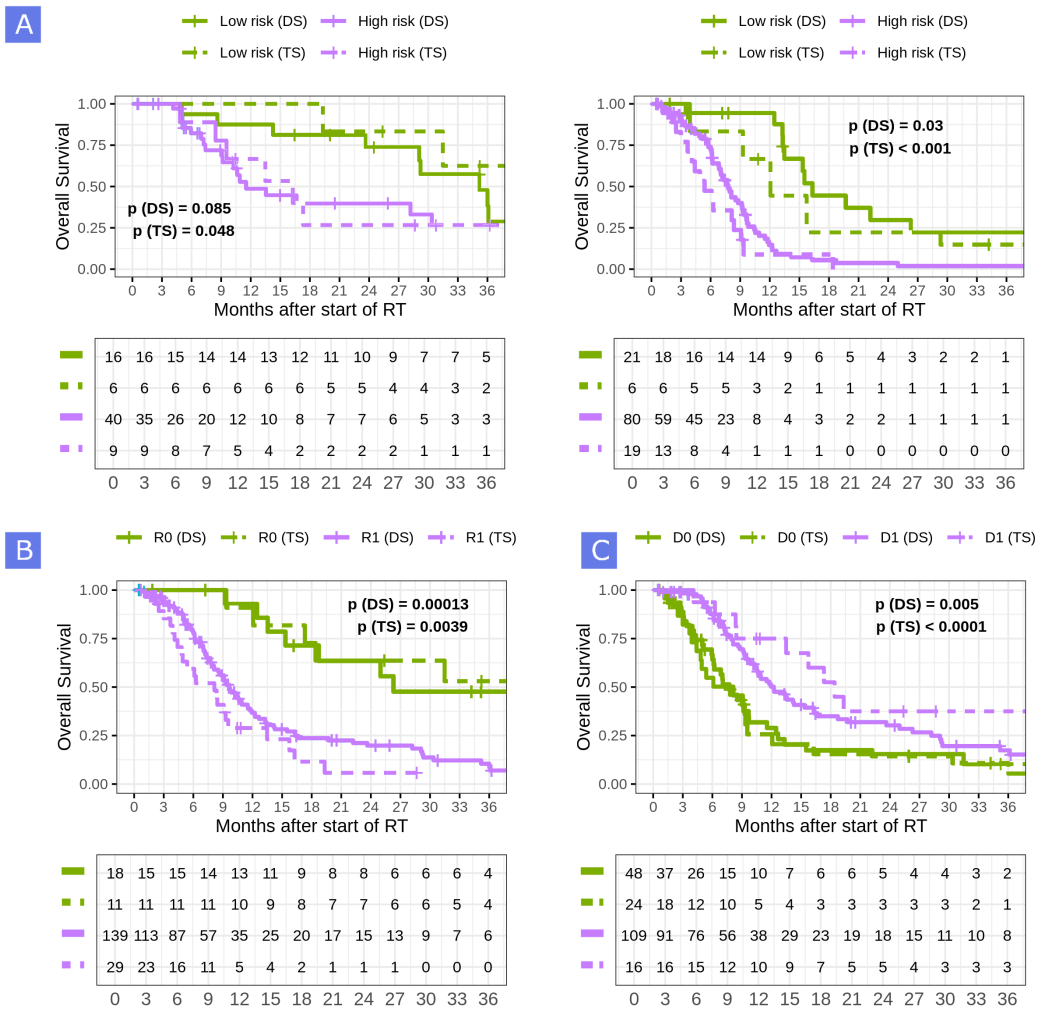


Figure 4.21: Kaplan-Meier overall survival curves for the rHGG cohort of the A) RDCS clusters for grade III (left) and IV (right) patients, B) RS clusters and C) DS cluster. Significant stratification into two risks groups was observed in all clusters considered in both the discovery set and the test set

RILF and XT NTCP models

The KM analysis of the NSCLC and HNC RILF, and XT models are shown in this section. Figure 4.23, and 4.22 show that the KM analysis of the RDCS-derived clusters significantly discriminated the patients into high and low risk of radiation-induced toxicity in both the discovery and test sets in both NTCP models considered. For cohort HNC, X-means yielded two clusters from both the radiomics (R0, R1) and dosiomics (D0, D1) features in the XT

RDCS models, and patients were assigned to the respective clusters (Figure D.3). Figure 4.22-B, C show significant stratification between the clusters. All 4 cluster combinations showed different risk predictions for XT and thus were treated separately.

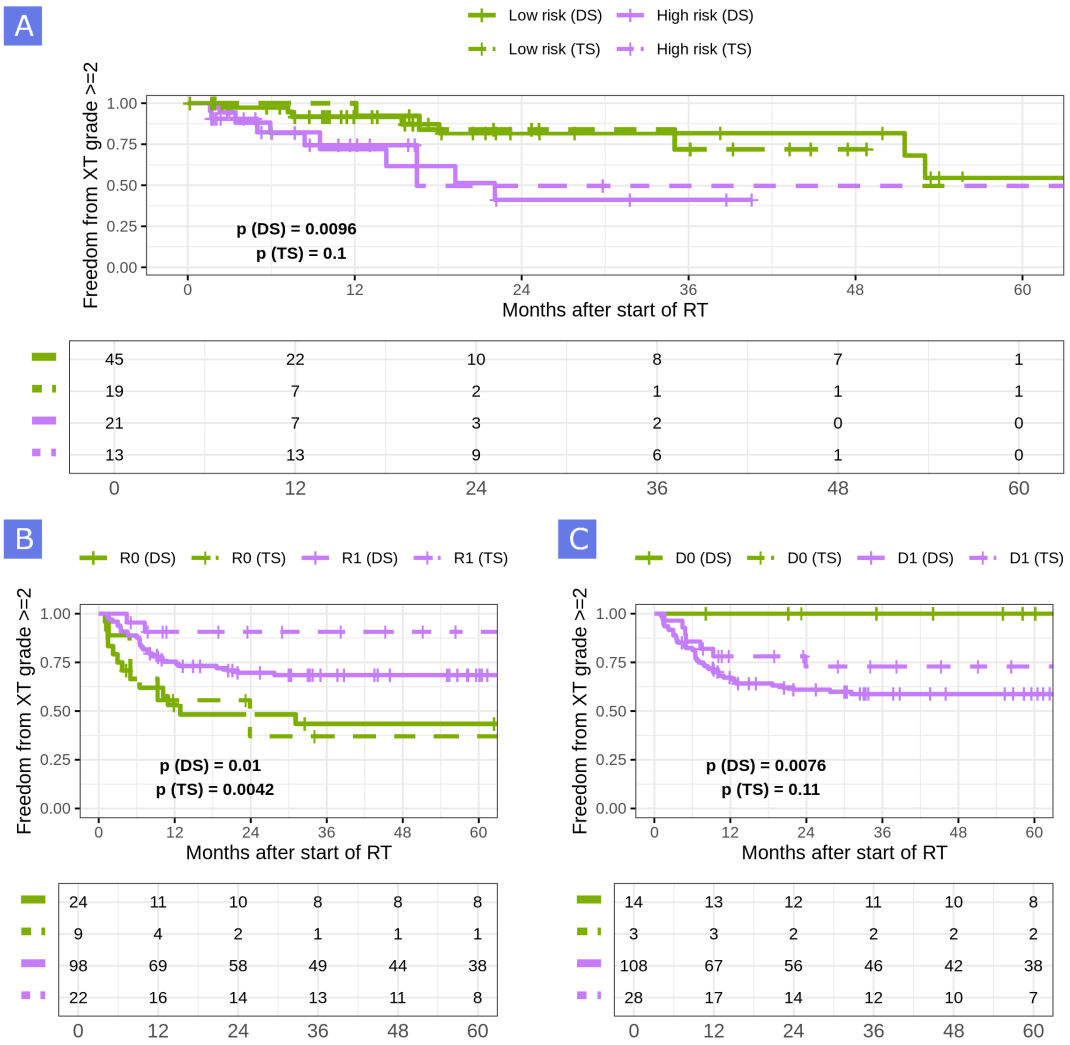


Figure 4.22: Kaplan-Meier XT curves for the HNC cohort of the A) RDCS clusters, B) RS cluster, and C) DS cluster

As for the NSCLC cohort, Figure 4.23 show significant patient stratification between the dosiomics clusters (D0, D1) and non-significant patient stratification between the radiomics clusters (R0, R1) (Figure D.2).

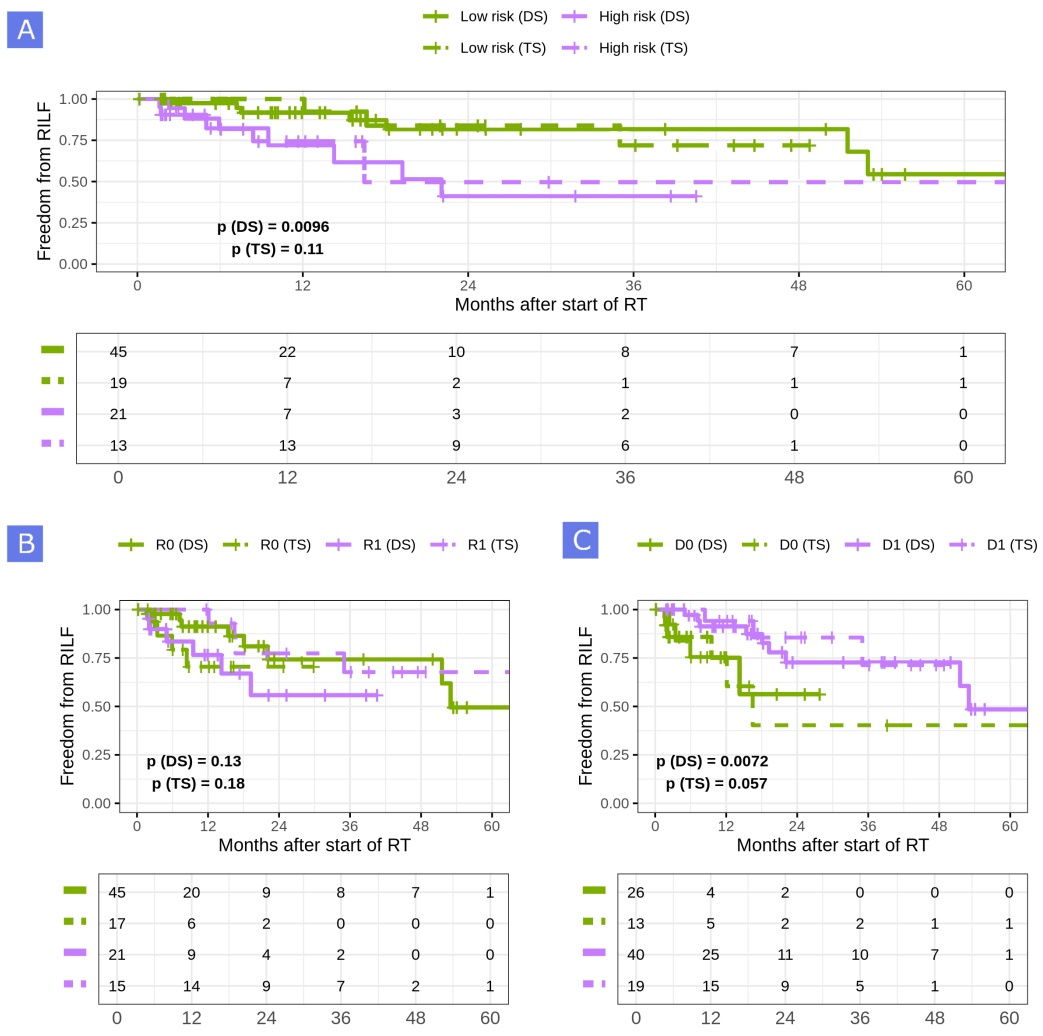


Figure 4.23: Kaplan-Meier RILF curves for the NSCLC cohort of the A) RDCS clusters, B) RS cluster and C) DS cluster

4.6 INTERPRETATION OF DOSIOMICS FEATURES: CORRELATION WITH DRBE AND LET

While textural dosiomics features have demonstrated good predictions, the interpretation of these features is still challenging. Therefore, correlations test with first-order DRBE and LET statistics and shape feature were performed to determine understandable features.

The different dosiomics features have demonstrated good prediction abilities, however, a challenge still occurs in explaining these features since they are not clearly understood as e.g. DVH and dose metrics, nor easily visually interpreted. With the aim of getting closer to understanding them, the DD after wavelet transformation was examined for both the high and low-risk groups. Image representation examples of the DD after wavelet transformations based on the wavelet decomposition that yielded the significant dosiomics features, subsetted based on tumour histology and patient risk group in the rHGG cohort, are shown in Figure 4.24. A clear difference is observed between high and low-risk groups' wavelet transformations of the dose distributions.

To identify a possible explanation as to why these specific features have shown to be predictive, correlations test (Spearman correlation, coefficients rs >0.80) with first order DRBE and LET statistics and shape feature were performed. The shape features, LET and DRBE first-order statistics, and DVH points that highly correlate with the significant dosiomics features are summarised in Tables 4.15, 4.14 and 4.13. Dose-averaged LET was derived using FRoG (Mein et al., 2018).

For the HNC cohort (Table 4.13), one feature of the wavelet decomposition - the Skewness (PTV) - was found to correlate with the significant dosiomics features F2, two features - gradient (ant-post)(ipsi-lung) and variance (ipsi-lung) - were found to correlate with the F5 feature and two features - gradient (right-left) and volume (contra-lung) were found to correlate with the F6 feature. F2, F5 and F6 features are presented in Figure 4.19.

For the NSCLC cohort (Table 4.14), two features - mean (ipsi-lung) and entropy (ipsi-lung) - were found to correlate with the F4 OS feature (Figure 4.16), one feature - variance (ipsi-lung) - was correlated with the F2 PFS feature (Figure 4.17) and one feature - range (PTV) - was correlated with F2 RILF (Figure 4.18).

In the rHGG cohort (Table 4.15), two features - LET D1 (PTV periphery) and LET variance x (PTV periphery) - were found to correlate with F5 OS (Figure 4.14), two features - compactness (CTV) and DVH skewness (GTV) - were found to correlate with F8 OS (Figure 4.14), and one feature - DRBE kurtosis (GTV) - was correlated with F6 PFS (Figure 4.15).

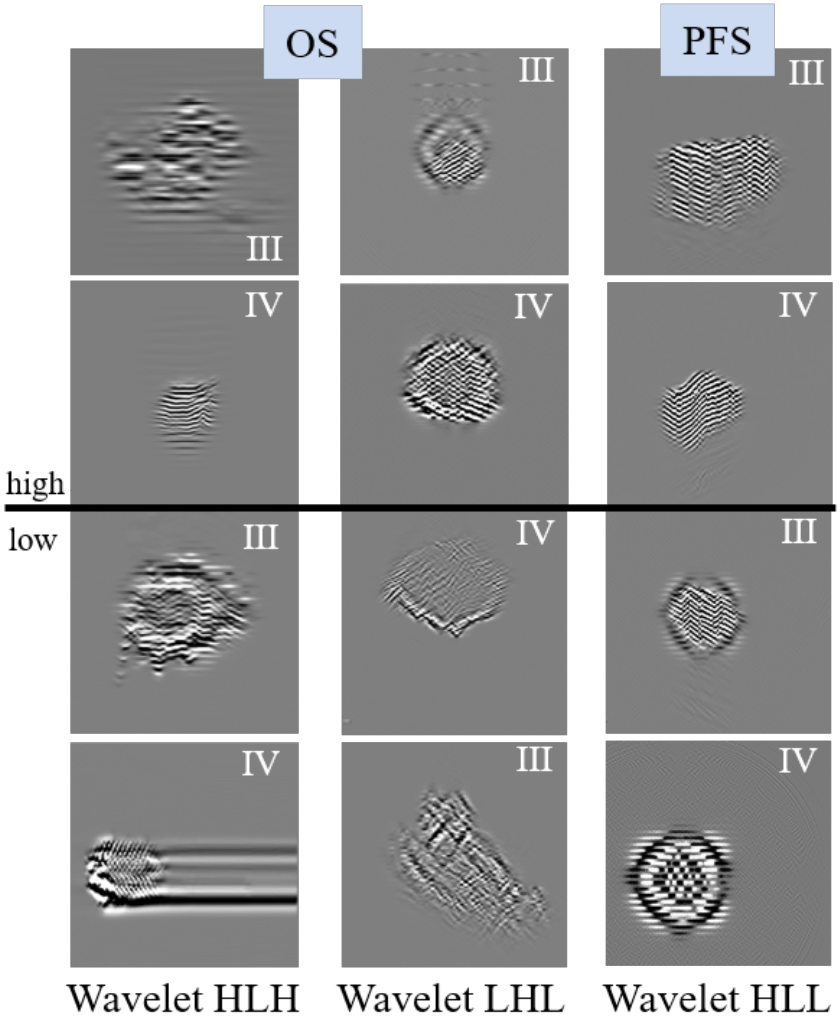


Figure 4.24: Visualization of the dosiomics significant features for the high (upper panel) and low (lower panel) risk group for the OS and PFS model for rHGG

Table 4.13: Shape and first-order statistics dosiomics correlated features with Spearman correlation coefficients rs >0.80 in HNC; ant-post: anterior-posterior; ipsi: ipsilateral; contra: contralateral

Feature	Dosiomics feature	rs
DD Skewness (PTV)	F2-XT	0.90
DD Gradient (ant-post) (Ipsi-lung)	F5-XT	0.85
DD Variance (Ipsi-lung)	F5-XT	0.89
DD Gradient (right-left)	F6-XT	0.91
Volume (Contra-lung)	F6-XT	0.91

Table 4.14: Shape and first-order statistics dosiomics correlated features with Spearman correlation coefficients rs >0.80 in NSCLC; ipsi: ipsilateral

Feature	Dosiomics feature	rs
DD Mean (Ipsi-lung)	F4-OS	0.87
DD Entropy (Ipsi-lung)	F4-OS	0.85
DD Variance (Ipsi-lung)	F2-PFS	0.89
DD Range (PTV)	F2-RILF	0.91

Table 4.15: Shape, DRBE, and LET dosiomics correlated features with Spearman correlation coefficients rs >0.80 in rHGG. Correlation checked for grade III and grade IV separately

Feature	Dosiomics feature	Tumor histology	rs
LET D1 (PTV periphery)	F5-OS	III, IV	0.92
LET variance x (PTV periphery)	F5-OS	III, IV	0.96
Compactness (CTV)	F8-OS	IV	0.95
DVH skewness (GTV)	F8-OS	IV	0.91
DRBE kurtosis (GTV)	F6-PFS	III	0.98

5 DISCUSSION

Currently, approximately half of cancer patients require radiotherapy (RT) throughout the course of disease. However, the main challenge remains, i.e. how can dose be optimally delivered to eradicate tumour cells while sparing normal healthy tissue. Based on cancer histology and organ type, different RT treatment plans are usually prescribed to destroy malignant cells. Noting that radiation-induced cell killing follows a deterministic and stochastic element (Wursthorn et al., 2022), statistical distribution models have been developed to model cell biological and fractionation effects in healthy and tumour cells. Specifically, taking into consideration mainly the dose delivered, fractionation schemes and volume effect, tumour control probability (TCP) models have been designed to measure the success rate of a given RT treatment, while normal tissue complication probability (NTCP) models have been developed to assess the risk of radiation-induced toxicity on normal healthy tissue (Bentzen et al., 2010; Emami et al., 1991). Nevertheless, different prognosis is still observed in patients receiving the same treatment.

