National Center for Radiation Research in Oncology – OncoRay
Direktorin: Frau Prof. Dr. Mechthild Krause

Radiomics risk modelling using machine learning algorithms for personalised radiation oncology

Dissertationsschrift

zur Erlangung des akademischen Grades
Doktor der Medizinischen Biometrie und Bioinformatik
Doctor rerum medicinalium (Dr. rer. medic.)
vorgelegt
der Medizinischen Fakultät Carl Gustav Carus
der Technischen Universität Dresden

von

Dipl.-Inf. (FH) Stefan Leger

aus Jena

Dresden 2018

1. Gutachter: PD Dr. Steffen Löck (TU Dresden)

2. Gutachter: Prof. Dr. Hans-Joachim Böhme (HTW Dresden)

3. Gutachter: Prof. Dr. David Craft (Harvard Medical School Boston/USA)

Tag der mündlichen Prüfung: 05. März 2019

gez.: Prof. Dr. Esther G. C. Troost
Vorsitzende der Promotionskommission

Abstract

One major objective in radiation oncology is the personalisation of cancer treatment. The implementation of this concept requires the identification of biomarkers, which precisely predict therapy outcome. Besides molecular characterisation of tumours, a new approach known as radiomics aims to characterise tumours using imaging data. In the context of the presented thesis, radiomics was established at OncoRay to improve the performance of imaging-based risk models. Two software-based frameworks were developed for image feature computation and risk model construction. A novel data-driven approach for the correction of intensity non-uniformity in magnetic resonance imaging data was evolved to improve image quality prior to feature computation. Further, different feature selection methods and machine learning algorithms for time-to-event survival data were evaluated to identify suitable algorithms for radiomics risk modelling. An improved model performance could be demonstrated using computed tomography data, which were acquired during the course of treatment. Subsequently tumour sub-volumes were analysed and it was shown that the tumour rim contains the most relevant prognostic information compared to the corresponding core. The incorporation of such spatial diversity information is a promising way to improve the performance of risk models.

Kurzfassung

Ein neuer Schwerpunkt in der Radioonkologie ist die Personalisierung der Krebsbehandlung, um beispielsweise die Strahlendosis individuell auf einen spezifischen Tumor anzupassen. Die Implementierung eines solchen Ansatzes erfordert die Identifizierung von Merkmalen zur präzisen Therapievorhersage. Statt der etablierten molekularen Tumorcharakterisierung verwendet radiomics, als neues Verfahren, bildbasierte Merkmale. In der vorliegenden Arbeit wurde radiomics am OncoRay etabliert, mit dem Ziel die Leistungsfähigkeit von bildbasierten Risikomodellen zu verbessern. Zunächst wurden zwei softwarebasierte Systeme für die Berechnung von Merkmalen und die Erstellung von Risikomodellen entwickelt. Darüber hinaus wurde ein neues Verfahren für die Korrektur von Signalinhomogenitäten in magnetresonanztomografischen Bilddaten implementiert, um die Bildqualität zu verbessern. Zudem erfolgte eine umfassende Analyse verschiedener Methoden zur Merkmalsauswahl und maschineller Lernverfahren für Überlebenszeitdaten, um geeignete Methoden zur bildbasierten Risikomodellierung zu identifizieren. Durch die Hinzunahme von Röntgen-Computertomographie Bilddaten, die während der Therapie erstellt wurden, konnte eine Verbesserung der Risikomodelle erreicht werden. Darüber hinaus wurde gezeigt, dass der Tumorrand, im Vergleich zum Tumorzentrum, die relevanten prognostischen Informationen enthält. Die Berücksichtigung der räumlichen Diversität ist daher ein vielversprechender Weg, um die Leistungsfähigkeit von Risikomodellen zu verbessern.

Contents

List of Figures

List of Tables

List of Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
ABC	artificial bee colony
BGLM	boosted gradient linear model
Bias	intensity non-uniformity
BT	boosted tree
C-Index	concordance index
CNN	convolutional neural networks
COV	coefficient of variation
Cox	Cox proportional hazard model
CSF	cerebro-spinal fluid
CT	computed tomography
CV	cross-validation
DKTK-ROG	German Cancer Consortium Radiation Oncology Group
DWT	discrete wavelet transform
FBP	filtered back projection
FDG	^{18}F-fluorodeoxyglucose
FMISO	^{18}F-fluoromisonidazole
FOV	field of view
GLCM	grey-level co-occurrence matrix
GLDZM	grey-level distance zone matrix

GLRLM	grey-level run length matrix
GLSZM	grey-level size zone matrix
GM	grey-matter
GTV	gross tumour volume
$\mathrm{GTV}_{\mathrm{entire}}$	entire gross tumour volume
Gy	Gray
HAC	hierarchical agglomerative clustering
HNSCC	head and neck squamous cell carcinoma
HPV	human papillomavirus
HU	Hounsfield units
ICC	intra-class correlation coefficient
ISBI	image biomarker standardisation initiative
LoG	laplacian of Gaussian
LRC	loco-regional tumour control
MIFS	mutual information feature selection
MIM	mutual information maximisation
MLM	multi-level model
MR	magnetic resonance
MRI	magnetic resonance imaging
MRMR	minimum redundancy maximum relevance
MSR-RF	maximally selected rank statistics random forest
MSR-RFVI	maximally selected rank statistics random forest variable importance
NET-Cox	regularised Cox proportional hazard model
NGLDM	neighbourhood grey level dependence matrix
NGTDM	neighbourhood grey tone difference matrix

NMR	nuclear magnetic resonance
NSCLC	non-small cell lung cancer
OS	overall survival
PCM	physical correction model
PDw	proton density-weighted
PET	positron emission tomography
PVI-RF	permutation variable importance random forest
RCDF	Radiotherapy Center Dresden-Friedrichstadt
RCT	radio-chemotherapy
RF	radio-frequency
RF-MD	random forest minimal depth
RF-VH	random forest variable hunting
RF-VI	random forest variable importance
RMF	risk modelling framework
ROI	region of interest
RSF	random survival forest
SD	standard deviation
SRM	survival regression model
SUV	standard uptake value
T1w	T1-weighted
T2w	T2-weighted
UDWT	decimated discrete wavelet transform
UKD	University Hospital Dresden
VIMP	variable importance
Voxel	volume element
WM	white-matter

1. Introduction

The personalisation of treatment is a major objective for improving modern cancer therapy, which for example tailors the prescribed type of chemotherapy, targeted drug or radiation dose to the specific tumour for each patient individually. Radiotherapy is prescribed for more than 50% of all cancer patients and attemps to eliminate all cancer stem cells of the primary tumour and regional lymph nodes using ionising radiation. In the last decade, the treatment have been improved by technological advances and the integration of (radio-) biological knowledge. Consequently, these improvements led to an increased loco-regional tumour control (LRC) describing the recurrence of the tumour in the same region or regional lymph nodes after a disease free period (Baumann et al., 2016). In general, evidence for novel treatment options are derived by clinical trials (e.g., Danish Head and Neck Cancer Group) comprising large patient cohorts with the same type of cancer stratified by clinical parameters, e.g., tumour stage or histology (Overgaard et al., 1998; Overgaard et al., 2003; Bentzen et al., 2015; Baumann et al., 2016). This population-based evidence, traditionally determined the treatment options for all patients with similar diagnoses and staging. However, personalisation of treatment offers the potential to further improve radiotherapy by considering the tumour characteristic of patients or subgroups individually. The implementation of this approach requires the identification of specific biomarkers, which are highly correlated with tumour radio-sensitivity and precisely predict therapy response. Their identification would enable the stratification of patients into smaller subgroups with the same tumour characteristic. For instance, patients at high risk of treatment failure and low risk of severe side effects may receive intensified treatment while for patients at high risk of severe toxicities and good chances for cure a lower radiation dose may be delivered (figure 1.1). One example for treatment individualisation is aspired for patients with head and neck squamous cell carcinoma (HNSCC) which comprise different types of tumours, e.g., larynx, oral cavity and hypopharynx. The worldwide incidence rate of HNSCC is about 7% and the overall survival (OS) rate at five years after treatment for early-stage tumours ranges between 70 and 90%. However, patients with advanced disease show poorer survival rates of only 30 and 50% (Jemal et al., 2011; Gatta et al., 2015). Tumours in the head and neck region comprise an additional challenge for radiotherapy because they are surrounded by various organs, which are fundamental for, e.g., swallowing, sensation and communication. Consequently, HNSCC treatment has to tackle two major challenges: achieving a high tumour control rate while limiting the amount of side effects, which degrade the patients' quality of life. Therefore, patients could benefit from individualised radiotherapy based on risk models, e.g., to adjust the applied total dose to the individual patient.

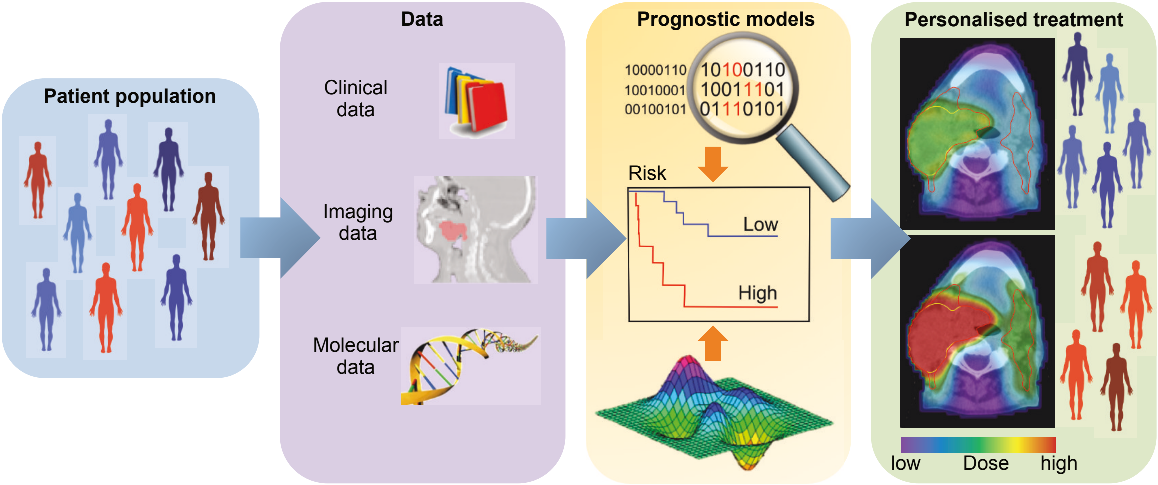


Figure 1.1.: Schematic representation of the workflow to individualise radiotherapy based on different steps: From a heterogeneous patient population clinical data, imaging data (e.g., X-ray computed tomography) and molecular data are combined to develop predictive or prognostic risk models for the prediction of treatment outcome and patient stratification. The risk model can be used to personalise treatment, e.g., by dose adaptation in clinical trials.

During the last years the identification of biomarkers for characterisation of the tumour phenotype has focused on molecular biomarkers using genomics data (Toustrup et al., 2011; Schmidt et al., 2018). One substiantial example of a radiobiological-associated biomarker is the human papillomavirus (HPV), which is related to different types of cancer, e.g., cervical cancer or HNSCC. For instance, Lohaus *et al.* (Lohaus et al., 2014) demonstrated that the HPV status is a strong prognostic biomarker for LRC, in particular for oropharyngeal tumours. As a result, patients with HPV-positive tumours may be qualified for dose de-escalation strategies in future. Furthermore, it was shown that hypoxia-associated genes and cancer stem cell markers are correlated with tumour recurrence for patients with HNSCC (Toustrup et al., 2011; Eustace et al., 2013; Linge et al., 2016a; Linge et al., 2016c).

Despite these promising results, the spatial and temporal heterogeneity between different patients and within tumours is a major challenge for development of molecular-based risk models to personalise therapy. For the personalisation of cancer therapy tissue extractions of the tumour by biopsies are required to analyse the molecular profiles for patients treated with primary radio-chemotherapy. Recently, Gerlinger *et al.* (Gerlinger et al., 2012) observed differences in gene expressions extracted from different parts of the tumour for patients with renal cell carcinoma. Consequently, the development of risk models using biomarkers obtained from a single biopsy may not represent the entire tumour and the outcome of the patient may not be predicted correctly. To overcome this challenge, multiple and repeated tumour biopsies would be needed, which seems to be limited due to the invasiveness of the procedures (Gerlinger et al., 2012).

However, medical imaging has the potential to circumvents these problems. Imaging data are non-invasive and can provide comprehensive information regarding the entire tumour. Furthermore, a variety of imaging data is acquired in clinical practice for diagnosis and treatment guidance: X-ray computed tomography (CT), positron emission tomography (PET) and magnetic resonance imaging (MRI). For example, it was shown that risk models based on quantitative imaging features, such as tumour volume determined on CT and MRI scans or uptake values of PET measurements, were able to predict treatment response (Partridge et al., 2005; Hutchings et al., 2006; Zschaeck et al., 2017). Therefore, it is hypothesised that detailed tumour phenotype characteristics may be extracted from imaging data. In the last years radiomics emerged as a new approach, which is based on the extraction and analysis of a large amount of quantitative biomarkers (features) from medical images by machine learning algorithms to predict patient specific outcome. Imaging features may provide complementary and interchangeable information compared to other patient data, e.g., clinical or genetic data and may have the potential to enhance the effort to individualise radiation oncology (Lambin et al., 2012). Radiomics features are based, e.g., on the one-dimensional histogram of the grey values, spatial relationships between various intensity levels or texture heterogeneity patterns extracted from medical imaging data and various transformed images. This results in a high variety of imaging features ($\gg$1000). Subsequently, machine-learning algorithms are applied to identify prognostic and non-redundant features, which show a high association with clinical outcome to build predictive or prognostic radiomics models.

The application of radiomics comprises mainly two different domains in clinical oncology: diagnosis and tumour prognosis. Several studies demonstrated the potential of radiomics in the field of cancer diagnosis. For instance, malignant and benign prostate tumours were differentiated based on texture features computed from MRI or the incidence of lung cancer was predicted using CT imaging (Wibmer et al., 2015; Hawkins et al., 2016). Further, in the field of tumour prognosis and outcome prediction, radiomics has been successfully applied using different image modalities such as CT, PET and MRI (Chicklore et al., 2013; Coroller et al., 2015; Nicolasjilwan et al., 2015; Wibmer et al., 2015; Coroller et al., 2016). For instance, loco-regional recurrences and distant metastases were predicted for risk assessment using PET-CT imaging of HNSCC patients. Moreover OS was predicted for patients with newly diagnosed glioblastoma multiforme based on a subset of radiomics features (signature) computed from MRI data (Kickingereder et al., 2016; Vallières et al., 2017). In addtion, several studies have shown the discriminating capabilities of radiomics features for the stratification of tumour histology, tumour grades or clinical outcomes (Ganeshan et al., 2010; Yamamoto et al., 2012; Aerts et al., 2014; Jain et al., 2014; Hatt et al., 2015; Kather et al., 2016). Besides the feasibility of radiomics-based risk models, the correlation between imaging features and gene expression profiles have been demonstrated (Segal et al., 2007; Panth et al., 2015; Grossmann et al., 2017).

Despite the potential of radiomics for the individulisation of cancer therapy, further research and developments are required, e.g., regarding the choice of suitable machine learning algorithms for risk modelling (Kumar et al., 2012; Gillies et al., 2015; Aerts, 2016; Yip and Aerts, 2016). The aim of this thesis was to establish and implement radiomics at the National Centre for Radiation Research in Oncology (Dresden, Germany) to improve the performance of imaging-based risk models. The presented results help to facilitate image-based treatment decisions and to gain a deeper understanding of radiomics.

Image-based risk modelling comprises a complex task consisting of the feature computation and extraction based on the imaging data and the development of predictive or prognostic models. In this thesis, two software-based frameworks for image feature computation and risk modelling were developed and introduced (chapter 2).

The feature computation framework provides different pre-processing algorithms, e.g., to enhance the image quality. Hence, a novel data-driven approach was developed to correct intensity non-uniformity in MRI data to improve the image quality prior to feature computation. The proposed approach, which is introduced in chapter 3 is motivated by the physical properties of a typical MRI coil system.

Feature reduction and selection of the most informative features are necessary for achieving high performance. Also the choice of the machine learning algorithm for risk modelling may be vital in achieving a good performance. Consequently, the identification of suitable methods are a integral component to develop highly accurate and reliable clinical risk models. An extensive evaluation of different feature selection methods and machine learning algorithms for time-to-event survival data was realised using a large multi-centre cohort of patients with locally advanced HNSCC (chapter 4).

In general, the improvement of the risk model performance is an essential factor to facilitate and to drive the success of radiomics applications for clinical decision making. In chapter 5 the potential of CT imaging during the course of treatment was investigated. In particular, an exploratory and an independent validation cohort of patients with locally advanced HNSCC were used to compare the performance of the risk models developed on pre-treatment scans with the models based on in-treatment CT images.

Moreover, in radiomics, the characterisation of the tumour phenotype is usually based on radiomics features extracted from the entire tumour. However, tumours are biologically complex and exhibit spatial variations. Therefore, in chapter 6, different tumour sub-volumes were investigated. In particular, risk models were developed and compared based on the rim of the tumour and the corresponding core in terms of their prognostic performance and their ability to stratify patients into low and high risk groups of recurrence. This analysis enables to gain a deeper understanding which parts of the tumour contain the most relevant prognostic informations and whether the incorporation of spatial diversity may improve the model performance.

Chapter 7 provides a general summary of the work presented in this thesis and related further perspectives.

2. Theoretical background

2.1. Basic physical principles of image modalities

2.1.1. Computed tomography

Computed tomography (CT) is a commonly imaging technique in diagnostic radiology and radiation oncology to provide three-dimensional (3D) information of anatomical structures using X-rays from various directions. In radiation oncology an important application of CT imaging is for treatment planning purposes. A CT scanner consists of an X-ray tube with an opposing detector array which rotates around the scan object, while the scan object is moved along the perpendicular axis to the rotation plane. The resulting 3D X-ray attenuation profile can be described by the Lambert-Beer law. The detected signal I_φ at a given angle φ is given by, (Bouguer, 1922):

$$I_\varphi(\tau) = I_0 \cdot \int_E \Omega(E) \cdot \mathrm{e}^{-\int\int \mu[E, r_\varphi(\tau)]\,\mathrm{d}^2 r}\,\mathrm{d}\,E. \tag{2.1}$$

Here, I_0 and $r_\varphi(\tau)$ defines the initial intensity emitted from the X-ray tube, which is attenuated exponentially in dependence of the mass attenuation coefficient μ. The mass attenuation coefficient μ depends on the photon energy E and the material at the distance r_φ. The total attenuation at the distance τ perpendicular from the rotation axis is calculated as the integral over all energies. The attenuation coefficient μ can be described as a product of the electron density ϱ_e of the material/tissue and the photon attenuation cross section per electron σ_e:

$$\mu(E) = \varrho_e \cdot \sigma_e(E). \tag{2.2}$$

Thereby, for typical energies in diagnostic radiology ($E < 200\ keV$) the cross section $\sigma_e(E)$ is dominated by three different interactions with matter: photoelectric effect σ_e^{ph}, coherent σ_e^{coh} which depend on the photon energy E and the atomic number Z and incoherent scattering σ_e^{inc}, depending on the photon energy E only (Rutherford et al., 1976; Jackson and Hawkes, 1981).

Based on the measured X-ray attenuation profile for different projection angles φ, image reconstruction algorithms, e.g., filtered back projection (FBP) provide the individual attenuation coefficient μ of each volume element (Voxel) within the field of view (FOV) (Kachelriess et al., 2004). The attenuation coefficient μ represents electron density of a material in relation to that of water and air, which is expressed as CT number H in Hounsfield units (HU):

$$H = \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \cdot 1000. \tag{2.3}$$

In the clinical domain, the Hounsfield scale ranges from -1024 HU to 1024 HU to differentiate human tissue with H=0 HU for water (figure 2.1). Some modern scanners have an extended Hounsfield scale up to 4000 HU representing materials with high effective atomic numbers (e.g., metallic implants).

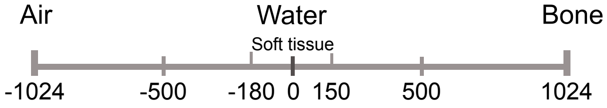


Figure 2.1.: Illustration of a commonly used Hounsfield scale, ranging from -1024 (Air) to 1024 Hounsfield unit (HU) (Bone). In radiomics studies the soft tissue range from -150 to 180 HU is of most interest.

2.1.2. Magnetic resonance imaging

MRI is another established and non-invasive imaging technique for diagnostic radiology and radiation oncology. MRI is used, e.g., for the assessment of characteristics of human tissue (Low, 2007), to support therapy planning (Nomden et al., 2013) and to monitor therapy response (Yamaner et al., 2012). The advantages of MRI are the excellent soft-tissue contrast without loss of spatial resolution and the lack of undesirable X-ray radiation dose compared to CT imaging. Furthermore, using MRI a high variety of images can be created reflecting different physical and physiologic phenomena such as tissue susceptibility variations (Haacke et al., 2009), diffusion (Carr and Purcell, 1954; Stejskal and Tanner, 1965), biomechanical properties (Muthupillai et al., 1995) and oxygen levels (Ogawa et al., 1990; Belliveau et al., 1991).

MRI is based on the basic physics of nuclear magnetic resonance (NMR), which was first experimentally demonstrated by Purcell *et al.* (Purcell et al., 1946) and Bloch *et al.* (Bloch, 1946). An atomic nucleus is comprised of protons and neutrons, and may have a non-zero nuclear magnetic moment. The nuclear magnetic moment is determined through the pairing of the constituent protons and neutrons. The proton (^1H) is very important for MRI as it is abundant in human tissue and possesses a non-zero nuclear magnetic moment. Typically, due to thermal movement, the nuclear magnetic moments are randomly orientated resulting in a non observable magnetisation (figure 2.2, left). However, when placed in a strong uniform magnetic field B_0, the nuclear magnetic moments within a scan object align with the magnetic field, resulting in a measurable net magnetic moment M_0 in the direction of B_0 (figure 2.2, right).

Furthermore, protons experience a torque in perpendicular to the direction of the applied magnetic field that causes precession. The precession occurs with an angular frequency ω_0

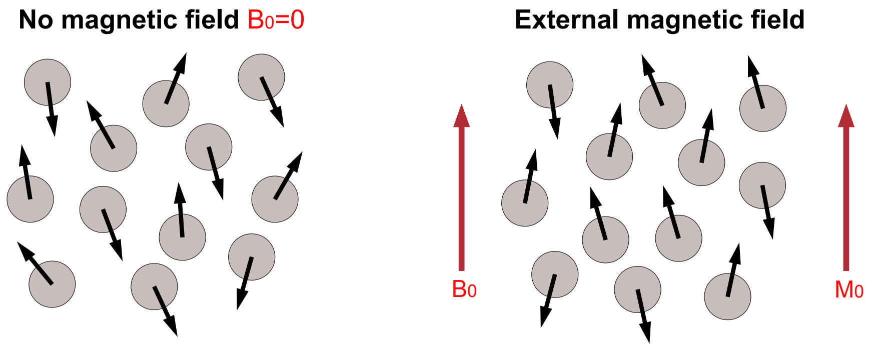


Figure 2.2.: Simplified distributions of "free" protons without (left) and with an external magnetic field B_0 (right). Without an external field, the magnetic moments of the protons are randomly orientated, producing an overall magnetic moment of zero. Under the influence of an applied external magnetic field B_0 the protons assume a align in parallel and anti-parallel orientation to the applied magnetic field. The alignment in parallel direction, results in a measurable net magnetic moment (magnetisation) M_0 in the direction of B_0. Picture is taken and modified from Bushberg *et al.* (Bushberg, 2002).

(Larmor frequency) that is proportional to the magnetic field strength B_0, described by the Larmor equation:

$$\omega_0 = \gamma B_0, \tag{2.4}$$

where γ defines the nuclei-specific gyromagnetic ratio. For the ^1H nucleus, γ is roughly equal to 42.58 MHz/Tesla. Through the application of a radio-frequency (RF) pulse with time-varying magnetic field B_1(t) tuned to the Larmor frequency ω_0 in the transverse plane (x, y-plane) and tips the net magnetisation in this plane. As a result, M_0 is excited into the transverse plane and will precess as long as $B_1(t)$ is applied. After the RF pulse, the magnetisation moment in the transverse plane M_{xy} will decay with a material-specific relaxation time $T2$ (spin-spin relaxation time constant). The elapsed time between the peak transverse signal (e.g., after a 90-degree RF pulse) and 37% of the peak level (1/e) is given by, (Bushberg, 2002):

$$M_{xy}(t) = M_0\, \mathrm{e}^{-t/T2}, \tag{2.5}$$

where $M_{xy}(t)$ is the transverse magnetic moment at time t for a sample that has the maximum transverse magnetisation M_0 at $t = 0$. The longitudinal magnetisation M_z will recover with the characteristic time $T1$ (spin-lattice relaxation time constant) to its previous equilibrium state governed by B_0. This occurs exponentially as:

$$M_z(t) = M_0(1 - \mathrm{e}^{-t/T1}), \tag{2.6}$$

where $T1$ is the time needed for the recovery to 63% of M_z after a 90-degree RF pulse.

A magnetic resonance (MR) image can be formed by the variation of the $T1$ and $T2$ relaxation times, where the contrast is based on the differences in $T1$ or $T2$ relaxation times of different tissues of the human body, due to their different molecular environments. For instance, the contrast in T1-weighted (T1w) images are chiefly based on the $T1$ characteristics of the tissue, while de-emphasis of $T2$ and proton density contributions to the signal. proton density-weighted (PDw) image relies mainly on differences in the number of magnetized protons per unit volume in tissue. For instance, tissue with a large proton density, e.g., fat and lipids have a corresponding large M_z compared to other soft tissues. The contrast in PDw images are achieved by reducing the contributions of $T1$ recovery and $T2$ decay. Furthermore, the contrast in a T2-weighted (T2w) image follows directly from the PDw image by reducing the $T1$ and emphasis of the $T2$ differences in tissue (Bushberg, 2002).

2.2. Basic principles of survival analyses

Survival analyses comprise a collection of statistical techniques used to describe and quantify time-to-event data. In survival analyses the term 'failure' defines the occurrence of an event of interest (e.g., death). The term 'survival time' specifies the period of time until the failure to occurs (e.g., time from entry into a clinical trial until death). Survival outcome is investigated in different studies, e.g., retrospective observational studies or animal experiments to compare two different treatments in terms of differences in overall survival time. The survival function defines the probability of surviving up to time t and is expressed as:

$$S(t) = P(T > t) = 1 - F(t), 0 < t < \infty, \tag{2.7}$$

where T defines a non-negative random variable representing the time until an event of interest and $F(t) = P(T \leq t)$ defines the cumulative distribution function. In survival analyses, the main objective is to calculate the instantaneous failure rate at time t, which is expressed in terms of the hazard function h:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} \\ &= \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}\,t} ln(S(t)), \end{aligned} \tag{2.8}$$

where $f(t) = -\frac{\mathrm{d}}{\mathrm{d}\,t}S(t) = \frac{\mathrm{d}}{\mathrm{d}\,t}F(t)$ defines the probability density function. The hazard function h at time t described the failure rate, i.e., the frequency of events at time t. Based on (2.8) it is possible to define the survival function by integrating both sides, yielding:

$$-\ln(S(t)) = \int_0^t h(s)\,\mathrm{d}\,s = H(t), \tag{2.9}$$

so that

$$S(t) = \mathrm{e}^{-H(t)}, \tag{2.10}$$

where $H(t)$ defines the cumulative hazard function and $S(0)$=1. For modelling of survival data different survival distributions are available, e.g., the exponential and the Weibull distributions, which are described in the following. The exponential survival distribution has a constant hazard, $h(t) = \lambda$. Using (2.9) its cumulative hazard function may be derived as:

$$H(t) = \int_0^t h(s)\,\mathrm{d}\,s = \lambda t. \tag{2.11}$$

Subsequently, the survival function $S(t)$ is given by:

$$S(t) = \mathrm{e}^{-\lambda t}, \tag{2.12}$$

and the probability density function:

$$f(t) = h(t)S(t) = \lambda \cdot \mathrm{e}^{-\lambda t}. \tag{2.13}$$

A more sophisticated survival distribution is the Weibull distribution, which offers more flexibility for modelling survival data. The hazard function $h(t)$ is given by:

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1} = \alpha\lambda^{\alpha}t^{\alpha-1}, \tag{2.14}$$

where α defines the shape and λ the scale of the hazard function. The cumulative hazard and survival function as well as the probability density function are respectively given by:

$$H(t) = (\lambda t)^{\alpha} \tag{2.15}$$

and

$$S(t) = \mathrm{e}^{-(\lambda t)^{\alpha}} \tag{2.16}$$

as well as

$$f(t) = \alpha\lambda^{\alpha}t^{\alpha-1}\,\mathrm{e}^{-(\lambda t)^{\alpha}}. \tag{2.17}$$

Figure 2.3 shows an example of a hazard and a survival function based on the exponential and the Weibull distributions. While for the exponential distribution, survival decreases

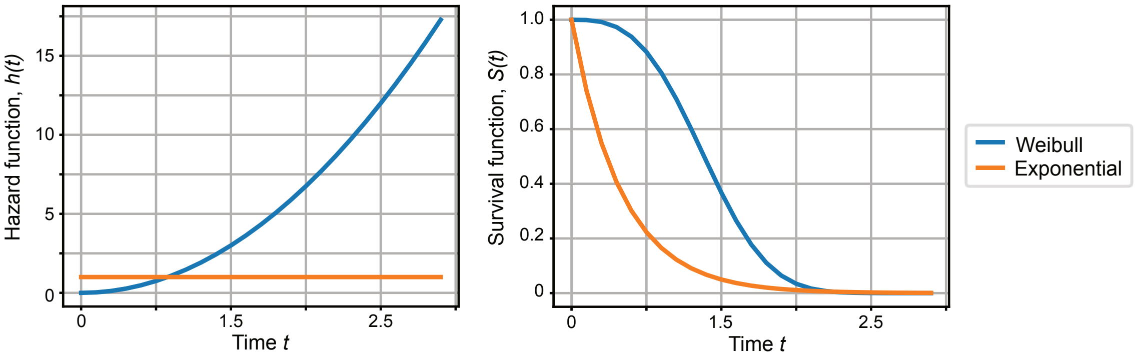


Figure 2.3.: Examples of the hazard (left) and the survival functions (right) for the exponential (λ=1) and the Weibull distributions (λ=1, α=3).

exponentially with time t, the Weibull function allows for a more complex behaviour. Based on the described survival theory, two general types of survival models, semi-parametric and full-parametric survival models are briefly described in the following section.

2.2.1. Semi-parametric survival models

Semi-parametric survival models have an unspecific dependence on time. Two different survival models, the Cox proportional hazard model (Cox) and the random survival forest (RSF) are described to demonstrate the different algorithm concepts. The mathematical nomenclature for this section is defined as follows. Given is an observed random variable $Z_i = (X, Y)$ where $X_i = (X_{i,1}, \ldots, X_{i,p})$ defines the p-dimensional covariate (feature) vector and Y_i denotes the time-to-event survival data for a subject $i = 1, ..., N$. The survival data $Y_i = (t_i, \delta_i)$ consists of survival time t_i at which an event occurred ($\delta_i = 1$, e.g., death) or the observation was censored ($\delta_i = 0$).