Ever since the first therapeutic application of radiation by Dr Leopold Freund in 1896, the dose prescribed to the region of interest has been of homogeneous nature (Freund, 1897). However, as inter- and intra-tumour heterogeneity are present, e.g. arising from different tumour oxygenation status, cell-cycle state and overall tumour biology, large spatial differences in response to the same RT treatment could be observed. Therefore, personalised radiotherapy became a matter of interest where treatments are tailored anatomically to patients, e.g. by using radiomics, which provides valuable information about the inter- and intra-tumour heterogeneity. Therefore, more interest has been seen in data-based NTCP/TCP modelling, where different layers of information can be analysed and integrated into the modelling process and thus might lead to an improved outcome because of the infinite complexity of the underlying biology. Data-based modelling currently mainly makes use of clinical/demographic data, radiomics data, and dosimetric information such as dose metrics and DVH points. A drawback of solely using the latter is the partial loss of spatial information from the 3D dose distribution (DD), which could contain valuable insights. In literature, there has been an exponential increase in recent years in the number of works that focused on applying radiomics to various clinical problems. Dosiomics is still at an early stage but has great potential to be adopted in modelling. Therefore, the first question addressed in this work is the assessment of the improvement of survival and radiation-induced toxicity prediction modelling through spatially analysing DD, i.e. dosiomics. Combined models, i.e. comprising all available informa-

tion from different modalities (clinical, radiomics, and dosiomics), were next built, and their performance was compared to single modality models.

While preparing the different datasets included in this work, DICOM metadata inconsistencies in the exported clinical cohorts were observed, which did not allow the proper curation of the image modalities. To quantify this inconsistency, a computational experiment was performed using the two internal cohorts, pHGG and rHGG, and the public cohort TCGA-GBM, where the MR image DICOM series description (SD) entries were used to classify the MR sequences. The results showed that around 10% discrepancies in each cohort existed when the SD was compared to the manually derived labels. MR image DICOM series description (SD) entries usually represent the MR sequence protocol applied, however, they are sometimes edited by clinical staff or missing. Therefore, going towards a content-based rather than a text-based classification would be beneficial. Since exported clinical cohorts usually include all sorts of available data, the open set recognition problem arises, i.e. if a classifier is trained to distinguish between a lung and a brain image, given either a lung or brain, the classifier should be able to classify the images correctly. However, if an unseen image is inferred to the classifier, e.g., a heart, then the classifier will have to assign the image to one of the known classes. To tackle this issue, a tool was developed where one-vs-all deep convolutional neural networks were employed to automatically learn the image modalities' intrinsic characteristics and enable automatic classification and curation. The tool was tested on multiple MR sequences leading to the development of MR-Class, a python-based tool for brain MR image classification.

Compared to metadata text-based image classification, the use of MR-Class for image classification is independent of inconsistencies between different image sources, not affected by human error, and less labour-intensive. Traditional machine learning techniques such as K-nearest Neighbor (kNN) or support vector machine have also been used in medical image classification; however, the design of a deep learning network, i.e. the use of cascading multiple layers, enables it to learn both simple and complex features thus forming a hierarchical feature representation which is useful when information needs to be extracted from a large amount of data collected from different sources (Zhang, Wang, and Liu, 2018). MR-Class can differentiate between T1w, T1wce, T2w, T2w-FLAIR, ADC, and SWI while handling unknown classes, with classification accuracies of 96.7% [95.8, 97.3, and 94.4% [95% CI: 93.6, 96.1] on two independent cohorts (rHGG, and TCGA-GBM). MR-Class consists of 5 one-vs-all DCNNs (one for each class), followed by a binary classifier for T1w images to determine whether a contrast agent was administered. This design enables MR-Class to handle unknown classes since each DCNN only classifies an image if it belongs to its respective class, and thus an image not labelled by any of the DCNNs is rendered as unknown. The multiclass vs multiple dual-class classification experiment was performed to

compare the performance of such a design to the basic multiclass classification approach. Both methods were observed to have comparable classification results (multiclass: 98.6% multiple one-vs-all: 98.1%) in the context of MR brain image classification. However, the latter can deal with the open-set recognition problem frequently encountered when handling data from clinical cohorts and thus can help reduce data-based driven TCP and NTCP modelling study designs, which include MRI.

Most of the incorrectly classified images by MR-Class had severe blurring or had different types of MR artefacts. These were observed in a higher prevalence in TCGA-GBM than in the rHGG cohort. A reason could be the time interval in which the cohort was collected, as the images from the TCGA cohort were acquired up to 15 years earlier. Most of these classifications were false negatives, i.e. they were labelled as unclassifiable by MR-Class. This can benefit radiomics modelling since any corrupted image would be automatically disregarded, and all images labelled as a specific class would have similar content. Another subset of the misclassified images showed brain tumour volumes overlapping the ventricles. Statistical analysis was performed between these misclassified images and a subset of the correctly labelled images, confirming altered anatomy as a possible reason for misclassification. More detailed studies are warranted to assess further the impact of surgery on alterations of overall anatomy (i.e. biopsy, partial resection, total resection), as well as on tumours (chemo/radiotherapy), as the latter might, e.g., change the pattern of contrast enhancement.

MR-Class is a helpful, ready-to-use python tool for the data preparation of MR-based research studies in brain MRI. It eliminates the need to manually sort out the images, a tedious task due to large amounts of data and different naming schemes. Furthermore, since MR-Class classifies images based on the content rather than the metadata, any corrupted image would be automatically disregarded, and all images labelled as a specific class would have related content. MR-Class is a useful and time-efficient tool for big data MR radiomics-based studies and has been integrated into PyCURT, a python-based curation tool for radiotherapy (Sforazzini et al., 2020). Future work includes the addition of modalities and sequences to MR-class for different anatomy sites, enabling fast deployment of data-based driven TCP and NTCP modelling studies, an important step in the gradual transition towards precision RT.

Another issue encountered while preparing the cohorts for TCP and NTCP modelling was the absence or corruption of certain DICOM RT data, which occurs due to the deployment of new RT systems or simply the loss of data during transfer. As ROIs were extracted in this work from the RT SS, a deep learning segmentation framework - nnU-Net was trained to learn the segmentations performed based on the institutional guidelines for RT treatment and was applied to the patient data with missing RT SS to complete the dataset.

Noting that around 16% of the RT SS in cohort NSCLC are missing or corrupted, nn-Unet was used to train GTV and heart segmentation DCNNs with the available structures while leaving 10% out for testing. While building the different TCP and NTCP models, it is important to have a dataset as clean and homogeneous as possible, therefore, only patients with non-missing data, i.e. a subset of 106 patients from the NSCLC cohort, were only included in the modelling. However, for training the different segmentation networks, all patients from the NSCLC cohort were used, as the performance of neural networks can continually improve as more data is provided. Noting that both 2D and 3D DCNNs were trained, the best performing DCNN for GTV segmentation is the 3D DCNN with a test dice score of 0.83 [0.77-0.86], while the 2D DCNN with a test dice score of 0.92 [0.88 -0.94] has the best performance for the 2D heart segmentation. Thus, a 2D approach yielded better results for larger ROIs. The most probable reason for the performance differences could be the volume difference between the two ROIs. Early-stage NSCLC GTVs do not provide many 2D training slices due to the tumour's small volume, which might result in the 3D approach yielding better results. As for the heart, since it is a relatively large organ which appears in multiple 2D slices, the network has more data in 2D and thus learns the 2D segmentation better. Similar results have been observed in different studies (Zettler and Mastmeyer, 2021). All patient CTs in the NSCLC cohort with a missing RT SS were inferred into the segmentation networks to automatically segmented the GTV and the heart. Manual checks and corrections of the ROI were next performed. After obtaining all NSCLC patient ROI segmentations, all automatically segmented ROI patient data were only included in the DS, while the TS only included patients with manually segmented structures following institutional guidelines for RT treatment. MR-class and nn-Unet are easy-to-use helpful tools that could facilitate data preparation and aid in dealing with missing and corrupted data in preparation for radiomics studies.

Robust radiomics models often require large amounts of data; therefore, medical images are usually collected from multiple centres, sites, and scanners. This leads to the need to implement different preprocessing methods to standardise all images and remove the effects arising from the different scanners and centres. Specifically in MR images, since MR intensities are acquired in scanner-dependent arbitrary units, scans from different scanners and subjects are not directly comparable, even when the same scanning protocol is implemented. While this intensity variation has no major effects on the clinical diagnosis, it drastically impacts the performance of subsequent MRI preprocessing, such as image registration, segmentation and, consequently, radiomics features calculation (Alam and Rahman, 2018; Collewet, Strzelecki, and Mariette, 2004). Multiple intensity normalisation algorithms have been developed over time; however, even though the image biomarker standardisation initiative (IBSI) has defined a more general standardised radiomics

image processing workflow, no specific guidelines on the proper choice of intensity normalisation methods are currently present (Zwanenburg et al., 2020). Therefore a computational experiment was conducted to determine the impact of MRI intensity normalisation algorithms on MR-based radiomics survival prediction models in the pHGG and rHGG cohorts.

Since multiple MR scanners were found in both cohorts, where some have been withdrawn from clinical practice, phantoms could not be applied to assess the impact of the intensity normalisation methods. Therefore, the hard endpoint OS was used in this study as a robust outcome surrogate. However, to standardise and generate reproducible radiomics models, specific phantoms and radiomics-specific tools need to be designed to quantify the impact of the different scanners, protocols and preprocessing steps since radiomics features are sensitive to all of these factors. Therefore, the documentation of the adopted normalisation approach and all implemented preprocessing steps are necessary to enable the reproducibility of radiomics models. Without proper image protocol harmonisation strategies across different centres and scanners, the field of quantitative image analysis will find little progress in the near future.

Performance assessment of the intensity normalisation method-specific CPH and POI survival prediction models showed an impact on the survival predictions between the different intensity normalisation methods and the different MR sequences. Therefore, it can be concluded that the MR intensity normalisation approach directly impacts the overall power of the radiomics-based MR predictive models. Moreover, considering the variability of the acquired results for the different MR sequences, it can be seen that the intensity normalisation algorithm performance is correlated with the MR sequence and that the problem cannot be simplified to one intensity normalisation method.

Due to these variations and for a better interpretation of the results, a ranking score was developed. The WS method showed promising results in T1wce models as it was ranked first in two independent multi-scanner datasets. Combat and the HM method showed consistent prediction results between the two cohorts for T2w models. These two methods were the top-performing methods for T1w in pHGG, however, only HM achieved high predictions in rHGG and not Combat. This might be due to the higher number of batches and the number of images per batch, as 22% of T1w images in rHGG were missing, making batch effect removal more challenging. As for T2w-FLAIR, the FCM showed favourable results in both cohorts; however, with different mask combinations, including the wm and csf or wm and mode. A narrower intensity range is observed in T2w-FLAIR than the other sequences, as csf signals are attenuated. These results might indicate that a mask-based normalisation approach might be more favourable when dealing with images with narrower intensity ranges. The application of both an image-based and feature-based normalisation method had little impact on the performance of

the CPH and POI models. Exceptions were observed in the dataset where combat was ranked first, i.e. T1w in pHGG and T2w in rHGG.

As CPH models were already part of the radiomics signature building pipeline, POI models were also trained to assess whether model performances were biased to CPH models. Compared to survival analysis using CPH, where time-to-event data is used for modelling, Poisson regression models consider the rate at which an event occurs. Comparably to CPH models, the impact of the intensity normalisation methods was also observed in the POI models. Furthermore, the performance of both models was similarly affected after the elimination of the intensity normalisation impacted significant features. A mean increase in the 10-CV C-I and a decrease in 10-CV MSE of 0.05 and 0.03, respectively, were observed in all four sequences across both cohorts.

This experiment included two independent HGG cohorts collected from a single university hospital, UKHD. However, since the data cohorts included data between 2008 and 2019, 19 different scanners from 3 vendors with a 0.5 to 3.0-Tesla range were identified. Noting that the OS models derived from the non-normalised datasets generally ranked low in both cohorts across the sequences, the application of intensity normalisation has indeed improved the OS prediction in radiomics survival models, demonstrating that the need for intensity normalisation is based on the number of scanners and image protocols identified in the cohorts and not necessarily only the number of centres. However, an exception is seen in the T1w dataset in pHGG. This might be because a high number of images in the T1w dataset from pHGG were reconstructed using NiftyMic (as mostly 2D MR scans were present) and therefore preprocessed before applying the intensity normalisation methods (Ebner et al., 2020).

Differences in the performance of the different IN methods across both cohorts can be due to the differences between tumour entities, or the structure of intratumoral heterogeneity, which differs between pHGG and rHGG, as well as the difference in the treatment of rHGG in comparison to pHGG since the therapy of rHGG is not standardised as for pHGG, i.e. incorporating surgery, postoperative adjuvant RT and adjuvant chemotherapy (Campos et al., 2016). In addition, validated HGG biomarkers, such as MGMT methylation, IDH1/2 mutation, and 1p/19q deletion, can lead to survival prediction differences (Boots-Sprenger et al., 2013).

Different preprocessing methods make it generally hard to seamlessly assess the impact of different normalisation methods. The changes in the radiomics values are as much affected by other preprocessing methods as image discretisation or delineating the region of interest. This suggests that the application of intensity normalisation alone may not be enough. This work aimed to limit the effect of intensity discretisation by applying five different bin counts and reporting the average score. Nevertheless, as demonstrated by several radiomics robustness studies, the discretisation approach affects the

overall performance and reproducibility of the radiomics models (Bologna, Corino, and Mainardi, 2019; Duron et al., 2019; Molina et al., 2016). Nevertheless, similarly to using correlation coefficient heatmaps between the different normalisation methods to determine stable radiomics features, the same can be implemented across different bin counts or widths.

In literature, multiple intensity normalisation methods have been reported in HGG radiomics studies, where all implemented the same method across all MR sequences (Fatania et al., 2022). However, as demonstrated in this work, the performance of different methods varies. The study showed that the variations are big and that if the radiomics model reproducibility is possible, the intensity normalisation method should be reported. Another way is to eliminate features impacted by the different normalisation methods. When unstable features are eliminated, the performance of the individual MR sequence prediction models is reduced, a necessary tradeoff for stable radiomics models. However, combining stable radiomics signatures from multiple MR sequences or modalities might mitigate that reduction and improve survival prediction models.

After completion of the radiomics and dosiomics preprocessing workflow, the second part of this thesis focused on the building of the different data-based driven TCP and NTCP models while analysing the improvement of patient prognostic stratification and survival and radiation-induced toxicity predictions through a multimodality approach that incorporates preRT radiomics features, dosiomics features from DD, and clinical features in three cohorts (rHGG, NSCLC, and HNC) with different entities, i.e. brain, HNC and lung, and outcomes. Across the three cohorts, a C-I increase of 10-20% in the DS across three resampling approaches and the TS in both the CPH and RSF models was observed when radiomics and dosiomics significant features were combined with the CS, i.e. improvements in OS, PFS, RILF, and XT prediction were observed when multiple modalities were integrated into the survival models. Furthermore, the combined signature achieved a high vs low risk significant prognostic separation on the TSs. These results indicate that certain interactions are present between the different modalities and that using a multimodality approach can improve patient risk stratification, leading to better treatment decisions. This agrees with other studies on various entities where combined models have been shown to have higher predictive power (Chopra et al., 2020; Gabryś et al., 2018; Lee et al., 2020; Murakami et al., 2022).