Cox proportional hazards model

The Cox proportional hazards model assumes linearity of the covariates X on the log hazard function to model the interactions between the covariates and the risk (Cox, 1972). It is given by:

$$h(t, \beta, X) = h_0(t)e^{\sum_{j=1}^{p} \beta_j X_j}, \tag{2.18}$$

where X_j is a vector of covariate values for subjects $i = 1, ..., N$ and β is a vector of coefficients, with one coefficient for each covariate. The coefficients β are fitted using the likelihood maximisation, which describes the probability to obtain the data with a given model. The total partial likelihood $L(\beta)$ is defined as:

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{N} \left[\frac{h_0(t_i)\, \mathrm{e}^{\sum_{j=1}^{p} \beta_j X_{ij}}}{\sum_{k \in R_i} h_0(t_i)\, \mathrm{e}^{\sum_{j=1}^{p} \beta_j X_{kj}}} \right]^{\delta_i} \\
&= \prod_{i=1}^{N} \left[\frac{\mathrm{e}^{\sum_{j=1}^{p} \beta_j X_{ij}}}{\sum_{k \in R_i} e^{\sum_{j=1}^{p} \beta_j X_{kj}}} \right]^{\delta_i},
\end{aligned}
\tag{2.19}
$$

where β_j is a coefficient, for each covariate j and $R_i = \{k : t_k \geq t_i\}$ denotes the risk set containing all subjects at risk at time t_i. In other words, the numerator is the hazard for the subjects who experienced the failure ($\delta = 1$). However, the denominator is the sum of all subjects ($\delta = 0$ and $\delta = 1$, respectively) in the risk set who are still at risk at time t_i. In practise, instead of using the partial likelihood $L(\beta)$ the logarithm transformed likelihood function $LL = \ln[L(\beta)]$ is used for the estimation of β, e.g., by the Newton Raphson algorithm. Using (2.19) the LL is given by:

$$
LL(\beta) = \sum_{i=1}^{N} \delta_i \left\{ \sum_{j=1}^{p} \beta_j X_{ij} - \ln \left[\sum_{k \in R_i} \mathrm{e}^{\sum_{j=1}^{p} \beta_j X_{kj}} \right] \right\}. \tag{2.20}
$$

One property of the Cox model is the unspecific baseline hazard function $h_0(t)$, which cancels out of the numerator and denominator in (2.19). Therefore, instead of calculating the hazard function directly the ratio of the hazard function (HR) is calculated, yielding to the proportional hazards assumption. This assumption fulfilled if the ratio of the hazard function is constant between two survival groups, e.g., a control and a treatment group. The undefined baseline hazard function is one advantageous of the Cox model since it is robust against misspecification and has fewer restrictions.

Random survival forest

The Cox model assumes linearity of the covariates X on the log hazard function. However, for covariates with a non-linear structure this assumption may lead to misspecifications due to not fitting the correct functional form. Therefore, for covariates with non-linear effects or higher order interactions, non-linear machine learning algorithms are appropriate, e.g., the Random forest (Breiman, 2001). For continuous time-to-event survival data, the RSF is an extension of the Random forest (Ishwaran and Kogalur, 2007; Ishwaran et al., 2008). The basic idea behind the random forest algorithm is to train several decision trees (nTree) by bootstrapping the sample data and subsequently estimate an ensemble prediction using all trees. For each bootstrap sample a tree is grown by selecting random covariates (mtry) splitting the data. Subsequently, a maximum number of split points (nSplit) is randomly chosen from all possible split points within a selected splitting candidate (covariate). Finally, a node is splitted on the particular split point which maximises the survival differences between the resulting daughter nodes. For RSF, two typical splitting rules for survival data are the log-rank test (LR) and the log-rank-score test (LRS).

Both methods split a node q in the tree using a given covariate $x \in X$ and a selected split point c, which forms two daughter nodes ($x \leq c$ and $x > c$, respectively) and two new sets of survival data. Let $T_1 < T_2 < ... < T_D, k = 1, ..., D$ be the distinct event times (e.g., death) in the parent node q, and let $d_{k,j}$ be the number of events and $\Upsilon_{k,j}$ be the number of individuals at risk at time T_k in the daughter nodes $j = 1, 2$. The number of subjects in daughter j that are alive or had an event at T_k, is defined by (Ishwaran et al., 2008):

$$\begin{aligned} \Upsilon_{k,1} &= \#\{t_i \geq T_k, x_i \leq c\}, \\ \Upsilon_{k,2} &= \#\{t_i \geq T_k, x_i > c\}, \end{aligned} \tag{2.21}$$

where x_i and t_i are the value of covariate $x \in X$ and the survival time for subject $i = 1, .., N$, respectively. The *LR* criteria for a split at the value c for covariate $x \in X$ is given by:

$$\text{LR}(x, c) = \frac{\sum_{k=1}^{D} (d_{k,1} - \Upsilon_{k,1} \frac{d_k}{\Upsilon_k})}{\sqrt{\sum_{k=1}^{D} \frac{\Upsilon_{k,1}}{\Upsilon_k} (1 - \frac{\Upsilon_{k,1}}{\Upsilon_k}) (\frac{\Upsilon_k - d_k}{\Upsilon_k - 1}) d_k}}, \tag{2.22}$$

where $\Upsilon_k = \Upsilon_{k,1} + \Upsilon_{k,2}$ and $d_k = d_{k,1} + d_{k,2}$ defines the total number of events and individuals at risk at time T_k. A large value for LR(x, c), indicates a large difference between the two groups.

Another useful splitting criterion is the LRS which computes the "ranks" for each survival time t_i based on the ordered covariate $x \in X$ such that $x_1 \leq x_2 \leq ... \leq x_N$ (Hothorn and Lausen, 2003):

$$a_i = \delta_i - \sum_{m=1}^{\gamma_i(t)} \frac{\delta_m}{n - \gamma_m(t) + 1}, \tag{2.23}$$

where

$$\gamma_m(t) = \sum_{i=1}^{N} 1 \{t_i \leq t_m\}, \tag{2.24}$$

is the number of observations which died or were censored before or at time t_i and $1 \{\cdot\}$ denotes the indicator function. The resulting log-rank score test, is given by

$$\text{LRS}(x, c) = \frac{\sum_{x_i \leq c} a_i - n_1 \bar{a}}{\sqrt{n_1 (1 - \frac{n_1}{N}) s_a^2}}, \tag{2.25}$$

where $\bar{a}$, s_a^2 and n_1 are the sample mean, the sample variance of a_i:$i = 1, ..., N$ and the number of samples in the group formed by the split point c. Maximisation of LRS yields to the optimal split point. The selection of splitting candidates and computation of split points is repeated until either the terminal nodes contain no less than a defined number of subjects (node-Size) with unique events or the maximal depth of the tree (max-Depth) is reached. The estimation of the cumulative hazard function for a subject i, which is propagated down the tree based on predictor x_i and stops in terminal node q is calculated by:

$$\hat{H}(t \mid x_i) = \sum_{t_{i,q} \leq t} \frac{d_{i,q}}{\Upsilon_{i,q}}, \tag{2.26}$$

where $t_{i,q}$ defines the distinct failures times, $d_{i,q}$ and $\Upsilon_{i,q}$ are the number of failures and individuals at risk at time $t_{i,q}$ in terminal node q. The estimation of the ensemble cumulative hazard function over all trees (nTree) is given by:

$$\hat{H}_e^*(t \mid x_i) = \frac{1}{\text{nTree}} \sum_{b=1}^{\text{nTree}} \hat{H}_b(t \mid x_i). \tag{2.27}$$

2.2.2. Full-parametric survival models

Full-parametric survival models allow to completely describe the structure (shape) of the time-dependent baseline hazard function. A further advantage is that the extrapolations of the survival function becomes possible. However, full-parametric models require to specify the shape of the expected baseline hazard function by a, e.g., Weibull, exponential or Gaussian distribution. For instance, the Weibull probability density function is given by:

$$f(t) = h(t) \cdot S(t) = \alpha\lambda^{\alpha} t^{\alpha-1} \cdot \mathrm{e}^{-(\lambda t)^{\alpha}}. \tag{2.28}$$

The distribution parameters λ and α defines the scale and the shape of the Weibull distribution are estimated during the training phase. The hazard function of the Weibull distribution in proportional hazard form is given by, (Therneau and Grambsch, 2000):

$$h(t, X, \beta, \lambda) = \lambda\alpha t^{\alpha-1} \cdot e^{\beta X}. \tag{2.29}$$

In this thesis, the survival regression model (SRM) is used as a full-parametric model, which provides different survival functions, e.g., Weibull, Exponential, Gaussian. In contrast to the Cox model, for which the baseline hazard is unknown, this full-parametric regression allows for predicting the time-dependent survival probability for each patient.

2.3. Radiomics risk modelling

Radiomics aims to characterise the tumour phenotype using quantitative imaging features computed and extracted from medical imaging data, e.g., CT imaging data. After feature computation, the resulting features are used to develop prognostic or predictive radiomics risk models, e.g., for predicting OS. Therefore, the radiomics process comprises two main tasks: (I) the feature computation and extraction from the imaging data and (II) the application of machine learning algorithms for risk modelling to predict patient specific outcome. In the following, both processing steps are described more in detail, defining the basis procedure for the further chapters in this thesis.

2.3.1. Feature computation framework

Feature computation consists of a sequence of different operations that are required to derive the imaging features. Figure 2.4 shows the different processing steps, including image pre-processing and feature computation. In general, two inputs are necessary to compute imaging features: the imaging data and the region of interest (ROI), which is delineated either by an human observer (e.g., physician) or generated by an (semi-)automatic segmentation algorithm. Prior to the feature computation the imaging data have to be pre-processed, e.g., to enhance the image quality. For instance, the intensity in MRI scans are often non-uniform, which may affect the expression of the radiomics features and introduce additional variance. Therefore, a reduction of image artefacts is required. In chapter 3 a novel data driven algorithm to correct intensity non-uniformity in MRI data is presented. Furthermore, imaging data are often acquired with different Voxel dimensions across different patients and institutions. Therefore, image interpolation is performed to down- or up-sample the original image to a uniform voxel spacing. In this thesis, trilinear interpolation is used to interpolate the images. Trilinear interpolation uses the intensities of the eight closest neighbouring Voxel in the original (base) image grid to interpolate the intensity using linear interpolation.

Subsequently, to quantify additional image characteristics such as edges and blobs, image transformations (e.g., wavelet transformation) are applied to the base image. This procedure generates additional images for feature computation. Details are described below. In case of image interpolation, the ROI is likewise interpolated to the uniform voxel spacing. Afterwards, a morphological and an intensity mask is generated, based on the ROI. The ROI is re-segmented to cover only soft tissue voxels using a specific defined intensity range to create the intensity mask. For instance, to remove voxels containing air and bone from the ROI in CT scans, an intensity range between -150 and 180 HU is reasonable. The morphological mask is identical to the original (interpolated) ROI. Subsequently feature computation is performed on the processed set of images and by using both generated ROI masks.

Prior to the computation of texture features, intensities are discretised to reduce the calculation time and to suppress image noise. Image intensities are discretised by assigning each intensity g_k of Voxel k to a bin number b in the range $[1, N_g]$:

$$g_{k,b} = \begin{cases} 1 & g_k = g_{\min} \\ \lceil N_g \frac{g_k - g_{\min}}{g_{\max}} - g_{\min} \rceil & g_k \geq g_{\min} \end{cases}, \tag{2.30}$$

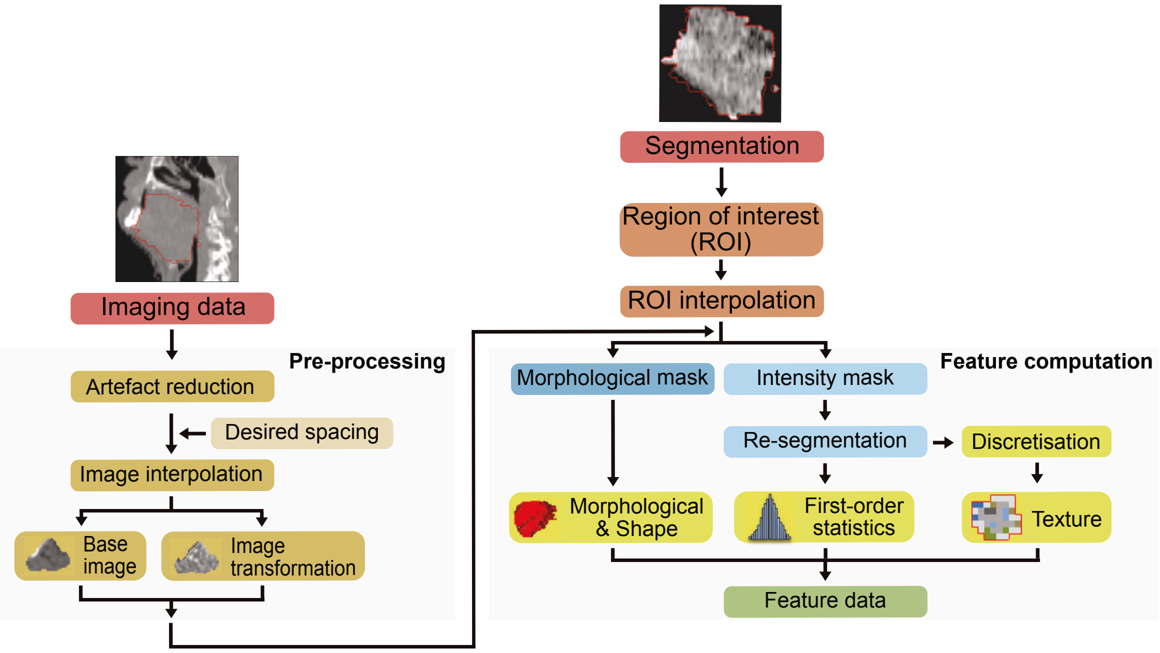


Figure 2.4.: Illustration of the feature computation framework consists of a sequence of different operations, which are required to derive the imaging features. For the computation of the imaging features the imaging data and the region of interest (ROI) are required.

where g_{min} and g_{max} defines the lowest and the highest grey value in the intensity mask.

Imaging features

The radiomics image features, used in this thesis, can be categorised into four major families, namely first-order statistical, morphological, shape and textural-based features. In general, features are extracted and computed from a defined ROI, e.g., from the gross tumour volume (GTV) using the base image or its transformations such as its decimated discrete wavelet transform (UDWT) or laplacian of Gaussian (LoG) images. In the following sections, the three feature groups are shortly described. Further details and the full mathematical description of all used imaging features can be found in Zwanenburg *et al.* (Zwanenburg et al., 2016).

Morphological and shape features

Morphological and shape based imaging features describe geometric aspects of the ROI using the morphological mask of the ROI. For instance, the volume V of the ROI may be approximated by:

$$V = \sum_{j=1}^{N_R} V_j, \tag{2.31}$$

where N_R equals the number of voxels in the morphological mask and V_j the volume of Voxel j. Another feature of this family is the surface area of the ROI. Prior to computing the surface area, the morphological mask is transformed into a triangle mesh, for instance using a meshing algorithm such as the Marching Cubes algorithm (Cubes, 1987; Lewiner et al., 2003). The resulting mesh consists of N_{fc} triangle faces, spanned by N_{vx} vertex points, and is used to calculate the surface area A by summing over the areas A_k of the triangle faces:

$$A = \sum_{k=1}^{N_{fc}} A_k = \sum_{k=1}^{N_{fc}} \frac{\mid ab_k \cdot ac_k \mid}{2}, \tag{2.32}$$

where the edge $ab = b - a$ defines the vector from vertex a to vertex b, and the edge $ac = c - a$ is the vector from vertex a to vertex c.

First-order statistical features

First-order statistical features provide information related to the distribution of intensities in the ROI. For instance, the mean value F_{mean} of a ROI is given by:

$$F_{\text{mean}} = \frac{1}{N_{xyz}} \sum_{i}^{N_{xyz}} I(i), \tag{2.33}$$

where I denotes the ROI mask with N_{xyz} voxels. Additional statistical features were computed based on the intensity histogram of the ROI after discretising intensities.

Texture features

The features computed from the first-order statistics provide information related to the intensity distribution of the ROI. However, they do not incorporate any information about the spatial positions of intensities within the ROI. For example, first-order statistics features are not able to measure whether low intensities are clustered together or are, rather, mixed with high intensities. Texture features enable the description of the tissue heterogeneity observable within the ROI by considering discretised intensities (grey levels) within neighbourhoods. Texture features are based on texture matrices, which can be calculated slice-by-slice (two-dimensional (2D)) or volumetrically 3D. In this thesis several distinct types of texture matrices have been considered: grey-level co-occurrence matrix (GLCM), grey-level run length matrix (GLRLM), neighbourhood grey tone difference matrix (NGTDM), grey-level size zone matrix (GLSZM), grey-level distance zone matrix (GLDZM) and neighbourhood grey level dependence matrix (NGLDM).

The GLCM describes how combinations of discretised grey levels of neighbouring pixels, or voxels in a 3D volume, are distributed along one of the specific image directions (Haralick et al., 1973). In a 3D image, the direct neighbourhood of a voxel consists of the 26 directly

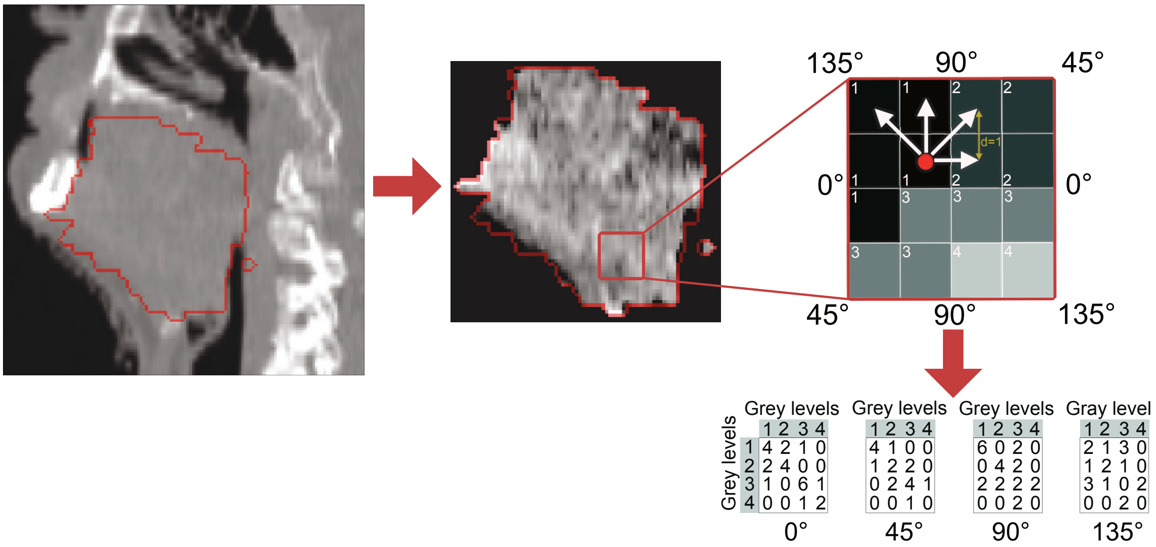


Figure 2.5.: Example for the computation of the the grey-level co-occurrence matrix for a two-dimensional segment (small red square) within a selected part of the tumour (patch) with a neighbourhood distance d = 1 and the resulting four image directions: 0°, 45°, 90° and 135°. Based on the resulting single texture matrices imaging features can be computed.

neighbouring voxels. Thus, there are 13 unique direction vectors within a neighbourhood distance d, e.g., d = 1. Figure 2.5 illustrates the computation of the GLCM texture matrix for a 2D segment within a selected part of the tumour (patch) and the resulting four image directions.

The GLRLM describes the distribution of discretised grey levels by assessing the run lengths within the ROI (Galloway, 1975; Dasarathy and Holder, 1991). A run length is defined as the length of a consecutive sequence of pixels or voxels with the same grey level along a given direction. For instance, in a coarse texture, relatively long gray-level runs would occur more often whereas a fine texture should contain primarily short runs. The NGTDM contains the sum of grey level differences of pixels or voxels with discretised grey level and the average discretised grey level of neighbouring pixels or voxels within a neighbourhood distance d (Amadasun and King, 1989). The GLSZM counts the number of groups (or zones) of linked voxels (Thibault et al., 2014). Voxels are linked when the neighbouring Voxel has an identical discretised grey level. A voxel classifies as a neighbour depends on its neighbourhood. In a 2D image slice a voxel has eight neighbourhood, whereas a voxel within a 3D volume consists of 26 neighbourhood voxels. The GLDZM counts the number of groups of connected voxels with a specific discretised grey level value and the distance to the ROI edge. The matrix captures the relation between location and grey level (Thibault et al., 2009). The NGLDM as an alternative to the GLCM, which quantifies the coarseness of the overall texture and is rotationally invariant (Sun and Wee, 1983).

The above described Calculation of the texture matrices yields to one or more different single matrices, e.g., thirteen matrices for the GLCM, due to specific image directions. There-

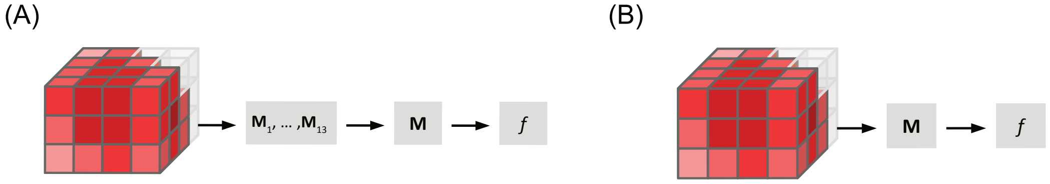


Figure 2.6.: Illustration of the computation of a feature *f* using different single texture matrices *M*. In (A) the different single matrices, e.g., in the case of the grey-level co-occurrence matrix the thirteen image directions are merged by summing the different matrix elements prior to feature calculation and (b) a single matrix is constructed, e.g., using all thirteen directions simultaneous. (figure taken from Zwanenburg *et al.* (Zwanenburg et al., 2016))

fore, texture-based features can be calculated either by averaging the values of the matrices computed for thirteen distinct directions (figure 2.6 (A)) or by a single matrix. These accounts for, e.g., co-occurrence information (GLCM) based on all thirteen directions simultaneous (figure 2.6 (B)) to improve rotational invariance (Depeursinge and Fageot, 2017).

Image transformations

Image transformations enable to emphasise additional image characteristics such as edges and blobs.

Discrete wavelet transform

A given signal or function $f(x)$ can be decomposed using a family of wavelet basis functions. These basis functions are generated from a mother wavelet by dilatation (or scale) and translation operations (Burrus et al., 1998). The decomposition of a given signal or function $f(x)$ by a wavelet system can be represented by the series:

$$f(x) = \sum_k \sum_j a^j_k \psi^j_k(x). \tag{2.34}$$

Here, k and j are integer indices describing the space location (translation) and the scale (resolution), respectively, a^j_k are a set of expansion coefficients called the discrete wavelet transform (DWT) of $f(x)$ and $\psi^j_k(x)$ is a set of real-valued functions of x called wavelet expansion set (Burrus et al., 1998; Usmanij et al., 2013). In contrast to the Fourier transformation, which decomposes a given signal according to its frequency content only, the wavelet expansion maps it into a two-dimensional array of coefficients. This two-dimensional representation allows for localising the signal in both time and frequency.

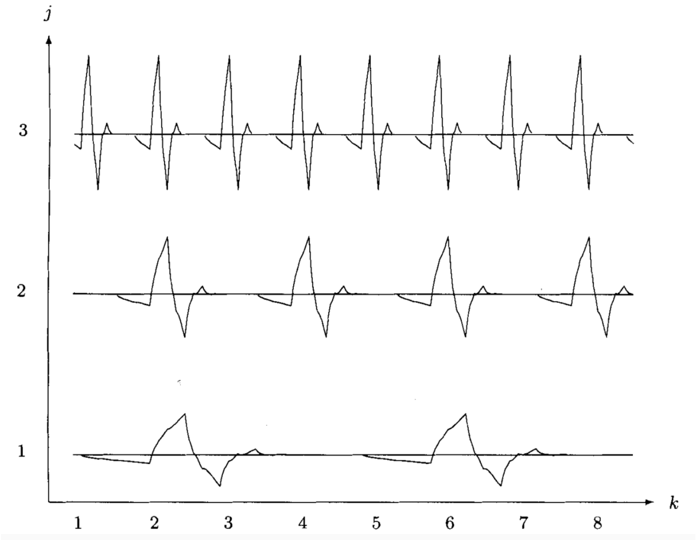


Figure 2.7.: Illustration of the translation (every fourth k) and scaling of a single Daubechies mother wavelet. (figure taken from Burrus *et al.* (Burrus et al., 1998))

The wavelet expansion functions $\psi(x)$ are generated from a single scaling function or wavelet by simple scaling and translation and is given by :

$$\psi_k^j(x) = 2^{j/2}\psi(2^j x - k), \tag{2.35}$$

where all wavelets $\psi_k^j(x)$ are scaled and translated versions of the mother wavelet $\psi(x)$ defined by the integers j and k, respectively. Figure 2.7 illustrates the effect of the translation and the scaling parameters k and j, respectively, of a single mother wavelet. A change of index k leads to a change of the location of the wavelet along the horizontal axis. This allows the expansion to explicitly represent the location of events in time or space. A change of index j changes, the shape of the wavelet in scale. This allows a representation of detail or resolution (Burrus et al., 1998).

Besides of the decomposition of a signal $f(x)$ by wavelet expansion functions generated from a single scaling function $f(x)$ can also be expressed at different scales and spatial locations. This allows for decomposing a signal into increasingly finer details, i.e., a cascade filter. The formulation of such multi-resolution analysis is made in terms of two closely related basis functions: a scaling function $\varphi(x)$ and a wavelet function $\psi(x)$. The scaling function $\varphi(x)$ can be expressed in terms of a weighted sum of translated versions of $\varphi(2x)$ such as, (Burrus et al., 1998):

$$\varphi(x) = \sqrt{2}\sum_{n\epsilon\mathbf{Z}} l(n)\varphi(2x - n), \tag{2.36}$$

where $l(n)$ defines the scaling (low-pass) coefficients. The wavelet function $\psi(x)$ can be represented by a weighted sum of shifted scaling functions $\varphi(2x)$ by:

$$\psi(x) = \sqrt{2}\sum_{n\epsilon\mathbf{Z}} h(n)\varphi(2x - n), \tag{2.37}$$

where $h(n)$ defines the wavelet (high-pass) coefficients. The relation between these coefficient is given by, (Burrus et al., 1998):

$$h(n) = (-1)^n l(1 - n). \tag{2.38}$$

Using both basis functions of (2.36) and (2.37), the decomposition of a signal $f(x)$ into a finite number of scaling levels J becomes:

$$f(x) = \sum_k a_k^{j_0} 2^{j_0/2}\varphi(2^{j_0}x - k) + \sum_k \sum_{j=j_0}^{J} d_k^j 2^{j/2}\psi(2^j x - k), \tag{2.39}$$

where the coefficients $a_k^{j_0}$ represent the approximation of the signal at the lowest level (or scale) J with the scaling function $\varphi(x)$. Thereby $\varphi(x)$ represents the coarse details of the signal, or its low-frequency components. The decomposition coefficients d_k^j are used to represent the fine details of the signal, or its high-frequency components. The coefficients a_k^j and d_k^j at scale j can be expressed in terms of the coefficients of the previous scale using the following recursive equations:

$$\begin{aligned} a_k^j &= \sum_{n\epsilon\mathbf{Z}} a_k^{j-1} l(n - 2k), \\ d_k^j &= \sum_{n\epsilon\mathbf{Z}} a_k^{j-1} h(n - 2k), \end{aligned} \tag{2.40}$$

where $l(n)$ and $h(n)$ defines the set of scaling and wavelet coefficients. The wavelet coefficients a_k^j and d_k^j are obtained by the convolution over space of the scaling and wavelet functions defined at each level j. Figure 2.8(A) shows an example of the coiflet-1 scaling $\psi(x)$ and wavelet $\phi(x)$ functions. The discrete wavelet decomposition leads to a down-sampling of the input signal by a factor of two after each decomposition level J. However, down-sampling of the input signal is sometimes undesired, e.g., in the case of medical imaging data analyses. UDWT represents an alternative to the DWT. It is based on same wavelet theory as previously described. However, instead of the incorporation of downsampling operations, the UDWT inserts zeros after the low- and the high-pass filtering operations (Holschneider et al., 1990), e.g., to preserve the original image size.

The wavelet theory described above has so far involved the transformation of one-dimensional (1D) signals. The 1D multiresolution wavelet decomposition can be extended to two or three dimensions. For instance, the 2D decomposition is performed by introducing 2D scaling and

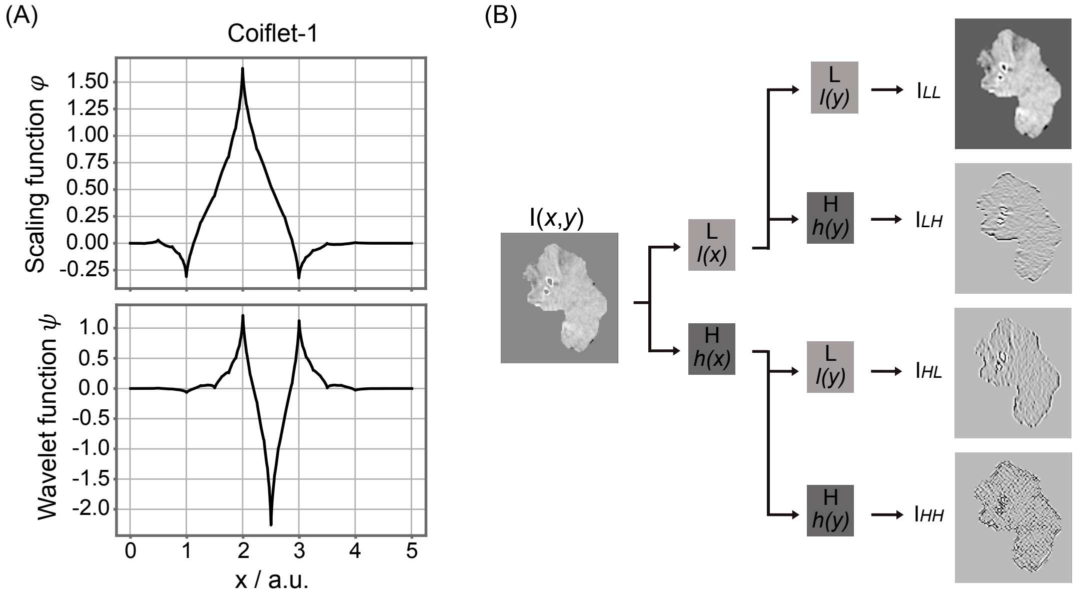


Figure 2.8.: (A) Illustration of the coiflet-1 scaling and wavelet functions. (B) Example of the wavelet transformation using coiflet-1 wavelet applied in *x* and *y* direction of the two-dimensional tumour patch. Coarse image characteristics are described by the low-frequency components, whereas fine details are represented by the high-frequency components.

wavelet functions as tensor product of their 1D complements (Stollnitz et al., 1995) such that:

$$\begin{aligned} I_{L,L}(x,y) &= \phi(x)\phi(y), \\ I_{L,H}(x,y) &= \phi(x)\psi(y), \\ I_{H,L}(x,y) &= \psi(x)\phi(y), \\ I_{H,H}(x,y) &= \psi(x)\psi(y). \end{aligned} \tag{2.41}$$

Performing one level of a 2D-UDWT consists of filtering an image $I(x,y)$ both horizontally and vertically with the 1D low-pass filter (L) ψ and the 1D high-pass filter (H) ϕ. As a result, the wavelet coefficients of four different sub-bands are generated: *LL*, *LH*, *HL* and *HH* (figure 2.8 (B)). In the case of 3D-UDWT the wavelet coefficients of eight different sub-bands are generated: *LLL*, *LHL*, *LHH*, *HLL*,*HHL*, *HLH* and *HHH*.

Laplacian of Gaussian transform

The LoG transformation is an isotropic filter and measures the 2$^{\text{nd}}$ spatial derivative of an image $I(x,y)$. The LoG transformation comprise two image filters: a smoothing operation by a Gaussian filter *G* to reduce the noise in the image, followed by applying the Laplace filter

∇^2 to calculate the second derivative of the image intensity. Subsequently, the LoG image I_{LoG} is the result of the convolution of the filter $\nabla^2 G$ with image I, (Marr and Hildreth, 1980):

$$I_{\text{LoG}} = \nabla^2 G * I. \tag{2.42}$$

The 2D-LoG function is given by:

$$\nabla^2 G(x, y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^4}\right] e^{\frac{-(x^2+y^2)}{2\sigma^2}}, \tag{2.43}$$

where σ defines the spatial scale of the Gaussian filter. A small value for σ may be used to emphasize fine image details, whereas larger σ values highlight coarser image details, respectively.

2.3.2. Risk modelling framework

The development of risk modelling framework (RMF) for decision support is a complex task. Different challenges arise in typical radiomics studies due to the high number of different radiomics features (typically $\gg 1000$) relative to the limited number of data points ($\ll 1000$ patients). In machine learning theory, this problem is called the curse of dimensionality (Bellman, 1961). With an increased dimensionality D the volume of the space increases exponentially, so that the available data become sparse. Therefore, it is reasonable to limit the number of radiomics features which are used to develop accurate and generalisable risk model due to the few number of data points (patients). A further challenge and major pitfall in radiomics risk modelling is model over-fitting. Model over-fitting occurs when the model is closely fit to a set of given data points, e.g., the model captures the noise of the data. This results in a poor generalisation error on new and unseen data. Model over-fitting usually arises when the model complexity is increased beyond an optimal complexity, e.g., by using a set of radiomics features (signature) which contains a high number of imaging features during model development, which is illustrated in figure 2.9.