A bottleneck in multimodality prediction studies is that if a single modality is missing for a certain patient, the individual patient data can usually not be included in the modelling. 56, 24, and 12 patients from the DSs of rHGG, NSCLC, and HNC, respectively, had at least one modality missing. Therefore, a multimodality imputation pipeline was applied where the observations missing significant features were imputed using all significant features

derived from the other modalities. Compared to the complete models, i.e. excluding imputed data resulting in fewer observations, a narrower DS CI was observed, indicating that the performance is more analogous across the different resampling approaches. Furthermore, the imputed models have smaller variations between the CPH and RSF TS performance. Therefore, the respective imputed signature was used for further analysis. Another approach could aim to create missing sequences from one another, as in (Conte et al., 2021). Nonetheless, the generalizability of the methods is still an issue.

To assess whether the different unimodality and multimodality signatures are only suitable for the semi-parametric CPH models, non-parametric RSF models were evaluated on the TS. Between all three cohorts, the largest C-I variation was observed in rHGG by the CT OS model (C-I CPH-RSF: 0.11) and the T2w PFS model (C-I CPH-RSF: 0.008), and the T2w-FLAIR PFS model (C-I CPH-RSF: 0.06). All remaining unimodality models had a C-I difference less than or equal to 0.05. Noting that a C-I below 0.5 indicates a very poor model, above 0.6 an average model, and above 0.7 a good model, the RSF T2w-FLAIR PFS model from rHGG had poor prediction performance with a C-I of 0.47 [0.46 0.49]. Performance of the CPH T2w-FLAIR PFS model was similarly poor, with a C-I of 0.53 [0.52 0.54] from rHGG, indicating that the model is slightly better at predicting an outcome than random chance. Similar observations were made in HNC by the CS models for both CPH: 0.55 [0.53 0.58] and RSF: 0.50 [0.48 0.54]. The variation between the CPH-RSF models fitted by the multimodality signatures, i.e. the RS and the RDCS, are very small, with a C-I CPH-RSF < 0.03 for all models. The CPH models were considered for further analysis since they had higher and more stable performance, which might be due to the study sample size since non-parametric models require more data due to the larger number of parameters.

The rHGG CS model comprised of the recurrent tumour histology and the reRT risk score (RRRS), previously reported by (Niyazi et al., 2018), achieved an OS C-I of 0.67 [0.65 0.68] and a PFS C-I of 0.60 [0.58 0.62] across all three resampling approaches on the DS. Similar results were observed on TS, where an OS C-Is of 0.66 [0.66 0.68]/0.64 [0.63 0.65] and PFS C-Is of 0.61 [0.60 0.62]/0.62 [0.61 0.63] were achieved by the CPH/RFS models. This result validates the RRRS score, comprised of the initial tumour histology, clinical performance status, and age, for rHGG patients treated with pHGG2. The NSCLC OS CS model includes the tumour location, with a C-I of 0.70 [0.69 0.72] on the DS and 0.67 [0.66 0.69] on the TS. The NSCLC PFS and RILF CS model include the tumour site with a CI 0.56 [0.55 0.58] and 0.59 [0.56 0.63] on the DS and 0.63 [0.61 0.66] and 0.67 [0.66 0.70] on the TS respectively. The HNC XT CS model includes the tumour site with a C-I of 0.51 [0.50 0.53] on the DS and 0.55 [0.53 0.58] on the TS. All clinical features were included in the RDCS multimodality signatures for rHGG and NSCLC. However, the HNC multimodality signature only includes radiomics and dosiomics features.

All nine rHGG unimodality models, i.e. from T1w, T1wce, T2w, T2w-Flair, ADC maps, SWI, CT, DD, and clinical demographic, had an average performance for OS and PFS prediction on the DS after resampling. Good OS predictions by the T1wce, T1w, and DD models and good PFS prediction by the T1w were observed on the TS. Considering the CPH models, T1w (mean discovery C-I across the three resampling approaches: 0.64 [0.63 0.66], test C-I: 0.73 [0.73 0.74]), T1wce (0.68 [0.66 0.70], 0.73 [0.72 0.73]), and DD (0.69 [0.67 0.70], 0.72 [0.71 0.72]) had the best overall performance for OS while T1w (0.63 [0.61 0.66], 0.70 [0.68 0.70]), T1wce (0.67 [0.64 0.69], 0.66 [0.66 0.67]) and CT (0.64 [0.62 0.66], 0.67 [0.66 0.68]) for PFS. The highest number of OS significant features identified by the feature significance search pipeline were from T1wce (n=4), while the lowest number was from SWI (n=1). As for PFS significant features, the CT signature included the highest number of features (n=5), while T2w-FLAIR and SWI had the lowest (n=1). As for the unimodality models in NSCLC, good OS prediction was achieved by the CS model (0.70 [0.69 0.72], 0.67 [0.66 0.69]). The DD model has good RILF prediction (0.70 [0.68 0.73], 0.71 [0.68 0.73]), which also had the highest number of features, together with the SM-CT signature (n=3). Similar performance was observed for HNC, where the DD model had a good XT prediction (0.70 [0.68 0.73], 0.71 [0.68 0.73]) and the highest number of significant features (n=5). DD showed a better prediction than CT in radiation-induced toxicity prediction, i.e. RILF and XT prediction in NSCLC and HNC. All remaining NSCLC and HNC models had an average performance.

The multimodality rHGG RDCS OS signature includes the two clinical features from the CS, the square root of the mean (RMS) of all T1wce squared intensity values in the CTV, 2 PTV texture features from the DD, and seven texture radiomics features, specifically one from the modality-ROI T1w-PTV, 1 T2w-CTV, 1 ADC-PTV, 1 CT-GTV, and 1 CT-PTV. As for the multimodality rHGG RDCS PFS signature, it includes the two clinical features from the CS, 1 PTV texture feature from the DD, and eight texture radiomics features, specifically one from the modality-ROI, 1 T1wce-GTV, 1 T1w-PTV, 1 SWI-CTV, 1 ADC-CTV, 1 ADC-PTV, 1 CT-GTV, and 1 CT-CTV. Other than SWI in the RDCS OS models, all modalities contributed to the signatures. The NSCLC RDCS OS signature includes the tumour location, one texture feature from the DD in the ipsilateral lung, and two texture radiomics, specifically one from the SH-CT-GTV and one from the SH-CT-heart muscle. The NSCLC RDCS PFS signature includes the tumour location, a texture feature from the DD and the ipsilateral lung, and a radiomics texture feature from the SH-CT-GTV. As for the radiation-induced toxicity signatures, the NSCLC RDCS RILF signature includes the tumour site, one texture feature from the DD and the PTV, and three texture radiomics features, specifically one from the SH-CT and the contralateral lung, 1 SH-CT-GTV, and 1 SM-CT-PTV. Lastly, the HNC RDCS signature includes three dosiomics texture features (1 from the

PTV, one from the ipsilateral, and one from the contralateral parotid gland) and four radiomics texture features (1 from the PTV, one from the ipsilateral, and two from the contralateral parotid gland). There was no significant correlation between the textural features and the prescribed dose and volume, with Spearman's correlation coefficient ranging between -0.68 to 0.51. Texture features are 3D spatial features that describe statistical relationships and interactions between voxels with similar or dissimilar contrast values, which can reflect the structure of intra-tumour heterogeneity. The texture descriptors are based on the Laplacian of Gaussian and wavelet transformations. Variability in these features can be attributed to the tumour response to RT, tumour aggressiveness, and extent of infiltration. The prediction capabilities of radiomics in the prediction of OS, PFS, XT, and RILF seen in this work approved with previous work for lung and HNC outcome prediction modelling following RT (Carbonara et al., 2021; Desideri et al., 2020). As T1wce extracted features were found significant in most analyses performed in this study, it can be recommended to use T1wce MR sequences in OS and PFS studies.

To assess the effect of radiomics and dosiomics features, patients' risk stratification, unsupervised clusters of radiomics, and dosiomics features were calculated, and KM analysis was performed on these clusters. The KM plots show that the radiomics features can generally better stratify patients in the DS and TS than the dosiomics features. However, combined, they provide a higher stratification power. It has been demonstrated that certain clinical features, e.g., the RRRS and tumour histology in rHGG or tumour location and site in NSCLC, are predictors and prognosticators. By performing KM analysis on subgroups of patients based on the significant clinical features, it was observed that combining radiomics and dosiomics features could further stratify patients, suggesting that when used together, they can hold valuable information on the outcomes considered. Furthermore, combined clinical, radiomics, and dosiomics nomograms, i.e. graphical representations of the built models, can be constructed, facilitating an individualised preRT identification of a higher risk of radiation-induced toxicity or lower chance of survival.

Dose-volume histogram (DVH) has been shown to have important insight into outcomes in multiple previous studies. However, the number of dosiomics features is significantly larger and thus leads to better chances of correlations. The dosiomics texture features identified in the three cohorts are features calculated from the DD wavelet transformation that mainly point to the heterogeneity of the dose distribution and the low dose level concentration areas. The wavelet filter can provide low and high-frequency signals in the spatial distributions of the dose, suggesting again that the dose inhomogeneities in target volumes and dose variation in organs at risk play an important role in local recurrences and radiation-induced toxicity outcomes. This agrees with other works finding (Buizza et al., 2021; Liang et al., 2019;

Wu et al., 2020). The different dosiomics features have demonstrated good prediction abilities, however, a challenge still occurs in interpreting these features, as they are not clearly understood as DVH and dose metrics. The process of transforming the different DD into quantitive texture features cannot be accurately described with analytic functions; thus, to identify a possible explanation for why these specific features have shown to be predictive, correlation tests with first-order DRBE and LET statistics and shape features were performed. Correlation of the identified significant dosiomics features with first-order statistics and shape features allowed for the identification of understandable features in contrast to dosiomics features which are usually hard to be observed by humans. This allows for treatment adjustment, e.g., by adjustment of TVs or inclusion of objectives constraints during treatment planning optimisation. However, the dosiomics features still have better prediction capabilities in all three cohorts.

Improvement in prediction accuracy could be performed with the addition of more patient data, more feature candidates, or other omics layers such as genomics - e.g. Bøvelstad, Nygård, and Borgan, 2009 combined genomics and clinical information and found that the combination outperforms genomic-only models - transcriptomics - e.g. Tang et al., 2022 discovered a gene signature associated invasion through transcriptome analysis - or proteomics, as each layer provides different information. Thus, complex interactions between the different layers could be learned, potentially representing information underlying various diseases. All these layers could be integrated into the model to improve prediction accuracy and model robustness.

The following limitations exist in this work. All TVs were segmented following institutional guidelines for RT treatment. Nonetheless, delineation variabilities impact radiomics and dosiomics features and, thus, model performances. However, as automatic tumour segmentation networks become more robust and popular, these inter-observer variabilities will be reduced, thus reducing another layer of uncertainty. Furthermore, external validation was not possible due to a lack of cohort availability, so prospective validation of the proposed models is warranted. However, the main purpose of this work was first to evaluate dosiomics and second, to a combined-modality approach to prediction power. My work proves that these approaches may improve radiation-induced toxicity and survival prediction models in the three cohorts considered. As discussed earlier, another limitation was the impact of different preprocessing steps on the different features, which was assessed partially in this work, however, not fully. An example of dosiomics features would be the dose calculation algorithms' impact and the choice of dose grid or RT modalities.

Compared to classical TCP and NTCP models, multimodality radiomics and dosiomics models could lead to advanced and complicated optimisation problems since objective functions are not as intuitive. Therefore, these mod-

els' behaviours during optimisation also need to be studied. Future work will focus on integrating the multimodality models into the RT chain to be able to make use of these models during treatment planning.

6

SUMMARY

Cancer is today the second leading cause of death worldwide. Surgery, radiotherapy, systemic therapy, e.g. chemotherapy, immunotherapy, or combinations thereof, are cornerstones of current cancer therapy. In the course of their treatment, more than half of cancer patients receive radiotherapy, either for curative or palliative purposes, providing it with a significant role in cancer treatment. Over the years, the delivery of radiotherapy has improved to optimize the radiation dose delivered to the tumour and spare as much healthy tissue as possible. Identifying factors affecting normal tissue complication and tumour control probabilities is highly important for improving treatment planning and, consequently, outcomes. This work incorporated the analysis of dose distribution - dosiomics - with imaging features - radiomics - and clinical information, and integrated the multiple medical information layers to identify relevant features associated with therapy outcome. To this end, an artificial intelligence-based workflow for data curation, preprocessing, and analysis has been developed. To enable accurate modelling, the datasets used must be adequately curated, complete, and structured in a standardized manner. Therefore, the first step was to implement a deep learning-based data curation tool that organizes medical imaging data from retrospectively assembled clinical cohorts into the desired structures, saving time in the data preparation step. The developed tool brought the superior performance of content-based brain MRI sequence classification compared to traditional text-based classification. The approach can be adapted to other image analysis studies with similar classification tasks. During magnetic resonance image preprocessing, an important step is image intensity normalization since magnetic resonance images are measured in arbitrary units and should have similar scales for adequate computer analysis. While attempting to find a suitable normalization method, it was observed that the intensity normalization directly impacted the overall power of the radiomics models. Furthermore, varying performance was observed between the different sequences. Therefore, no one-fits-all method can be advised as the intensity normalization algorithm's performance correlates with the magnetic resonance sequence. Consequently, a methodology was developed that can be employed for any magnetic resonance image dataset that requires intensity normalization, facilitating the method search and highlighting the need to report the applied normalization method in future studies. After completion of the radiomics and dosiomics workflow, the second part of this work focused on the improvement of tumour control and normal tissue complication probability estimations, and patient prognostic stratification through

a multimodality approach that incorporates pre-radiotherapy radiomics features, dosiomics features from dose distributions, clinical features, and treatment outcomes in three cohorts with different entities, i.e., brain, head and neck, and lung. Across the three considered cohorts, i.e., recurrent high-grade glioma, early-stage non-small cell lung cancer, and head and neck cancer, a concordance-index increase of 10-20% was observed for tumour control and normal tissue control probability endpoints when radiomics and dosiomics significant features were combined with the clinical signature, suggesting that multimodality models can lead to advanced and complex treatment response estimations. Dose distribution spatial features showed to be associated with the development of normal tissue complications (xerostomia and fibrosis) and might serve as means to optimize treatment plans. Similarly, the evaluation of tumour control probability endpoints as progression-free survival and overall survival showed associations with different sources of medical information, highlighting the high degree of inter- and intra-tumoral heterogeneity and the need to adapt treatment regimens. The combined signature identified high versus low-risk prognostic groups for endpoints' overall survival, progression-free survival, and radiation-induced toxicity. These results indicate that certain interactions are present between the different modalities and that the integration of multimodal information outperforms the unimodal prognostic separation. Therefore, multimodal prognosticators may improve treatment decision support and highlight the relevance of considering biological, physical and morphological data for patient stratification. As the proposed models showed promising performance for both tumour control and normal tissue complication probability estimations, prospective studies of the proposed multimodality models are warranted. Furthermore, future work could focus on integrating and comparing deep learning with the existing machine learning models developed in this work, as deep learning models can uncover hidden features. For this purpose, a large amount of data is required, therefore, additional studies for collecting new datasets are necessary. In conclusion, this work presented a complete multimodality prediction modelling methodology and could pave the way to integrating complex tumour control and normal tissue complication probability estimations models into the treatment planning chain, bringing personalized radiotherapy one step closer.