Therefore, a reduction of the high-dimensional feature vector by feature selection methods and an optimisation of the hyper-parameters of the learning algorithms are an essential steps in risk modelling. The RMF was developed to perform such unbiased and automated radiomics analyses. The RMF can be used to create radiomics signatures, to optimise the hyper-parameters of machine learning algorithms, to train prognostic/predictive models and subsequently to assess the model performance by a variety of different metrics. Figure 2.10 shows the RMF and its five major processing steps: (I) feature pre-processing, (II) feature selection, (III) hyper-parameter optimisation, (IV) model development and (V) model validation. These steps are described in the following paragraphs.

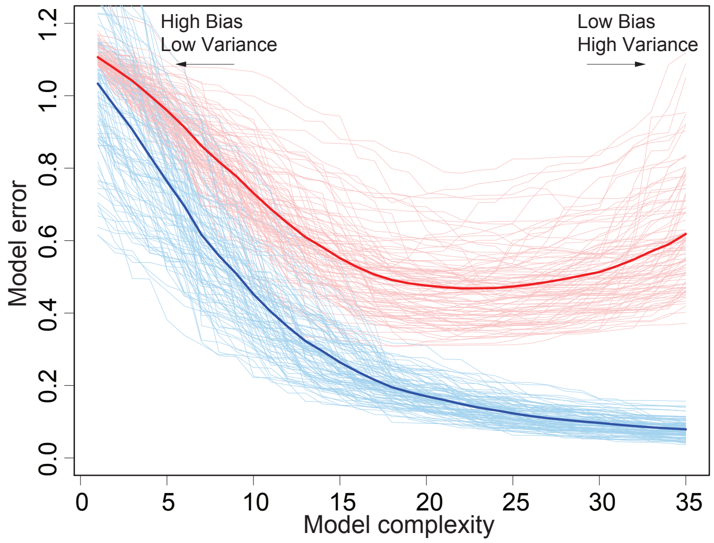


Figure 2.9.: Illustration of the behaviour of training and test prediction error as the model complexity is varied, e.g., by different signature sizes. The light blue curves show the training error, while the light red curves show the test error of bootstrap samples as the model complexity is increased. The solid curves show the expected training and test error. (figure taken from Friedman *et al.* (Friedman et al., 2001))

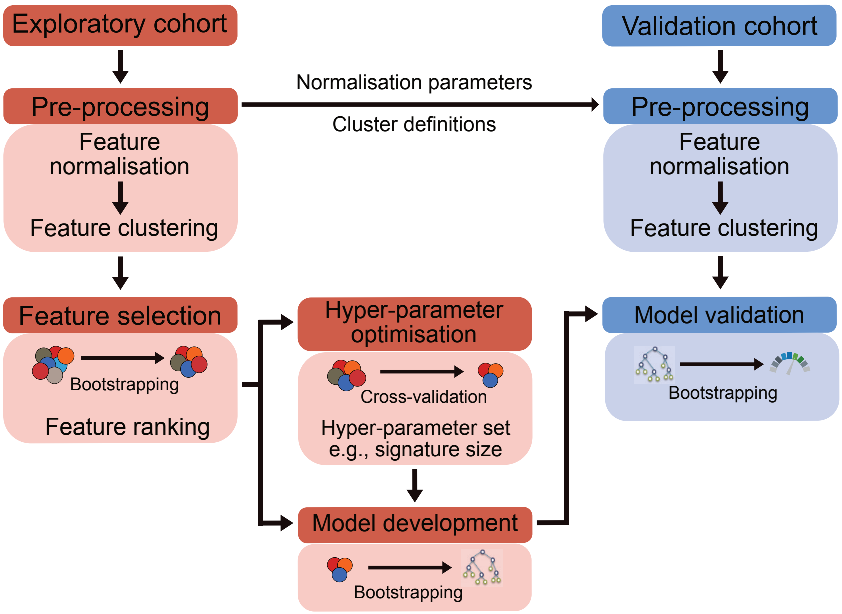


Figure 2.10.: Illustration of the major radiomics processing chain within the risk modelling framework (RMF): (I) feature pre-processing, (II) feature selection, (III) hyper-parameter optimisation, (IV) model development and (V) model validation. Steps I-IV are performed only on the exploratory cohort. Subsequently, the normalisation parameters and the cluster definitions were transferred and applied to the independent validation cohort unchanged. Finally the trained models are validated.

Risk modelling framework architecture

(I) Feature pre-processing

Based on the computed imaging features, the first steps comprise the normalisation of the feature values to a defined range and the clustering of the radiomics features. The range of imaging features may vary widely and a majority of machine learning algorithms will not perform properly without normalisation. Therefore, a widely used normalisation strategy is to standardise the radiomics feature values $x \in X$ to have a zero-mean and a unit-variance, given by:

$$x^{'} = \frac{x - \mu(x)}{\sigma(x)}, \tag{2.44}$$

where $\mu(x)$ and $\sigma(x)$ are the sample mean and the sample standard deviation of feature x, respectively (Jain and Dubes, 1988).

Furthermore, several radiomics features may be highly correlated. Highly correlated features do not provide additional information to a model and may moreover lead to numerically unstable risk models as well as an increase in the overall computational time. Therefore, within feature pre-processing one further objective is to identify an initial non-redundant set of radiomics features (Parmar et al., 2015c). In the proposed RMF, unsupervised clustering is performed using hierarchical agglomerative clustering (HAC) (Langfelder and Horvath, 2012). The HAC is an iterative algorithm which defines each data point to be a cluster. Subsequently, the distance between two clusters is estimated by computing the average distance (or similarity) between the data points in the first cluster and the data points in the second cluster. Based on the distance between the clusters, in each iteration those two clusters will combined which have the smallest average linkage distance. The similarity can be measured, e.g., either by the Pearson or Spearman correlation coefficient. The definition of clusters depends on the clustering height h, which is a threshold that describes the minimum correlation which is required to cluster features into a single cluster. The resulting clusters may be represented by: (a) a new meta-feature calculated by averaging over all features within the cluster, (b) one selected feature which shows the highest correlation with the outcome measured by mutual information criteria and (c) a central feature represented by the feature which is the closest to the cluster centre. Figure 2.11 shows the effect of different clustering heights ($h = 0.9, 0.8, 0.5, 0.3$) and different cluster representations as a function of training and validation performance. The cluster height $h = 0.8$ and the method generating a new meta-feature as cluster representation yields to a good performance on the training and validation cohort. In contrast, no clustering leads to a reduced model performance on the training and validation cohort, indicating that feature clustering is recommended.

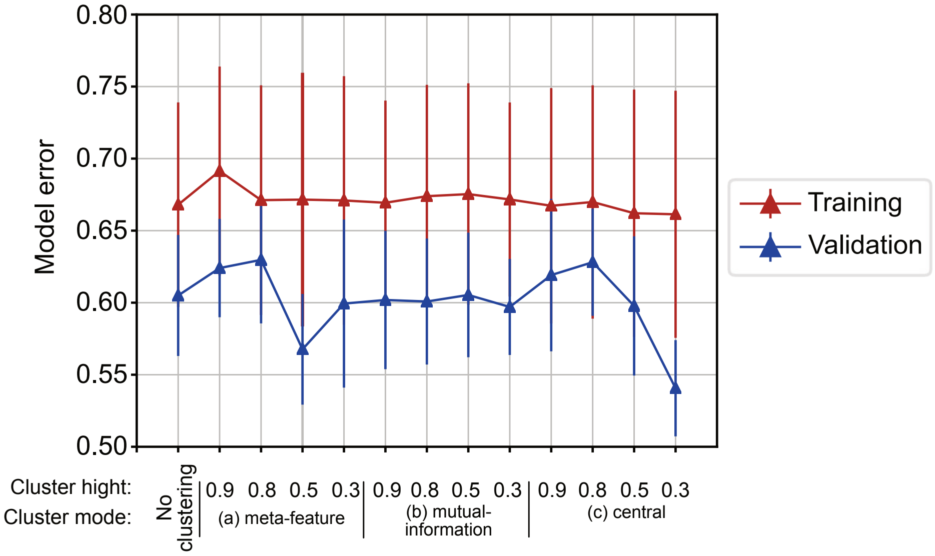


Figure 2.11.: Illustration of the effect of different clustering heights and different cluster representations as a function of training and validation error.

(II) Feature selection

The objective of feature selection is to identify a subset of biomarkers (signature) which is strongly related to the endpoint and robust against perturbations in the data. Several methods may be used to identify features, which are described in chapter 2.5.1.

To identify biomarkers that are stable and robust against data perturbations, feature selection is performed using a bootstrap sampling strategy. Bootstrap sampling randomly selects samples from the training data set, with replacement, and results in the creation of a new data set of the same size as the original data set, but with different contents (Friedman et al., 2001). Bootstrapping is repeated *B* times (e.g., *B*=1000), producing *B* bootstrap data sets where feature selection is performed.

Within each bootstrap data set, features are ranked according to their perceived importance. Subsequently, these results are aggregated to identify a small feature subset. For this purpose, the top *k* best ranking imaging features are selected from each bootstrap sample $b = 1, ..., B$. Afterwards, the selected top biomarkers *f* are aggregated over the bootstraps by calculating an importance score F_f, given by:

$$F_f = \frac{\sum_{b=1}^{B} \sqrt{R_{b,f}}}{occ_f^2}. \tag{2.45}$$

Here, $R_{b,f}$ defines the rank within the *b*-th bootstrap sample (low ranks for important features) and occ_f is the frequency of occurrence of feature *f* over all bootstrap samples. The feature rank aggregation score above is based on the enhanced Borda score (Wald et al., 2012), with the difference that feature occurrence receives a greater weight.

(III) Hyper-parameter optimisation

After feature selection and rank aggregation, hyper-parameters of the machine learning algorithms, such as the signature size and other algorithm-specific settings require optimisation. A major objective of hyper-parameter optimisation is to limit model over-fitting and to adjust the model parameters to the prediction task. The individual hyper-parameter set Ω of each learning algorithm A is tuned using an internal cross-validation performed on the training data. Cross-validation uses a part of the available data to train the model and the remaining part to test the performance (Friedman et al., 2001). The basic idea is to split the training data into K equal-sized sub-samples. Subsequently, a single sub-sample is retained as a validation sample for testing the model, and the remaining $K-1$ sub-samples are used to train the model. The cross-validation (CV) procedure is then repeated K times, with each of the K sub-samples being used exactly once for validation. Furthermore, the whole cross-validation process is repeated N_{cv} times using the trainings data. In this thesis, hyper-parameter optimisation is performed using a grid search through a pre-defined hyper-parameter space with k = 2 and N_{cv}=40. The objective of the hyper-parameter optimisation is to find a hyper-parameter configuration Ω^* which minimises the performance differences $L_H(A(X),A(Y))$ between the internal training and validation folds, X and Y, respectively, by a trained learning algorithm A, given by:

$$\Omega^* = \arg\min_{\omega \in \Omega} L_H(A_\omega(X), A_\omega(Y)). \tag{2.46}$$

The minimisation criterion L_H is defined by three penalties P,

$$L_H(\mu_X, \mu_Y) = P_1(\mu_X, \mu_Y) + P_2(\mu_{\text{bal}}) + P_3(\mu_X, \mu_Y). \tag{2.47}$$

The term P_1 defines a penalty for the training and test errors, which is given by:

$$P_1(\mu_X, \mu_Y) = \max\left(\frac{1-\alpha}{(\mu_X-\alpha)^2} - 1, \frac{1-\alpha}{(\mu_Y-\alpha)^2} - 1\right), \tag{2.48}$$

where μ_X and μ_Y define the average prediction accuracy of the internal training (X) and validation (Y) folds. α=0.5 is a correction factor, representing the performance of a random experiment. The term P_2 penalises differences between training and test error, given by:

$$P_2(\mu_{\text{bal}}) = 50\left(\frac{1}{(1-\mu_{\text{bal}})^2} - 1\right), \tag{2.49}$$

where μ_{bal} is the difference in average prediction performance between training (X) and validation (Y) folds. This leads to hyper-parameter sets where training and test error are more similar (balanced), which may increase model generalisability. The penalty term P_3 accounts for the discordance in train μ_X and test μ_Y errors, to avoid selecting hyper-parameters where the predictions on the training set are concordant with the outcome, and discordant on the validation set neither. P_3 is given by:

$$P_3 = \begin{cases} 1000 & (\mu_X > \alpha \vee \mu_Y < \alpha) \wedge (\mu_X < \alpha \vee \mu_Y > \alpha) \\ 0 & \text{otherwise} \end{cases}. \tag{2.50}$$

(IV) Model development

Model development is performed using the imaging features, ordered by aggregated importance, and the optimised hyper-parameter set. Model training is conducted using the entire training data set once. In addition, it is performed multiple times (e.g., m=1000) using bootstrap samples (i.e.,.632 bootstrap method with replacement) of the training data to enhance model robustness. Afterwards, an ensemble prediction is made by averaging the predicted risk scores for every model using the validation data (Dietterich, 2000) . In this thesis different types of machine learning algorithms are used. Therefore, chapter 2.5.2 briefly describes the basic concepts of the used machine learning algorithms.

2.4. Performance assessments

The RMF offers different performance metrics to assess the quality of the risk models using internal or external validation data. In the case of continuous time-to-event survival data, the model performance is usually measured by: I) the concordance index (C-Index) and II) Kaplan-Meier analyses. Both metrics are described shortly in the following.

Concordance Index

The C-Index is a commonly used to assess the performance of time-to-event survival models (Harrell et al., 1996b; Pencina and D'Agostino, 2004). For the C-Index calculation, only those pairs of subjects are included, where at least one had an event, resulting in either event and event or event and non-event comparisons. The C-Index is measured as the proportion of all usable pairs in which the risk predictions and outcomes are concordant. It denotes by *r* the predicted risk of the model and *t* the given survival time for a pair of subjects *i* and *j*, the C-Index is given by:

$$C = \frac{\pi_c + 0.5 \cdot \pi_t}{\pi_c + \pi_d + \pi_t}, \tag{2.51}$$

where

$$\begin{aligned}\pi_c &= P(r_i > r_j \wedge t_i < t_j) + P(r_i < r_j \wedge t_i > t_j),\\ \pi_d &= P(r_i > r_j \wedge t_i > t_j) + P(r_i < r_j \wedge t_i < t_j),\\ \pi_t &= P(t_i = t_j),\end{aligned} \tag{2.52}$$

define the probability of concordance and of discordance as well as proportion of pairs that are undesired, respectively. The value of the C-Index ranges between 0 and 1: a value of 0.5 indicates that the model has no ability to discriminate between low and high risk subjects, whereas the values 0 and 1 indicate that the model can perfectly discriminate between these subjects.

In the majority of radiomics studies the C-Index ranges approximately between 0.5 and 0.7 (Aerts et al., 2014; Coroller et al., 2015). Therefore, in this thesis the following rating scale is defined: a C-Index$\leq$0.55 indicates a close to random, a C-Index between 0.55 and 0.6 a moderate, a value between 0.6 and 0.65 a good and a C-Index$>$0.65 a high prognostic performance.

Risk-based patient stratification

Besides assessing model performance using the C-Index, model performance may be assessed by patient stratification into risk groups (e.g., low and high risk). These risk groups are based on the estimated log hazard ratios or predicted survival times and formed by setting a cut-off value. This value is usually determined on the training data. One straightforward method for setting the cut-off is by computing the median risk value. Alternatively, more complex schemes may be used, for instance by bootstrapping. In the latter case different bootstrap samples are generated from the training data. Afterwards, each potential cut-off is applied to the bootstrap samples and the statistical differences between the risk groups is measured. Subsequently, the fraction of significant stratification results (power) is calculated for each cut-off, leading to the optimal value which has the largest power (Linge et al., 2016b). After patient stratification, survival curves are estimated using the Kaplan-Meier method. The Kaplan-Meier estimator is non-parametric statistic to estimate the survival function (Kaplan and Meier, 1958). The estimator is the product over the failure times of the conditional probabilities of surviving to the next failure time and is given by,

$$\hat{S}(t) = \prod_{t_i \leq t} \left[1 - \frac{d_i}{n_i} \right], \tag{2.53}$$

where n_i and d_i defines the number of individuals at risk at time t_i and the number of individuals who fail at this time, respectively. Subsequently, to quantify the statistical differences between the resulting risk groups, the log-rank test is applied (Mantel, 1966; Cox, 1972).

2.5. Feature selection methods and machine learning algorithms

This section contains a short description of the different feature selection methods and machine learning algorithms that were used in this thesis. All described methods and algorithms are able to process continuous time-to-event survival data.

2.5.1. Feature selection methods

The feature selection methods described in this section comprise simple linear methods as well as advanced non-linear algorithms. The different hyper-parameters of the individual methods are listed and defined in the appendix A. However, these parameters were kept fixed and were not optimised during hyper-parameter optimisation, to reduce the overall computational time for a Radiomics analysis.

Pearson correlation coefficient

The Pearson correlation coefficient is a measure of linear dependency between two random variables. The correlation coefficient ρ for a feature $x \in X$ and the corresponding outcome Y is defined as (RJ and Nicewander, 1988),

$$\rho(x, Y) = \frac{Cov(x, Y)}{\sigma_x \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{Y})^2}}. \tag{2.54}$$

Here, $Cov(x, Y)$ defines the covariance matrix and σ_x and σ_Y the standard deviations of covariate x and outcome Y, respectively. In the case of time-to-event survival data, only those times Y for which an event occurred ($\delta = 1$) are taken into account. The correlation coefficient ρ ranges from -1 and +1, where 1 signifies perfect linear correlation, 0 no correlation and -1 perfect anti-correlation.

Spearman correlation coefficient

The Spearman correlation coefficient provides a non-parametric measure of rank correlation between two variables (Spearman, 1910). Spearman correlation is based on the Pearson correlation coefficient in (2.54) using ranked variables $x \epsilon X$ and Y which is given by,

$$\rho(x, Y) = \frac{Cov(rank(x), rank(Y))}{\sigma_{rank(x)}\sigma_{rank(Y)}}. \tag{2.55}$$

Like the Pearson correlation, Spearman correlation for time-to-event survival data is calculated using only those observations where an event occurred.

Mutual information maximisation

The mutual information maximisation (MIM) method estimates the relevance of feature $x \epsilon X$ for the corresponding outcome Y using a linear approximation based on the correlation ρ. The mutual information MI is defined as (Gelfand and Iaglom, 1959),

$$\mathrm{MI}(x, Y) = -\frac{1}{2} \ln(1 - \rho(x, Y)^2). \tag{2.56}$$

In the case of survival outcome data $\rho(x, Y) = 2(\text{C-Index} - 0.5)$ is based on the C-Index, including a correction for C-Index<0.5. For mutual information of continuous data, the Spearman correlation coefficient is used.

Mutual information feature selection

The mutual information feature selection (MIFS) algorithm (Battiti, 1994) is based on a greedy search and selects a subset of features $S \epsilon X$ which maximises the objective function f:

$$f(X, Y) = \arg \max_{x \epsilon X} \left(MI(x, Y) + \beta \sum_{s_j \epsilon S} I(x, s_j) \right). \tag{2.57}$$

Here $I(x, s_j)$ is the mutual correlation between features x and s_j, as before. The parameter β was set to 1 in this thesis (Battiti, 1994).

Minimum redundancy maximum relevance

The minimum redundancy maximum relevance (MRMR) algorithm (Peng et al., 2005) combines two constraints: maximal mutual information between the features in feature subset $S \epsilon X$ and the outcome Y, combined with minimal redundancy between the features in S. This is done by selecting the feature that maximises f using an incremental search method that is based on the mutual information MI,

$$f(X, Y) = \arg \max_{x \epsilon X \setminus S} \left(\mathrm{MI}(x, Y) - \frac{1}{\mid S \mid} \sum_{s_j \epsilon S} MI(x, s_j) \right). \tag{2.58}$$

Univariate- and multivariate-Cox-regression

The Cox proportional hazard regression model is trained for each feature (univariate) or a subset of features (multivariate) to predict outcome using a k-fold cross validation scheme which was repeated n times in the exploratory cohort. The resulting features are ranked according to the C-Index of the prognostic performance of the univariate or multivariate model.

Random forest minimal depth

The random forest minimal depth (RF-MD) is a variable importance algorithm that assesses feature importance by assessing the distance (depth) of each covariate relative to the root node over all trees (Ishwaran et al., 2010; Ishwaran et al., 2011). The algorithm assumes that covariates which occur at low depths, are more important for the model than those in distant nodes.

Random forest variable hunting

The random forest variable hunting (RF-VH) algorithm uses training data from a stratified k-fold subsampling to fit a random forest by m randomly selected covariates (Ishwaran et al., 2008). The m features are ordered by increasing minimal depth and are added sequentially until the joint variable importance (VIMP) no longer increases. The VIMP is calculated by permuting a covariate (i.e., noising it up) and then calculating the change in prediction error, between the original forest and the noised-up forest predictor. The process is repeated n times and features are ranked by average minimal depth.

Random forest variable importance

The random forest variable importance (RF-VI) algorithm is similar to the RF-VH. However, features are ranked by the VIMP score, described above (Ishwaran et al., 2008).

Maximally selected rank statistics random forest variable importance

The maximally selected rank statistics random forest variable importance (MSR-RFVI) algorithm computes the maximally selected rank statistics for each candidate covariate as follows (Wright et al., 2016). A split point is considered optimal if the separation of the survival curves in two groups is maximised. The linear rank statistic for a split point μ is the sum of all log-rank scores $a_1, \ldots, a_n$ in the group with $x_i \leq \mu$, $x \epsilon X$,

$$\begin{aligned} S_{n\mu} &= \sum_{i=1}^{n} 1_{x_i \leq \mu} a_i, \\ a_i &= \delta_i - \sum_{j=1}^{\gamma_i(T)} \frac{\delta_j}{(n - \gamma_j(T) + 1)}. \end{aligned} \tag{2.59}$$

Here $T = (t_1, \dots, t_n)$ are the survival times, δ is the censoring indicator and $\gamma_j(T) = \sum_{i=1}^{n} 1_{T_i \leq T_j}$ is the number of observations with survival time up to T_j. To compare different splits, the score test statistic is given by,

$$T_{n\mu} = \frac{S_{n\mu} - E_{H_0}(S_{n\mu} \mid a, X)}{\sqrt{\mathrm{Var}_{H_0}(S_{n\mu} \mid a, X)}}, \tag{2.60}$$

where E_{H_0} and Var_{H_0} define the expected value and the variance under the null hypothesis. The null hypothesis assumes there is no influence of a split by the cut point μ on the distribution of Y is given by: $H_0 : P(Y \leq y \mid X \leq \mu) = P(Y \leq y \mid X > \mu)$ for all μ and all y. The obtained p-value for the maximally selected rank statistic is used to rank each covariate.

Permutation variable importance random forest

The permutation variable importance random forest (PVI-RF) algorithm partitions the data randomly into two sets of equal size. Each set is used to construct a random forest (Janitza et al., 2015). The two forests are used to compute the importance of the hold-out observations for each covariate. The null distribution $\hat{F}$ is constructed afterwards based on variables that are likely non-relevant (i.e., with negative or zero importance scores). Based on $\hat{F}$, a p-value for a covariate $x \epsilon X$ is derived as,

$$p_x = 1 - \hat{F}(x). \tag{2.61}$$

Finally the VIMP are ranked according their corresponding p-value.

2.5.2. Machine learning algorithms

this paragraphs presents further basic concepts of the machine learning algorithms which were used in this thesis besides of the Cox and RSF algorithms previously described in section 2.2.1. These algorithms comprise semi- and full-parametric survival algorithms. Furthermore, the model-specific hyper-parameters of the individual algorithms are listed and defined in the appendix B. These hyper-parameters are used during the hyper-parameter optimisation, previously described in section 2.3.2.

Regularised Cox proportional hazard model

The regularised Cox proportional hazard model (NET-Cox) is based on the Cox model and uses additional penalisation terms (Simon et al., 2011). This penalised constraint $P_\alpha(\beta)$ is a mixture of the L_1 (lasso) and L_2 (ridge regression) penalty, which is used to maximise the scaled log partial likelihood

$$\hat{\beta} = \arg, \max_{\beta} \left[\frac{2}{N} \left(\sum_{i=1}^{N} \delta_i \left(\sum_{j=1}^{p} \beta_j X_{ij} - \log \left(\sum_{k \epsilon R_i} e^{\sum_{j=1}^{p} X_{kj}\beta_j} \right) \right) \right) - \lambda P_{\alpha}(\beta_j) \right], \qquad (2.62)$$

where

$$\lambda P_{\alpha}(\beta) = \lambda \left(\alpha \sum_{j=1}^{p} \mid \beta_j \mid + \frac{1}{2}(1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \right). \qquad (2.63)$$

Here, β_j defines the coefficients of the covariates X_j of subject i. The lasso term penalise the model optimisation, e.g., when two covariates are very correlated by picking one and entirely ignore the other. The ridge regression term scales all the coefficients towards zero but sets none to exactly zero. This helps to regularise the model training when the number of covariates are larger than the number of samples, but does not give a sparse solution. Furthermore, in the case that two covariates are correlated predictors than ridge regression will tend to give them equal weights (Simon et al., 2011).

Boosted gradient linear and boosted tree models

The motivation of boosting is to produce a prediction model $G(X)$ based on a sequence of weak prediction models $G_m(X), n = 1, ..., M$ in an iterative fashion. The final prediction model $G(X)$ is given by,

$$G(X) = \text{sign}(\sum_{m=1}^{M} \alpha_m G_m(X)). \qquad (2.64)$$

Here $\alpha_1, ..., \alpha_M$ are weights computed by the boosting algorithm which weights the contribution of each respective $G_m(X)$ (Friedman et al., 2001). The weights give a higher influence to the more accurate weak prediction model G_m. The initialisation step trains the model on the data in the usual manner. In each further iteration $m = 2, ..., M$ the weights are individually modified. At step m the weights of those observations that were incorrectly predicted by the model $G_{m-1}(X)$ at the previous step are increased, whereas the weights are decreased for those that were correctly predicted. A more general formulation using basis function expansions leads to the form,

$$G(X) = \sum_{m=1}^{M} \beta_m b(X, \Theta_m), \qquad (2.65)$$

where $\beta_m, m = 1, ..., M$ are the expansion coefficients and $b(X, \Theta_m)$ are the weak (or base) learners of the covariates X, characterised by a set of parameters Θ. Typically, the training of these models (2.65) are fitted by minimising a loss function L averaged over the training data,

$$\min_{\{\beta_m,\Theta_m\}_1^M} \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M} \beta_m b(x_i, \Theta_m)). \tag{2.66}$$

In this thesis, different types of boosted gradient linear model (BGLM) and boosted tree (BT) models were used: BGLM-Cox, BGLM-CIndex and BT-Cox, BT-CIndex, respectively as well as the full-parametric models: BGLM-Weibull and BT-Weibull.

Maximally selected rank statistics random forest

The maximally selected rank statistics random forest (MSR-RF) algorithm is based on an improved split point criterion to reduce split point selection bias (Wright et al., 2016). In MSR-RF a split point is considered optimal if the separation of survival curves in the two groups is maximised. The standard split criterion for the RSF is the logrank or the logrankscore test statistic. In contrast, the MSR-RF uses either the maximally selected rank statistics (maxstat) or the Harrell's C statistics (C) for split point selection (Harrell et al., 1996a). As described above, the covariate with the lowest *p*-value is selected as splitting candidate. If the adjusted *p*-value is not smaller than the threshold α, no splitting is performed.

3. A physical correction model for automatic correction of intensity non-uniformity in magnetic resonance imaging

Medical imaging is one of the major tasks in medical science and treatment. In particular, MRI is an established non-invasive imaging technique for clinical diagnostics and in radiation oncology. However, MR images may be influenced by artefacts caused by different sources. One of the most frequent artefact is intensity non-uniformity (Bias) induced by a number of factors, such as magnetic field inhomogeneity caused by the choice of the RF coil or patient anatomy (Bellon et al., 1986; Condon et al., 1987; Simmons et al., 1994; Herrick et al., 1997; Krupa and Bekiesińska-Figatowska, 2015).

Intensity non-uniformity occurs as a smooth intensity variation across the image, such that the intensity of the same tissue changes within the image region. This degrades the quality of acquired data. For a human observer it is usually difficult to perceive, whereas automatic image segmentation or registration algorithms are very sensitive to such variations of image intensities, as shown in figure 3.1. Likewise, the prediction results of radiomics risk models may be negatively influenced. In particular, the disturbed tissue intensities may lead to large variations in the expression of radiomics features. Consequently, the risk models may not predict the patient outcome correctly. Therefore, a reduction of intensity non-uniformity prior to performing automated and quantitative image analyses is required (Kickingereder et al., 2016; Milletari et al., 2016; Lao et al., 2017; Li et al., 2017). Hence, a new data-driven approach motivated by the physical properties of a typical MRI coil system (e.g., head coil) was developed to correct intensity non-uniformity. One advantage of the proposed approach is the application of only smooth and gradual intensity corrections due to the derived correction model and the introduced penalty concept. This helps to enhance the image quality and preserves a too strong intensity correction compared to other methods. The work presented within this chapter has been published in an international journal (Leger et al., 2017a) and was presented at several international conferences (Leger et al., 2014; Leger et al., 2015).

3.1. Intensity non-uniformity correction methods

Several methods have been developed to correct Bias in MRI during the last years. These methods can be classified into two major groups: prospective and retrospective methods (Vovk et al., 2007). Prospective methods assume that intensity non-uniformity is caused by a systematic error of the MRI acquisition process. Thus, additional information about the non-uniformity is acquired by either measuring homogeneous objects or additional images of the same object using different coils (Liney et al., 1998; Collewet et al., 2002). Nevertheless, a disadvantage of the prospective methods is that they do not consider patient dependent inhomogeneities.

Retrospective correction methods are applied after the image acquisition and can be categorised as: (I) filtering methods, (II) surface fitting methods, (III) segmentation methods and (IV) histogram based methods. (I) Filtering methods are based on the assumption that the intensity non-uniformity is a low-frequency artefact, which can be extracted by a low-pass filter from the high-frequency content (Cohen et al., 2000; Zhou et al., 2001; Chang et al., 2017). For instance, George *et al.* proposed a 2D non-iterative multi-scale approach using Log-Gabor filter bank (George et al., 2017). (II) Surface fitting methods fit a parametric surface, typically modelled by a polynomial or spline function to different image features which contain information about intensity non-uniformity (Dawant et al., 1993; Zhuge et al., 2002; Milles et al., 2007). (III) Segmentation based intensity non-uniformity correction methods perform the segmentation and the correction task simultaneously, which benefit from each other to yield a better segmentation and a corrected image (Wells et al., 1996; Van Leemput et al., 1999). For instance, Ivanovska *et al.* presented a level-set based approach for simultaneous intensity non-uniformity correction and segmentation of MR images (Ivanovska et al., 2016). Segmentation-based Bias correction methods usually depend on the segmentation accuracy which may lead to good correction results in the case of homogeneous tissue structures. For instance, MR scans of the human brain consist mainly of grey-matter (GM), white-matter (WM) and cerebro-spinal fluid (CSF). However, for the correction of more heterogeneous tissue regions, e.g., in human abdominal scans, the correction may lead to poor results due to an imprecise segmentation. (IV) Histogram-based methods estimate the correction function directly from the image intensity histograms. A typical strategy is based on an iterative deconvolution approach which attempts to maximise the high frequency content of the tissue intensity distribution (Sled et al., 1998; Vovk et al., 2006; Dzyubachyk et al.,

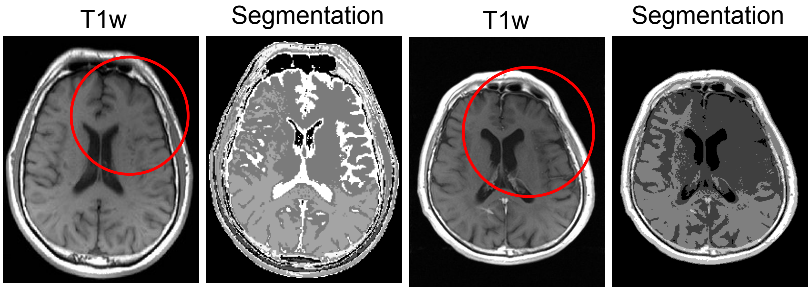


Figure 3.1.: Examples of two uncorrected T1-weighted (T1w) image slices and the corresponding segmentation results using the fuzzy c-means segmentation algorithm (Bezdek et al., 1984). The influence of intensity non-uniformity is apparent in the direction from anterior to posterior (red circle), which degrades the quality of acquired data and the segmentation results.