7 ZUSAMMENFASSUNG

Krebs ist heute die zweithäufigste Todesursache weltweit. Chirurgischen Eingriff, einer Strahlentherapie, einer systemischen Therapie (z.B. Chemotherapie bzw. Immuntherapie) oder aus deren Kombination sind Eckpfeiler der derzeitigen Krebstherapie. Mehr als die Hälfte der Krebspatienten erhält im Laufe ihrer Behandlung eine Strahlentherapie, entweder zu kurativen oder palliativen Zwecken, wodurch sie eine wichtige Rolle in der Krebsbehandlung spielt. Im Laufe der Jahre wurde die Durchführung der Strahlentherapie verbessert, um die auf den Tumor abgegebene Strahlendosis zu optimieren und möglichst viel gesundes Gewebe zu schonen. Die Ermittlung von Faktoren, die sich auf die Wahrscheinlichkeit von Komplikationen im Normalgewebe und die Tumorkontrolle auswirken, ist für die Verbesserung der Behandlungsplanung und folglich des Behandlungsergebnisses von großer Bedeutung. In dieser Arbeit wurde die Analyse der Dosisverteilung (Dosiomics), mit Bildgebungsmerkmalen (Radiomics) und klinischen Informationen kombiniert. Die medizinischen Informationsschichten wurden integriert, um relevante Merkmale zu identifizieren, welche das Therapieergebnis beeinflussen. Zu diesem Zweck wurde ein auf künstlicher Intelligenz basierender Arbeitsablauf für die Datenkuration, Datenvorverarbeitung und Datenanalyse entwickelt. Um eine genaue Modellierung zu ermöglichen, müssen die verwendeten Datensätze adäquat kuratiert sowie vollständig und standardisiert strukturiert sein. Daher wurde im ersten Schritt ein auf Deep Learning basierendes Tool zur Datenkuratierung implementiert. Dieses organisiert medizinische Bildgebungsdaten aus retrospektiv zusammengestellten klinischen Kohorten in die gewünschten Strukturen und spart somit Zeit bei der Datenaufbereitung. Das entwickelte Tool erbrachte eine überlegene Leistung bei der inhaltsbasierten Klassifizierung von MRT-Sequenzen des Gehirns im Vergleich zur herkömmlichen textbasierten Klassifizierung. Ein Ansatz, welcher auch bei anderen Studien zur Bildanalyse mit Klassifizierungsaufgaben eingesetzt werden kann, die diesem ähnlich sind. Ein wichtiger Schritt bei der Vorverarbeitung von Magnetresonanzbildern ist die Normalisierung der Bildintensität, da Magnetresonanzbilder in willkürlichen Einheiten gemessen werden, jedoch für eine angemessene Computeranalyse ähnliche Skalen haben sollten. Bei der Suche nach einer geeigneten Normalisierungsmethode wurde festgestellt, dass sich die Intensitätsnormalisierung direkt auf die Gesamtleistung der radiometrischen Modelle auswirkt. Zudem wurden zwischen den verschiedenen Sequenzen unterschiedliche Leistungen festgestellt. Daher empfiehlt sich ein differenzierter Ansatz, denn die Leistung des Intensitätsnormalisierungsalgorithmus hängt von der jeweiligen Magnetresonanzse-

quenz ab. Folglich wurde eine Methodik entwickelt, die für jeden beliebigen Magnetresonanzbilddatensatz, der eine Intensitätsnormalisierung erfordert, eingesetzt werden kann. Dies erleichtert die Suche nach einer Methode und unterstreicht die Notwendigkeit, die angewandte Normalisierungsmethode in zukünftigen Studien anzugeben. Nach der Fertigstellung des Radiomics- und Dosiomics-Arbeitsablaufs konzentrierte sich der zweite Teil dieser Arbeit auf die Verbesserung der Schätzung der Tumorkontroll- und Normalgewebskomplikationswahrscheinlichkeit sowie auf die prognostische Stratifizierung von Patienten durch einen multimodalen Ansatz. Dieser beinhaltet Radiomics-Merkmale basierend auf Bilddaten, die vor der Strahlentherapie erhoben wurden, Dosiomics-Merkmale aus der Dosisverteilung selbst, klinische Parameter und Behandlungsergebnisse für drei Kohorten. Die Kohorten beinhalten jeweils das rezidivierende hochgradige Gliom, das nichtkleinzellige Lungenkarzinom im Frühstadium und Kopf-Hals-Tumoren mit den entsprechenden organischen Entitäten, d. h., Gehirn, Lunge sowie Kopf und Hals. In allen drei Kohorten wurde ein Anstieg des Konkordanzindexes von 10-20% für die Wahrscheinlichkeitsendpunkte Tumorkontrolle und Normalgewebskontrolle beobachtet, wenn signifikante radiomische und dosiomische Merkmale mit der klinischen Signatur kombiniert wurden. Dies deutet darauf hin, dass multimodale Modelle zu fortschrittlichen und komplexen Schätzungen des Behandlungsansprechens führen können. Es zeigte sich, dass die räumlichen Merkmale der Dosisverteilung mit der Entwicklung von Komplikationen im Normalgewebe (Xeroistomie und Fibrose) in Zusammenhang stehen und als Mittel zur Optimierung der Behandlungspläne dienen könnten. Auch bei der Bewertung der Tumorkontrollwahrscheinlichkeit (progressionsfreies Überleben und Gesamtüberleben) zeigten sich Zusammenhänge mit die medizinischen Informationsquellen, die das hohe Maß an inter- und intratumoraler Heterogenität und die Notwendigkeit der Anpassung von Behandlungsschemata unterstreichen. Durch die kombinierte Signatur konnten prognostische Gruppen mit hohem bzw. niedrigem Risiko für die Endpunkte Gesamtüberleben, progressionsfreies Überleben und strahleninduzierte Toxizität ermittelt. Diese Ergebnisse deuten darauf hin, dass es gewisse Wechselwirkungen zwischen den verschiedenen Modalitäten gibt und dass die Integration multimodaler Informationen gegenüber der unimodalen prognostischen Trennung überlegen ist. Daher können multimodale Prognostikatoren die Entscheidungsfindung bei der Behandlung verbessern und unterstreichen die Bedeutung der Berücksichtigung biologischer, physikalischer und morphologischer Daten für die Patientenstratifizierung. Da die vorgeschlagenen Modelle eine vielversprechende Leistung sowohl für die Schätzung der Tumorkontrolle als auch der Wahrscheinlichkeit von Komplikationen im Normalgewebe zeigten, sind prospektive Studien zu den vorgeschlagenen multimodalen Modellen angezeigt. Da haben sie das Potential die personalisierte Strahlentherapie weiter voranzubringen.

BIBLIOGRAPHY

Adachi, Takanori, Mitsuhiro Nakamura, Ryo Kakino, Hideaki Hirashima, Hiraku Iramina, Yusuke Tsuruta, Tomohiro Ono, Nobutaka Mukumoto, Yuki Miyabe, Yukinori Matsuo, et al. (2022). **Dosiomic feature comparison between dose-calculation algorithms used for lung stereotactic body radiation therapy**. In: *Radiological Physics and Technology* 15.1, pp. 63–71.

Adachi, Takanori, Mitsuhiro Nakamura, Takashi Shintani, Takamasa Mitsuyoshi, Ryo Kakino, Takashi Ogata, Tomohiro Ono, Hiroaki Tanabe, Masaki Kokubo, Takashi Sakamoto, et al. (2021). **Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy**. In: *Medical Physics* 48.4, pp. 1781–1791.

Alam, Fakhre and Sami Ur Rahman (2018). **Medical image registration: Classification, applications and issues**. In: *JPMI* 32.4, p. 300.

Amaldi, Ugo and Gerhard Kraft (July 2005). **Radiotherapy with beams of carbon ions**. en. In: *Rep. Prog. Phys.* 68.8. Publisher: IOP Publishing, pp. 1861–1882. ISSN: 0034-4885. DOI: 10.1088/0034-4885/68/8/R04. URL: https://doi.org/10.1088/0034-4885/68/8/r04 (visited on 06/23/2022).

Atun, Rifat, David A Jaffray, Michael B Barton, Freddie Bray, Michael Baumann, Bhadrasain Vikram, Timothy P Hanna, Felicia M Knaul, Yolande Lievens, Tracey YM Lui, et al. (2015). **Expanding global access to radiotherapy**. In: *The lancet oncology* 16.10, pp. 1153–1186.

Avants, Brian B, Nick Tustison, Gang Song, et al. (2009). **Advanced normalization tools (ANTS)**. In: *Insight j* 2.365, pp. 1–35.

Ayyachamy, Swarnambiga, Varghese Alex, Mahendra Khened, and Ganapathy Krishnamurthi (2019). **Medical image retrieval using Resnet-18**. In: *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 10954. SPIE, pp. 233–241.

Baştanlar, Yalin and Mustafa Özuysal (2014). **Introduction to machine learning**. In: *miRNomics: MicroRNA biology and computational analysis*, pp. 105–128.

Bentzen, Søren M, Louis S Constine, Joseph O Deasy, Avi Eisbruch, Andrew Jackson, Lawrence B Marks, Randall K Ten Haken, and Ellen D Yorke (2010). **Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues**. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3, S3–S9.

Bezdek, James C, Robert Ehrlich, and William Full (1984). **FCM: The fuzzy c-means clustering algorithm**. In: *Computers & geosciences* 10.2-3, pp. 191–203.

Bhide, SA and CM Nutting (Apr. 2010). **Recent advances in radiotherapy**. In: *BMC Medicine* 8.1, p. 25. ISSN: 1741-7015. DOI: 10.1186/1741-7015-8-25. URL: https://doi.org/10.1186/1741-7015-8-25 (visited on 06/23/2022).

Bologna, Marco, Valentina Corino, and Luca Mainardi (2019). **Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain**. In: *Medical physics* 46.11, pp. 5116–5123.

Boots-Sprenger, Sandra HE, Angelique Sijben, Jos Rijntjes, Bastiaan BJ Tops, Albert J Idema, Andreana L Rivera, Fonnet E Bleeker, Anja M Gijtenbeek, Kristin Diefes, Lindsey Heathcock, et al. (2013). **Significance of complete 1p/19q co-deletion, IDH1 mutation and MGMT promoter methylation in gliomas: use with caution**. In: *Modern Pathology* 26.7, pp. 922–929.

Bortfeld, Thomas and Wolfgang Schlegel (1996). **An analytical approximation of depth-dose distributions for therapeutic proton beams**. In: *Physics in Medicine & Biology* 41.8, p. 1331.

Bøvelstad, Hege M, Ståle Nygård, and Ørnulf Borgan (2009). **Survival prediction from clinico-genomic models-a comparative study**. In: *BMC bioinformatics* 10.1, pp. 1–9.

Breslow, Norman E (1975). **Analysis of survival data under the proportional hazards model**. In: *International Statistical Review/Revue Internationale de Statistique*, pp. 45–57.

Buizza, Giulia, Chiara Paganelli, Emma D'ippolito, Giulia Fontana, Silvia Molinelli, Lorenzo Preda, Giulia Riva, Alberto Iannalfi, Francesca Valvo, Ester Orlandi, et al. (2021). **Radiomics and dosiomics for predicting local control after carbon-ion radiotherapy in skull-base chordoma**. In: *Cancers* 13.2, p. 339.

Burman, C., G. J. Kutcher, B. Emami, and M. Goitein (May 1991). **Fitting of normal tissue tolerance data to an analytic function**. en. In: *International Journal of Radiation Oncology*Biology*Physics*. Three-Dimensional Photon Treatment Planning Report of the Collaborative Working Group on the Evaluation of Treatment Planning for External Photon Beam Radiotherapy 21.1, pp. 123–135. ISSN: 0360-3016. DOI: 10.1016/0360-3016(91)90172-Z. URL: https://www.sciencedirect.com/science/article/pii/036030169190172Z (visited on 06/22/2022).

Cai, Lei, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji (2018). **Deep adversarial learning for multi-modality missing data completion**. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1158–1166.

Campos, B, Lars Rønn Olsen, T Urup, and HS Poulsen (2016). **A comprehensive profile of recurrent glioblastoma**. In: *Oncogene* 35.45, pp. 5819–5825.

Carbonara, Roberta, Pierluigi Bonomo, Alessia Di Rito, Vittorio Didonna, Fabiana Gregucci, Maria Paola Ciliberti, Alessia Surgo, Ilaria Bonaparte, Alba Fiorentino, and Angela

Sardaro (2021). **Investigation of radiation-induced toxicity in head and neck cancer patients through radiomics and machine learning: a systematic review**. In: *Journal of Oncology* 2021.

Chen, Zhou, Mingwu Jin, Yue Deng, Jing-Song Wang, Heng Huang, Xiaohua Deng, and Chun-Ming Huang (2019). **Improvement of a deep learning algorithm for total electron content maps: Image completion**. In: *Journal of Geophysical Research: Space Physics* 124.1, pp. 790–800.

Chopra, N, T Dou, G Sharp, E Sajo, and RH Mak (2020). **A Combined Radiomics-Dosiomics Machine Learning Approach Improves Prediction of Radiation Pneumonitis Compared to DVH Data in Lung Cancer Patients**. In: *International Journal of Radiation Oncology, Biology, Physics* 108.3, e777.

Choubey, Dilip K, Manish Kumar, Vaibhav Shukla, Sudhakar Tripathi, and Vinay Kumar Dhandhania (2020). **Comparative analysis of classification methods with PCA and LDA for diabetes**. In: *Current diabetes reviews* 16.8, pp. 833–850.

Clark, Taane G, Michael J Bradburn, Sharon B Love, and Douglas G Altman (2003). **Survival analysis part I: basic concepts and first analyses**. In: *British journal of cancer* 89.2, pp. 232–238.

Collewet, Guylaine, Michal Strzelecki, and François Mariette (2004). **Influence of MRI acquisition protocols and image intensity normalization methods on texture classification**. In: *Magnetic resonance imaging* 22.1, pp. 81–91.

Combs, Stephanie E, Iris Burkholder, Lutz Edler, Stefan Rieken, Daniel Habermehl, Oliver Jäkel, Thomas Haberer, Renate Haselmann, Andreas Unterberg, Wolfgang Wick, et al. (2010a). **Randomised phase I/II study to evaluate carbon ion radiotherapy versus fractionated stereotactic radiotherapy in patients with recurrent or progressive gliomas: the CINDERELLA trial**. In: *BMC cancer* 10.1, pp. 1–8.

Combs, Stephanie E, Meinhard Kieser, Stefan Rieken, Daniel Habermehl, Oliver Jäkel, Thomas Haberer, Anna Nikoghosyan, Renate Haselmann, Andreas Unterberg, Wolfgang Wick, et al. (2010b). **Randomized phase II study evaluating a carbon ion boost applied after combined radiochemotherapy with temozolomide versus a proton boost after radiochemotherapy with temozolomide in patients with primary glioblastoma: the CLEOPATRA trial**. In: *BMC cancer* 10.1, pp. 1–9.

Conte, Gian Marco, Alexander D Weston, David C Vogelsang, Kenneth A Philbrick, Jason C Cai, Maurizio Barbera, Francesco Sanvito, Daniel H Lachance, Robert B Jenkins, W Oliver Tobin, et al. (2021). **Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model**. In: *Radiology* 299.2, pp. 313–323.

Cox, DR (1979). **A note on the graphical analysis of survival data**. In: *Biometrika* 66.1, pp. 188–190.

Demirer, Mutlu, Sema Candemir, Matthew T Bigelow, M Yu Sarah, Vikash Gupta, Luciano M Prevedello, Richard D White, S Yu Joseph, Rainer Grimmer, Michael Wels, et al. (2019). **A user interface for optimizing radiologist engagement in image data curation for artificial intelligence**. In: *Radiology. Artificial intelligence* 1.6.

Desideri, Isacco, Mauro Loi, Giulio Francolini, Carlotta Becherini, Lorenzo Livi, and Pierluigi Bonomo (2020). **Application of radiomics for the prediction of radiation-induced toxicity in the IMRT era: current state-of-the-art**. In: *Frontiers in oncology* 10, p. 1708.

Detterbeck, Frank C, Daniel J Boffa, Anthony W Kim, and Lynn T Tanoue (2017). **The eighth edition lung cancer stage classification**. In: *Chest* 151.1, pp. 193–203.

Ding, Haoran, Chenzhou Wu, Nailin Liao, Qi Zhan, Weize Sun, Yingzhao Huang, Zhou Jiang, and Yi Li (2021). **Radiomics in oncology: a 10-year bibliometric analysis**. In: *Frontiers in oncology*, p. 3677.

Drozdzal, Michal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal (2016). **The importance of skip connections in biomedical image segmentation**. In: *Deep learning and data labeling for medical applications*. Springer, pp. 179–187.

Duron, Loïc, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Jean-Claude Sadik, Isabelle Thomassin-Naggara, Laure Fournier, and Augustin Lecler (2019). **Gray-level discretization impacts reproducible MRI radiomics texture features**. In: *PLoS One* 14.3, e0213459.

Ebner, Michael, Guotai Wang, Wenqi Li, Michael Aertsen, Premal A Patel, Rosalind Aughwane, Andrew Melbourne, Tom Doel, Steven Dymarkowski, Paolo De Coppi, et al. (2020). **An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI**. In: *NeuroImage* 206, p. 116324.

Ellingson, Benjamin M, Martin Bendszus, Jerrold Boxerman, Daniel Barboriak, Bradley J Erickson, Marion Smits, Sarah J Nelson, Elizabeth Gerstner, Brian Alexander, Gregory Goldmacher, et al. (2015). **Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials**. In: *Neuro-oncology* 17.9, pp. 1188–1198.

Emami, Bahman, J Lyman, A Brown, L Cola, M Goitein, JE Munzenrider, B Shank, LJ Solin, and M Wesson (1991). **Tolerance of normal tissue to therapeutic irradiation**. In: *International Journal of Radiation Oncology* Biology* Physics* 21.1, pp. 109–122.

Fatania, Kavi, Farah Mohamud, Anna Clark, Michael Nix, Susan C Short, James O'Connor, Andrew F Scarsbrook, and Stuart Currie (2022). **Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review**. In: *European Radiology*, pp. 1–12.

Ford, Eric C and Stephanie Terezakis (2010). **How safe is safe? Risk in radiotherapy**. In: *International journal of radiation oncology, biology, physics* 78.2, pp. 321–322.

Freund, Leopold (1897). **Ein mit Rontgen-Strahlen behandelter Fall von Naevus pigmentosus piliferus**. In: *Wien. Med. Wochenschr.* 47, pp. 428–434.

Fröhlich, Holger, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H Maathuis, Yves Moreau, Susan A Murphy, Teresa M Przytycka, et al. (2018). **From hype to reality: data science enabling personalized medicine**. In: *BMC medicine* 16.1, pp. 1–15.

Gabryś, Hubert S, Florian Buettner, Florian Sterzing, Henrik Hauswald, and Mark Bangert (2018). **Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia**. In: *Frontiers in oncology* 8, p. 35.

Gabrys, Hubert (2020). *Machine learning using radiomics and dosiomics for normal tissue complication probability modeling of radiation-induced xerostomia*. eng. Dissertation. Place: Heidelberg. DOI: 10.11588/heidok.00026589. URL: https://archiv.ub.uni-heidelberg.de/volltextserver/26589/ (visited on 06/20/2022).

Glatzer, Markus, Cedric Michael Panje, Charlotta Sirén, Nikola Cihoric, and Paul Martin Putora (2020). **Decision making criteria in oncology**. In: *Oncology* 98.6, pp. 370–378.

Gudivada, Venkat N and Vijay V Raghavan (1995). **Content based image retrieval systems**. In: *Computer* 28.9, pp. 18–22.

Gueld, Mark Oliver, Michael Kohnen, Daniel Keysers, Henning Schubert, Berthold B Wein, Joerg Bredno, and Thomas Martin Lehmann (2002). **Quality of DICOM header information for image categorization**. In: *Medical imaging 2002: PACS and integrated medical information systems: design and evaluation*. Vol. 4685. SPIE, pp. 280–287.

Gulliford, Sarah (2015). **Modelling of Normal Tissue Complication Probabilities (NTCP): Review of Application of Machine Learning in Predicting NTCP**. en. In: *Machine Learning in Radiation Oncology: Theory and Applications*. Ed. by Issam El Naqa, Ruijiang Li, and Martin J. Murphy. Cham: Springer International Publishing, pp. 277–310. ISBN: 978-3-319-18305-3. DOI: 10.1007/978-3-319-18305-3_17. URL: https://doi.org/10.1007/978-3-319-18305-3_17 (visited on 06/22/2022).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). **Deep residual learning for image recognition**. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hervey-Jumper, Shawn L and Mitchel S Berger (2014). **Reoperation for recurrent high-grade glioma: a current perspective of the literature**. In: *Neurosurgery* 75.5, pp. 491–499.

Holthusen, Hermann (1936). **Erfahrungen über die Verträglichkeitsgrenze für Röntgenstrahlen und deren Nutzanwendung zur Verhütung von Schäden**. In: *Strahlentherapie* 57, pp. 254–269.

Isensee, Fabian, Jens Petersen, Simon AA Kohl, Paul F Jäger, and Klaus H Maier-Hein (2019a). **nnu-net: Breaking the spell on successful medical image segmentation**. In: *arXiv preprint arXiv:1904.08128* 1.1-8, p. 2.

Isensee, Fabian, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. (2019b). **Automated brain extraction of multisequence MRI using artificial neural networks**. In: *Human brain mapping* 40.17, pp. 4952–4964.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic (2007). **Adjusting batch effects in microarray expression data using empirical Bayes methods**. In: *Biostatistics* 8.1, pp. 118–127.

Joiner, Michael C and Albert J van der Kogel (2018). **Basic clinical radiobiology**. CRC press.

Jovčevska, Ivana, Nina Kočevar, and Radovan Komel (2013). **Glioma and glioblastoma-how much do we (not) know?** In: *Molecular and clinical oncology* 1.6, pp. 935–941.

Jović, Alan, Karla Brkić, and Nikola Bogunović (2015). **A review of feature selection methods with applications**. In: *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, pp. 1200–1205.

Kaliyadan, Feroze and Vinay Kulkarni (2019). **Types of variables, descriptive statistics, and sample size**. In: *Indian dermatology online journal* 10.1, p. 82.

Källman, Patrick, A Ågren, and Anders Brahme (1992). **Tumour and normal tissue responses to fractionated non-uniform dose delivery**. In: *International journal of radiation biology* 62.2, pp. 249–262.

Kang, John, Russell Schwartz, John Flickinger, and Sushil Beriwal (Dec. 2015). **Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective**. en. In: *International Journal of Radiation Oncology*Biology*Physics* 93.5, pp. 1127–1135. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2015.07.2286. URL: https://www.sciencedirect.com/science/article/pii/S0360301615030783 (visited on 06/22/2022).

Kaplan, Edward L and Paul Meier (1958). **Nonparametric estimation from incomplete observations**. In: *Journal of the American statistical association* 53.282, pp. 457–481.

Knoll, Maximilian, Jennifer Furkel, Jürgen Debus, Amir Abdollahi, André Karch, and Christian Stock (2020). **An R package for an integrated evaluation of statistical approaches to cancer incidence projection**. In: *BMC medical research methodology* 20.1, pp. 1–11.

Kursa, Miron B and Witold R Rudnicki (2010). **Feature selection with the Boruta package**. In: *Journal of statistical software* 36, pp. 1–13.

Lambin, Philippe, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG

Even, Arthur Jochems, et al. (2017). **Radiomics: the bridge between medical imaging and personalized medicine**. In: *Nature reviews Clinical oncology* 14.12, pp. 749–762.

Lambin, Philippe et al. (Mar. 2012). **Radiomics: Extracting more information from medical images using advanced feature analysis**. en. In: *European Journal of Cancer* 48.4, pp. 441–446. ISSN: 0959-8049. DOI: 10.1016/j.ejca.2011.11.036. URL: https://www.sciencedirect.com/science/article/pii/S0959804911009993 (visited on 06/18/2022).

Lang, Michel, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, and Bernd Bischl (2019). **mlr3: A modern object-oriented machine learning framework in R**. In: *Journal of Open Source Software* 4.44, p. 1903.

Laskar, Sarbani Ghosh and Sangeeta Kakoti (2022). **Modern Radiation Oncology: From IMRT to Particle Therapy—Present Status and the Days to Come**. In: *Indian Journal of Medical and Paediatric Oncology* 43.01, pp. 047–051.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). **Deep learning**. In: *Nature* 521.7553, pp. 436–444.

Lee, Kyubum, Maria Livia Famiglietti, Aoife McMahon, Chih-Hsuan Wei, Jacqueline Ann Langdon MacArthur, Sylvain Poux, Lionel Breuza, Alan Bridge, Fiona Cunningham, Ioannis Xenarios, et al. (2018). **Scaling up data curation using deep learning: an application to literature triage in genomic variation resources**. In: *PLoS computational biology* 14.8, e1006390.

Lee, Sang Ho, Peijin Han, Russell K Hales, K Ranh Voong, Kazumasa Noro, Shinya Sugiyama, John W Haller, Todd R McNutt, and Junghoon Lee (2020). **Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy**. In: *Physics in Medicine & Biology* 65.19, p. 195015.

Li, Linchao, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan (2020). **Estimation of missing values in heterogeneous traffic data: application of multimodal deep learning model**. In: *Knowledge-Based Systems* 194, p. 105592.

Liang, Bin, Hui Yan, Yuan Tian, Xinyuan Chen, Lingling Yan, Tao Zhang, Zongmei Zhou, Lvhua Wang, and Jianrong Dai (2019). **Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis**. In: *Frontiers in oncology* 9, p. 269.

Lin, Danyu Y and Lee-Jen Wei (1989). **The robust inference for the Cox proportional hazards model**. In: *Journal of the American statistical Association* 84.408, pp. 1074–1078.

Louis, David N, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvet, Bernd W Scheithauer, and Paul Kleihues (2007). **The 2007 WHO classification of tumours of the central nervous system**. In: *Acta neuropathologica* 114.2, pp. 97–109.

Lyman, John T (1985). **Complication probability as assessed from dose-volume histograms**. In: *Radiation Research* 104.2s, S13–S19.

Maaten, Laurens Van der and Geoffrey Hinton (2008). **Visualizing data using t-SNE.** In: *Journal of machine learning research* 9.11.

Maulud Dastan, Adnan M (2020). **A review on linear regression comprehensive in machine learning**. In: *Journal of Applied Science and Technology Trends* 1.4, pp. 140–147.

Mein, Stewart, Kyungdon Choi, Benedikt Kopp, Thomas Tessonnier, Julia Bauer, Alfredo Ferrari, Thomas Haberer, Jürgen Debus, Amir Abdollahi, and Andrea Mairani (2018). **Fast robust dose calculation on GPU for high-precision 1H, 4He, 12C and 16O ion therapy: the FRoG platform**. In: *Scientific reports* 8.1, pp. 1–12.

Mildenberger, Peter, Marco Eichelberg, and Eric Martin (2002). **Introduction to the DICOM standard**. In: *European radiology* 12.4, pp. 920–927.

Mohan, Radhe, GS Mageras, B Baldwin, LJ Brewster, GJ Kutcher, S Leibel, CM Burman, CC Ling, and Z Fuks (1992). **Clinically relevant optimization of 3-D conformal treatments**. In: *Medical physics* 19.4, pp. 933–944.

Molina, David, Julián Pérez-Beteta, Alicia Martínez-González, Juan Martino, Carlos Velásquez, Estanislao Arana, and Víctor M Pérez-García (2016). **Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images**. In: *Computers in biology and medicine* 78, pp. 49–57.

Molinelli, Silvia et al. (Aug. 2016). **Dose prescription in carbon ion radiotherapy: How to compare two different RBE-weighted dose calculation systems**. en. In: *Radiotherapy and Oncology* 120.2, pp. 307–312. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2016.05.031. URL: https://www.sciencedirect.com/science/article/pii/S0167814016311732 (visited on 06/25/2022).

Munro, TR and CW Gilbert (1961). **The relation between tumour lethal doses and the radiosensitivity of tumour cells**. In: *The British journal of radiology* 34.400, pp. 246–251.

Murakami, Yu, Takashi Soyano, Takuyo Kozuka, Masaru Ushijima, Yuuki Koizumi, Hikaru Miyauchi, Masahiro Kaneko, Masahiro Nakano, Tatsuya Kamima, Takeo Hashimoto, et al. (2022). **Dose-Based Radiomic Analysis (Dosiomics) for Intensity Modulated Radiation Therapy in Patients With Prostate Cancer: Correlation Between Planned Dose Distribution and Biochemical Failure**. In: *International Journal of Radiation Oncology* Biology* Physics* 112.1, pp. 247–259.

Muthukrishnan, R and R Rohini (2016). **LASSO: A feature selection technique in predictive modeling for machine learning**. In: *2016 IEEE international conference on advances in computer applications (ICACA)*. IEEE, pp. 18–20.

Newhauser, Wayne, Timothy Jones, Stuart Swerdloff, Warren Newhauser, Mark Cilia, Robert Carver, Andy Halloran, and Rui Zhang (2014). **Anonymization of DICOM**

electronic medical records for radiation therapy. In: *Computers in biology and medicine* 53, pp. 134–140.

Niyazi, Maximilian, Sebastian Adeberg, David Kaul, Anne-Laure Boulesteix, Nina Bougatf, Daniel F Fleischmann, Arne Grün, Anna Krämer, Claus Rödel, Franziska Eckert, et al. (2018). **Independent validation of a new reirradiation risk score (RRRS) for glioma patients predicting post-recurrence survival: a multicenter DKTK/ROG analysis**. In: *Radiotherapy and Oncology* 127.1, pp. 121–127.

Nutting, C, DP Dearnaley, and S Webb (2000). **Intensity modulated radiation therapy: a clinical review.** In: *The British journal of radiology* 73.869, pp. 459–469.

Nyúl, László G and Jayaram K Udupa (1999). **On standardizing the MR image intensity scale**. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.6, pp. 1072–1081.

Panje, Cédric M, Markus Glatzer, Charlotta Sirén, Ludwig Plasswilm, and Paul M Putora (2018). **Treatment options in oncology**. In: *JCO Clinical Cancer Informatics* 2, pp. 1–10.

Pelleg, Dan, Andrew W Moore, et al. (2000). **X-means: Extending k-means with efficient estimation of the number of clusters.** In: *Icml*. Vol. 1, pp. 727–734.

Pérez-García, Fernando, Rachel Sparks, and Sébastien Ourselin (2021). **TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning**. In: *Computer Methods and Programs in Biomedicine* 208, p. 106236.

Placidi, L, D Cusumano, J Lenkowicz, L Boldrini, and V Valentini (2021a). **On dose cube pixel spacing pre-processing for features extraction stability in dosiomic studies**. In: *Physica Medica* 90, pp. 108–114.

Placidi, Lorenzo et al. (Jan. 2021b). **A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity**. en. In: *Cancers* 13.15. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 3835. ISSN: 2072-6694. DOI: 10.3390/cancers13153835. URL: https://www.mdpi.com/2072-6694/13/15/3835 (visited on 06/18/2022).

Puttanawarut, Chanon, Nat Sirirutbunkajorn, Narisara Tawong, Suphalak Khachonkham, Poompis Pattaranutaporn, and Yodchanan Wongsawat (2022). **Impact of Interfractional Error on Dosiomic Features**. In: *Frontiers in oncology* 12.

Qayyum, Adnan, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid (2017). **Medical image retrieval using deep convolutional neural network**. In: *Neurocomputing* 266, pp. 8–20.

Rackwitz, Tilmann and Jürgen Debus (2019). **Clinical applications of proton and carbon ion therapy**. In: *Seminars in oncology*. Vol. 46. 3. Elsevier, pp. 226–232.

Rai, Robba, Shivani Kumar, Vikneswary Batumalai, Doaa Elwadia, Lucy Ohanessian, Ewa Juresic, Lynette Cassapi, Shalini K Vinod, Lois Holloway, Paul J Keall, et al.

(2017). **The integration of MRI in radiation therapy: collaboration of radiographers and radiation therapists**. In: *Journal of Medical Radiation Sciences* 64.1, pp. 61–68.

Reda, Moataz, Alexander F. Bagley, Husam Y. Zaidan, and Wassana Yantasee (Sept. 2020). **Augmenting the therapeutic window of radiotherapy: A perspective on molecularly targeted therapies and nanomaterials**. In: *Radiother Oncol* 150, pp. 225–235. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2020.06.041. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8493937/ (visited on 06/22/2022).

Reinhold, Jacob C, Blake E Dewey, Aaron Carass, and Jerry L Prince (2019). **Evaluating the impact of intensity normalization on MR image synthesis**. In: *Medical Imaging 2019: Image Processing*. Vol. 10949. SPIE, pp. 890–898.

Remedios, Samuel, Dzung L Pham, John A Butman, and Snehashis Roy (2018). **Classifying magnetic resonance image modalities with convolutional neural networks**. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. SPIE, pp. 558–563.

Reynolds, Douglas A (2009). **Gaussian mixture models.** In: *Encyclopedia of biometrics* 741.659-663.

Rossi, Linda, Rik Bijman, Wilco Schillemans, Shafak Aluwini, Carlo Cavedon, Marnix Witte, Luca Incrocci, and Ben Heijmen (2018). **Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy**. In: *Radiotherapy and Oncology* 129.3, pp. 548–553.

Scapicchio, Camilla, Michela Gabelloni, Andrea Barucci, Dania Cioni, Luca Saba, and Emanuele Neri (2021). **A deep look into radiomics**. In: *La radiologia medica* 126.10, pp. 1296–1311.

Scarpace, L. et al. (2016). *Radiology Data from The Cancer Genome Atlas Glioblastoma Multiforme [TCGA-GBM] collection [Data set]*. https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9. DOI: 10.7937/K9/TCIA.2016.RNYFUYE9.

Scheirer, Walter J, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult (2012). **Toward open set recognition**. In: *IEEE transactions on pattern analysis and machine intelligence* 35.7, pp. 1757–1772.

Schork, Nicholas J (2015). **Personalized medicine: time for one-person trials**. In: *Nature* 520.7549, pp. 609–611.

Seppenwoolde, Yvette, Joos V. Lebesque, Katrien de Jaeger, José S. A. Belderbos, Liesbeth J. Boersma, Cees Schilstra, George T. Henning, James A. Hayman, Mary K. Martel, and Randall K. Ten Haken (Mar. 2003). **Comparing different NTCP models that predict the incidence of radiation pneumonitis**. en. In: *International Journal of Radiation Oncology*Biology*Physics* 55.3, pp. 724–735. ISSN: 0360-3016. DOI: 10.1016/S0360-3016(02)03986-X. URL: https://www.sciencedirect.com/science/article/pii/S036030160203986X (visited on 06/22/2022).

Sforazzini, F, P Salome, A Kudak, M Ulrich, N Bougatf, J Debus, M Knoll, and A Abdollahi (2020). **pyCuRT: An Automated Data Curation Workflow for Radiotherapy Big**

Data Analysis using Pythons' NyPipe. In: *International Journal of Radiation Oncology, Biology, Physics* 108.3, e772.

Sforazzini, Francesco, Patrick Salome, Andreas Kudak, Matthias Ulrich, Laila König, Rolf Warta, Nina Bougatf, Juergen Debus, Christel Herold-Mende, Maximilian Knoll, et al. (2021). *PD-L1-R: A MR based surrogate for PD-L1 expression in Glioblastoma multiforme.*

Shah, Mohak, Yiming Xiao, Nagesh Subbanna, Simon Francis, Douglas L Arnold, D Louis Collins, and Tal Arbel (2011). **Evaluating intensity normalization on MRIs of human brain with multiple sclerosis**. In: *Medical image analysis* 15.2, pp. 267–282.