2013). For instance, a well-known and widely established correction approach is the N4 algorithm (Tustison et al., 2011).

3.2. Physical correction model

The proposed physical correction model (PCM) is based on the assumption that the effect of intensity non-uniformity in MRI occurs due to the image signal emitted by the tissue which is slowly decreasing towards the center of the coil array. Furthermore, we hypothesise that this decrease is basically caused by damping of the RF intensity emitted by the coil and the tissue response. To support this hypothesis we performed an MRI experiment using a cylindrical water phantom, which is supposed to have uniform intensity in the image region. The image volume was acquired with a 1.5 Tesla MRI scanner (Siemens Avanto) using a typical MRI receiver head coil array with eight single coil segments. Figure 3.2(A) illustrates the 3D-view of the used water phantom. Due to the fixed geometry of the head coil the phantom lies close to the lower coil segments, while there exists an intermediate space between the upper segments and the phantom. This leads to a higher image signal around the lower coil segments (figure 3.2(A), orange arrows) and a lower image signal close to the upper coil segments (figure 3.2(A), blue arrow). Furthermore, it is observable that the image signal is slowly exponentially decreasing in the direction towards the centre of the coil array. We quantified this assumption by plotting the intensity values along the y- and in longitudinal z- direction, which are shown in figure 3.2(B-C). The signal decrease from the lower coil segments to the image centre can be described by an exponential function (figure 3.2(B)) with sufficient accuracy, which is motivated by the damping effects of body tissue, whereas in longitudinal z-direction we assume a Gaussian like intensity profile (figure 3.2(C)).

3.2.1. Correction strategy and model definition

To correct intensity non-uniformity in MRI scans, an established formation model is a simplified multiplicative approach (Axel et al., 1987; Dawant et al., 1993). According to this approach, the acquired image $I(x, y, z)$ is obtained by:

$$I(x, y, z) = I'(x, y, z) \cdot f(x, y, z) + \xi(x, y, z), \tag{3.1}$$

where (x, y, z) is the spatial position, I' is the desired uniform image emitted by the tissue, f is an unknown non-uniformity function and ξ describes independent additive noise. The noise will be neglected in the following considerations. The multiplicative model (3.1) can be used to obtain the uniform image I' which is emitted by the tissue,

$$I'(x, y, z) = \frac{I(x, y, z)}{f(x, y, z)}. \tag{3.2}$$

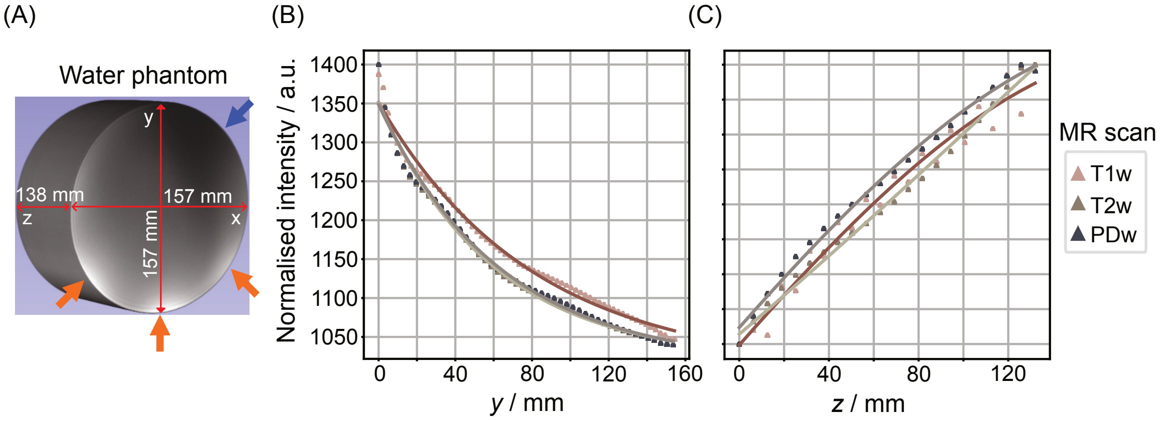


Figure 3.2.: (A) Three-dimensional view of the used water phantom. The phantom lies close to the lower coil segments of the head coil leading to a higher image signal around the lower coil segments (orange arrows) and a lower image signal close to the other coil segments (blue arrow). Intensity profiles are shown for T1- (T1w), T2- (T2w) and PD-weighted scans of the water phantom plotted in *y*- (B) and longitudinal *z*-direction (C). The measured data (triangles) are fitted by an exponential (B) and a Gaussian function (C).

The derivated physical correction model *f* is based on the experimental results as previously described. The gradually decreasing image signal to the coil centre (x_0, y_0) is modelled by an exponential base function for each coil segment $i = 1...n$,

$$f(x,y,z) = f_z(z) \cdot \sum_{i=1}^{n} \mathrm{e}^{-a_i\sqrt{(f_{ix}(z)+x)^2+(f_{iy}(z)+y)^2}}. \tag{3.3}$$

The exponential decay rate of coil segment *i* is described by a_i. The functions f_{ix} and f_{iy} describe the geometric location of coil segment *i* and are given by

$$f_{ix} = \cos(i\alpha+\omega)\cdot d_i - x_0 - S_x(z),$$
$$f_{iy} = \sin(i\alpha+\omega)\cdot d_i - y_0 - S_y(z),$$

where d_i is the distance from image centre to coil *i*, α is a constant angle between the coil segments *i* and ω is the angular shift, describing a global rotation angle of the entire coil array. Furthermore, linear shifts S_x and S_y in *z*-direction have been included in the correction function (3.3) to compensate horizontal and vertical shifts of the patient due to the positioning on the scanner table,

$$S_x(z) = s_x \cdot z + v_x,$$
$$S_y(z) = s_y \cdot z + v_y,$$

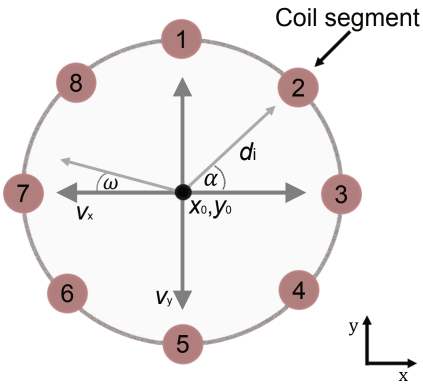


Figure 3.3.: Illustration of the geometric parameters of the physical correction model (PCM) for a typical magnetic resonance imaging head coil consisting of eight single coil segments.

where s_x and s_y define the slope and v_x and v_y the intercept in x- and y-direction, respectively. The intensity non-uniformity in longitudinal z-direction is described by a Gaussian base function f_z,

$$f_z(z) = \frac{q_1}{\sqrt{2\pi\sigma_1^2}} \mathrm{e}^{-\frac{(z-v_z)^2}{2\sigma_1^2}}, \tag{3.4}$$

in which v_z describes the shift in z-direction and σ_1 the standard deviation of the Gaussian function. The parameter q_1 is a global pre-factor to scale the whole correction function. Figure 3.3 shows a schematic 2D-view of a typical MRI head coil as well as the geometric parameters of the introduced model. For this head coil array consisting of eight single coil segments ($n = 8$), the proposed correction model has 27 degrees of freedom in total. Three of these parameters, x_0, y_0 and $\alpha = 360 \cdot n^{-1}$ are given by the geometry of the coils which can be extracted, e.g., from the meta information of the image file.

For the compensation of intensity non-uniformity im MR images using (3.2), parameters of the correction function f (3.3) are determined by maximisation of the image information. This is expressed by the fitness function F,

$$F(I) = E(I') + \epsilon, \tag{3.5}$$

where $E(I')$ is the Shannon entropy and ϵ is an additional penalty term. The Shannon entropy $E(I')$ is given by (Shannon, 1948):

$$E(I') = -\frac{1}{X} \cdot \sum_{b=1}^{B} p_b \cdot \log_2(p_b). \tag{3.6}$$

It is based on the intensity distribution of the MRI scan I' computed by a grey value histogram. Furthermore, it contains the volume X of the spatial domain as normalisation factor, the number of bins B of the histogram and $p_b \cdot \log_2(p_b)$ represents the joint probability of bin value b. The Shannon entropy is a non-negative value and reaches its maximum if all grey levels are equally distributed. The penalty term ϵ prevents an excessive image correction and consists of two different penalty functions,

$$\epsilon = P_1 + P_2, \tag{3.7}$$

which will be explained in the next section.

The entire workflow of the correction process is depicted in figure 3.4. First, the fore- and background regions of the uncorrected image I, are identified using the Otsu threshold algorithm, which sets a threshold based on the image histogram (Otsu, 1975). Since the background regions do not provide additional information concerning intensity non-uniformity, they are subsequently removed to reduce computation time. For further reduction of the computation time, the uncorrected image is down-sampled (e.g., by factor 4) prior to the image correction. Subsequently, the image is corrected through an iterative process that optimises and fits the model parameters according to the optimisation criteria in (3.5). Due to the high dimensionality of the proposed physical correction model (PCM) in (3.3), a swarm intelligence optimisation algorithm is used, called artificial bee colony (ABC) (Karaboga, 2005). The ABC algorithm is inspired by the collective behaviour of a honey bee swarm and is suitable for the numerical optimisation of high dimensional functions(Karaboga and Basturk, 2007; Karaboga and Basturk, 2008). The whole optimisation process is executed multiple times, e.g., $p_{\mathbb{N}}$=10, to reduce the negative influence of non-optimal random initialisations of the model parameters. The optimisation process stops if one of the following stopping criteria are reached: (I) the ratio of the estimated optimum values between subsequent optimisation iterations is smaller than a defined convergence threshold (e.g., 0.0001) or (II) the maximal number of iterations is reached. Subsequently, the final model parameters are selected according to the best global optimum determined by the optimisation process. In order to further improve the image quality, the entire correction process starts again using the corrected image I' until the maximal number of correction iterations K are reached.

3.2.2. Model parameter constraints

In order to reduce the high dimensional parameter space of the correction function (3.3), valid ranges for each model parameter were defined. These ranges were derived from the coil geometry and from results of the initial water phantom experiments. The geometric parameters of the model (3.3) are depend on the size of the voxel grid of the MRI scan. The limits for a $1 \times 1 \times 1$ mm^3 voxel grid size are defined as follows: v_x = v_y = [-75.0, 75.0] mm, d_i = [100.0, 300.0] mm, ω = [-360.0n^{-1}, 360.0n^{-1}], s_x = s_y = [-0.5, 0.5] mm, q_1 = [6.0, 3000.0] mm, a_i = [0.0, 0.05] mm^{-1}, σ_1 = [50.0, 900.0] mm and v_z = [0.0, 140.0] mm, where n describes the number of coil segments. During the iterative correction process an excessive increase or reduction of the image signal intensity has to be avoided. Therefore, additional

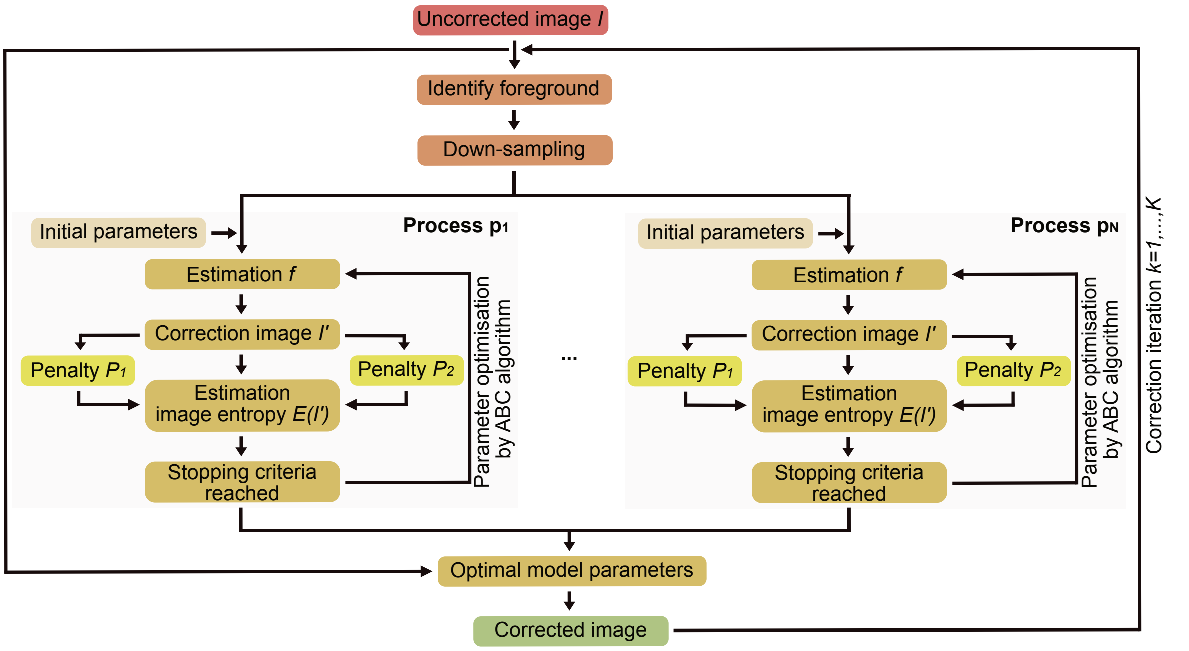


Figure 3.4.: Illustration of the correction process for the proposed physical correction model (PCM). The correction process consists of two major steps: image pre-processing (i.e., identify background) and the estimation of the correction function f through an iterative optimisation performed by the artificial bee colony (ABC) algorithm. The result of the entire correction process is the corrected image I'.

model constraints were defined, which penalise such undesirable effects. The optimisation process is penalised when the image signal increases over a defined threshold, expressed by the penalty function P_1,

$$P_1 = 1 + \operatorname{erf}\left(\frac{r}{\sigma_2 \cdot \sqrt{2}}\right) \cdot 1 + \bar{r}. \tag{3.8}$$

Here $r = \frac{\max(I'^{(k)})}{\max(I^{(0)})} - 1$ is the ratio between the maximum grey value of the corrected image $I'^{(k)}$ within the k-th correction step and the maximum grey value of the original image $I^{(0)}$, σ_2 is a constant parameter and $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-x^2}\,\mathrm{d}x$ is the error function, which allows for a continuous penalisation of the optimisation criteria. The second penalty function P_2 penalises the geometric misbehaviour of (3.3) to prevent an excessive intensity correction by single coils, which is illustrated in figure 3.5. Therefore, it prevents single coils from moving within a defined margin window (figure 3.5, red lines) around the scan object such as patient body. The size and location of the margin window is derived from the resulting

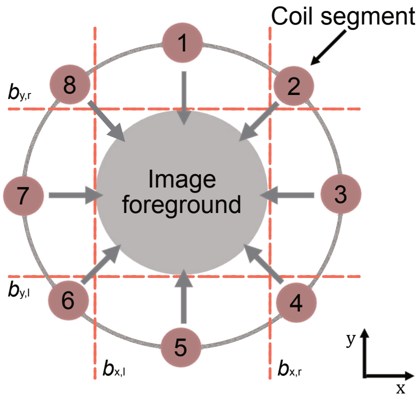


Figure 3.5.: Illustration of the second penalty term P_2 for a head coil array, which penalises the geometric misbehaviour of the correction function (3.3). The left and the right border of the margin window ($b_{x,r}$, $b_{y,r}$, $b_{x,l}$ and $b_{y,l}$, respectively) prevents that single coil segments move into the window to avoid an excessive intensity correction by single coils.

foreground image, which is determined during the correction process. The second penalty term is defined as

$$P_2 = \begin{cases} C = 1000, & \text{if } \min(v_x) \geq b_{x,l} \vee \min(v_y) \geq b_{y,l} \vee \min(v_x) \leq b_{x,r} \vee \min(v_y) \leq b_{y,r} \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

Here, $b_{x,l}$, $b_{y,l}$, $b_{x,r}$ and $b_{y,r}$ define the left (l) and the right (r) border of the margin window, respectively, and v_x and v_y the geometric shift of the coils in x, y-direction of the correction function (3.3).

3.3. Experiments

The performance and applicability of the PCM algorithm were demonstrated on three different MRI data sets. The first experiment uses the water phantom and a simulated MRI data set based on the Brain-Web-Simulator to test and to optimise the PCM algorithm (data set I) (Cocosco et al., 1997). In a second experiment, the PCM was evaluated on two real human brain MRI data sets acquired on a 1.5 Tesla and a 3.0 Tesla MRI scanner, respectively (data set II). In the third experiment an abdominal MRI data set was used, which was acquired on a 1.5 Tesla MRI instrument (data set III).The assessment of the correction quality was performed using the sum of the coefficient of variation (COV) over different tissue classes T (Wicks et al., 1993), which is defined by,

$$COV = \sum_{t \in T} \frac{\sigma_t}{\mu_t}. \tag{3.10}$$

Here σ_t and μ_t are the standard deviation and the mean of the intensities of the generated tissue classes. Different tissue classes were generated for the specific performed experiments. In theory, a small COV value indicates a more uniform intensity distribution and indicates a better image quality.

Furthermore, the achieved COV of the PCM algorithm was compared with the COV of the original data and the established bias correction algorithm N4 for data sets II and III (Tustison et al., 2011). The N4 algorithm based on the assumption that the corruption of the low frequency bias field can be modelled by a convolution of the intensity histogram with a Gaussian function. The correction is performed during an iterative deconvolution of the intensity histogram with the Gaussian function followed by a spatially smoothing of the resulting correction function using a B-spline function (Tustison et al., 2011). For the N4 correction algorithm two different parameter configurations were used. The first configuration ($N4_{100}$) is the default configuration suggested in Tustison *et al.* (Tustison et al., 2011). The second configuration ($N4_{50}$) is based on the custom configuration with the following settings: full width at half maximum = 0.15, convergence threshold = 0.0001, maximum number of iterations = 500, 400, 300, bins = 250 and spline distances 50 mm. Finally, differences in the achieved COV were statistically evaluated using a two-sided Wilcoxon signed-rank test with a significance level of α = 0.05.

3.3.1. Phantom and simulated brain data set

Experimental design

Data set I consists of three scans of a water phantom acquired on a 1.5 Tesla MRI scanner and of nine different simulated MRI brain scans created by the BrainWeb-MRI simulator (Cocosco et al., 1997) with different intensity non-uniformity levels: 0%, 20% and 40%, respectively. The used MR sequences were T1w, T2w and PDw. The considered MR sequence parameters are summarised in table 3.1. For quantitative evaluation of the correction performance, the phantom scans were automatically segmented using the Otsu threshold method to identify the foreground of the images (Otsu, 1975). For the simulated data set, the segmented tissue classes: GM, WM and CSF were available for all MR scans.

Characterisation of hyper-parameters for the PCM algorithm

Aside from the model parameters of the proposed correction function (3.3), other algorithm-specific hyper-parameters of the PCM need to be selected, e.g., the bee swarm colony size for the ABC optimisation algorithm and the maximal number of correction iterations *k*. Therefore, the phantom data set was used to evaluate the effect of the colony size and number of correction iterations *k* on the correction performance. Figure 3.6 shows the results of the experiment for different colony sizes (20, 50, 100, 200) and different numbers of

Table 3.1.: Magnetic resonance sequence parameters of the water phantom and the simulated brain data sets.

Sequence parameter	Water phantom scans			Simulated brain scans		
	T1w	T2w	PDw	T1w	T2w	PDw
Sequence type	spin echo	turbo spin echo		spin echo	turbo spin echo	
Magnetic field strength in Telsa	1.5					
Echo time in ms	7.7	81	13	10	35	
Repetition time in ms	500	3330		18	3300	
Flip angle in degrees	90	150		30	90.0	
Acquisition matrix size for *x*, *y*, *z*	384 × 512 × 23	288 × 384 × 23		181 × 217 × 181		
Voxel size (*x*, *y*, *z*) in mm	0.6 × 0.6 × 6.0			1.0 × 1.0 × 1.0		

Abbreviations: T1w, T1-weighted; T2w, T2-weighted;
PDw, PD-weighted MRI scans

correction iterations (k = 1 − 50) for all MR sequences. The results showed that the COV converged to a steady solution after about k=15 iterations for the T1w, T2w and PDw scans. Furthermore, a colony size of 200 leads to the lowest COV for the T2w and PDw scans and a comparable performance in the case of the T1w scan. Therefore, the following hyper-parameters of the PCM were used for all experiments: colony size = 200, maximum number of iterations k = 15, convergence threshold = 0.001, down-sampling factor = 10 and the penalty parameter σ_2 = 0.1.

Intensity non-uniformity correction performance

After the characterisation of the algorithm specific hyper-parameters, the phantom and simulated brain data set were used to demonstrate the applicability of the developed PCM algorithm. An example of the correction using the PCM algorithm for the initial phantom data set is shown in figure 3.7. The PCM improved image quality and created a more uniform intensity distribution of the water compared to the original data (figure 3.7, red circle). Furthermore, the applied correction function is very smooth and gradually which is one property of the PCM algorithm (figure 3.7, right). Steep correction gradients are not possible, leading to the fact that intensity correction of high intensity changes within small image regions cannot be performed (figure 3.7, orange arrows). The improved image quality of the phantom data is demonstrated by the COV analysis, leading to a reduced COV in comparison to the original data, which is summarised in table 3.2. Also, the results of the COV analysis for the simulated brain data set are shown in table 3.2. The simulated data could be successfully corrected by the PCM, leading to a similar COV value as the non-Bias (0%) images.

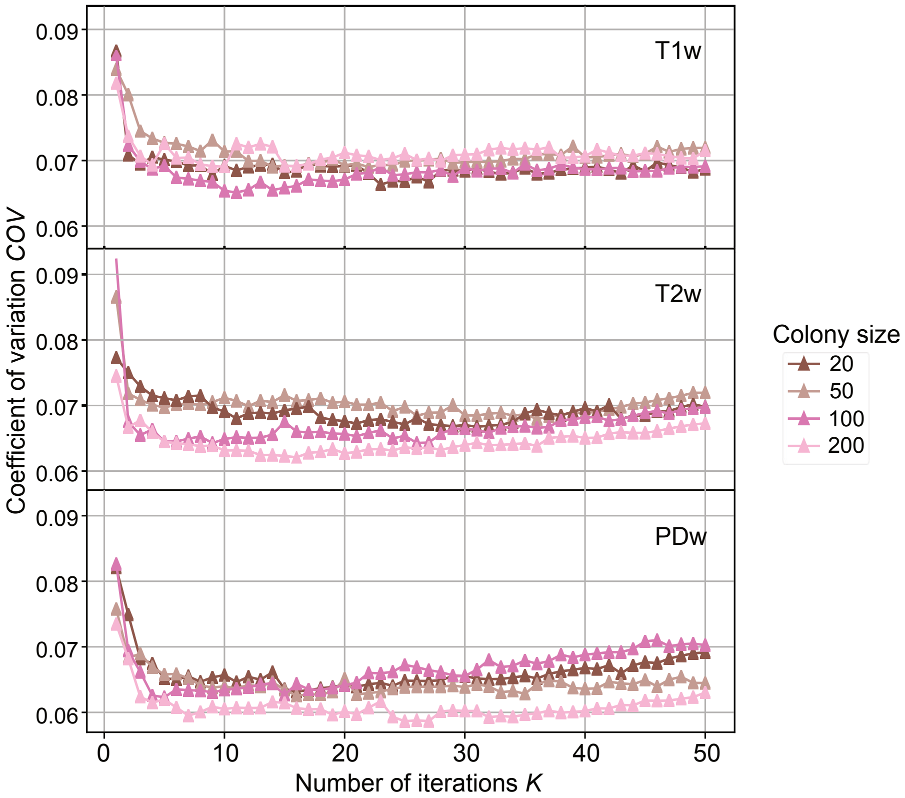


Figure 3.6.: Correction performance as a function of the number of correction iterations *k* for different colony sizes on T1-weighted (T1w), T2-weighted (T2w) and PD-weighted (PDw) scans of the water phantom. The correction performance reached an optimum after about 15 correction iterations in combination with a colony size of 200.

Furthermore, the correction performance by the N4 algorithm led to results comparable with the PCM. However, in some cases (e.g., T1w scan) the N4 led to a too strong correction of intensity non-uniformity.

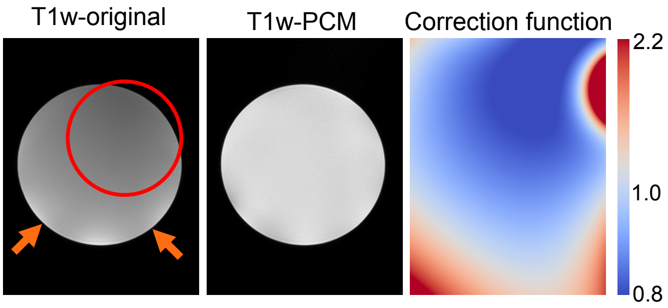


Figure 3.7.: Example of an uncorrected (left) and corrected (middle) T1-weighted (T1w) scan of the water phantom data set as well as the estimated correction function (right) by the physical correction model (PCM). The image quality could be improved through the reduction of the intensity non-uniformity regions (high signal: orange arrows, low signal: red circle).

Table 3.2.: Coefficient of variation (COV) for the phantom and the simulated magnetic resonance brain scans (data set I) for different bias levels.

MRI sequence	Water phantom scans		Simulated brain scans				
	Original	PCM	Bias (%)	Original	PCM	$N4_{50}$	$N4_{100}$
T1w	0.22	0.07	0	0.39	0.39	0.40	0.38
			20	0.42	0.39	0.39	0.38
			40	0.49	0.39	0.40	0.38
T2w	0.23	0.06	0	0.43	0.43	0.43	0.42
			20	0.45	0.43	0.43	0.42
			40	0.51	0.43	0.45	0.44
PDw	0.22	0.06	0	0.17	0.17	0.20	0.19
			20	0.21	0.17	0.20	0.19
			40	0.30	0.17	0.20	0.19

Abbreviations: T1w, T1-weighted; T2w, T2-weighted;
PDw, PD-weighted MRI scans

3.3.2. Clinical brain data set

Experimental design

Data set II consists of 93 multi-channel images from the human brain of 27 patients which were consecutively acquired in the third quarter of 2013 on a 1.5 Tesla MR Siemens Avanto system during clinical routine. The data set comprises 24 T1w, 16 T2w and 14 PDw MR images without contrast agent as well as 21 T1w, 9 T2w and 9 PDw MR images with contrast agent (gadolinium). In addition, 54 MR scans (27 T1w and 27 T2w), from a publicly available database of the Human Connectome Project consortium were investigated (Van Essen et al., 2013). The MR scan were acquired using a 3.0 Tesla scanner. The used MR sequence parameters are shown in table 3.3. The required tissue classes (i.e., GM, WM and CSF) for the quantitative analysis were automatically segmented using the Expectation-Maximisation algorithm of the Slicer 3D Software for all modalities (Fedorov et al., 2012).

Intensity non-uniformity correction performance

Figure 3.8 shows examples of corrected T1w and T2w 1.5 Tesla MR brain scans. The PCM image correction led to an improved image quality through reduction of intensity non-uniformity regions (figure 3.8, red circles) compared to the original images. This is also demonstrated by an improved segmentation result using fuzzy c-means segmentation algorithm (Bezdek et al., 1984). Furthermore, the correction result of the $N4_{100}$ method shows in some cases (e.g., figure 3.8, orange arrows) a too strong intensity correction for the T1w image through steep correction gradient which influences the image quality and the segmentation result.

The result of the quantitative evaluation are summerised in table 3.4 for the 1.5 and 3.0 Tesla MR images. For all 1.5 Tesla MR scans the proposed PCM could significantly enhance

Table 3.3.: Magnetic resonance sequence parameters for the 1.5 Tesla and 3.0 Tesla brain data sets.

Sequence parameter	1.5 Tesla brain scans			3.0 Tesla brain scans		
	T1w	T2w	PDw	T1w	T2w	PDw
Sequence type	spin echo	turbo spin echo		spin echo	turbo spin echo	
Magnetic field strength in Tesla		1.5			3.0	
Echo time in ms	7.7	81	13	10	35	
Repetition time in ms	450	3330		18	3300	
Flip angle in degrees	90	150		30	90	
Acquisition matrix size for x, y, z	384 × 512 × 27			181 × 217 × 181		
Voxel size (x, y, z) in mm	0.5 × 0.5 × 6.0			1.0 × 1.0 × 1.0		

Abbreviations: T1w, T1-weighted; T2w, T2-weighted;
PDw, PD-weighted MRI scans

the image quality compared to the original data ($p \leq 0.001$). Furthermore, the PCM algorithm outperformed the N4$_{50}$ approach for the T1w and PDw scans (T1w: p<0.001, PDw: p=0.038, respectively). For the 1.5 Tesla T2w the PCM algorithm achieved no better image correction performance than the N4$_{50}$ approach (T2w: p=0.047). Also the N4$_{100}$ algorithm showed a similar or slightly better correction performance for 1.5 Tesla MR images compared to the PCM algorithm (T1w: p>0.05, T2w: p<0.001 and PDw: p=0.019, respectively).

For the 3.0 Tesla MR scans the PCM approach could significantly enhance the image quality for all MRI sequences compared to the original data (p<0.001). In addition, the PCM outperformed the N4$_{100}$ algorithm (p=0.002) for T2w images. For the T2w images the PCM was not able to achieved a better correction performance compared to the N4$_{100}$ algorithm. In comparison to the N4$_{50}$ the PCM showed a slightly better or a similar correction performance (T1w: p<0.001 and T2w: p>0.05, respectively)

Table 3.4.: Results of the coefficient of variation (COV) analysis as well as the statistical results for data set II consisting of 1.5 and 3.0 Tesla MRI brain scans. The mean and standard deviation (SD) of the differences in COV between the physical correction model (PCM) and the original images as well as two configurations of the N4 algorithm are shown.