Shinohara, Russell T, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. (2014). **Statistical normalization techniques for magnetic resonance imaging**. In: *NeuroImage: Clinical* 6, pp. 9–19.

Steck, Harald, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar (2007). **On ranking in survival analysis: Bounds on the concordance index**. In: *Advances in neural information processing systems* 20.

Suit, HD (1965). **Radiation response of C3H mouse mammary carcinoma evaluated in terms of cellular radiation sensitivity**. In: *Cellular radiation biology*, pp. 514–530.

Tang, Zili, Ivana Dokic, Maximilian Knoll, Federica Ciamarone, Christian Schwager, Carmen Klein, Gina Cebulla, Dirk C Hoffmann, Julian Schlegel, Philipp Seidel, et al. (2022). **Radioresistance and transcriptional reprograming of invasive glioblastoma cells**. In: *International Journal of Radiation Oncology* Biology* Physics* 112.2, pp. 499–513.

Teoh, May, CH Clark, K Wood, S Whitaker, and A Nisbet (2011). **Volumetric modulated arc therapy: a review of current literature and clinical use in practice**. In: *The British journal of radiology* 84.1007, pp. 967–996.

Thirumuruganathan, Saravanan, Nan Tang, Mourad Ouzzani, and AnHai Doan (2020). **Data Curation with Deep Learning.** In: *EDBT*, pp. 277–286.

Tucker, Susan L, H Helen Liu, Zhongxing Liao, Xiong Wei, Shulian Wang, Hekun Jin, Ritsuko Komaki, Mary K Martel, and Radhe Mohan (2008). **Analysis of radiation pneumonitis risk using a generalized Lyman model**. In: *International Journal of Radiation Oncology* Biology* Physics* 72.2, pp. 568–574.

Tustison, Nicholas J, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee (2010). **N4ITK: improved N3 bias correction**. In: *IEEE transactions on medical imaging* 29.6, pp. 1310–1320.

Van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). **mice: Multivariate imputation by chained equations in R**. In: *Journal of statistical software* 45, pp. 1–67.

Van Griethuysen, Joost JM, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve

Pieper, and Hugo JWL Aerts (2017). **Computational radiomics system to decode the radiographic phenotype**. In: *Cancer research* 77.21, e104–e107.

Voort, Sebastian R van der, Marion Smits, and Stefan Klein (2021). **DeepDicomSort: an automatic sorting algorithm for brain magnetic resonance imaging data**. In: *Neuroinformatics* 19.1, pp. 159–184.

WHO (2022). *New WHO/IAEA publication provides guidance on radiotherapy equipment to fight cancer*. URL: https://www.who.int/news/item/05-03-2021-new-who-iaea-publication-provides-guidance-on-radiotherapy-equipment-to-fight-cancer (visited on 07/14/2022).

Wallis, Sean (2013). **Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods**. In: *Journal of Quantitative Linguistics* 20.3, pp. 178–208.

Webb, S. (Jan. 1993). **The Physics of Three Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning**. en. Google-Books-ID: 2kKStkqe4UUC. CRC Press. ISBN: 978-1-4200-5036-3.

Wissler, Clark (1905). **The Spearman correlation formula**. In: *Science* 22.558, pp. 309–311.

Wu, Aiqian, Yongbao Li, Mengke Qi, Xingyu Lu, Qiyuan Jia, Futong Guo, Zhenhui Dai, Yuliang Liu, Chaomin Chen, Linghong Zhou, et al. (2020). **Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases**. In: *Oral Oncology* 104, p. 104625.

Wursthorn, Anne, Christian Schwager, Ina Kurth, Claudia Peitzsch, Christel Herold-Mende, Jürgen Debus, Amir Abdollahi, and Ali Nowrouzi (2022). **High-Complexity cellular barcoding and clonal tracing reveals stochastic and deterministic parameters of radiation resistance**. In: *International journal of cancer* 150.4, pp. 663–677.

Zettler, Nico and Andre Mastmeyer (2021). **Comparison of 2D vs. 3D U-Net Organ Segmentation in abdominal 3D CT images**. In: DOI: 10.48550/ARXIV.2107.04062. URL: https://arxiv.org/abs/2107.04062.

Zhang, Lei, Shuai Wang, and Bing Liu (2018). **Deep learning for sentiment analysis: A survey**. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, e1253.

Zhang, Yongyue, Michael Brady, and Stephen Smith (2001). **Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm**. In: *IEEE transactions on medical imaging* 20.1, pp. 45–57.

Zwanenburg, Alex, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. (2020). **The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping**. In: *Radiology* 295.2, p. 328.

A

COHORT DESCRIPTIONS AND SPECIFICATIONS

Table A.1: MR scanner models found in the cohorts

Dataset	Manufacturer	Tesla	Model
pHGG	Siemens	0.35	Open
		1	Allegra, Harmony
		1.5	Aera, Amira, Avanto, Espree, Sonata Symphony, Vision
		3	Prisma fit, Skyra, Trio, TrioTim,Verio
	Philips	1	Panorama
		1.5	Achieva, Ingenia, Intera,
		3	NT
	GE	1.5	Signa, Signa Excite-HDxt
rHGG	Siemens	1	Harmony
		1.5	Avanto, Aera, Espree, Sonata, Symphony
		3	Prisma fit, Skyra, TrioTim,Verio
	Philips	1.5	Achieva, Ingenia, Intera
	GE	1.5	Optima MR450w, Signa HDxt
TCGA-GBM	Siemens	1.5	Avanto, Espree, Sonata, Symphony
		3	Verio, Trio, TrioTrim
	Philips	0.5	T5
		1.5	Achieva, Intera
	GE	1.5	Signa, Signa Excite-HDx-HDxt
	Hitachi	0.3	Airis II

Table A.2: Primary high-grade glioma cohort overview. RT TP: planning RT time-point with MR

	pHGG (n=320)	subset-pHGG (n=141)
Age at RT		
<50	104 (31%)	47 (33%)
50-69	167 (52%)	73 (52%)
$\geqslant$ 70	49 (17%)	21(15%)
Gender		
Male	196 (61%)	86 (61%)
Female	124 (39%)	55 (39%)
Tumor grade		
III	73 (36%)	34(24%)
IV	247 (64%)	65 (46%)
MR sequence at RT TP		
T1wce	309 (100%)	141 (100%)
T1w	285 (94%)	110 (78%)
T2w-FLAIR	272 (85%)	128 (91%)
T2w	205(71%)	112 (80%)

Table A.3: TCGA-GBM cohort overview

	TCGA-GBM (n=256)
Gender	
Male	155 (60%)
Female	101 (40%)
Age	
<50	50 (20%)
50-69	145 (57%)
$\geqslant$ 70	61 (23%)
Tumor grade	
III	0 (0%)
IV	256 (100%)
Radiation therapy	
No	113 (44%)
Yes	143 (64%)

Table A.4: NSCLC with missing information cohort overview

	NSCLC (n=74)
Age at RT	
<60	10 (14%)
60-74	40 (54%)
⩾75	22 (32%)
Gender	
Male	46 (62%)
Female	28 (36%)
Karnofsky performance score (KPS)	
⩾70	51 (69%)
<70	9 (12%)
unknown	14 (19%)
Cigarettes pack/year	
0	40 (54%)
<40	9 (12%)
>=40	15 (20%)
Smoker but unknown	10 (14%)
Tumor site	
upper lobe	40 (54%)
middle lobe	5 (7%)
lower lobe	23 (31%)
unknown	4 (8%)
Total dose [Gy]	
30	14 (19%)
45	31 (42%)
50	2 (3%)
54	3 (4%)
60	24 (32%)

Table A.5: MR image protocols found in the cohorts. TI: Inversion time, TE: Echo time, TR: Repetition time, FA: Flip angle

	Protocol	%	
T1w	Sag T1w 3D MPRAGE- with tra reconstruction	48	TI: 800-1000 ms, TE: 2.27-3.7 ms, TR: 1,740-2200 ms, and FA: 8-15° In-plane resolution: 0.42 x 0.42 – 1 x 1 mm, Slice thickness: 0.9-1.3 mm
	T1w 3D MPRAGE	25	TI: 1100 ms, TE: 2.44-4 ms, TR: 1,680-2,000 ms, and FA: 7-15° In-plane resolution: 0.5 x 0.5- 1x1 mm, Slice thickness: 1.3 mm
	T1w Spin Echo	12	TE: 8-17 ms, TR: 350-744 ms, and FA: 70-90° In-plane resolution: 0.45 x 0.45 – 1.05 x 1.05 mm, Slice thickness: 3-6 mm
	T1w 3D Turbo Field Echo	4	TE: 3.17-4.7 ms, TR: 6.5-8.2 ms, and FA: 8° In-plane resolution: 0.5 x 0.5 – 0.93 x 0.93 mm, Slice thickness: 0.9-2 mm
	T1w 3D FLASH	4	TE: 3.56-.9 ms, TR: 7.3-15 ms, and FA: 10-30° In-plane resolution: 0.93x 0.93 mm, Slice thickness: 1.2 mm
	T1w Turbo Spin Echo	2	TE: 11 ms, TR: 400-439 ms, and FA: 150° In-plane resolution: 0.45 x 0.45 – 0.75 x .78 mm, Slice thickness: 3-6 mm
	T1w FLASH	2	TE: 2.48-4 ms, TR: 220-355 ms, and FA: 70-90° In-plane resolution: 0.4 x 0.4- 0.6 x 0.6 mm, Slice thickness: 4-5 mm
	T1w Fast Field Echo	1	TE: 1.69-2.41 ms, TR: 143-187 ms, and FA: 80° In-plane resolution: 0.36 x 0.36- 0.9 x 0.9 mm, Slice thickness: 5-6 mm
	T1w 3D spoiled gradient echo (SPGR)	1	TE: 3.47 ms, TR: 8.8 ms, and FA: 12° In-plane resolution: 0.47 x 0.47 mm, Slice thickness: 1.2 mm
	T1w 3D SPGR BRAVO	1	TI: 300 ms, TE: 5.2 ms, TR: 12.38 ms, and FA: 20° In-plane resolution: 0.94 x 0.94 mm, Slice thickness: 1.2 mm
T2w	Turbo Spin Echo	86	TE: 64-125 ms, TR: 2488-6680 ms, and FA: 90-180° In-plane resolution: 0.22 x 0.22 – 0.97 x 0.97 mm, Slice thickness: 3-6 mm
	Multiple Spin Echo	7	TE: 10 ms, TR: 3000 ms, and FA: 140° In-plane resolution: 0.88 x 0.88 mm, Slice thickness: 5 mm
	Fast Spin Echo	3	TE: 10 ms, TR: 3000 ms, and FA: 140° In-plane resolution: 0.88 x 0.88 mm, Slice thickness: 5 mm
	Turbo Spin Echo/Propeller	2	TE: 100 ms, TR: 5150-6100 ms, and FA: 160° In-plane resolution: 0.46 x 0.46 mm, Slice thickness: 5.mm
	Turbo Spin Echo/Blade	2	TE: 100 ms, TR: 4000 ms, and FA: 150° In-plane resolution: 0.71 x 0.71 mm, Slice thickness: 5.5 mm
T2w-FLAIR	T2w 3D Fast Flair	69	TI: 1700 ms, TE: 95 ms, TR: 8000 ms, and FA: 90° In-plane resolution: 0.35 x 0.35 mm, Slice thickness: 3 mm
	Turbo Dark Fluid	17	TI: 1950 ms, TE: 110 ms, TR: 9000 ms, and FA: 1500° In-plane resolution: 0.94 x 0.94 mm, Slice thickness: 5 mm
	Tra T2w FLAIR Fs	14	TI: 2500 ms, TE: 135 ms, TR: 10000 ms, and FA: 1800° In-plane resolution: 1.05 x 1.05 mm, Slice thickness: 6 mm
ADC maps	-	100	TE: 55-137 ms, TR: 2440-8100ms, FA: 90° and 180°, In-plane resolution: 0.71 x 0.71 – 1.95 x 1.95 mm, Slice thickness: 4-6 mm. b-value 0 and 1000-1200 s/mm
SWI	SWI images	100	TE: 19.7 ms, TR: 27 ms, and FA: 15° In-plane resolution: 0.71 x 0.71 mm, Slice thickness: 2.5- mm

B RADIOMICS AND DOSIOMICS FEATURES SUMMARY

Table B.1: The identifiable radiomics and dosiomics features for all considered cohorts and endpoints represented as image modality/ROI/image transformation-parameter/feature name

	Feature	Cohort	Endpoint
Radiomics	CT/GTV/Wavelet-LHL/Glszm-SAE	rHGG	OS
	ADC/PTV/Log-3mm/Glrlm-LGLRE		
	CT/PTV/Wavelet-LLL/Glcm-ldmn		
	T1wce/CTV/Original/Firstorder-RMS		
	T1w/PTV/Wavelet-LLL/Glcm-lmc1		
	T2w/CTV/Wavelet-LHL/Glszm-SZNU		
	ADC/CTV/Log-2mm/Firstorder-Skewness		
	SWI/CTV/Wavelet-LHL/Glszm-ZoneEntropy	rHGG	PFS
	T1w/PTV/Wavelet-HLL/Glrlm-LRLGLE		
	CT/GTV/Wavelet-LHL/Firstorder-Skewness		
	T1wce/GTV/Logarithm/Firstorder-RMS		
	CT/CTV/Wavelet-HHL/Glszm-SZNUN		
	ADC/PTV/Wavelet-LHL/Glrlm-SRLGLE		
	SH-CT/GTV/Log-3mm/Glszm-SAE	NSCLC	OS
	SH-CT/Heart/Log-3mm/Glszm-LRHGLE		
	SH-CT/Ipsi/Original/Glszm-ZoneEntropy	NSCLC	PFS
	SH-CT/Contra/Log-4mm/Glszm-SAHGLE		
	SH-CT/GTV/Wavelet-HLH/Glszm-SZNUN	NSCLC	RILF
	SM-CT/PTV/Wavelet-LHH/Glszm-SZNUN		
	CT/Contra/Wavelet-HHL/Glszm-LRHGLE		
	CT/Ipsi Wavelet-LHL/Glszm-ZoneEntropy	HNC	XT
	CT/Contra/Log-3mm/Glrlm-LGLZE		
	CT/PTV/Wavelet-LHL/Firstorder-Median		
	Feature	**Cohort**	**Endpoint**
Dosiomics	DD/PTV/Wavelet-HLH/Firstorder-Maximum	rHGG	OS
	DD/PTV/Wavelet-LHL/Glrlm-SRLGLE		
	DD/PTV/Wavelet-HHL/Firstorder-Median	rHGG	PFS
	DD/Ipsi/wavelet-LHL/Firstorder-Median	NSCLC	OS
	DD/Ipsi/Wavelet-LLL/Gldm-DV	NSCLC	PFS
	DD/PTV/Wavelet-HLH/Glszm-SAE	NSCLC	RILF
	DD/PTV/Wavelet-HHH/Firstorder-Mean		
	DD/Contra/Wavelet-LHL/Gldm-DE	HNC	XT
	DD/Ipsi/Wavelet-LHH/Glcm-ClusterShade		

C

DISCOVERY SET AVERAGE CONCORDANCE INDEX

Table C.1: Summary of the NSCLC average discovery set C-Is across the three different resampling approaches in the prediction of OS and PFS. Confidence interval (CI) represents the minimum and maximum C-I achieved. C-I: Concordance index, SM-CT: smooth-kernel CT, SH-CT: sharp-kernel CT, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature

	OS		**PFS**	
	Complete	Imputed	Complete	Imputed
CS	0.70 [0.69 0.72]		0.56 [0.55 0.58]	
SM-CT	0.68 [0.65 0.71]		0.61 [0.59 0.63]	
SH-CT	0.70 [0.65 0.72]	0.69 [0.68 0.71]	0.67 [0.65 0.68]	0.66 [0.65-0.68]
DD	0.73 [0.69 0.76]	0.66 [0.63 0.68]	0.69 [0.68 0.71]	0.64 [0.62-0.66]
RS	0.70 [0.66 0.73]	0.71 [0.69 0.73]]	0.62 [0.61 0.64]	0.62 [0.60-0.64]
RDCS	0.74 [0.69 0.78]	0.78 [0.76 0.80]	0.72 [0.69 0.74]	0.66 [0.64-0.68]

Table C.2: Summary of the NSCLC average discovery set C-Is across the three different resampling approaches in the prediction of RILF. Confidence interval (CI) represents the minimum and maximum C-I achieved. C-I: Concordance index, SM-CT: smooth-kernel CT, SH-CT: sharp-kernel CT, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature

	RILF	
	Complete	Imputed
CS	0.59 [0.56 0.63]	
SM-CT	0.66 [0.65 0.68]	
SH-CT	0.69 [0.67 0.71]	0.69 [0.68 0.70]
DD	0.68 [0.64 0.72]	0.70 [0.68 0.73]
RS	0.70 [0.68 0.73]	0.70 [0.68 0.73]
RDCS	0.72 [0.69 0.72]	0.75 [0.73 0.77]

Table C.3: Summary of the rHGG average discovery set C-Is across the three different resampling approaches in the prediction of OS and PFS. Confidence interval (CI) represents the minimum and maximum C-I achieved. C-I: Concordance index, CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature

	OS		**PFS**	
	Complete	Imputed	Complete	Imputed
CS	0.67 [0.65 0.68]		0.60 [0.58 0.62]	
CT	0.65 [0.63 0.67]		0.64 [0.62 0.66]	
DD	0.69 [0.67 0.70]		0.63 [0.61 0.66]	
T1wce	0.68 [0.66 0.70]		0.67 [0.64 0.69]	
T1w	0.64 [0.63 0.66]	0.64 [0.63 0.66]	0.65 [0.63 0.67]	0.63 [0.61 0.66]
T2w	0.65 [0.62 0.67]	0.65 [0.63 0.67]	0.62 [0.60 0.65]	0.62 [0.60 0.64]
T2w-FL	0.65 [0.63 0.66]	0.66 [0.64 0.67]	0.57 [0.55 0.59]	0.59 [0.57 0.60]
ADC	0.64 [0.63 0.66]	0.63 [0.61 0.65]	0.64 [0.62 0.66]	0.61 [0.59 0.63]
SWI	0.64 [0.61 0.66]	0.61 [0.59 0.64]	0.63 [0.60 0.66]	0.62 [0.60 0.63]
RS	0.74 [0.70 0.76]	0.73 [0.70 0.75]	0.71 [0.68 0.74]	0.69 [0.67 0.72]
RDCS	0.79 [0.76 0.81]	0.77 [0.75 0.78]	0.74 [0.70 0.77]	0.72 [0.71 0.74]

Table C.4: Summary of the HNC average discovery set C-Is across the three different resampling approaches in the prediction of XT. Confidence interval (CI) represents the minimum and maximum C-I achieved. CS: clinical signature, DD: dosiomics signature, RS: radiomics signature, RDCS: combined radiomics, dosiomics and clinical signature

	XT	
	Complete	Imputed
CS	0.51 [0.50 0.53]	
DD	0.72 [0.70 0.74]	
RS	0.69 [0.67 0.71]	0.70 [0.68 0.72]
RDCS	0.75 [0.73 0.77]	0.76 [0.74 0.78]

D

CLUSTER VISUALIZATION

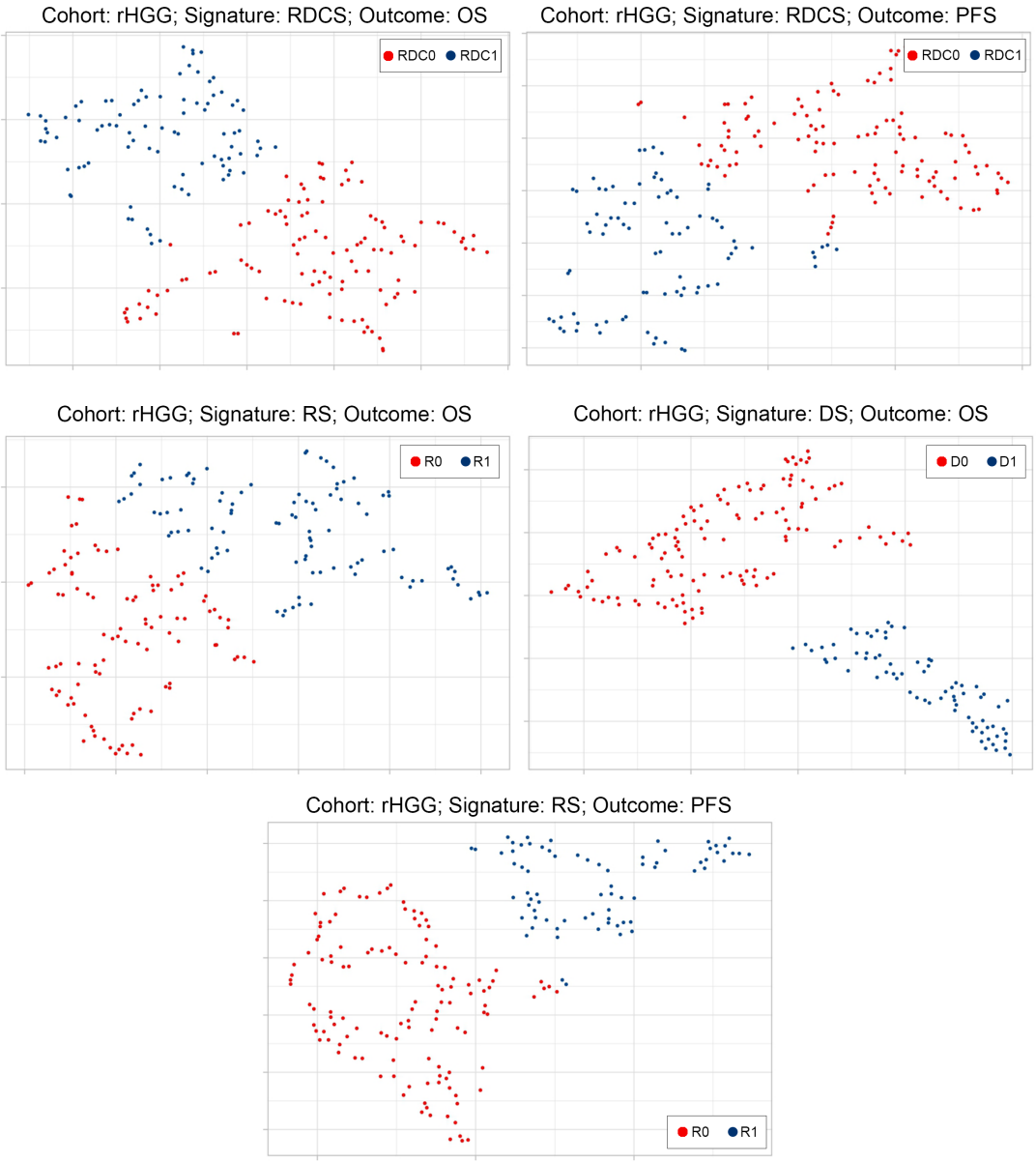


Figure D.1: t-SNE plots of the derived clusters from the radiomics, dosiomics and RDCS signatures in the rhGG cohort

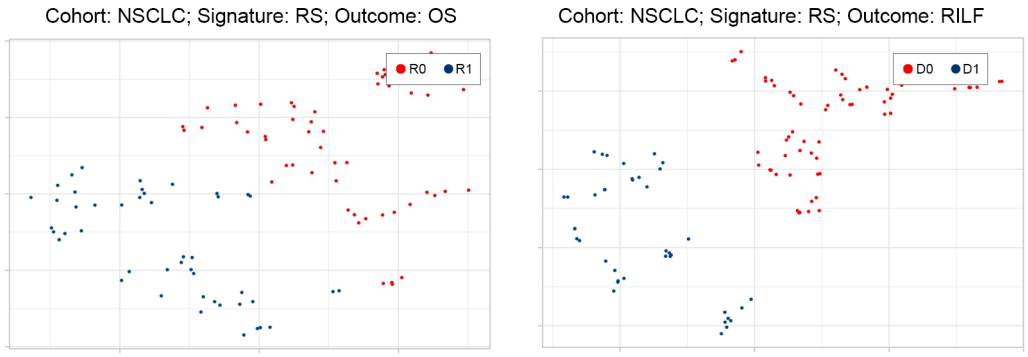


Figure D.2: t-SNE plots of the derived clusters from the radiomics, dosiomics and RDCS signatures in the NSCLC cohort

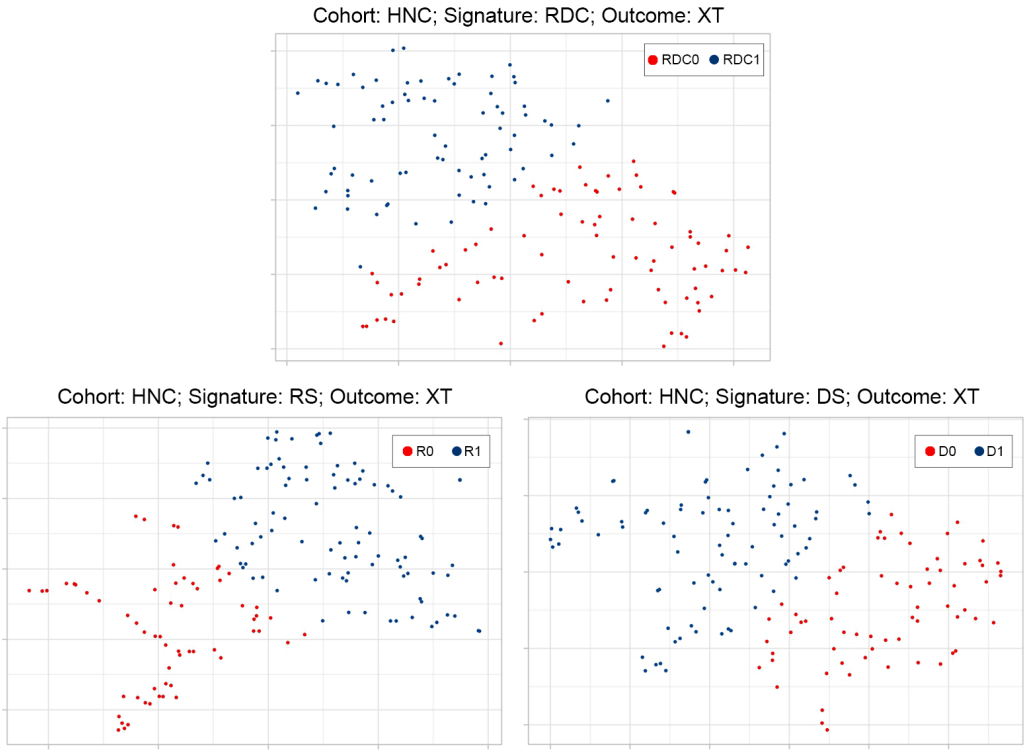


Figure D.3: t-SNE plots of the derived clusters from the radiomics, dosiomics and RDCS signatures in the HNC cohort

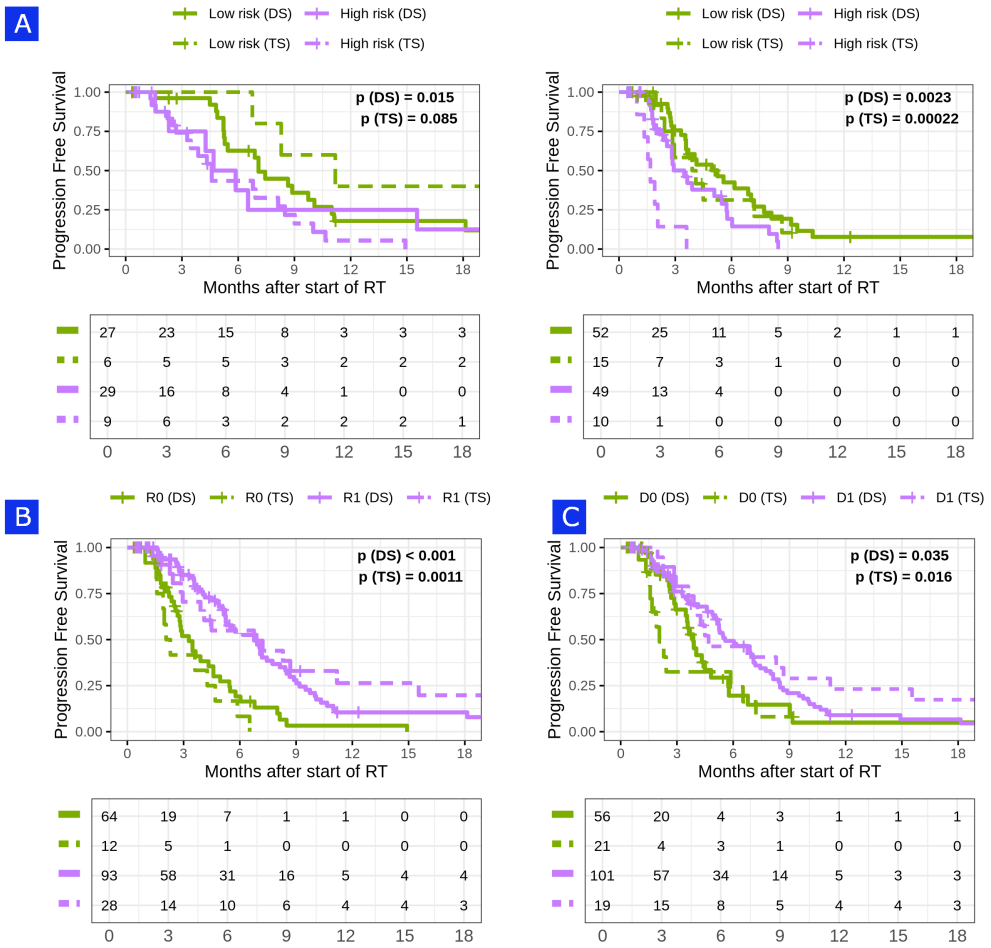


Figure E.1: Kaplan-Meier overall survival curves for rHGG of the A) signature clusters for grade III (left) and IV (right) patients, B) radiomics clusters and C) Dosiomics cluster. Significant stratification into 2 risks groups was observed in all clusters considered in both the discovery set (DS) and the test set (TS)

Figure E.2: Kaplan-Meier overall survival curves for combination of clusters for rHGG that showed similar risk. A) subsetted on RRRS=good; B) subsetted on RRRS=intermediate; C) subsetted on RRRS=poor; D) subsetted on tumor histology: grade III; E) subsetted on tumor histology: grade IV, GBM

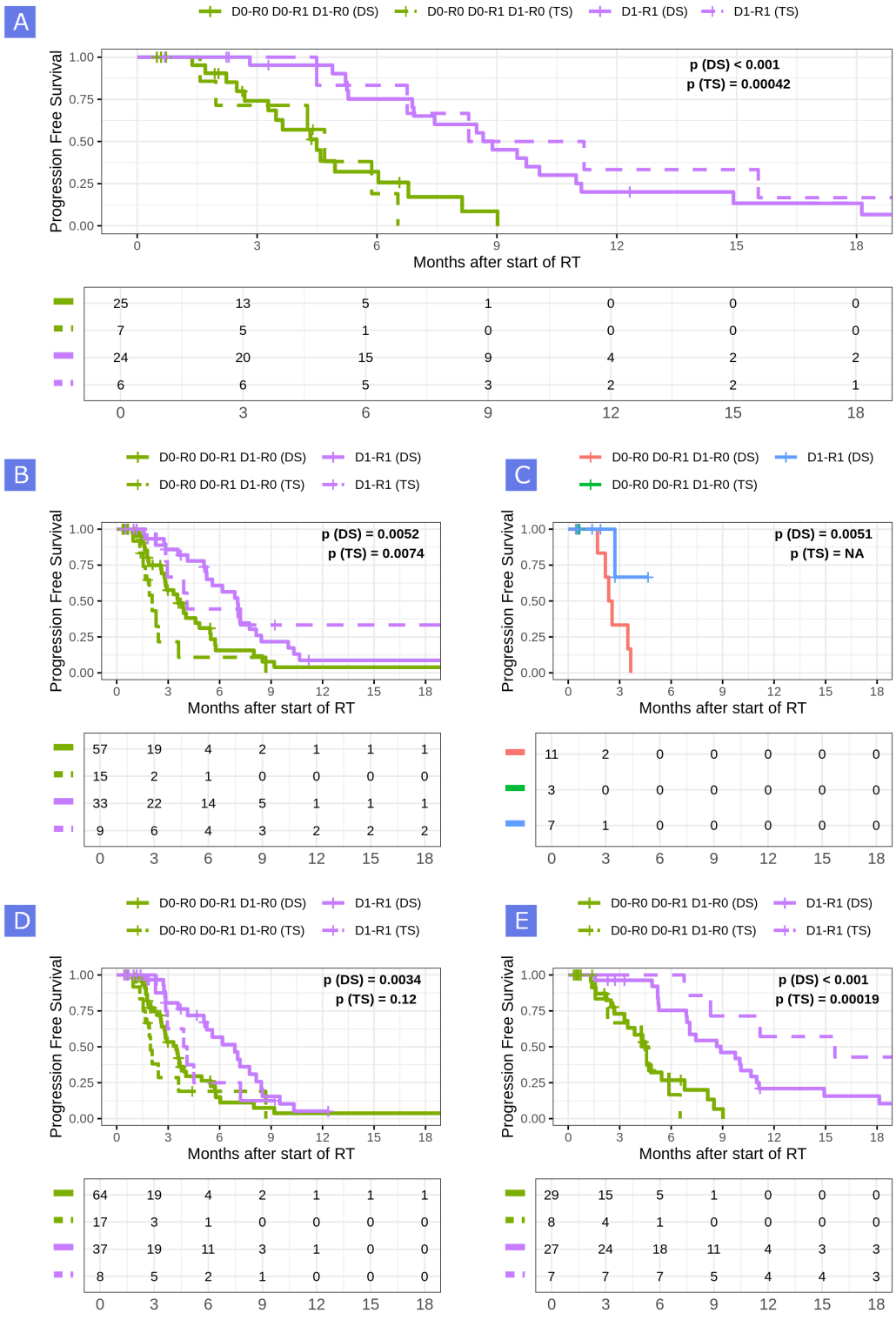


Figure E.3: Kaplan-Meier progression-free survival curves for combination of clusters for rHGG. A) subsetted on RRRS=good; B) subsetted on RRRS=intermediate; C) subsetted on RRRS=poor; D) subsetted on tumor histology: grade III; E) subsetted on tumor histology: grade IV, GBM