	MRI-Sequence	PCM-Original		PCM-N4$_{50}$		PCM-N4$_{100}$	
		Mean/SD	p-value	Mean/SD	p-value	Mean/SD	p-value
1.5 T	T1w	-0.09/0.06	<0.001	-0.14/0.34	<0.001	-0.02/0.14	>0.05
	T2w	-0.07/0.04	<0.001	0.01/0.02	0.047	0.02/0.02	<0.001
	PDw	-0.13/0.07	<0.001	-0.01/0.03	0.038	0.01/0.02	0.019
3.0 T	T1w	-0.33/0.08	<0.001	-0.04/0.02	<0.001	0.03/0.02	<0.001
	T2w	-0.20/0.08	<0.001	-0.02/0.07	>0.05	-0.04/0.06	0.002

Abbreviations: T1w, T1-weighted; T2w, T2-weighted;
PDw, PD-weighted MRI scans

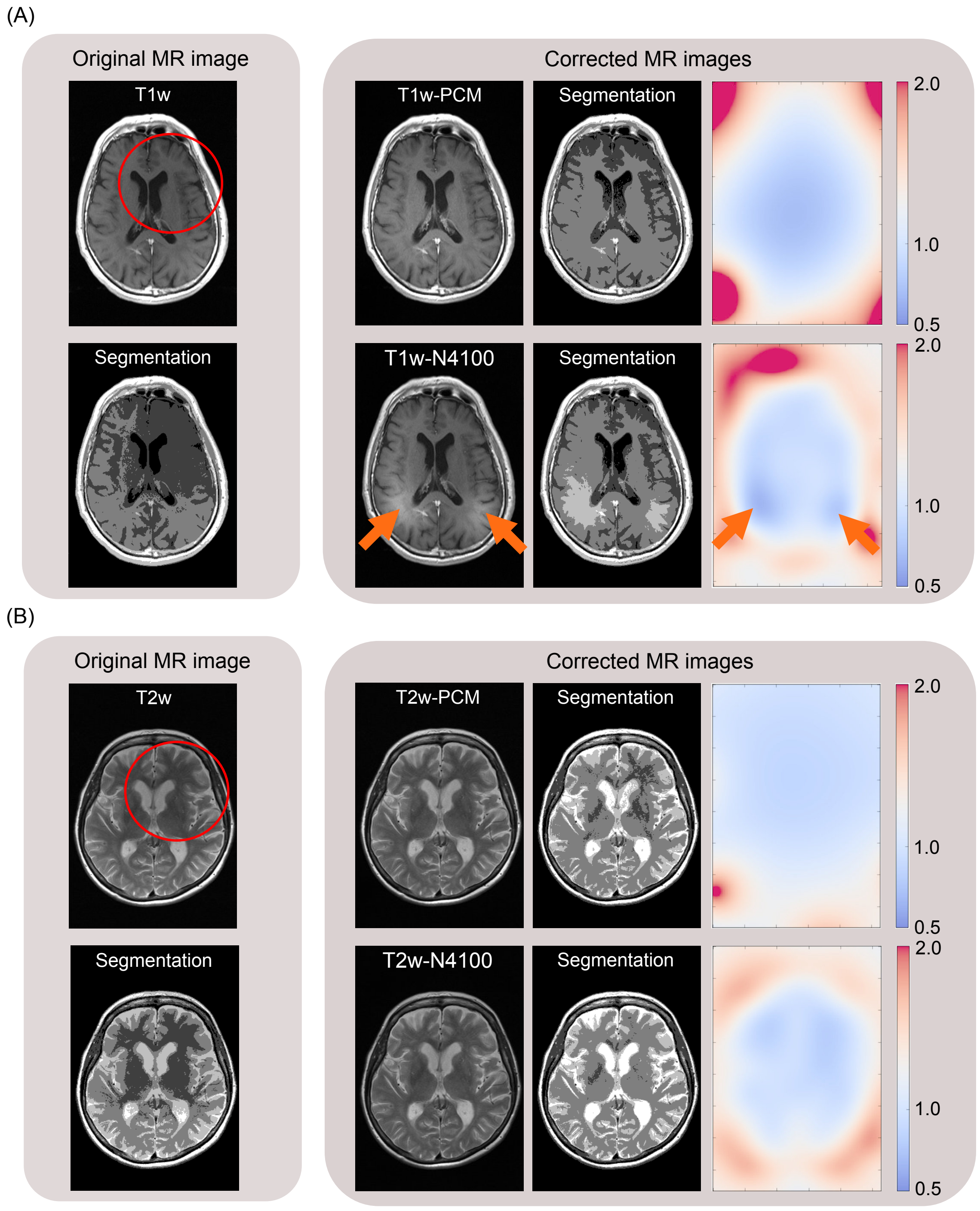


Figure 3.8.: Examples of uncorrected and corrected (A) T1-weighted (T1w) and (B) T2-weighted (T2w) images of the human brain as well as the estimated correction functions (right column) by the physical correction model (PCM) and N4$_{100}$ algorithm. Furthermore, the segmentation results by the fuzzy c-means algorithm are depicted (Bezdek et al., 1984). The red circles indicate the regions of intensity non-uniformity. The orange arrows in (A) show an example of the N4$_{100}$ in which a excessive intensity correction was applied.

3.3.3. Abdominal data set

Experimental design

Data set III comprises 14 T1w and 14 T2w MRI abdominal scans with and without contrast agent (gadolinium), which were acquired between 2011 and 2014. All data sets were acquired on a 1.5 Tesla MR Siemens Avanto system. The imaging parameters of the MR sequences are shown in table 3.5. The tissue classes for the quantitative evaluations consist of: kidney, liver and spleen. Those three tissue classes were generated semi-automatically with the Slicer 3D software for all MR scans (Fedorov et al., 2012).

Table 3.5.: Magnetic resonance sequence parameters for the 1.5 Tesla abdominal data set.

Sequence parameter	Abdominal scans	
	T1w	T2w
Sequence type	gradient echo	turbo spin echo
Magnetic field strength in Tesla	1.5	
Echo time in ms	2.2	86.0
Repetition time in ms	4.9	4101
Flip angle in degrees	10	140
Acquisition matrix size for *x*, *y*, *z*	512 × 416 × 80	
Voxel size (*x*, *y*, *z*) in mm	0.8 × 0.8 × 3.0	

Abbreviations: T1w, T1-weighted; T2w, T2-weighted MRI scans

Intensity non-uniformity correction performance

Figure 3.9 show examples of a corrected T1w and T2w MR abdominal scan. PCM improved the image quality compared to the original data, which could also be demonstrated by the improved segmentation result. In both examples, the $N4_{100}$ method corrected the intensity non-uniformity for the T1w image too strongly through steep correction gradients (figure 3.9(A), orange arrows). This led to the fact that the whole surrounding tissue had nearly the same signal intensity after the correction and that the contrast enhanced region was almost not visible anymore (e.g., kidney).

The results of the COV analysis as well as the statistical results are shown in table 3.6. For all considered 1.5 Tesla MR scans, the PCM algorithm could significantly enhance the image quality compared to the original data (T1w: p=0.001 and T2w: p=0.002, respectively).

Furthermore, the PCM algorithm showed reduced or similar correction performance for the T1w and T2w images, respectively, compared to the results of the $N4_{50}$ (T1w: p=0.001 and T2w: p=0.054, respectively). For the 1.5 Tesla T1w and T2w scans, the $N4_{100}$ algorithm achieved better correction results and showed similar correction performance compared to the PCM method (T1w: p=0.001 and T2w: p>0.05, respectively).

However, the N4 correction usually led to a too strong intensity correction which cannot be necessarily expressed by the COV. Figure 3.9 shows such an example. The calculated COV value is lower for the N4$_{100}$ than for the PCM algorithm, 0.49 and 0.51, respectively, although the obtained result of the PCM is preferable. Therefore, an additional tissue contrast analysis was performed by evaluating the ratio of the mean intensity values for the different

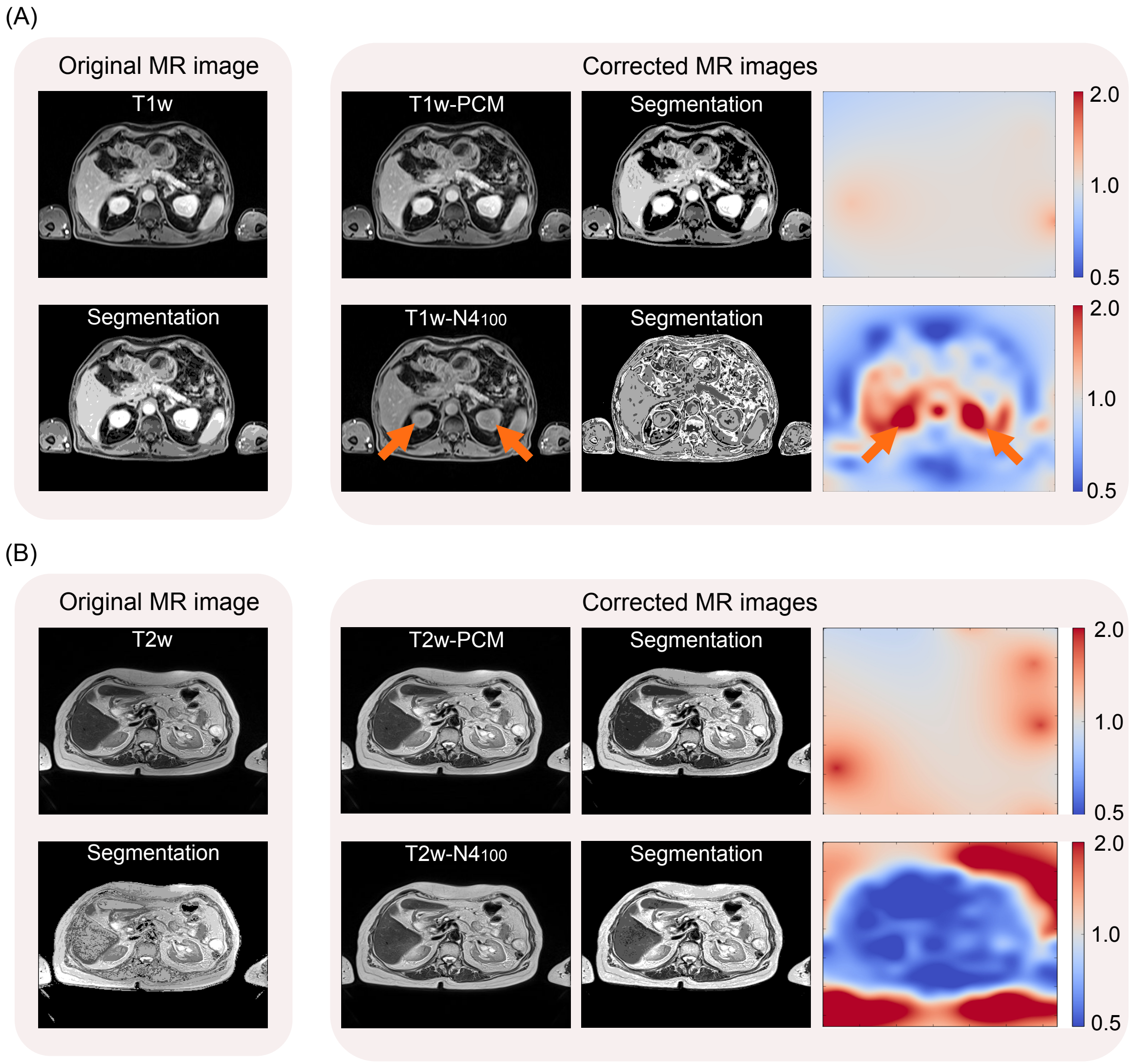


Figure 3.9.: Examples of uncorrected and corrected (A) T1-weighted (T1w) and (B) T2-weighted (T2w) images of the abdominal data set as well as the estimated correction functions (right column) by the physical correction model (PCM) and N4$_{100}$ algorithm. Furthermore, the segmentation results by the fuzzy c-means method are depicted (Bezdek et al., 1984). The orange arrows in (A) show an example of the N4$_{100}$ in which a excessive intensity correction was applied. The different tissue classes (liver or kidney) as well as the surrounding tissue had almost the same tissue signal after correction and the contrast agent (e.g., kidney) was almost not visible anymore.

tissue types liver, spleen and kidney in the abdominal MR data set. Only T1w images were considered due to the higher signal to noise ratio between the different tissue classes. The results are depicted in figure 3.10. While a correction using the PCM method preserves the tissue intensity differences between the tissue types, the correction using the N4 algorithm leads to similar signal intensities for all three tissue types. This is not preferable especially in the case of contrast-enhanced images.

Table 3.6.: Results of the coefficient of variation (COV) analysis as well as the statistical results for data set III consisting of 1.5 Tesla MRI abdominal scans. Shown are the mean and standard deviation (SD) of the differences in (COV) between the physical correction model (PCM) and the original images as well as two configurations of the N4 algorithm.

	MRI-Sequence	PCM-Original		PCM-N4$_{50}$		PCM-N4$_{100}$	
		Mean/SD	*p*-value	Mean/SD	*p*-value	Mean/SD	*p*-value
1.5 T	T1w	-0.06/0.03	0.001	0.07/0.02	0.001	0.02/0.02	0.004
	T2w	-0.08/0.03	0.002	-0.05/0.08	0.054	-0.01/0.05	>0.05

Abbreviations: T1w, T1-weighted; T2w, T2-weighted MRI scans

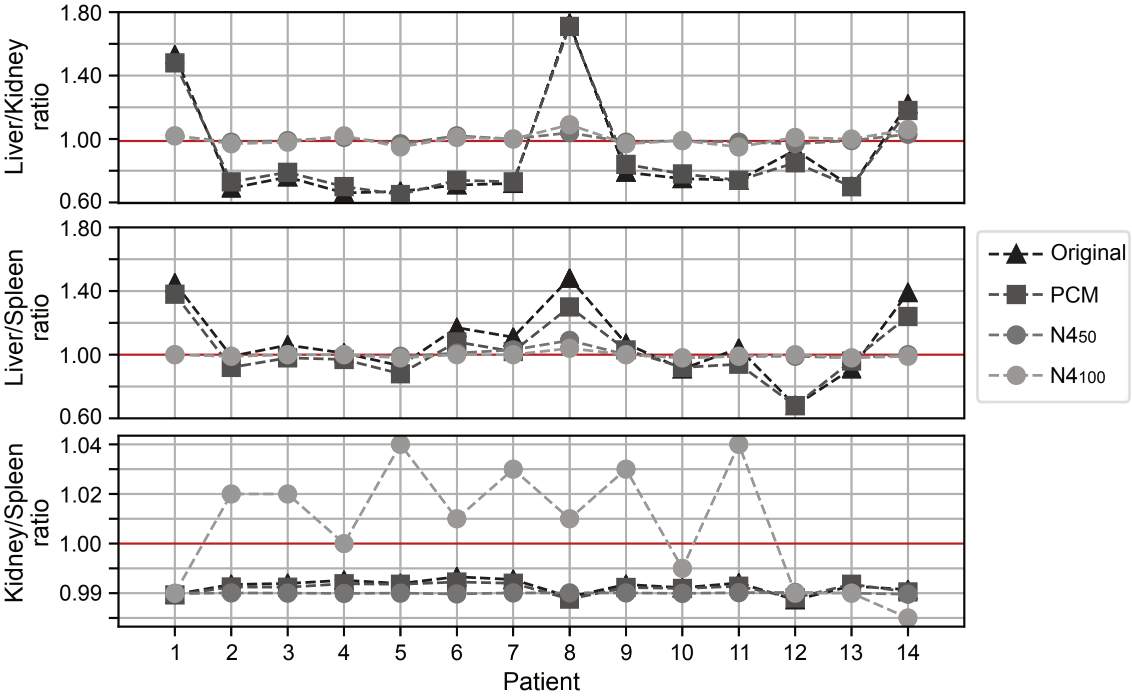


Figure 3.10.: Tissue intensity analysis for the three tissue types of the T1-weighted MR abdominal images. Shown are the ratio of the mean intensity values between the different tissue classes in the original and the corrected images by the physical correction model (PCM) and N4. A ratio of one means almost no difference between the two tissue types.

3.4. Summary and discussion

A typical phenomenon in MRI is intensity non-uniformity, which influences the image quality and interferes with automated and quantitative image analysis. Therefore, we present a novel fully automatic intensity non-uniformity correction approach for MR images.

The main characteristic of our algorithm is the physically motivated correction approach based on the geometry and physical properties of the MRI coil array. In general, the proposed PCM algorithm significantly improved the image quality for all considered data sets in comparison to the original MRI scans. Furthermore, the PCM algorithm often outperformed or achieved similar results compared to the N4 algorithm. While the N4 generally performs well in the case of simple tissue structures, e.g., in brain images, a strong intensity correction can occur when the tissue structures are more complex, like in MR abdominal scans. Because abdominal images, usually contain heterogeneous tissues, in terms of the tissue intensity, e.g., due to blood arteries in the liver or in the case of contrast agents, the N4 algorithm could reduce the COV significantly more than the PCM approach. However, the N4 algorithm often changed the soft tissue signal intensities in an undesired way, such that the intensity between different tissue types appeared similar, leading to lower COV values. Our tissue intensity analysis confirmed this observation and figure 3.9(A) shows an example of such a correction result by the N4 approach. The strong intensity correction of the liver and spleen affects the relative tissue intensity. Especially the strong intensity signal of the tissue surface caused by the contrast agent is almost completely removed. In contrast, an advantage of the PCM is the application of only smooth and gradual intensity corrections due to the exponential base functions and the introduced penalty concept. This prevents a spot-like correction, which is possible for the N4. Especially a non-optimal algorithm configuration, e.g., a low spline (knots) distance in combination with a high number of iterations, high intensity changes caused by steep gradients within small image regions cannot occur.

The derivation of the proposed model (3.3) was based on the water phantom experiment using a 1.5 Tesla MR scanner. However, for MR scans acquired with other magnetic field strengths our assumption, that the intensity is exponential decreasing to the centre of the coil array, may not be valid. Intensity non-uniformity in 3.0 Tesla or 7.0 Tesla MR images may occur due to inhomogeneous excitations caused by interactions between RF waves and electromagnetic properties of the tissues (Van de Moortele et al., 2009). Therefore, the cause of intensity non-uniformity is not only dependent on the coil array, which may influence the correction performance of the PCM. Furthermore, with modern MRI scanners an automatic Bias correction during the image acquisition is in principle possible. For instance, in the case of the 3.0 Tesla Philips Ingenuity TF scanner (Eindhoven, The Netherlands) a constant level appearance (CLEAR) correction can be performed which may significantly affect the performance of the PCM (figure 3.11). Further investigations to correct MR images with higher magnetic field strength are required to improve the performance of the PCM.

Still, the PCM was able to significantly reduce the intensity non-uniformity in 1.5 Tesla, but also in 3.0 Tesla MR images, which is essential for retrospective image evaluation such as for radiomics studies.

The PCM is based on a high-dimensional correction function which is iteratively optimised during the optimisation process. The iterative update of the large number of model parameters leads to a high computation time of the correction process, especially if the physical MRI coil array consists of a large number of coil segments (e.g., up to 2 minutes). The computation time also depends on the voxel grid size, which will further decrease in future. Therefore, a reduction of computation time might be required by further parallelisation of the PCM, e.g., by transferring the algorithm to graphics processing units. Furthermore, the high number of model parameters, especially by a large number of coil segments, is challenging for the optimisation algorithm to find a globally optimal solution. This requires an optimisation algorithm which is able to handle high-dimensional functions, such as the employed ABC optimisation algorithm. To further improve the optimisation of the model parameters, other optimisation algorithms may be investigated. Furthermore, the lower and upper limits of the model parameters were derived from the coil geometry and from the phantom experiment. However, for scans from other vendors the parameter limits of the PCM might differ, e.g., the decay parameters a_i or the global pre-factor q_1.

The quantitative evaluation was performed by measuring the COV of different tissue classes, which requires a segmentation of each class. For brain data sets, the segmentations were automatically created. However, e.g., in case of tumour structures, the resulting automatic segmentations may not be necessarily correct, which may affect the COV value. Furthermore, the COV value is based on the assumption that the spatial intensity distribution of a tissue of interest is piecewise constant. That assumption is not optimal, especially in the case of tumour regions which are usually more heterogeneous. Therefore, it is recommend

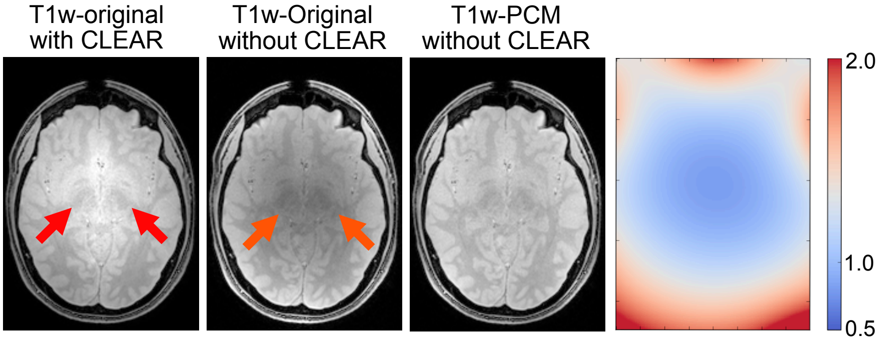


Figure 3.11.: Example of 3.0 Tesla T1-weighted (T1w) image, acquired with and without automatic constant level appearance correction (CLEAR) as well as the corrected image using physical correction model (PCM) and the correction function. The red and orange arrows show the region of intensity non-uniformity. In the case of the automatic CLEAR correction the physically motivated correction assumption becomes invalid.

to perform a smooth Bias correction, as done by the PCM, which is important, e.g., for advanced image analyses such as radiomics studies.

4. Comparison of feature selection methods and machine learning algorithms for radiomics time-to-event survival models

4.1. Motivation

The development of radiomics risk models employs various machine learning algorithms to characterise and quantify the tumour phenotype using advanced imaging features (Lambin et al., 2012). Different studies have investigated various radiomics features in terms of their prognostic or predictive abilities and their reliability across several tumour entities such as lung, head and neck as well as brain tumours (Aerts et al., 2014; Vallières et al., 2015; Hatt et al., 2016; Kickingereder et al., 2016; Song et al., 2016).

In such radiomics studies, feature selection is used to identify prognostic radiomics features and to reduce the dimensionality of the feature space (Guyon and Elisseeff, 2003). Machine learning algorithms subsequently use a subset of selected radiomics features (signature) to construct prognostic models by learning the decision boundaries of the underlying data distribution.

A variety of different feature selection methods and machine learning approaches have been developed as described previously in section 2.5. Most radiomics studies consider a combination of only one feature selection method and one learning algorithm. For instance, a univariate feature selection using the Cox followed by a multivariable Cox model to predict clinical endpoints such as OS (Aerts et al., 2014; Fave et al., 2017). Kickingereder *et al.* (Kickingereder et al., 2016) used a Cox regression model combined with a supervised principle component analysis based on the model coefficients as feature selection method to develop the radiomics signature and a Cox model for the prediction of OS. L. van Dijk *et al.* (van Dijk et al., 2017a) used Pearson correlation coefficients to identify relevant image features in combination with Lasso regularisation to develop a multivariable logistic regression model. It is uncertain whether these methodological choices led to the most accurate and reliable radiomics risk models. The usage of only one combination may also increase the risk of incidental findings and requires prior knowledge of the underlying data structure. For example, in the case of the Cox model the selected features have to show a linear correlation with the considered outcome, otherwise this may lead to wrong model predictions. Therefore, the identification of suitable feature selection methods and learning algorithms is a important integral step to develop highly accurate and reliable radiomics risk models.

In the field of radiomics only few studies have performed such an extensive analysis. Recently, Parmar *et al.* (Parmar et al., 2015a; Parmar et al., 2015b) investigated different algorithms in two different studies on patients with non-small cell lung cancer (NSCLC) as well as HNSCC. However, in these studies the outcome of interest, OS, was transformed to a binary endpoint. While dichotomisation of the endpoint is a common method for stratifying patient groups, it incurs the risk of biasing prognostic accuracy (Dupuy and Simon, 2007). Therefore, this study compared the prognostic ability of twelve feature selection methods and eleven learning algorithms, which are able to deal with continuous time-to-event survival data. The work in this chapter has been published in a peer-reviewed journal (Leger et al., 2017b) and was presented at the international conference of the European Society for Radiotherapy and Oncology (Leger et al., 2016).

4.2. Patient cohort and experimental design

4.2.1. Characteristics of patient cohort

The systematic evaluation is based on data from a multi-centre cohort consisting of 293 HNSCC patients. All patients suffered from histologically confirmed loco-regionally advanced HNSCC and received primary radio-chemotherapy. The cohort was divided into an exploratory and a validation cohort by an ratio of a approximately 2:1 based on the different included studies rather than on the treatment places. The exploratory cohort included 213 patients from which 152 patients were treated at one of the seven partner sites of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG) between 2005 and 2011 (Linge et al., 2016b). The remaining 61 patients of the exploratory cohort were treated at the University Hospital Dresden (UKD) between 1999 and 2006. The validation cohort consisted of 80 patients. 50 of these patients received their treatment between 2006 and 2012 within a prospective clinical trial (NCT00180180) at the UKD (Zips et al., 2012; Löck et al., 2017). The remaining 30 patients were treated at the UKD or at the Radiotherapy Center Dresden-Friedrichstadt (RCDF) between 2005 and 2009. The details of the patient characteristics for both cohorts are summarised in table 4.1.

Radiomics risk models for the prediction of the primary endpoint LRC and secondary endpoint OS were developed using non-contrast enhanced CT scans. Both clinical endpoints LRC and OS were calculated from the first day of radio-chemotherapy to the date of the event or censoring. The number of events for LRC and OS were 86 and 120 for the exploratory cohort, and 26 and 51 for the validation cohort, respectively. The median follow-up time was 28.8 months (range 1.3-70.3 months) for the exploratory cohort and 21.5 months (1.4-107.2 months) for the validation cohort. The two year LRC rate was 63.0% for the exploratory and 56.0% for the validation cohort (log-rank: p=0.61). Overall survival after two years was 50.0% for the exploratory and 53.0% for the validation cohort (log-rank: p=0.56). The Kaplan-Meier curves are are shown for both endpoints in figure 4.1.

Table 4.1.: Patient characteristics of the exploratory cohort and the validation cohort.

Clinical variable	Exploratory cohort	Validation cohort	*p*-value
Number of patients	213	80	-
Gender			
male	181	70	0.58^1
female	32	10	
Age in years			
median	59	54	0.005^3
range	39.0 - 81.9	37 - 74	
TN staging			
T stage 1 / 2 / 3 / 4 / missing	3 / 24 / 53 / 133 / 0	3 / 9 / 27 / 40 / 1	0.19^2
N stage 0 / 1 / 2 / 3 / missing	25 / 8 / 166 / 14 / 0	10 / 8 / 57 / 4 / 1	0.19^2
UICC stage 2010			
I / II / III / IV / missing	0 / 0 / 13 / 139 / 61	1 / 2 / 9 / 68 / 0	0.096^1
Tumour volume in cm^3			
median	27.6	34.8	0.19^3
range	0.3 - 276.3	2.7 - 244.8	
Dose in Gy			
median	72	72	<0.001^3
range	68 - 77	69 - 77	
HPV–16 DNA			
negative / positive / missing	164 / 27 / 22	39 / 5 / 36	0.63^2
Number of events			
LRC	86	26	-
OS	120	51	
Follow up time of patients alive in months			
median	52.6	52.7	-
range	4.2 - 131.9	7.8 - 107.2	

Abbreviations: T, clinical tumour stage; N, clinical nodal stage; UICC, Union internationale contre le cancer
Gy, Gray; HPV, human papillomavirus; DNA, deoxyribonucleic acid
LRC, loco-regional tumour control; OS, overall survival
^{1} exact Fisher test; textsuperscript2 χ^2 test; ^{3} Wilcoxon-Mann-Whitney test

4.2.2. Experimental design

The experimental design is depicted in figure 4.2. The feature computation and the risk modelling were performed within the two developed software frameworks previously described in chapter 2.3. Radiomics risk models were developed using the exploratory cohort for LRC and OS. Prognostic model performance and patient risk group stratification were evaluated on the validation cohort. Furthermore, the feature stability of the radiomics signatures were assessed by applying image perturbations such as image rotations and translations in *x*-*y*-directions.

Feature computation and risk modelling

The GTV of the primary tumour was manually delineated by a radiation oncologist on each CT scan separately. The voxel spacing was resampled using trilinear image interpolation to an isotropic voxel size of 1.0×1.0×1.0 mm^3 to correct for differences in voxel spacing and

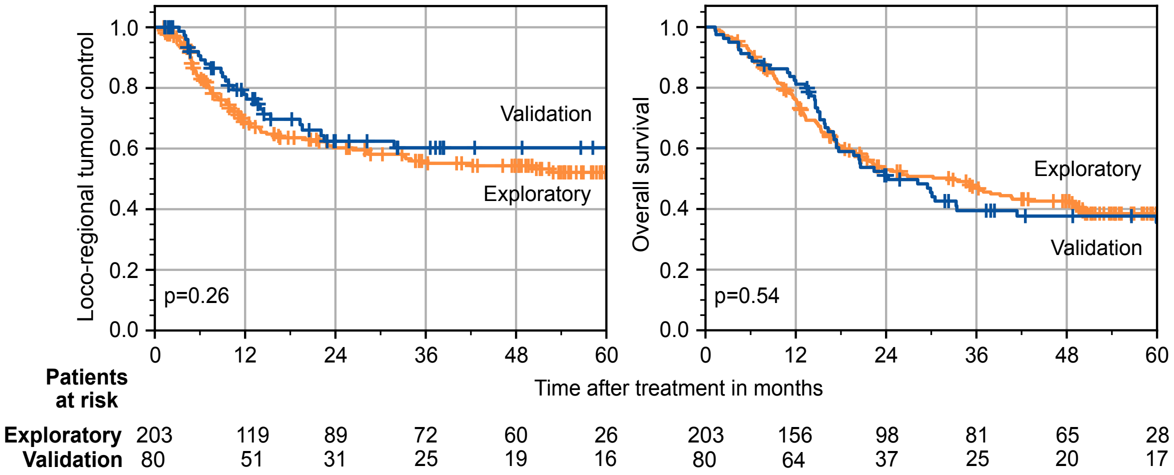


Figure 4.1.: Kaplan-Meier curves of the exploratory and validation cohort for (left) the primary endpoint loco-regional tumour control and (right) the secondary endpoint overall survival.

slice thicknesses between cohorts. Additional images were created by applying spatial filtering on the base image to emphasise image characteristics such as edges and blobs. Eight additional images were created by applying a stationary coiflet-1 wavelet high-/low-pass fil-

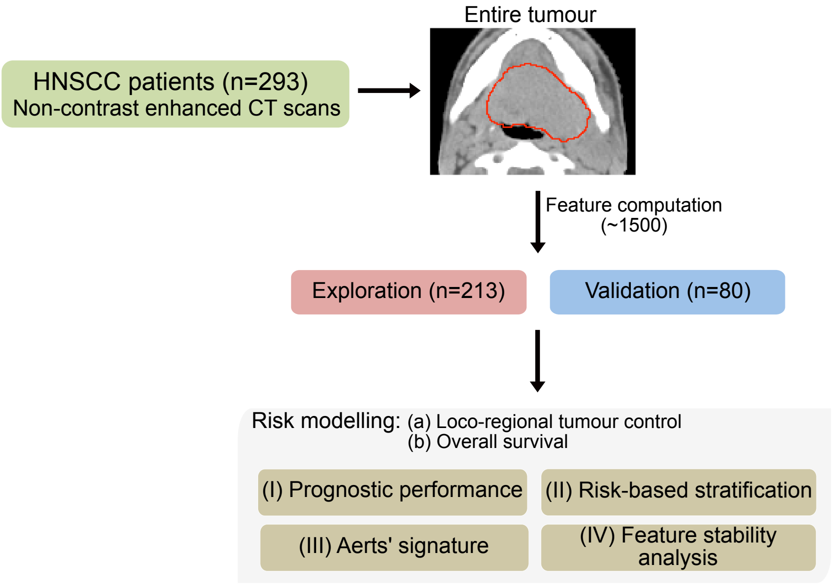

Entire tumour
HNSCC patients (n=293)
Non-contrast enhanced CT scans
Feature computation
(~1500)
Exploration (n=213)
Validation (n=80)
Risk modelling: (a) Loco-regional tumour control
(b) Overall survival
(I) Prognostic performance
(II) Risk-based stratification
(III) Aerts' signature
(IV) Feature stability analysis

Figure 4.2.: Experimental design of the systematic evaluation. Radiomics risk models were developed from the exploratory cohort to predict loco-regional tumour control (LRC) and overall survival (OS). Prognostic models were also trained using the radiomics signature obtained by Aerts *et al.* (Aerts et al., 2014). The performance of the prognostic models and patient risk group stratification were assessed on the validation cohort. Furthermore, the feature stability of the radiomics signatures were assessed by applying image perturbations such as image rotations and translations in *x*-*y*-directions.

ter along each of the three spatial dimensions. One further image was created by applying a LoG filter consisting of five different filter kernel widths (0.5, 1.0, 2.0, 3.0, 5.0 mm). Subsequently, the GTV mask was re-segmented to cover only soft tissue voxels between -150 and 180 HU, thereby removing voxels containing air and bone, which may affect feature expression. Finally, 18 statistical, 38 histogram-based and 95 texture features were extracted from the GTV within each image set (base and transformed images). 28 morphological features were computed within the base image only, leading to 1538 computed features in total.

Prior to the risk model development, the computed image features were normalised on the exploratory cohort using z-score normalisation in (2.44). The resulting scale and shift constants were applied to the independent validation cohort. Subsequently, feature clustering was performed on the exploratory cohort to obtain an initial non-redundant set of features, as described in section 2.3.2. A total of 229 non-singular clusters were created. The same clusters and meta-features were generated for the validation cohort. After clustering, the resulting feature set of the exploratory cohort was used to identify the most relevant features using twelve different feature selection algorithms. Feature selection was repeated 100 times using bootstrap samples (i.e.,.632 bootstrap method with replacement) of the exploratory cohort to ensure the selection of stable features and to increase the model generalisability. During feature selection each feature is ranked according to the weighted importance score introduced in section 2.3.2. Also, the training of the eleven risk models was performed 100 times using bootstrap samples (i.e.,.632 bootstrap method with replacement) of the exploratory cohort. Prior to model training, hyper-parameters of the machine learning algorithms, such as signature size or algorithm-specific settings were optimised using the balanced selection strategy (section 2.3.2) for each combination of feature selection and machine learning algorithm. The learning algorithms were trained on the generated bootstrap samples based on the best ranked features as well as the optimised hyper-parameter set. Finally, an ensemble prediction was made by averaging the predicted risk scores of each model for both the exploratory and the independent validation cohort.

Feature selection methods and machine learning algorithms

The comprehensive analyses comprised, twelve feature selection methods and eleven machine learning algorithms for the prediction of continuous time-to-event survival data to cover a wide range of different algorithm concepts.

The feature selection methods can be divided into three groups based on (a) statistical correlations, (b) mutual information optimisations and (c) model-based approaches. The group (a) comprised the Pearson and the Spearman correlation coefficient methods. The feature selection methods in group (b) are the MIM, the MIFS and the MRMR. The model based approaches in group (c) consisted of: a univariable (uni)- and a multivariable (multi)-Cox-regression model, a RF-MD, a random forest variable importance (RF-VI), a random forest based on variable hunting (RF-VH), a random forest based on maximally selected rank statistics (MSR-RFVI) and a random forest based on permutation variable importance (PVI-RF). Additionally, imaging features were selected at random (RAND) and no feature selection (None) was performed.

The investigated machine learning algorithms can be divided in mainly two groups: (i) non-/semi-parametric and (ii) full-parametric models. The non-/semi-parametric models (i) comprise the Cox model and the NET-Cox with lasso and elastic-net regularisation. Furthermore, group (i) contains of the BT-Cox and the BT-CIndex models as well as of the BGLM-Cox and the BGLM-CIndex models. In addition, this group comprised random forest based methods: RSF and MSR-RF. The second group (ii) consists of the full-parametric SRM, BT and BGLM models based on the Weibull distribution (Weibull).

Performance assessments

The systematic evaluation consists of four analyses (figure 4.2), which are described in more detail in the following paragraph.

(I) Prognostic performance of all combinations of feature selection methods and machine learning algorithms was evaluated based on the validation C-Index. Furthermore the median and standard deviation (SD) of the validation C-Index of a feature selection method over all machine learning algorithms and vice versa was assessed to measure the variance induced by the respective algorithms.

(II) Risk-based patient stratification is an important application of radiomics models for treatment individualisation. Patients were stratified based on the median risk ($\text{median}_{\text{risk}}$) cut-off value and the cut-off value using the bootstrapped method ($\text{boot}_{\text{risk}}$; section 2.4) determined on the exploratory cohort. Cut-off values were applied unchanged to the validation cohort. Survival curves were estimated by the Kaplan-Meier method and differences between the two risk groups were compared by log-rank tests and p-values$\leq$0.05 were considered as statistically significant.

(III) Feature stability analysis is an important aspect to build accurate and generalisable radiomics risk models. For instance, due to motion of the patient during the image acquisition radiomics imaging features may change their expression values leading to incorrect predictions. Therefore, to assess the stability of the selected features within the developed signatures six image rotations ($\pm 2^\circ$, $\pm 6^\circ$, $\pm 10^\circ$) and two image translations in x-y-direction (0.25 mm, 0.75 mm) were applied on the original images of the exploratory cohort as well as all combinations thereof. Subsequently, the feature stability was measured by calculating the intra-class correlation coefficient (ICC) for the various feature selection methods using the different image rotations and translations. The ICC is given by, (Shrout and Fleiss, 1979):

$$ICC = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + \text{WMS}}, \tag{4.1}$$

where BMS and WMS are the between-subjects and within-subjects mean squares, respectively. The ICC ranges from -1 (perfect anti-agreement) to +1 (perfect agreement), where +1 meaning that the feature values are identical and therefore robust against those image manipulations. An ICC$\geq$0.9 indicates a high feature stability, a ICC value between 0.9 and 0.8 a good and a value between 0.8 and 0.6 a moderate stability.

(IV) Assessment of the prognostic performance using the Aerts' signature may help to confirm the presented results to reduce the risk of incidental findings. For this purpose, the machine learning algorithms were built using the previously developed radiomics signature by Aerts *et al.* (Aerts et al., 2014) for both endpoints. The Aerts' signature consists of four imaging features derived from CT scans: a first order statistic feature (energy) describing the overall density of the GTV, a shape based feature (compactness) quantifying the compactness of the GTV volume relative to a sphere, a texture feature and a wavelet based feature (gray level non-uniformity) measuring the intra-tumour heterogeneity. These features were selected based on their stability across test-retest image scans and multiple tumour delineations by different observers as well as over 100 bootstrap samples on the exploratory cohort of 422 lung cancer patients. Subsequently, the signature was independently validated on 225 lung cancer and 231 HNSCC patients (Aerts et al., 2014). The differences between the signatures were quantified by a non-parametric analytical (NPA) approach (Kang et al., 2015) and *p*-values$<$0.05 were considered as statistically significant.

4.3. Results of feature selection methods and machine learning algorithms evaluation

(I) Prognostic performance

The prognostic performance of twelve feature selection methods combined with eleven machine learning algorithms was evaluated for the primary clinical endpoint LRC and the secondary endpoint OS.

For LRC, the considered learning algorithms achieved in general a good prognostic performance on the validation cohort (figure 4.3 (A)). The best performances were obtained by the semi-parametric models: MSR-RF (C-Index: 0.71, 95% confidence interval [0.62–0.83]), the BT-CIndex (C-Index: 0.71, [0.62–0.82]) and the BT-Cox (C-Index: 0.70, [0.59–0.81]) algorithms as well as by the full-parametric BT-Weibull model (C-Index: 0.70, [0.60–0.82]) , all in combination with the Spearman feature selection method.

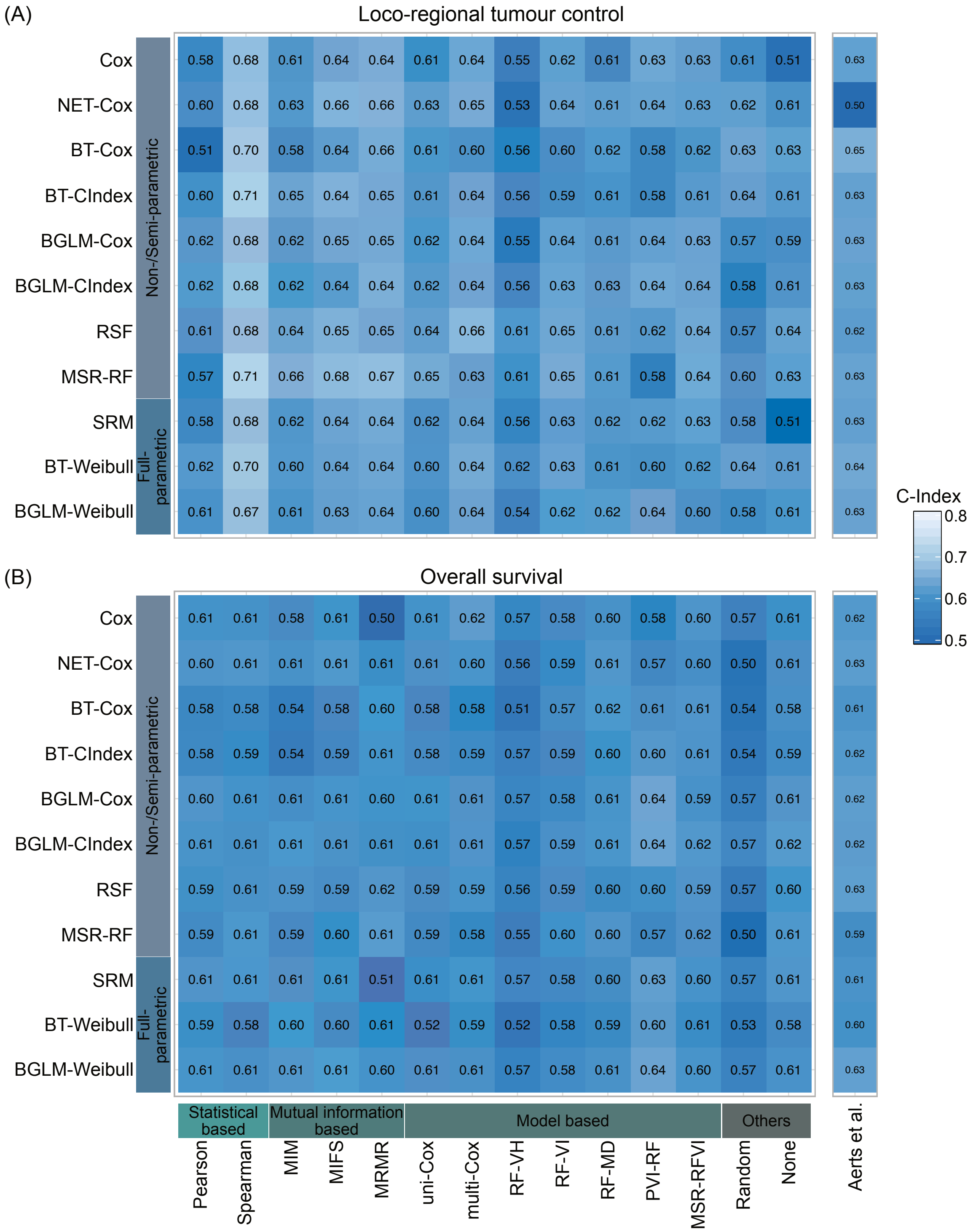
For OS, the performance was in general lower in comparison to LRC and similar between the different learning algorithms (figure 4.3 (B)). The best prognostic performances were obtained by the semi-parametric BGLM-CIndex and BGLM-Cox models as well as the full-parametric BGLM-Weibull algorithm (C-Index: 0.64, [0.53–0.71], 0.64, [0.52–0.70] and 0.64, [0.51–0.68], respectively), all in combination with the random feature selection. The resulting concordance indices for the exploratory cohort for both clinical endpoints are shown in appendix C.1.

For LRC, the median performance of the learning algorithms over all feature selection methods was generally similar (figure 4.4 (A)). The highest median performances were obtained by the RSF (C-Index: median±SD, 0.64±0.03), the MSR-RF (C-Index: 0.64±0.04) followed by the NET-Cox (C-Index: 0.63±0.04) and the BGLM-CIndex (C-Index: 0.63±0.03). For the feature selection methods the differences of the concordance indices (C-Index) were in general larger between over the machine learning algorithms (figure 4.4 (A)). The highest median performance of the feature selection methods was achieved by the Spearman correlation coefficient (C-Index: 0.68±0.01). Furthermore, the mutual information optimisation based MRMR and MIFS methods showed a good performance on the validation cohort (C-Index: 0.65±0.01 and 0.64±0.01, respectively). Also the model-based approaches multi-Cox, PVI-RF and RF-VI displayed a good performance (C-Index: 0.64±0.01, 0.63±0.01 and 0.63±0.01, respectively).

For OS, the highest median performances were revealed by BGLM-Cox (C-Index: 0.61±0.02) and SRM algorithm (C-Index: 0.61±0.03, figure 4.4 (B)). The performance of the feature selection methods was in general similar among the methods and lower compared to the LRC (figure 4.4 (B)). The highest median performances were achieved by the feature selection methods based on mutual information optimisation: MIM and MRMR (both C-Index: 0.61±0.01) and the model-based approach: uni-Cox (C-Index: 0.61±0.01) as well as by None feature selection (C-Index: 0.61±0.04).

(II) Risk-based patient stratification

For each combination of a feature selection method and a learning algorithm, patients were stratified into low and high risk groups using cut-off value based on the predicted risk determined on the exploratory cohort. The stratification results are shown in table 4.2 for the highest performing models on the validation cohort for LRC. These particular models were able to stratify the patients into low and high risk groups with a statistically significant difference in LRC using both cut-off selection methods, confirming the applicability of each model. The Kaplan-Meier curves for the BT-Weibull model on the exploratory and validation cohort are shown as an example in figure 4.5(A). However, several models with a high C-Index in the validation cohort were not able to separate the patients into two groups with significantly different in LRC on the validation cohort. For instance, the Cox model in combination with Spearman feature selection method achieved a high C-Index of 0.68, however, it was not able to stratify patients with a significant differences in LRC patients into two risk groups using both cut-off values ($\text{median}_{\text{risk}}$: p=0.058 and $\text{boot}_{\text{risk}}$: p=0.27, respectively). Figure 4.5(B) shows the Kaplan-Meier curves as an example. On the contrary, some models with a moderate prognostic performance could stratify patient into two significantly different groups, e.g.,



(A) Loco-regional tumour control

	Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Cox	0.58	0.68	0.61	0.64	0.64	0.61	0.64	0.55	0.62	0.61	0.63	0.63	0.61	0.51	0.63
NET-Cox	0.60	0.68	0.63	0.66	0.66	0.63	0.65	0.53	0.64	0.61	0.64	0.63	0.62	0.61	0.50
BT-Cox	0.51	0.70	0.58	0.64	0.66	0.61	0.60	0.56	0.60	0.62	0.58	0.62	0.63	0.63	0.65
BT-CIndex	0.60	0.71	0.65	0.64	0.65	0.61	0.64	0.56	0.59	0.61	0.58	0.61	0.64	0.61	0.63
BGLM-Cox	0.62	0.68	0.62	0.65	0.65	0.62	0.64	0.55	0.64	0.61	0.64	0.63	0.57	0.59	0.63
BGLM-CIndex	0.62	0.68	0.62	0.64	0.64	0.62	0.64	0.56	0.63	0.63	0.64	0.64	0.58	0.61	0.63
RSF	0.61	0.68	0.64	0.65	0.65	0.64	0.66	0.61	0.65	0.61	0.62	0.64	0.57	0.64	0.62
MSR-RF	0.57	0.71	0.66	0.68	0.67	0.65	0.63	0.61	0.65	0.61	0.58	0.64	0.60	0.63	0.63
SRM	0.58	0.68	0.62	0.64	0.64	0.62	0.64	0.56	0.63	0.62	0.62	0.63	0.58	0.51	0.63
BT-Weibull	0.62	0.70	0.60	0.64	0.64	0.60	0.64	0.62	0.63	0.61	0.60	0.62	0.64	0.61	0.64
BGLM-Weibull	0.61	0.67	0.61	0.63	0.64	0.60	0.64	0.54	0.62	0.62	0.64	0.60	0.58	0.61	0.63

(B) Overall survival

	Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Cox	0.61	0.61	0.58	0.61	0.50	0.61	0.62	0.57	0.58	0.60	0.58	0.60	0.57	0.61	0.62
NET-Cox	0.60	0.61	0.61	0.61	0.61	0.61	0.60	0.56	0.59	0.61	0.57	0.60	0.50	0.61	0.63
BT-Cox	0.58	0.58	0.54	0.58	0.60	0.58	0.58	0.51	0.57	0.62	0.61	0.61	0.54	0.58	0.61
BT-CIndex	0.58	0.59	0.54	0.59	0.61	0.58	0.59	0.57	0.59	0.60	0.60	0.61	0.54	0.59	0.62
BGLM-Cox	0.60	0.61	0.61	0.61	0.60	0.61	0.61	0.57	0.58	0.61	0.64	0.59	0.57	0.61	0.62
BGLM-CIndex	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.57	0.59	0.61	0.64	0.62	0.57	0.62	0.62
RSF	0.59	0.61	0.59	0.59	0.62	0.59	0.59	0.56	0.59	0.60	0.60	0.59	0.57	0.60	0.63
MSR-RF	0.59	0.61	0.59	0.60	0.61	0.59	0.58	0.55	0.60	0.60	0.57	0.62	0.50	0.61	0.59
SRM	0.61	0.61	0.61	0.61	0.51	0.61	0.61	0.57	0.58	0.60	0.63	0.60	0.57	0.61	0.61
BT-Weibull	0.59	0.58	0.60	0.60	0.61	0.52	0.59	0.52	0.58	0.59	0.60	0.61	0.53	0.58	0.60
BGLM-Weibull	0.61	0.61	0.61	0.61	0.60	0.61	0.61	0.57	0.58	0.61	0.64	0.60	0.57	0.61	0.63

Figure 4.3.: Concordance indices for the prediction of loco-regional tumour control (top) and overall survival (bottom) depending on the feature selection method (columns) and learning algorithm (rows) for the validation cohort are depicted. Furthermore the performance using the Aerts *et al.* (Aerts et al., 2014) signature is shown.

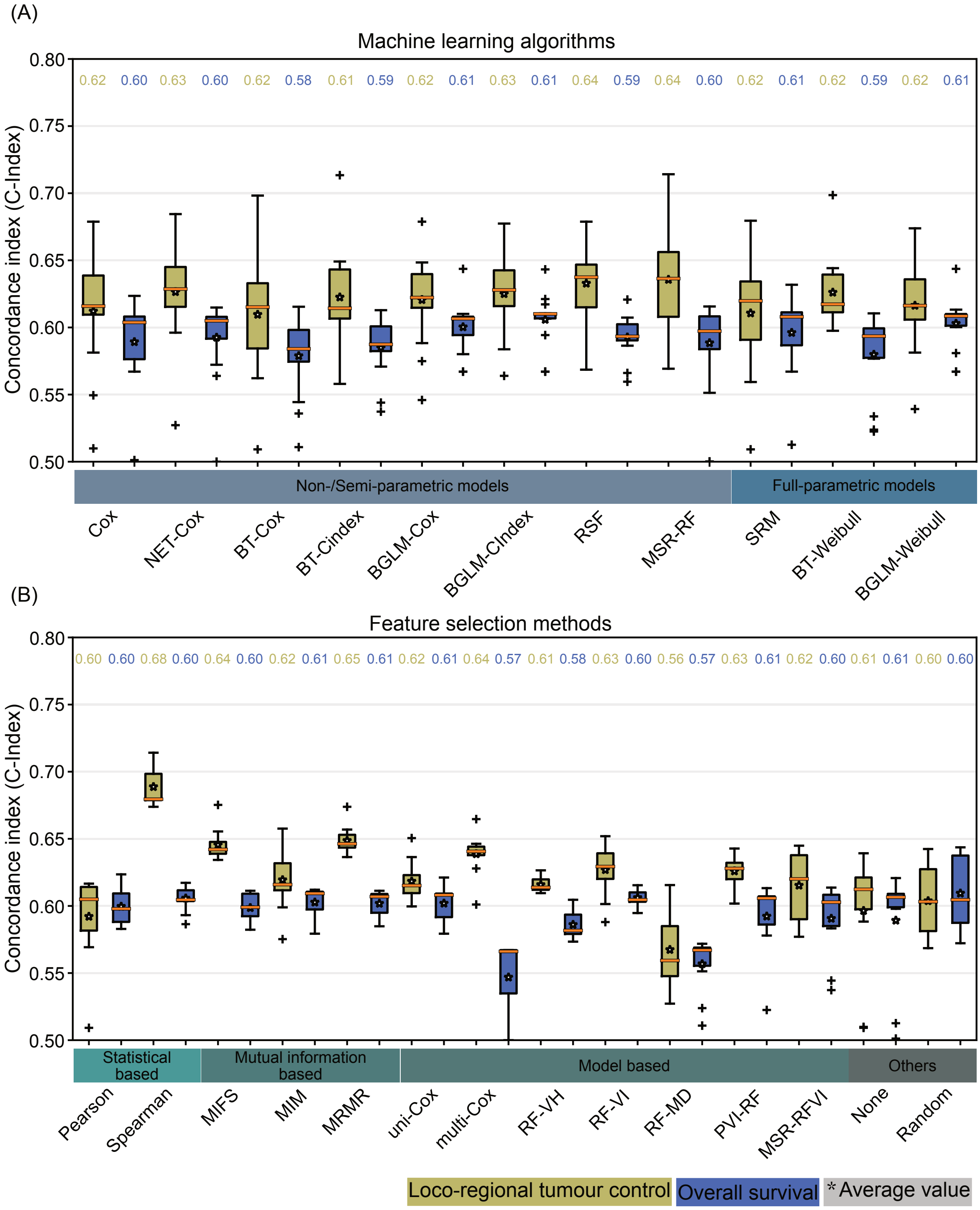


Figure 4.4.: Boxplots of the concordance indices (C-Index) on the validation for the different machine learning algorithms over all feature selection methods (top) and vice-versa (bottom) for the prediction of loco-regional tumour control (green) and overall survival (blue). Furthermore, the median values (orange line and numbers) and the average prognostic performances (asterisk) are shown.

the RSF model trained with the feature determined by the RF-MD feature selection method (C-Index: 0.61, both median$_{\text{risk}}$ and boot$_{\text{risk}}$: p=0.011, respectively).

For OS, the stratification results for the highest performing models are summarised in table 4.2. Stratification using the median$_{\text{risk}}$ method was not able to stratify the patients into low and high risk groups with a statistically significant difference in OS. However, the separation using the boot$_{\text{risk}}$ method led to significantly different OS between the patients in the low and high risk groups on the validation cohort. Kaplan-Meier curves of the BT-Weibull model in combinaiton with the random feature selection methods are shown as an example in figure 4.5(A). Furthermore, several models with a low C-Index in the validation cohort were able to separate the patients into two groups with significantly different in OS on the validation cohort. For instance, the RSF algorithm trained with the features obtained by the RF-VI feature selection method achieved only a moderate prognostic performance but a good stratification result using the bootstrapped cut-off method (C-Index: 0.60, median$_{\text{risk}}$: p=0.13, boot$_{\text{risk}}$: p=0.008), which is depicted as an example in figure 4.6(B). The p-values of the log-rank tests for all model combinations and both clinical endpoints on the validation cohort using the median$_{\text{risk}}$ and the boot$_{\text{risk}}$ cut-off calculations methods are depicted in appendix C.2 and appendix C.3, respectively.

(III) Radiomics feature stability

The stability analysis showed generally a high feature stability against the different image rotations and translations for the signatures selected for the endpoints LRC and OS. Figure 4.7 shows the boxplots of the obtained ICC values including the average ICC value over the feature selection methods in combination with all machine learning algorithms.

For LRC, radiomics features selected by the feature selection methods with the highest median performances showed high ICC values on average (MIM: 0.95, MRMR: 0.82, multi-

Table 4.2.: For the best performing model combinations, consisting of one feature selection method and one machine learning algorithm, the concordance index (C-Index) on the exploratory and validation cohort for loco-regional tumour control (LRC) and overall survival (OS) are summarised. Furthermore the p-values of the log-rank tests on the validation cohort are shown.

Endpoint	Model combination	C-Index		p-value	
		Exploratory	Validation	median$_{\text{risk}}$	boot$_{\text{risk}}$
	Spearman				
LRC	MSR-RF	0.84	0.71	0.008	0.008
	BT-Weibull	0.80	0.70	0.036	0.004
	BT-Cox	0.80	0.70	0.016	0.023
	BT-CIndex	0.80	0.71	0.022	0.032
	Random				
OS	BGLM-CIndex	0.59	0.64	0.052	0.009
	BGLM-Cox	0.59	0.64	0.052	0.015
	BGLM-Weibull	0.61	0.64	0.083	0.008

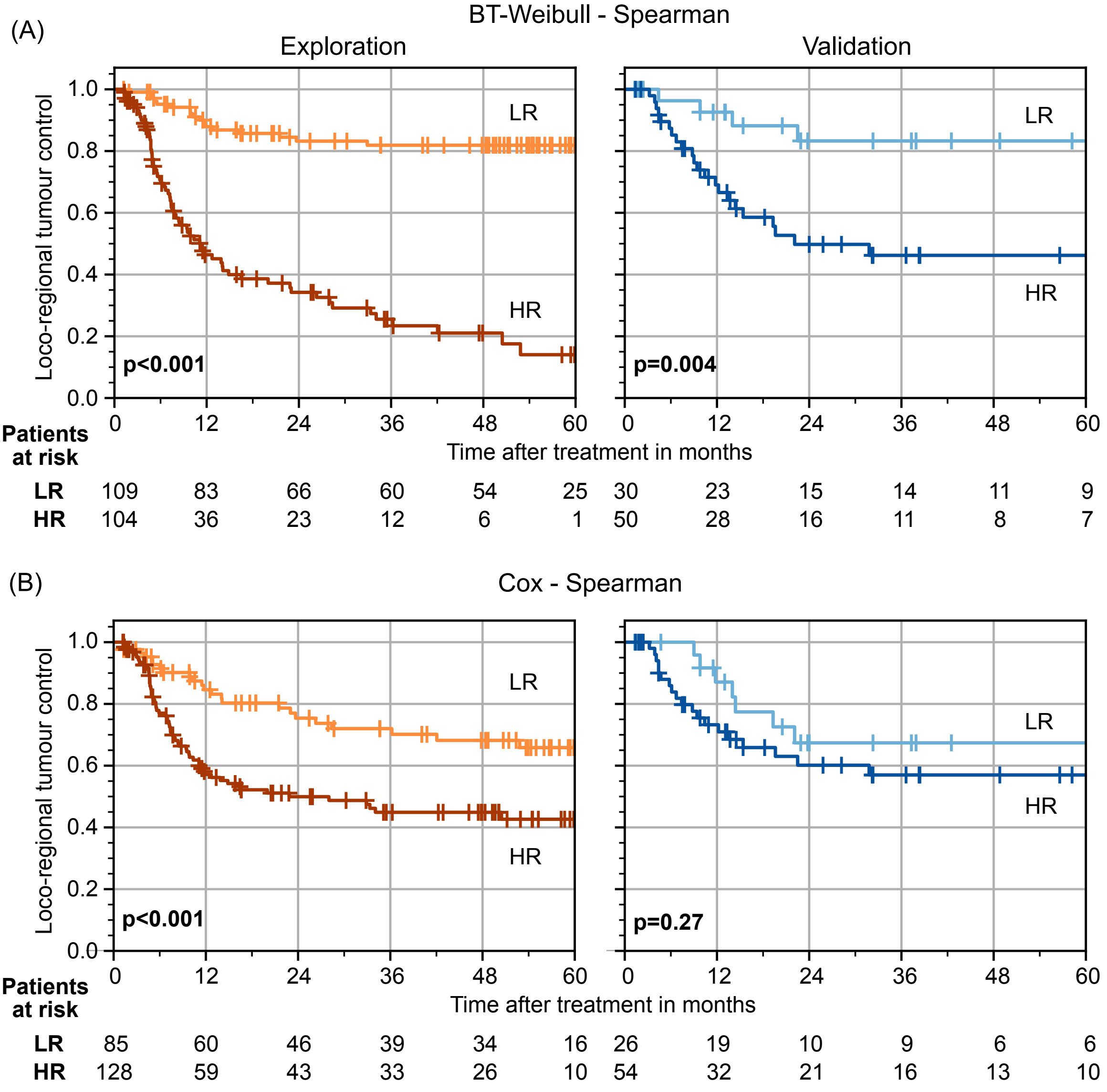


Figure 4.5.: Examples of Kaplan-Meier curves for the prediction of loco-regional tumour control (LRC) for patients of the validation cohort stratified into a low (LR) and a high (HR) risk groups using the boot$_{\text{risk}}$ cut-off value method determined on the exploratory cohort. (A) The BT-Weibull model in combination with Spearman feature selection showed a patient stratification with significantly differences as well as a high predictive performance (C-Index: 0.71). (B) The Cox model in combination with Spearman feature selection achieved a high prognostic performance (C-Index: 0.68) but the difference in LRC between low and high risk groups was not significant.

Cox: 0.97, PVI-RF: 0.92 and RF-VI: 0.92, respectively). The features selected by Spearman feature selection method achieved a moderate feature stability (ICC: 0.65). For OS, the signatures of the feature selection methods with the highest median performances achieved a high feature stability for the different image rotations and translations operations (MIM: 0.96, MRMR: 0.96 and uni-Cox: 0.96, respectively). The variability of the ICC for both endpoints are low within a feature selection method and the different learning algorithms, since the learning algorithms were developed with different signature sizes determined by the hyper-parameter optimisation.

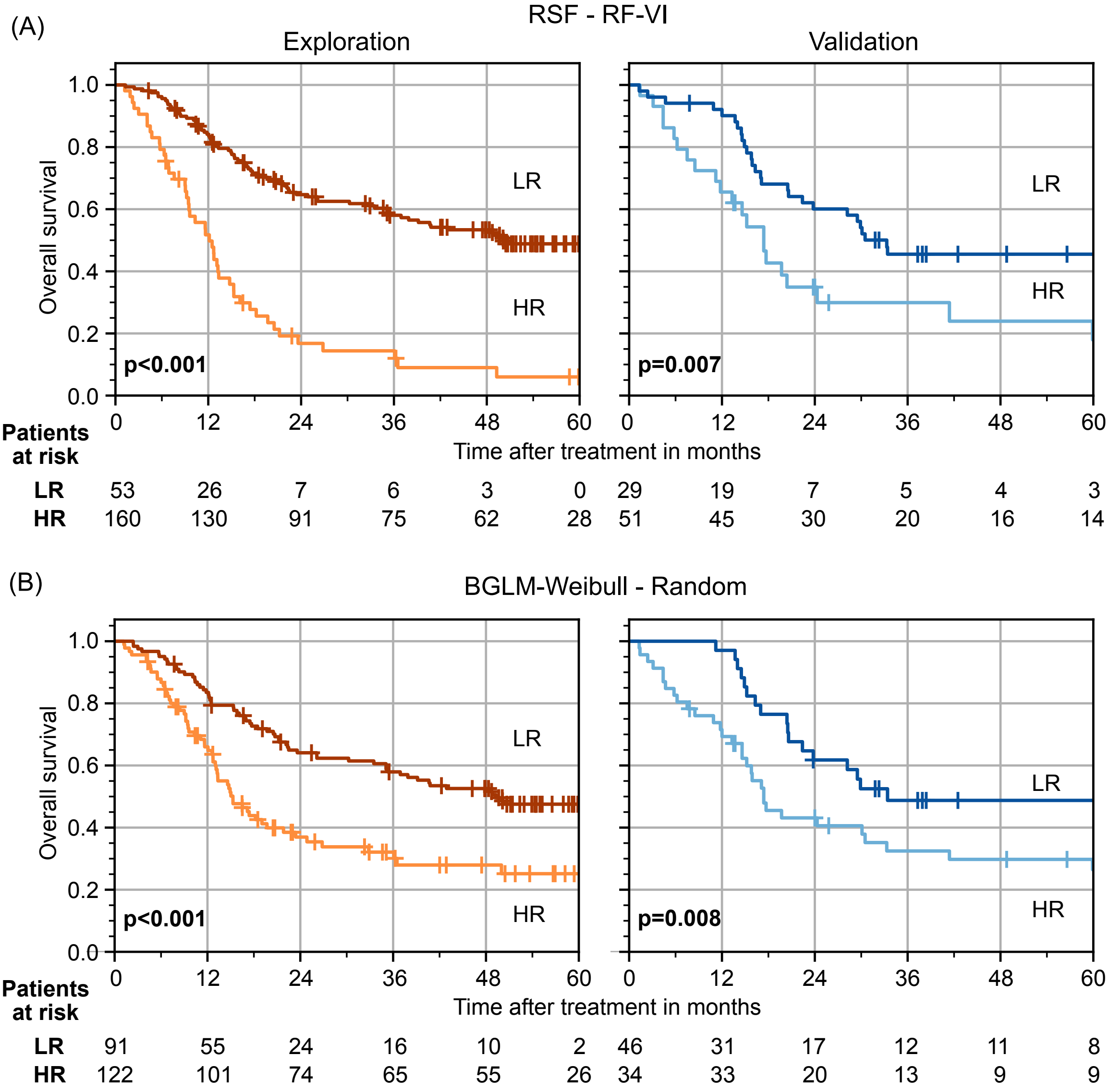


Figure 4.6.: Examples of Kaplan-Meier curves for overall survival for patients of the validation cohort stratified into a low (LR) and a high (HR) risk groups based on the boot$_{\text{risk}}$ cut-off method determined on the exploratory cohort. (A) The boosted gradient linear-based Weibull model (BGLM-Weibull) in combination with random feature selection achieved a high prognostic performance (C-Index: 0.64) as well as a significant patient stratification.(B) The random survival forest model (RSF) in combination with random forest variable importance feature (RF-VI) selection method achieved also a patient separation result with significantly differences between both groups although the prognostic performance was only moderate (C-Index: 0.60).