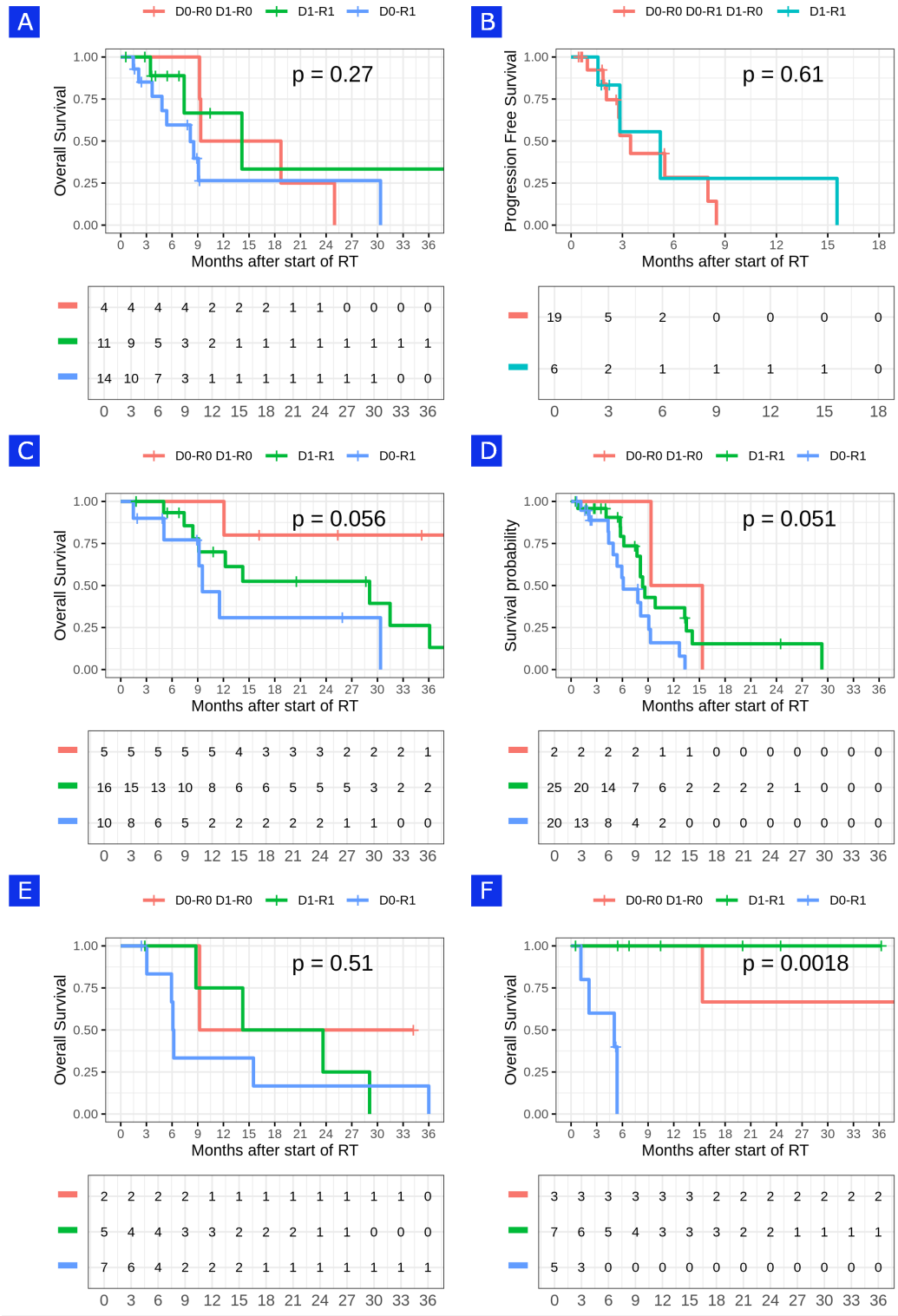


Figure E.4: Kaplan-Meier overall survival curves for patient subsets based on validated biomarkers in rHGG. A) MGMT Hypermethylated, B) MGMT Not hypermethylated, C) IDH1 Mutant, D) IDH1 Wildtype, E) 1p/19q Codeletion, and F) no 1p/19q Codeletion

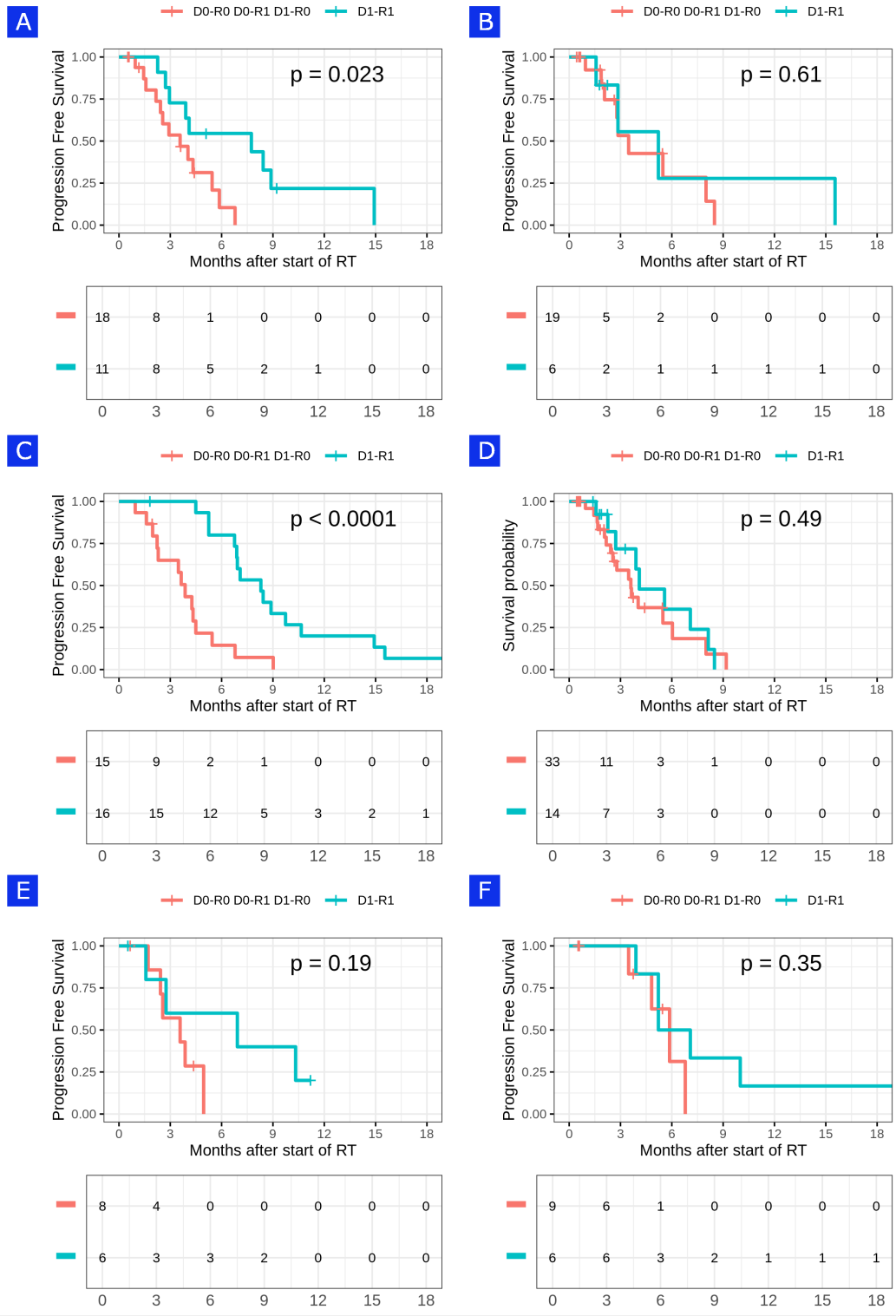


Figure E.5: Kaplan-Meier progression-free survival curves for patient subsets based on validated biomarkers in rHGG. A) MGMT Hypermethylated, B) MGMT Not hypermethylated, C) IDH1 Mutant, D) IDH1 Wildtype, E) 1p/19q Codeletion, and F) no 1p/19q Codeletion

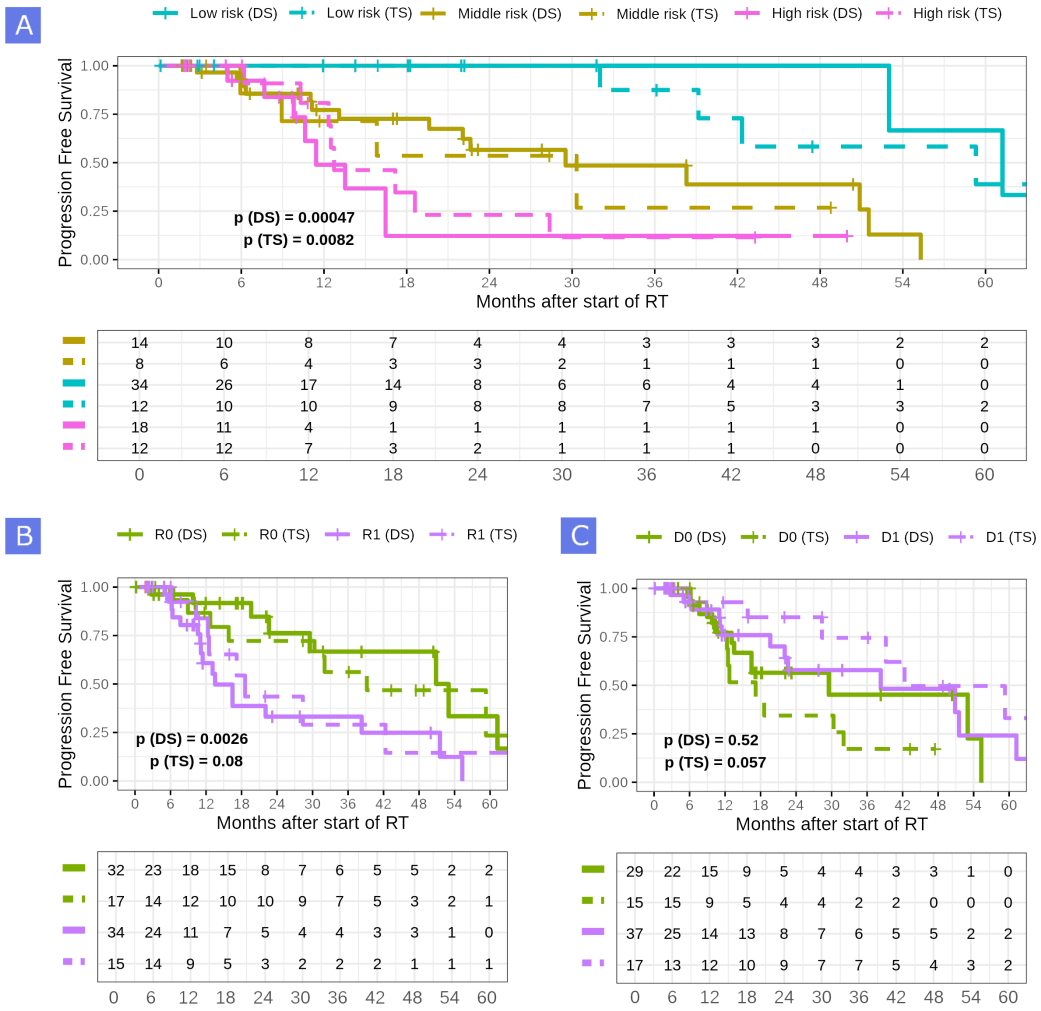


Figure E.6: Kaplan-Meier progression-free survival curves for NSCLC. A) of the hazard ratios stratified risk groups; B) of the Rad.cluster; C) of the Dos.cluster;

DISCLOSURE

This work was funded by the European Union's Horizon 2020 research and Marie Sklodowska-Curie Actions Innovative Training Network (MSCA-ITN-PREDICT), intramural funds of National Center for Tumor Diseases (NCT) Radiation Oncology Program as well as German Cancer Consortium Core Center, the German Cancer Research Center (DKFZ) and Heidelberg University Hospital.

The following manuscript will be in revision during the thesis evaluation process:

1. Salome, P., Sforazzini, F., Kudak, A., Abdollahi, A., Knoll, M. **MR-Class: MR Image Classification using one-vs-all Deep Convolutional Neural Network.**

2. Salome, P., Sforazzini, F., Kudak, A., Abdollahi, A., Knoll, M. **Impact Of MR Intensity Normalization For Different MR Sequences In MRI Based Radiomics Studies.**

I contributed to the following publications, mainly through the application of the tools built in this work:

1. Bennan, A.B.A., Unkelbach, J., Wahl, N., Salome, P. and Bangert, M. (2021). **Joint optimization of photon–carbon ion treatments for Glioblastoma.** Int J Radiat Oncol Biol Phys., 111(2), pp.559-572.

2. Sforazzini, F., Salome, P., Moustafa, M., Zhou, C., Schwager, C., Rein, K., Bougatf, N., Kudak, A., Woodruff, H., Dubois, L. and Lambin, P. (2022). **Deep Learning–based Automatic Lung Segmentation on Multiresolution CT Scans from Healthy and Fibrotic Lungs in Mice.** Radiology: Artificial Intelligence, 4(2), p.e210095.

3. Waltenberger, M., Furkel, J., Röhrich, M., Salome, P., Debus, C., Tawk, B., Gahlawat, A.W., Kudak, A., Dostal, M., Wirkner, U. and Schwager, C. (2022). **The impact of tumor metabolic activity assessed by 18 F-FET amino acid PET imaging in particle radiotherapy of high-grade glioma patients.** Frontiers in oncology, 12, pp.901390-901390.

Scientific output directly related to my thesis work, i.e., peer reviewed abstracts, paper and poster presentation at international conferences include:

1. Salome, P., Walz, D., Sforazzini, F., Kudak, A., Dostal, M., Regnery, S., Schlamp, K., Thomas, M., Felix Herth, F., Jäkel, O., Peter Heußel,

C., Hörner-Rieber, J. Debus, J., Knoll, M., Abdollahi, A. (2022). **Multi-Omics Classifier of Tumor Recurrence vs. Radiation-Induced Lung Fibrosis in NSCLC Patients Treated with SBRT**. Int J Radiat Oncol Biol Phys., 114(3), e388-e389.

2. Salome, P., Sforazzini, F., Kudak, A., König, L., Kickingereder, P., Bougatf, N., Wick, W., Jäkel, O., Debus, J., Knoll, M. and Abdollahi, A. (2021). **Improved risk stratification via integration of radiomics and dosiomics features in patients with recurrent high-grade glioma undergoing carbon ion radiotherapy (CIRT).** J Clin Oncol., 39(15), 2043.

3. Salome, P., Sforazzini, F., Kudak, A., Abdollahi, A., Knoll, M. **MR-Class: MR Image Classification using one-vs-all Deep Convolutional Neural Network.** ISMRM 2022.

4. Salome, P., Sforazzini, F., Kudak, A., Abdollahi, A., Knoll, M. **Impact Of MR Intensity Normalization For Different MR Sequences In MRI Based Radiomics Studies.** ISMRM 2022.

5. Sforazzini, F., Salome, P., Kudak, A., Ulrich, M., Bougatf, N., Debus, J., Knoll, M. and Abdollahi, A. (2020). **pyCuRT: An Automated Data Curation Workflow for Radiotherapy Big Data Analysis using Pythons' NyPipe.** Int J Radiat Oncol Biol Phys., 108(3), p.e772.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Dr. Maximillian Knoll for his guidance and support during the period of this work. Furthermore, I would like to thank Prof. Dr. Dr. Amir Abdollahi and Prof. Dr. Dr. Jürgen Debus for the opportunity to conduct my thesis at the Clinical Cooperation Unit Translational Radiation Oncology at the German Cancer Research Center (DKFZ) and the Heidelberg University Hospital and for being part of my thesis advisory committee together with Prof. Dr. Andrea Mairani. Their constructive input has been helpful during my PhD years. Next, I would like to acknowledge my colleagues, especially Dr Francesco Sforazzini, David Walz and Ahmad Neishabouri, for their help and input. To continue with, I want to mention my family George Salome, Maha Salome, Christophe Salome and Dr Mariam Hamwi Mella for their support and encouragement. Last but not least, I would like to thank Diana Mîndroc-Filimon for all of the brainstorming sessions and valuable input throughout the completion of this work.

EIDESSTATTLICHE VERSICHERUNG

1. Bei der eingereichten Dissertation zu dem Thema:

   Dosiomics-enhanced prediction modeling: an artificial intelligence-based workflow in radiation oncology

   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort, Datum

Unterschrift Doktorand