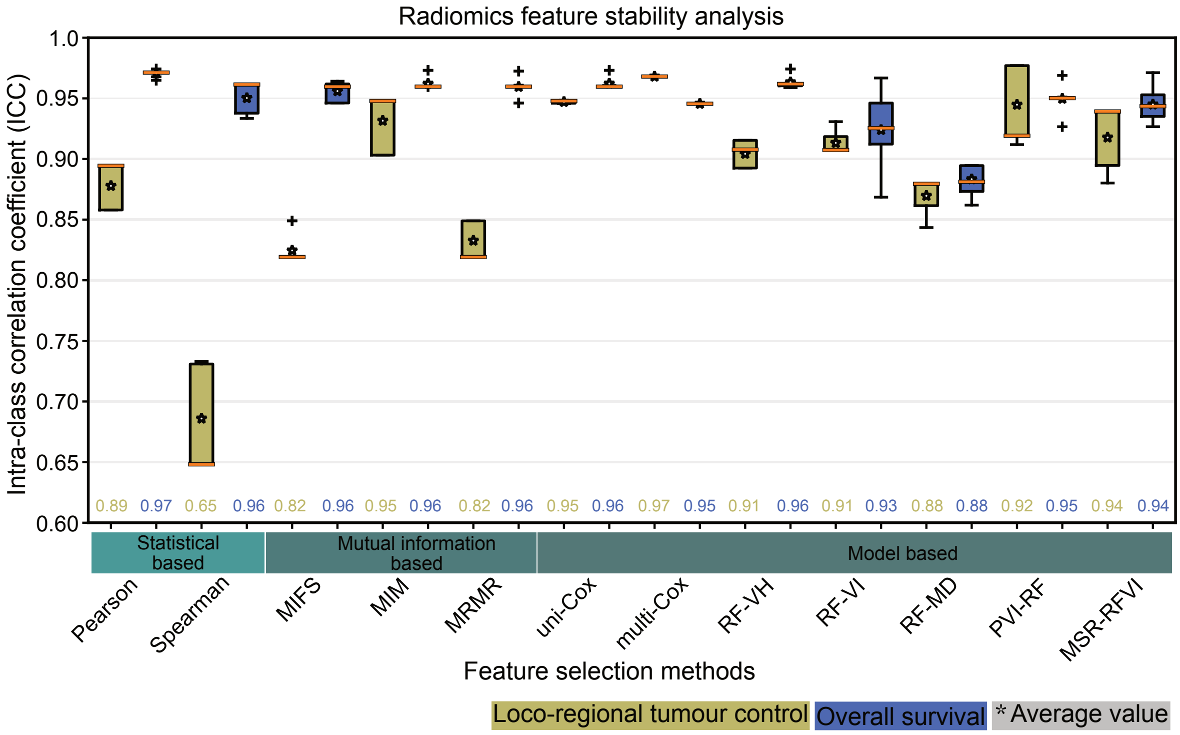


Figure 4.7.: Boxplots of the resulting intra-class correlation coefficient (ICC) of the radiomics feature stability analysis against different image rotations and translations for the developed signatures based on the features determined by the feature selection methods. The feature stability was measured of the signatures developed for the prediction of loco-regional tumour control (blue) and overall survival (green). Furthermore, the median ICC (orange line and numbers) and the average values (asterisk) are shown.

(IV) Assessment of the prognostic performance using the Aerts' signature

The previously developed radiomic signature by Aerts *et al.* (Aerts et al., 2014) showed a good performance on the validation cohort for LRC in combination with different learning algorithms (figure 4.3 (A)). The highest prognostic performance could be achieved by the BT-COX (C-Index: 0.65, [0.56–0.76]) and by the BT-Weibull algorithm (C-Index: 0.64, [0.55–0.75]). From the best performing models described above only the BT-CIndex model (C-Index: 0.71) achieved a significantly improved performance compared to the BT-Cox model (C-Index: 0.65) trained with the Aerts' signature (NPA test: p<0.001).

For OS, the highest prognostic performance was achieved by the RSF (C-Index: 0.63, [0.54–0.70]), the BGLM-Weibull (C-Index: 0.63, [0.55–0.72]), and the NET-Cox algorithm (C-Index: 0.63, [0.54–0.71]) (figure 4.3 (B)). Using the best performing models developed in this thesis no model achieved a significantly improved performance compared to the best performing models trained with the Aerts' signature (NPA-test: p>0.05).

The patient stratification into low and high risk groups based on the predicted risk of the BT-Cox model for LRC and the BGLM-Weibull model for OS showed significant differences

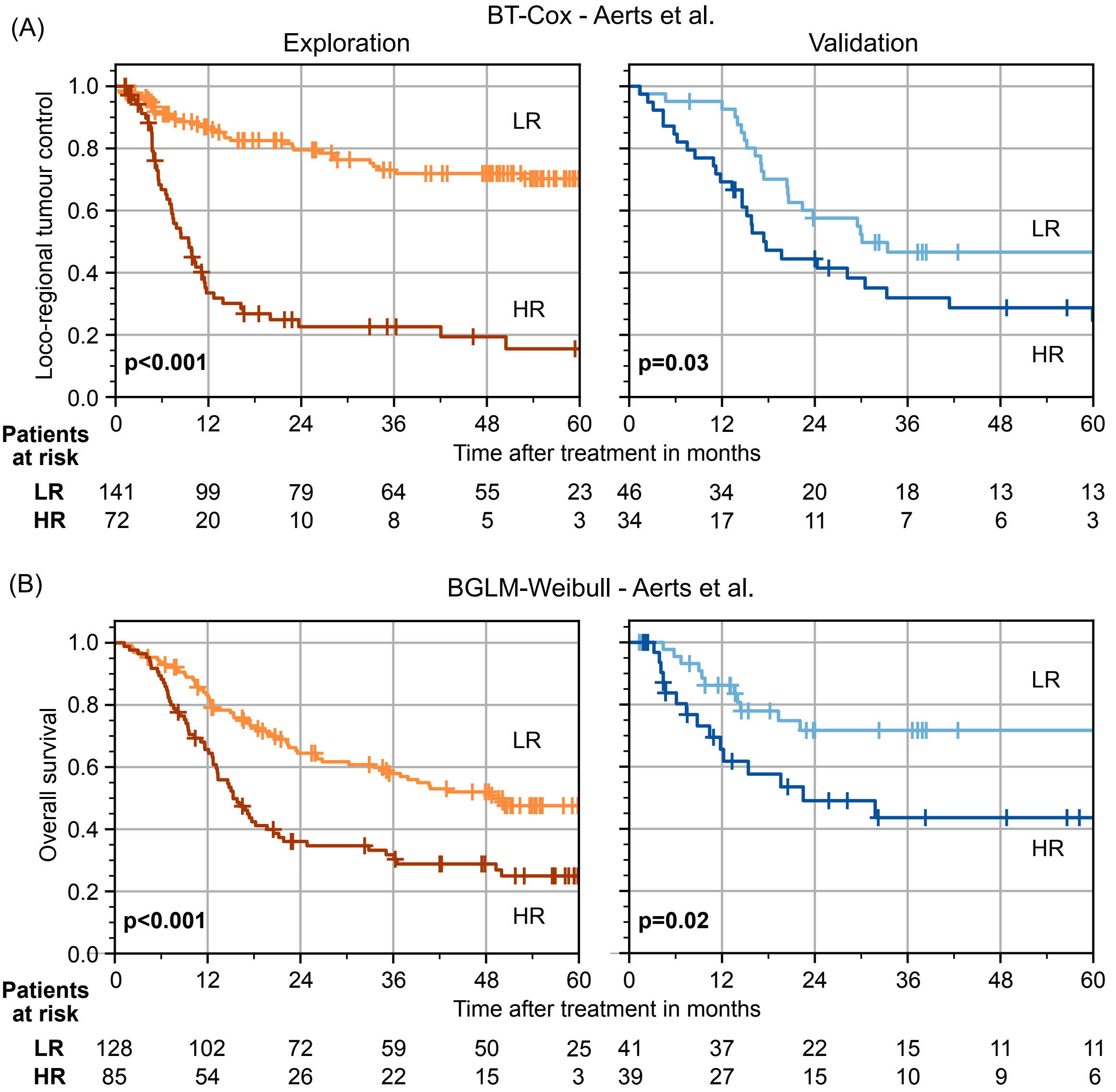


Figure 4.8.: Examples of Kaplan-Meier curves for (A) loco-regional tumour control and (B) overall survival for patients of the validation cohort stratified into low and high risk groups by the cut-off determined on the exploratory cohort. The Aerts *et al.* (Aerts et al., 2014) signature in combination with the boosted tree-based Cox model and the boosted gradient linear-based Weibull model showed a significant patient stratification as well as a high prognostic performance (concordance index: 0.65 and 0.63, respectively)

between the risk groups in the validation cohort (LRC: p=0.019 and OS: p=0.026, respectively). The Kaplan-Meier curves of the BT-Cox model for LRC and the BGLM-Weibull model for OS are shown as an example in figure 4.8.

4.4. Summary and discussion

In general, the systematic evaluation of comparing twelve feature selection methods and eleven learnings algorithms are showed a good prognostic performance with a C-Index up to 0.7 for LRC. However, there was no method which noticeably outperformed all others. Instead, a subset of different feature selection methods and learning algorithms led to similar results. This indicates that a wide range of different methods are useful and should be considered for further radiomics analyses. Moreover, applying multiple methods may decrease the influence of incidental findings which may occur in selecting one single approach. For instance, a typical choice in radiomics studies could be the Cox model combined with the Pearson feature selection method, which showed a low prognostic performance (C-Index<0.6) for the primary endpoint LRC (figure 4.3 (A)).

Furthermore, the evaluation showed that the performance differences between the learning algorithms were smaller than between the feature selection methods. This result is in line with the findings of published data. For instance, Parmar *et al.* (Parmar et al., 2015a; Parmar et al., 2015b) showed that the feature selection is an important aspect in the process of developing accurate radiomics model.

In contrast to LRC, the prognostic performance for OS was generally lower. This may occur since the cause of death does not necessarily have to be related to cancer, which led to additional bias and increases the noise on the outcome data. The best performances for OS were achieved by the BGLM-CIndex, BGLM-Weibull and BGLM-Cox models in combination by the Random feature selection method. One explanation for the good performance of Random feature selection may be that the hyper-parameter optimisation selected larger signature sizes, which were subsequently reduced by feature selection, which is performed internally by several of the machine learning algorithms, leading to this high model performance. However, for the particular risk models with such a high prognostic performance on the validation cohort showed an actually lower C-Index on the exploratory cohort. For instance, the BGLM-Cox model in combination with the Random feature selection method achieved a C-Index of 0.59 on the exploratory cohort and a higher C-Index on the validation cohort (C-Index: 0.64). This indicates that the resulting validation performance may be a statistical coincidence.

A significant difference in LRC was found between patients stratified into low and high risk groups using the best performing models, which confirms their clinical relevance. However, the results strongly depend on the selection process of the cut-off value, and not necessarily on the performance of the risk model. This was the case for multiple models which predicted risk well, yet did not lead to a stratification of patients to risk groups with significant differences in LRC. Therefore, the applied cut-off selection process, which is based on different bootstrap samples, could be an alternative for patient stratification compared to the median cut-off selection. For instance, the cut-off determination based on bootstrap samples led to more significant results (*n*=30) than the median cut-off selection process (*n*=20) on the validation cohort for LRC. Furthermore, it may be conceivable to stratify patients using unsupervised learning techniques, e.g., fuzzy c-means clustering algorithms (Bezdek et al., 1984). This has the advantage that the stratification could be learned based on the predicted risk using the exploratory cohort and the patients could be stratified in more than two risk groups (e.g., a low, middle and high risk group).

The signature obtained by the Spearman feature selection method achieved a higher validation performance than the published signature by Aerts *et al.* (Aerts et al., 2014), which comprises the best features from each of the four feature groups (statistical, shape, texture and wavelet features). In contrast, our radiomics signature for the Spearman method consisted mainly of features extracted from transformed images (e.g., wavelet), which showed an increased sensitivity against image perturbations, e.g., image rotations and translations nut still achieved the highest validation performance. This indicates a limited effect on the prognostic performance of those image perturbations. Signatures obtained by different feature selection methods were generally robust against image rotations and translations. One explanation for this behaviour could be the included patients from different institutions, resulting in a highly heterogeneous data set, which captures the variability between different CT settings and reconstruction parameters (Kim et al., 2016; Zhao et al., 2016). Therefore, the obtained signatures might be less biased from single-centre selection effects and thereby more generalisable and stable against such image perturbations. Furthermore the stability of feature selection methods was not assessed directly during the feature selection process, since the selected features in a particular bootstrap varied greatly from one bootstrap to the next. The features were aggregated according to their ranks and selected based on the rank and the occurrence using an adaptation of the enhanced Borda score (Wald et al., 2012), which may also increase the feature stability. To further enhance the stability of radiomics signatures, feature stability information, e.g., derived from test re-test datasets, could be included in the future. Furthermore, to improve the comparability and applicability of radiomics signatures, image processing should be done according to the recommendations of the imaging biomarker standardisation initiative (Zwanenburg et al., 2016).

Based on the systematic evaluation of feature selection methods and machine learning algorithms for continuous time-to event survival data we conclude that the Cox model can be used as a baseline prognostic model. The Cox model, despite its simplicity was able to achieve results comparable to the more complex models. In additional, the tree based methods (BT-Cox, RSF, MSR-RF) or the full parametric models like BT-Weibull and the BGLM-Cox should be considered. In the case of feature selection methods the Spearman correlation coefficient and mutual information based methods (MRMR, MIM and MIFS) are recommended. Multivariate-Cox feature selection as well as random forest baseds method (RF-VI, PVI-RF) led to an acceptable performance and may also be evaluated. In conclusion, a wide range of available machine learning methods appear useful for future radiomics studies. The application of suitable feature selection methods and learning algorithms is an important step to develop highly accurate and reliable clinical risk models and to reduce the risk of incidental findings.

5. Characterisation of tumour phenotype using computed tomography imaging during treatment

5.1. Motivation

Most radiomics models are based on pre-treatment imaging, which has shown promising results in several studies using different image modalities, such as CT, PET or MRI to predict survival outcome data (El Naqa et al., 2009; King et al., 2013; Aerts et al., 2014; Parmar et al., 2015c; Kickingereder et al., 2016; Song et al., 2016). Imaging during treatment may be of additional value, since it may reflect biological processes associated with therapy response, such as re-oxygenation and/or tumour shrinkage (Dietz et al., 2003; Ljungkvist et al., 2007; Linge et al., 2016b). Therefore, the consideration of CT imaging data acquired during the course of treatment may enhance the prognostic performance of radiomics risk models.

Several studies investigated the prognostic value of specific image feature over time, e.g., using PET imaging (van Putten, 1968; Dietz et al., 2003; Ljungkvist et al., 2007; Yaromina et al., 2011; Zips et al., 2012). For patients with locally advanced HNSCC, Hentschel *et al.* (Hentschel et al., 2011) showed that the decrease of the maximum standard uptake value (SUV) extracted from ^{18}F-fluorodeoxyglucose (FDG)-PET imaging in treatment weeks one or two had a higher prognostic value than at baseline. Furthermore, Zips *et al.* (Zips et al., 2012) demonstrated the strong prognostic value of ^{18}F-fluoromisonidazole (FMISO)-PET imaging parameters after week one and two of radiotherapy, which was recently validated (Löck et al., 2017).

In the field of radiomics, so far only few studies have assessed the change or the prognostic value of radiomics features during the course of treatment. Cunliffe *et al.* (Cunliffe et al., 2015) investigated the relationship between radiation dose characteristics and the change of CT-based radiomics features with the development of radiation pneumonitis using imaging before and after treatment. Recently, Fave *et al.* (Fave et al., 2017) showed that quantitative radiomics features derived from CT significantly change during treatment. However, they also found that these changes contain only limited prognostic value for patients with non-small cell lung cancer (delta radiomics). Van Timmeren *et al.* (van Timmeren et al., 2017a) described a feature selection methodology using cone beam CT (CBCT) to select reproducible delta radiomics features that are informative due to their change during treatment. However, the prognostic value of those features was not investigated. With those limited data available, additional studies are required to evaluate the possible improvement of prognostic radiomics models on CT imaging data acquired during treatment. Therefore, the main objective of this study was to investigate the potential of radiomics risk models (for LRC and OS) trained on pre-treatment planning CT imaging in comparison to CT imaging in the second week of radio-chemotherapy and to the combination of both data sets for patients with locally advanced HNSCC.

5.2. Patient cohort and experimental design

5.2.1. Characteristics of patient cohort

Radiomics risk models were developed and validated on two different patient cohorts with 78 patients in total. All patients were diagnosed with histologically confirmed locally advanced HNSCC and received primary radio-chemotherapy (RCT). The study design is presented in figure 5.1. The exploratory cohort consisted of 48 patients, treated within a prospective clinical trial (NCT00180180, (Zips et al., 2012; Löck et al., 2017)) at the UKD between 2006 and 2012. The imaging data comprises an FDG-PET/CT scan ($\text{CT}_{\text{W0-FDG}}$), which was used for treatment planning. Furthermore, FMISO-PET/CT scans were acquired between two and four days after the planning CT scan, but prior to initiation of RCT ($\text{CT}_{\text{W0-FMISO}}$), and after a dose of 18-20 Gray (Gy) (CT_{W2}, end of week two of RCT).

The validation cohort consisted of 30 patients, who were treated at the UKD and the Hospital Dresden-Friedrichstadt between 2005 and 2009. Imaging in this cohort contained an FDG-PET/CT ($\text{CT}_{\text{W0-FDG}}$) scan for treatment planning and a subsequent CT scan after a dose of 18-20 Gy (CT_{W2}, end of week two or three) during RCT. All imaging data were acquired with treatment masks in supine radiotherapy position.

Due to the differences between $\text{CT}_{\text{W0-FDG}}$ and the CT_{W2} in terms of the CT acquisition parameters, e.g., CT exposure feature stability was assessed prior to model development. Therefore, an additional cohort of 18 patients with HNSCC was included to assess the stability of radiomics features. These patients were treated within a prospective clinical trial at the UKD between 2014 and 2016 (DRKS00006007). Imaging data and time points in this cohort were comparable to the exploratory cohort. This cohort was excluded from further analyses due to insufficient follow-up for the evaluation of LRC and OS.

The patient characteristics of the exploratory and the validation cohorts are summarised in table 5.1. The clinical endpoints LRC (primary) and OS (secondary) were calculated from the first day of RCT to the date of event or censoring. Binary variables were compared between the patient cohorts using exact Fisher or χ^2 tests, while differences in continuous variables were evaluated by Mann-Whitney-U tests. Median follow-up time was 28.8 months (range: 1.3–70.3 months) for the exploratory cohort and 21.5 months (range: 1.4–107.2 months) for the validation cohort. The two-year LRC rate was 63.0% for the exploratory and 56.0% for the validation cohort (log-rank test: p=0.61). Overall survival after two years was 50.0% for the exploratory and 53.0% for the validation cohort (log-rank test: p=0.56). The corresponding Kaplan-Meier curves for both endpoints are shown in figure 5.2. Patients in the validation cohort had a significantly lower clinical T stage (χ^2 test: p<0.001).

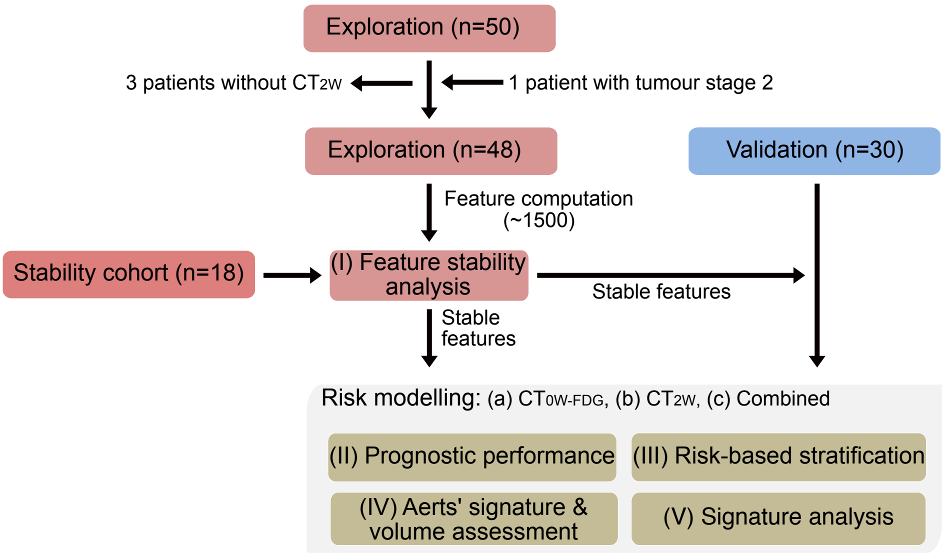

Exploration (n=50)
3 patients without CT2W
1 patient with tumour stage 2
Exploration (n=48)
Validation (n=30)
Feature computation (~1500)
Stability cohort (n=18)
(I) Feature stability analysis
Stable features
Stable features
Risk modelling: (a) CT0W-FDG, (b) CT2W, (c) Combined
(II) Prognostic performance
(III) Risk-based stratification
(IV) Aerts' signature & volume assessment
(V) Signature analysis

Figure 5.1.: Representation of the study design. Three cohorts were included. Computed tomography (CT) images from the exploratory cohort and stability cohort were used to identify a stable feature set. Subsequently, pre-treatment ($CT_{W0\text{-}FDG}$) and in-treatment (CT_{W2}) images from the exploratory cohort as well as their combination were used to train prognostic radiomics models. Prognostic models were also trained using the radiomics signature obtained by Aerts *et al.* (Aerts et al., 2014) and the tumour volume. Prognostic model performance and patient risk group stratification were assessed on the validation cohort. In addition, correlation of the signatures with ^{18}F-fluoromisonidazole positron emission tomography parameters was investigated using the exploratory cohort.

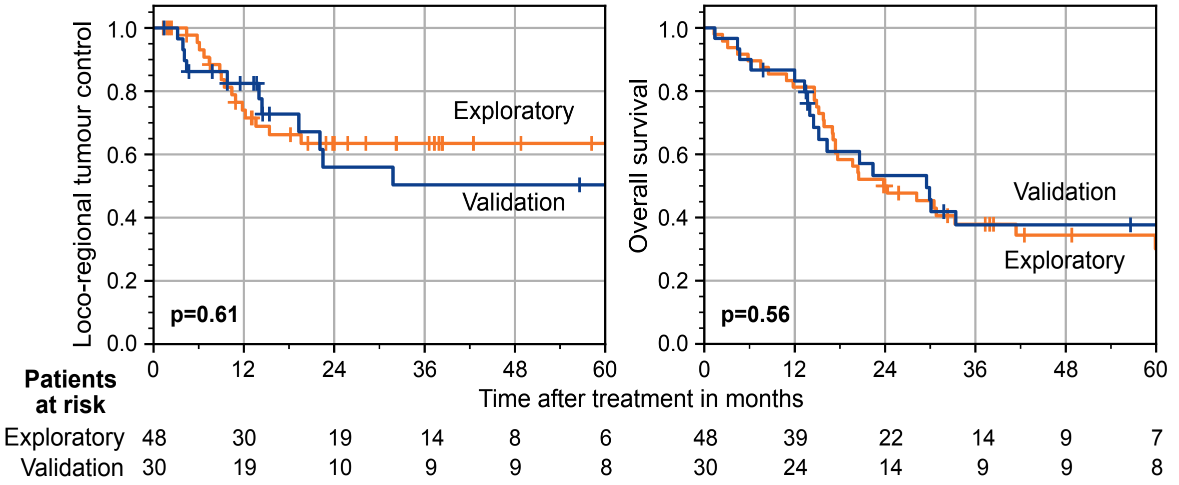


Figure 5.2.: Kaplan-Meier curves of the exploratory and validation cohort for (left) the primary endpoint loco-regional tumour control and (right) the secondary endpoint overall survival.

Table 5.1.: Patient characteristics of the exploratory and the validation cohort.

Clinical variable	Exploratory cohort	Validation cohort	*p*-value
Number of patients	48	30	-
Gender			0.47^1
male	41	27	
female	7	3	
Age in years			
median	53.5	54.5	0.79^3
range	42 - 74	37 - 74	
TN staging			
T stage 1 / 2 / 3 / 4 / missing	0 / 1 / 17 / 30 / 0	3 / 8 / 9 / 9 / 1	$<$0.001^2
N stage 0 / 1 / 2 / 3 / missing	4 / 6 / 37 / 1 / 0	4 / 2 / 21 / 2 / 1	0.43^2
UICC stage 2010			
I / II / III / IV / missing	0 / 0 / 7 / 41 / 0	1 / 2 / 1 / 26 / 0	0.069^1
Tumour volume in cm^3			
median	40.7	23.4	0.94^3
range	7.3 - 239	2.7 - 183	
In-treatment tumour volume in cm^3			
median	39.5	18.1	1.0^3
range	6.8 - 248	2.8 - 173.6	-
Prescribed dose in Gy			
median	72	72	0.32^3
range	69 - 72	71 - 77	
HPV16 DNA status			
negative / positive / missing	36 / 5 / 7	0 / 0 / 30	-
Number of events			
LRC	15	11	-
OS	33	17	-
Follow up time of patients alive in months			
median	38.4	61.7	-
range	23.8 - 70.3	7.8 - 107.2	

Abbreviations: T, clinical tumour stage; N, clinical nodal stage; UICC, Union internationale contre le cancer
Gy, Gray; HPV, human papillomavirus; DNA, deoxyribonucleic acid
LRC, loco-regional tumour control; OS, overall survival
^{1} exact Fisher test; ^{2} χ^2 test; ^{3} Wilcoxon-Mann-Whitney test

5.2.2. Experimental design

Feature computation and radiomics risk modelling

The GTV of the primary tumour was manually delineated by a radiation oncologist and independently validated on each CT scan separately. The voxel spacing was resampled using trilinear image interpolation to an isotropic voxel size of 1.0×1.0×1.0 mm^3 to correct for differences in voxel spacing and slice thicknesses between the cohorts (Aerts et al., 2014; Shafiq-Ul-Hassan et al., 2017). Additional images were created by applying spatial filtering to the base image to emphasise image characteristics such as edges and blobs. Eight additional images were created by applying a stationary coiflet-1 wavelet high-/low-pass filter along each of the three spatial dimensions (section 2.3.1). One mean LoG image was additional created by applying a LoG filter consisting of five different filter kernel widths (1.0, 2.0, 3.0, 5.0, 6.0 mm, section 2.3.1). Subsequently, the GTV mask was re-segmented to cover only soft tissue voxels between -150 and 180 HU, thereby removing voxels containing air and bone, which may affect feature expression. Finally, according to the image biomarker standardisation initiative (ISBI) 18 statistical, 38 histogram-based and 95 texture features were extracted from the GTV within each image set (base image and nine transformed images). 28 morphological features were computed using the base image only, leading to 1538 features in total (Zwanenburg et al., 2016).

The radiomics risk models were developed within the RMF (section 2.3) consisting of five major processing steps: (I) feature pre-processing, (II) feature selection, (III) hyper-parameter optimisation, (IV) model development and (V) model validation. Feature selection was repeated using 1000 bootstrap samples (i.e.,.632 bootstrap method with replacement) of the exploratory cohort to reduce randomness in the selection of relevant features. Afterwards, the hyper-parameter optimisation was performed followed by model training. The models were trained using 1000 bootstrap samples (i.e.,.632 bootstrap method with replacement) of the exploratory cohort for each combination of feature selection method and machine learning algorithm. Subsequently, an ensemble prediction was made by averaging the predicted risk scores for each model using data of the independent validation cohort.

Feature selection methods and machine learning algorithms

For the risk modelling six different feature selection methods and learning algorithms were used to avoid incidental findings. The considered methods were found as most reliable according to the previous chapter 4. The following feature selection methods were applied: Spearman correlation, MIM, MIFS, MRMR, RF-VI and a forward feature selection based on Cox regression model (multi-Cox). For model building the following algorithms were used: Cox, BT-Cox, BGLM-Cox, RSF and MSR-RF. Additionally, we investigated the full-parametric BT-Weibull model.

Performance assessments

The systematic evaluation consists of five analyses (figure 5.1):

(I) The stability of radiomics image features was assessed prior to model development and validation to reduce the influence of different CT acquisition parameters on the prognostic models (table 5.2). The in-treatment scans ($\mathrm{CT_{W2}}$) were acquired with a lower exposure than the pre-treatment images ($\mathrm{CT_{W0\text{-}FDG}}$) in the exploratory and validation cohorts to limit patient radiation dose, while the acquisition parameters between the $\mathrm{CT_{W0\text{-}FMISO}}$ pre-treatment and the $\mathrm{CT_{W2}}$ scans were similar. Therefore, feature stability was assessed using the $\mathrm{CT_{W0\text{-}FDG}}$ and $\mathrm{CT_{W0\text{-}FMISO}}$ images of the exploratory cohort and the additional cohort of 18 patients, leading to 66 patients in total. The $\mathrm{CT_{W0\text{-}FDG}}$ and the $\mathrm{CT_{W0\text{-}FMISO}}$ scans were rigidly registered with RayStation (version 6.0, RaySearch Laboratories AB,

Table 5.2.: Image acquisition parameters of the different cohorts.

Imaging parameter	Exploratory cohort			Stability cohort		Validation cohort	
	CT$_{\text{W0-FDG}}$	CT$_{\text{W0-FMISO}}$	CT$_{\text{W2}}$	CT$_{\text{W0-FDG}}$	CT$_{\text{W0-FMISO}}$	CT$_{\text{W0-FDG}}$	CT$_{\text{W2}}$
x, *y* spacing in mm							
(0.85, 0.85)	0	1	0	0	0	0	0
(0.97, 0.97)	12	1	0	18	0	1	0
(1.36, 1.36)	36	46	48	0	18	29	30
z spacing in mm							
2.0	0	0	1	6	18	0	0
3.0	12	1	0	12	0	1	0
5.0	36	47	47	0	0	29	30
Image reconstruction kernel							
B10s	1	0	0	0	0	0	0
B19f	0	10	12	0	18	0	0
B20f	32	37	36	0	0	28	30
B20s	0	0	0	0	0	2	0
B31f	12	0	0	18	0	0	0
NA	3	47	1	0	0	0	0
Mean exposure in mAs	30.8	7.7	7.9	39.0	9.6	39.1	8.1
Mean exposure time in ms				420–500			
Tube voltage in kV				120			

Abbreviations: NA, not available; mAs, milliampere second; kv, kilovolt

Stockholm, Sweden). Afterwards, the GTV was manually transferred from the CT$_{\text{W0-FDG}}$ to the CT$_{\text{W0-FMISO}}$ images. Imaging features with a Spearman rank correlation coefficient (SCC)$\geq$0.8 between CT$_{\text{W0-FDG}}$ and CT$_{\text{W0-FMISO}}$ were considered stable and used for feature selection and model building (Leijenaar et al., 2013).

(II) The prognostic performance was evaluated and compared based on (a) CT$_{\text{W0-FDG}}$, (b) CT$_{\text{W2}}$ and (c) the combination of CT$_{\text{W0-FDG}}$, CT$_{\text{W2}}$ using the corresponding delta features (ΔCT=CT$_{\text{W2}}$/CT$_{\text{W0-FDG}}$), i.e. the ratio of feature values derived from CT$_{\text{W2}}$ and CT$_{\text{W0-FDG}}$ for every feature to investigate the potential of the single time points and their complementary information, using the exploratory and the validation cohort. The prognostic performance of radiomics models (a)-(c) was assessed using the C-Index. Furthermore, the average model performances were compared by a multi-level model (MLM) to quantify statistical differences. Multi-level models are regression models that incorporate group-specific effects, i.e., the effect of treatment time, feature selection methods and machine learning algorithms, measured on different levels (Demidenko, 2013; Brown and Prescott, 2014). Here, a MLM was developed for assessing the differences in C-Index between the CT$_{\text{W0-FDG}}$ images and CT$_{\text{W2}}$ scans independent from the effects of the feature selection methods and learning algorithms. The MLM is described in appendix D.A. The model was fitted using Markov chain Monte Carlo in STAN software, using seven chains with 250 warm-up iterations and 250 sample iterations each (Carpenter et al., 2016).

Subsequently, models (a)-(c) generated by the feature selection method and the learning algorithm with the highest mean performance in pre-treatment using the exploratory cohort were selected as single representative models and analysed in more detail.

In addition to the external validation of the models (a) and (b), internal cross-validation experiments using the combined cohorts for the primary endpoint LRC was performed. The models were developed with RMF using a three-fold cross validation scheme with 33 repetitions. At each fold, feature selection and model training were performed 20 times.

(III) Risk-based patient stratification is an important application of radiomics risk models to show their clinical reliability, e.g., for treatment individualisation. Therefore, patients were stratified into low and high risk groups based on their predicted risk according to the radiomics models. For the cut-off calculation two different methods were applied: cut-off based on the median predicted risk value ($\text{median}_{\text{risk}}$) and based on the predicted risk value computed using bootstrap samples ($\text{boot}_{\text{risk}}$)(section 2.4). Cut-off values were applied unchanged to the validation cohort. Survival curves were estimated by the Kaplan-Meier method and differences between the two risk groups were compared by log-rank tests and p-values<0.05 were considered as statistically significant.

(IV) Assessment of the prognostic performance using a established signature may further validate the results and reduce the risk of incidental findings. The risk models were trained and validated based on the $\text{CT}_{\text{W0-FDG}}$ images and CT_{W2} scans using the Aerts' signature consisting of four imaging features as previously described in section 4.2.2. Furthermore, the prognostic performance of the tumour volume was evaluated for models (a) and (b) reflecting the clinical importance of this parameter.

(V) The developed radiomics signatures for the representative models for LRC were analysed in detail. Features within the signatures and their expression values are depicted as heatmaps for the exploratory and the validation cohort to represent the level of expression and to show possible differences between the time points. For this purpose, patients were ordered according to their predicted risk and to their risk group membership. Furthermore, the association between the radiomics features within the signatures and LRC was measured by the univariate Cox model based on bootstrap samples using the entire patient cohort to quantify their overall importance.

Moreover, Spearman correlation coefficient ρ between the developed signatures and hypoxia-specific FMISO-PET imaging features for LRC of models (a) and (b), as well as those in Aerts' signature were investigated to assess the link between the signatures and tumour hypoxia. The enhancement of tumour hypoxia and the capacity for re-oxygenation during radiotherapy vary amongst patients. If tumours partially or fully re-oxygenate early during the course of treatment they have a more favourable prognosis. Therefore, two FMISO-PET imaging features were analysed: (1) the hypoxic volume ($\text{HV}_{1.6}$), calculated using 1.6 as threshold relative to the background and (2) the peak tumour-to-background-ratio (TBR_{peak}) representing the maximum ratio between FMISO up-take in the tumour and the background, i.e., in deep neck muscles (Zips et al., 2012; Löck et al., 2017). This analysis was performed in the exploratory cohort only.

5.3. Results of computed tomography imaging during treatment

(I) Radiomics feature stability

In order to develop radiomics risk models, 1538 imaging features were extracted per CT scan. Feature stability assessment between the $CT_{W0\text{-}FDG}$ and $CT_{W0\text{-}FMISO}$ scans reduced the feature set to 269 stable imaging features (SCC$\geq$0.8), consisting of 12 statistical, 18 morphological, 26 histogram-based and 213 texture features, entering into the following analyses.

(II) Prognostic performance

The prognostic performance of different feature selection methods combined with machine learning algorithms was evaluated for the clinical endpoints LRC (primary) and OS (secondary) using imaging data acquired (a) pre-treatment ($CT_{W0\text{-}FDG}$), (b) after the second week of treatment (CT_{W2}) and (c) their combination including delta features. Figure 5.3 shows the resulting concordance indices for LRC using the validation cohort and the exploratory cohort.

The validation C-Index averaged over all learning algorithms and feature selection methods was significantly higher on the in-treatment CT_{W2} images (C-Index: 0.73±0.04, mean±SD) than on the pre-treatment $CT_{W0\text{-}FDG}$ images (C-Index: 0.62±0.04) (MLM: p=0.005). Using the combined feature set also led to improved results with a mean C-Index of 0.70±0.05 compared to the pre-treatment $CT_{W0\text{-}FDG}$ scans (MLM: p=0.06).

According to the introduced selection strategy (chapter 5.2.2), a representative model for LRC was selected based on the baseline pre-treatment $CT_{W0\text{-}FDG}$ scans, namely the BT-Cox algorithm in combination with the Spearman feature selection method. This particular combination showed a prognostic performance of 0.95 (C-Index) on the exploratory cohort (95% confidence interval [0.92–1.00]). At baseline, the model achieved a good prognostic performance of C-Index=0.65 ([0.51–0.79]) on the validation cohort. The model based on CT_{W2} scans achieved a higher performance (C-Index: 0.79, [0.77–0.96]), while the model based on the combined feature set (ΔCT) performed similar to the baseline model (C-Index: 0.65, [0.49–0.88]).

The internal cross validation experiments confirmed that the in-treatment CT_{W2} images were more prognostic on average than pre-treatment imaging using the entire data set for LRC ($CT_{W0\text{-}FDG}$: 0.61 and CT_{W2}: 0.70, respectively, MLM: p=0.16, appendix D.1).

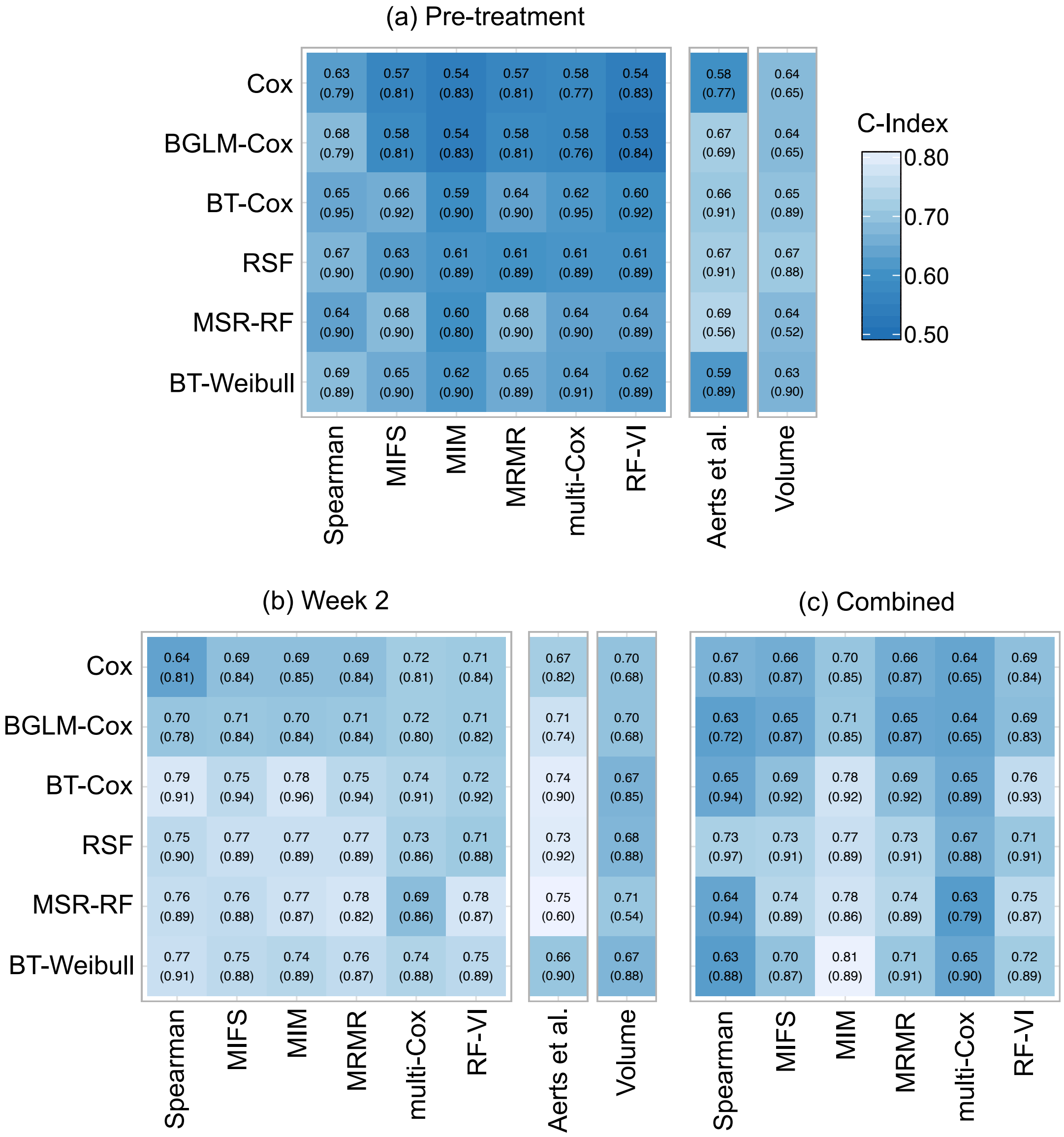


Figure 5.3.: Concordance indices for the prediction of loco-regional tumour control for the exploratory cohort (in parentheses) and the validation cohort. Radiomics models were developed using a feature selection method (columns) and a learning algorithm (rows), based on (a) pre-treamtent computed tomography (CT) images ($CT_{W0\text{-}FDG}$), (b) in-treatment CT scans (CT_{W2}) and (c) the combined feature set. Furthermore, the performance of the Aerts *et al.* (Aerts et al., 2014) signature and the tumour volume is shown.

For OS, the resulting C-Index for the validation and exploratory cohort are shown in figure 5.4. The validation C-Index averaged over all feature selection methods and learning algorithms was slightly higher on the in-treatment CT_{W2} images (C-Index: 0.62±0.04) than on the pre-treatment $CT_{W0\text{-}FDG}$ images (C-Index: 0.59±0.04), which was not statistically significant different (MLM: p=0.28). Using the combined feature set also did not led to improved results with a mean C-Index of 0.54±0.05 compared to the pre-treatment $\mathrm{CT_{W0\text{-}FDG}}$ scans (MLM: p=0.73).

The selected representative model for the baseline $\mathrm{CT_{W0\text{-}FDG}}$ scans was obtained by MIM feature selection combined with the BT-Cox model with an average performance of 0.85 on the exploratory cohort. In validation, a slightly improved performance of the $\mathrm{CT_{W2}}$ scans was observed compared to pre-treatment $\mathrm{CT_{W0\text{-}FDG}}$ images (C-Index: 0.59 and 0.61, respectively).

(III) Risk-based patient stratification

For each combination of feature selection methods and learning algorithms, patients were stratified into low and high risk groups based on the predicted risk of the radiomics risk models on the exploratory cohort. Patients were stratified into low and high risk groups according to the median risk ($\mathrm{median_{risk}}$) and to the bootstrapped-based method ($\mathrm{boot_{risk}}$). The representative BT-Cox models trained on the pre-treatment $\mathrm{CT_{W0\text{-}FDG}}$ scans and on the combined feature set were not able to stratify patients of the validation cohort with a significant difference in LRC using both cut-off calculation methods ($\mathrm{median_{risk}}$: p=0.06 and p=0.19 as well as $\mathrm{boot_{risk}}$: p=0.18 and p=0.46, respectively), whereas the model trained on the $\mathrm{CT_{W2}}$ scans led to a significant stratification for LRC based on both cut-off values ($\mathrm{median_{risk}}$: p=0.002 and $\mathrm{boot_{risk}}$: p<0.001). Figure 5.5 exemplary shows Kaplan-Meier curves for models (a)-(c) using the median cut-off values for LRC.

For OS, the representative models based on the $\mathrm{CT_{W0\text{-}FDG}}$, the $\mathrm{CT_{W2}}$ and the combined feature set were not able to stratify patients of the validation cohort with a significant difference in OS using both cut-off calculation methods (both $\mathrm{median_{risk}}$: p=0.61, p=0.68, p=0.64 and $\mathrm{boot_{risk}}$: p=0.61, p=0.68 and p=0.64, respectively). The resulting p-values of the log-rank tests using both cut-off value methods for LRC and OS are depicted in appendix D.2 and appendix D.3, respectively.

(IV) Assessment of the Aerts' radiomics signature and tumour volume

For LRC the in-treatment BT-Cox model based on Aerts' signature also led to improved prognostic performance on the validation cohort compared to the pre-treatment model confirming the achieved result using the newly developed signatures (C-Index: $\mathrm{CT_{W2}}$: 0.74, [0.61–0.91] and $\mathrm{CT_{W0\text{-}FDG}}$: 0.66, [0.51–0.89], respectively).

However, the resulting cut-off values of the proposed methods ($\mathrm{median_{risk}}$/$\mathrm{boot_{risk}}$) from both models were not able to stratify patients of the validation cohort into low and high risk groups with a significant difference in LRC ($\mathrm{CT_{W0\text{-}FDG}}$: p=0.53/p=0.23 and $\mathrm{CT_{W2}}$: p=0.30/p=0.97, respectively).

For OS the pre-treatment model based on Aerts' signature achieved a higher performance (C-Index: 0.66, [0.52–0.82]) than the in-treatment model (C-Index: 0.62, [0.48–0.80]).

Kaplan-Meier analysis of the representative model based on the $CT_{W0\text{-}FDG}$ scans could not separate the patients into two risk groups with a significant difference in OS using both cut-off methods ($\text{median}_{\text{risk}}/\text{boot}_{\text{risk}}$: p=0.13/p=0.05). However the CT_{W2} based model was able to stratify patients of the validation cohort into low and high risk groups with a significant difference in OS using the median cut-off value ($\text{median}_{\text{risk}}/\text{boot}_{\text{risk}}$: p <0.001/p=0.30). The

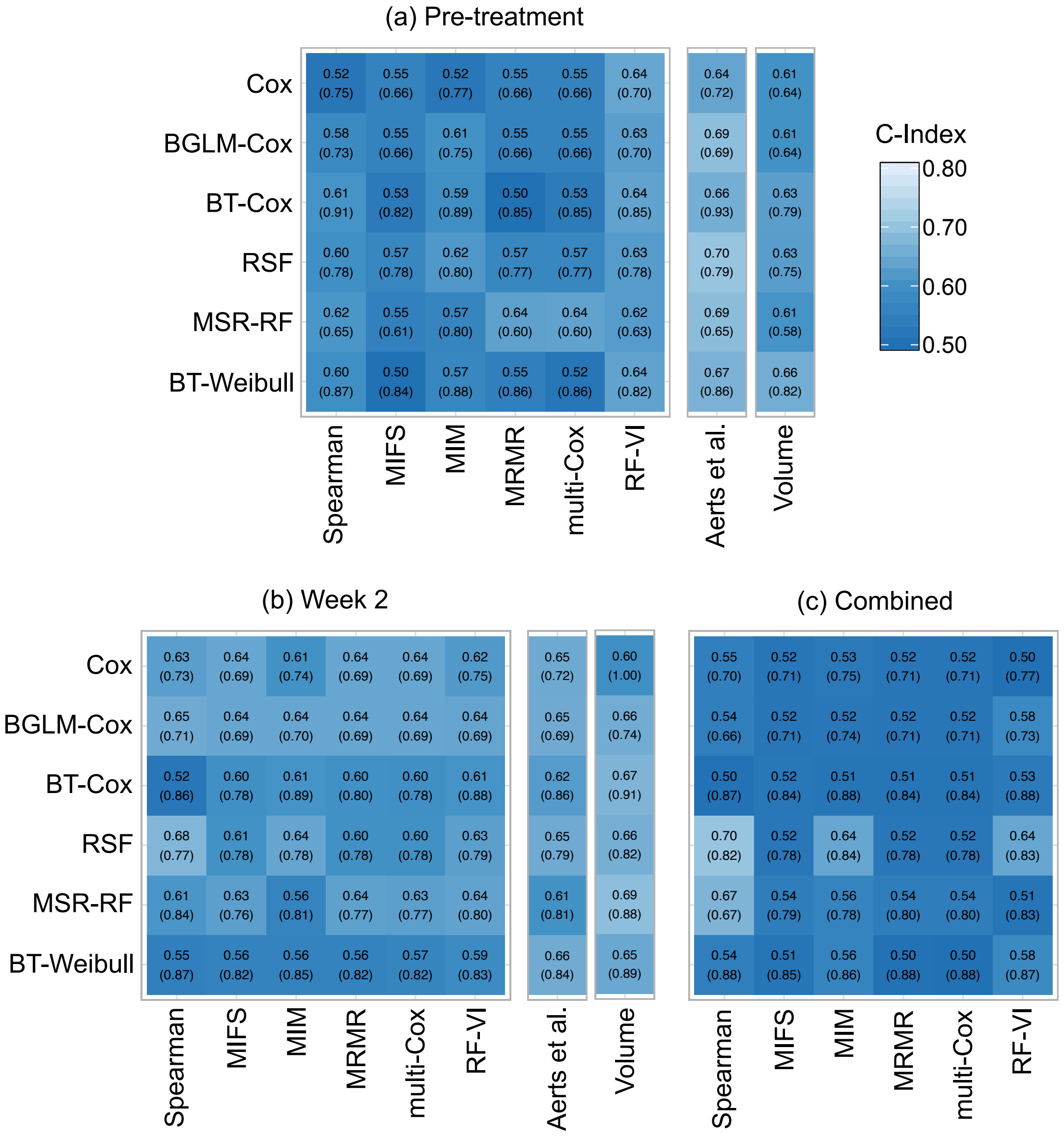


Figure 5.4.: Concordance indices for the prediction of overall survival for the the exploratory cohort (in parentheses) and validation cohort. Radiomics models were trained using a feature selection method (columns) and a learning algorithm (rows), based on (a) pre-treamtent computed tomography (CT) images ($CT_{W0\text{-}FDG}$), (b) in-treatment CT scans (CT_{W2}) and (c) the combined feature set. Furthermore, the performance of the Aerts *et al.* (Aerts et al., 2014) signature and the tumour volume is shown.

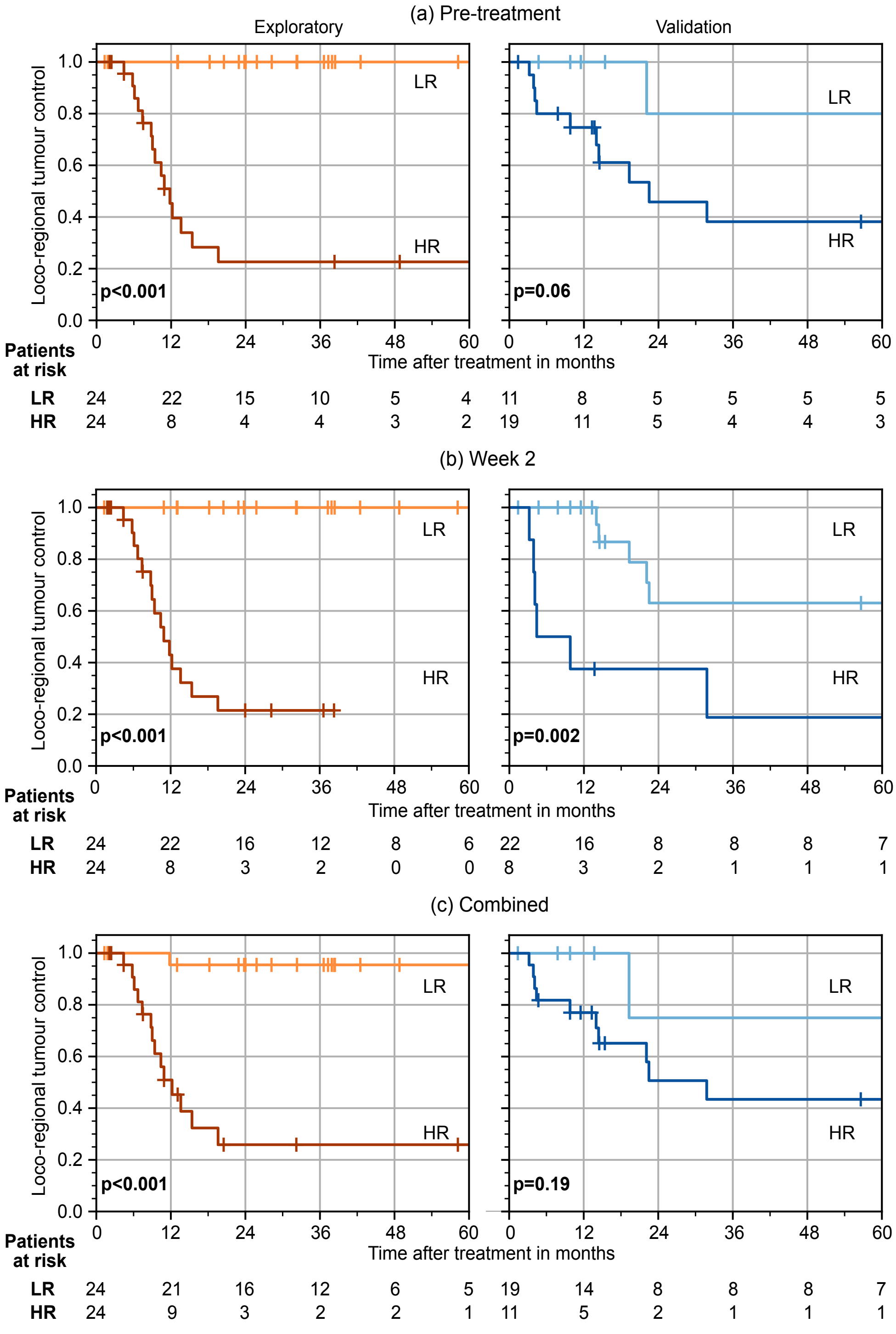


Figure 5.5.: Kaplan-Meier curves for loco-regional tumour control for patients of the exploratory (left) and the validation cohort (right) stratified into a low (LR) and a high (HR) risk group based on the median risk value determined on the exploratory cohort based on the boosted tree-based Cox model in combination with Spearman feature selection.

resulting *p*-values of the log-rank tests using both cut-off value methods for LRC and OS are depicted in appendix D.2 and appendix D.3, respectively.

Additional BT-Cox models were built using the tumour volume for both clinical endpoints. For LRC, these models achieved a good performance of C-Index=0.65 [0.51–0.84] and a high prognostic performance C-Index=0.67 [0.53–0.88] on the validation cohort for pre-treatment and in-treatment images, respectively. The representative models for OS using the tumour volume showed a good performance of C-Index=0.63 [0.50–0.80] using pre-treatment CT scans and a high prognostic performance C-Index=0.67 [0.56–0.83] for in-treatment scans.

(V) Signature analysis

Radiomics signatures were investigated for the representative models (a)-(c) for LRC. Figure 5.6 shows the feature expressions of the developed signatures for LRC. For OS, feature expressions of the developed signatures are show in appendix D.4. The names of the selected radiomics features within the signatures are summarised in appendix D.1. The signature for the pre-treatment based model consists of a first-order statistical, two texture based and a morphology based radiomics feature. One of four selected radiomics features within the signature were significant associated with LRC using univariate Cox analysis on the entire patient cohort ($\mathrm{F1}_{\mathrm{T}}$: p=0.003). The signature for the in-treatment model consists of four radiomics features, which were mainly texture based. For instance, the radiomics feature $\overline{\mathrm{F1}}_{\mathrm{T}}$ is cluster based and comprises mainly GLRLM features measuring the intra-tumour heterogeneity by assessing the gray level run lengths. Three of the four selected radiomics features within the signature showed a significant association with LRC on the entire patient cohort ($\overline{\mathrm{F1}}_{\mathrm{T}}$: p=0.001, $\mathrm{F2}_{\mathrm{T}}$: p=0.001 and $\overline{\mathrm{F4}}_{\mathrm{T}}$: p=0.02, respectively). The representative model based on the combined signature contains only delta radiomics features (ΔCT), which are texture based. Furthermore, two of the five selected features were significant associated with LRC based on the univariate Cox analyses for the entire patient cohort ($\Delta\mathrm{F2}_{\mathrm{T}}$: p=0.01 and $\overline{\Delta\mathrm{F5}}_{\mathrm{T}}$: $p<0.001$, respectively).

The correlation analysis between the radiomics signature of the representative models based on pre-treatment and in-treatment images and the FMISO-PET parameters are depicted in figure 5.7. The selected features of the radiomics signature using the pre-treatment scans were weakly correlated to both FMISO-PET parameters $\mathrm{HV}_{1.6}$ and $\mathrm{TBR}_{\mathrm{peak}}$ (ρ<0.50), with the exception of the texture-based features ($\overline{\mathrm{F2}}_{\mathrm{T}}$), which were moderately correlated to $\mathrm{HV}_{1.6}$ (0.50$\leq$ ρ<0.70). For the signature based on the in-treatment scans, a moderate correlation with $\mathrm{HV}_{1.6}$ and $\mathrm{TBR}_{\mathrm{peak}}$ exists for the $\overline{\mathrm{F1}}_{\mathrm{T}}$ ($\mathrm{HV}_{1.6}$: ρ=0.54 and $\mathrm{TBR}_{\mathrm{peak}}$: ρ=0.55, respectively) and for the $\mathrm{F2}_{\mathrm{T}}$ (both ρ=-0.56). The remaining features were only weakly correlated (ρ<0.5).

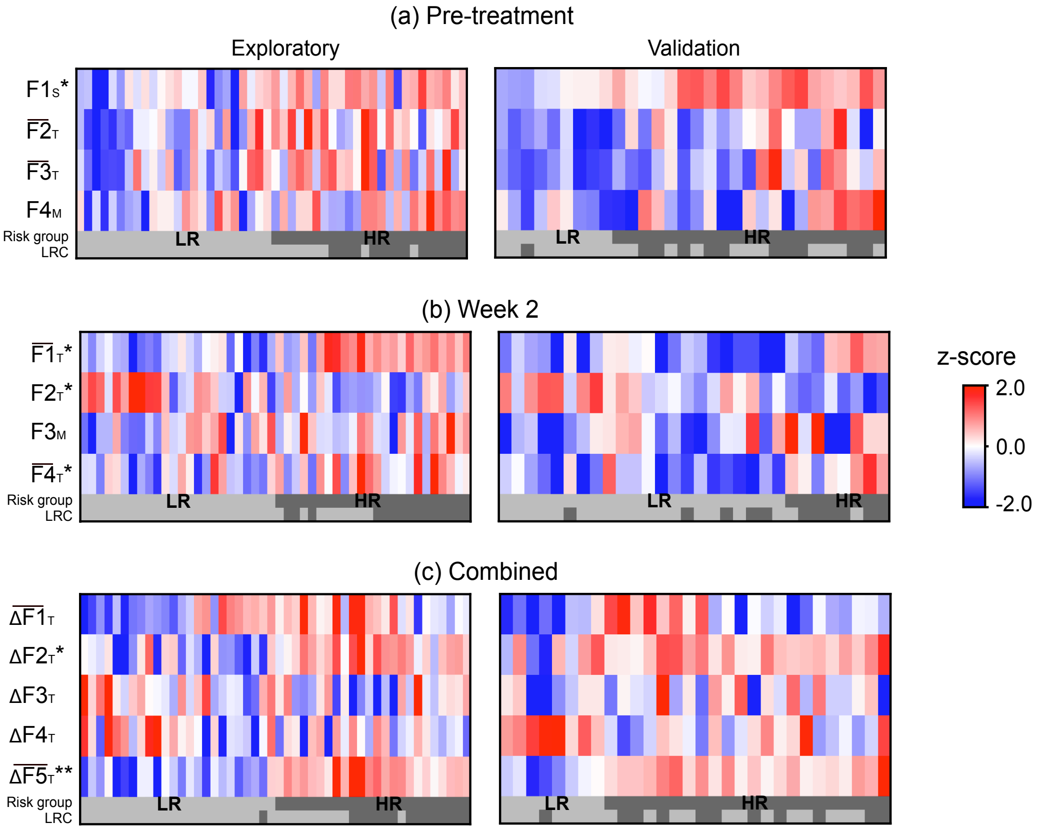
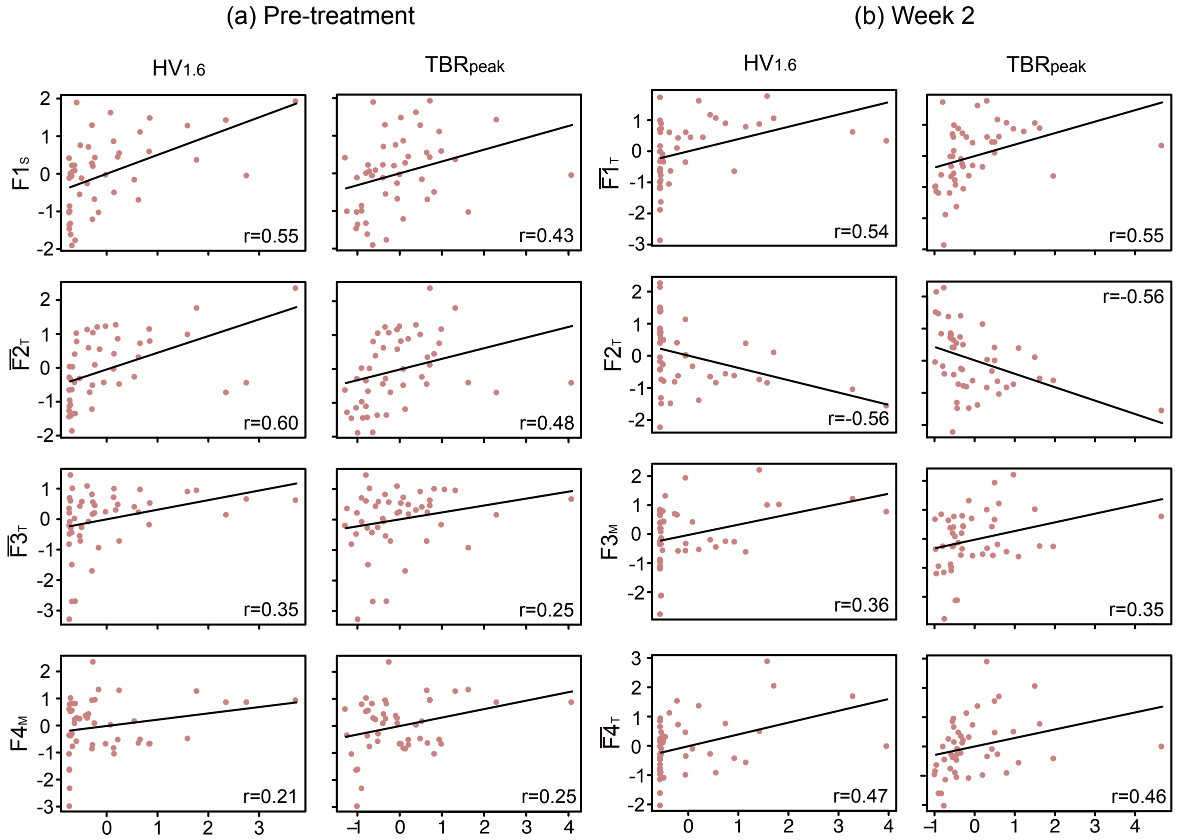


Figure 5.6.: Feature expressions of developed signatures for the representative models, boosted tree-based Cox model in combination with the Spearman feature selection method, trained on the $\mathrm{CT_{W0\text{-}FDG}}$, $\mathrm{CT_{W2}}$ and combined feature set to predict loco-regional tumour control (LRC). LRC during follow-up (yes, light; no, dark) and features with a significant correlation with LRC are shown (**p*<0.05 and ***p*<0.001). A detailed description of the feature abbreviations can be found in appendix D.1. Abbreviations: $\overline{\mathrm{F}}$ cluster feature consisting of several features represented by the mean value as a new meta-feature, $\mathrm{F_S}$ first order statistical feature, $\mathrm{F_M}$ morphological feature, $\mathrm{F_T}$ texture feature, Δ F delta feature.

The Aerts' signature contains a first-order statistical ($\mathrm{F1_S}$), a morphological ($\mathrm{F2_M}$) and two texture based features ($\mathrm{F3_T}$ and $\mathrm{F4_T}$). Figure 5.8 shows the results of the correlation analysis. The features $\mathrm{F1_S}$, $\mathrm{F3_T}$ and $\mathrm{F4_T}$ based on the $\mathrm{CT_{W0\text{-}FDG}}$ showed a moderate correlation with $\mathrm{HV_{1.6}}$ (ρ=0.64, ρ=0.68 and ρ=0.58, respectively) and a weak correlation with $\mathrm{TBR_{peak}}$ (all: ρ<0.5). The morphological feature ($\mathrm{F2_M}$) showed a weak Spearman correlation with both hypoxia FMISO-PET imaging features (both ρ=0.13). Furthermore, Aerts' signature based on the $\mathrm{CT_{W2}}$ showed similar or slightly higher correlations for the features $\mathrm{F1_S}$, $\mathrm{F3_T}$ and $\mathrm{F4_T}$ with the $\mathrm{HV_{1.6}}$ (ρ=0.64, ρ=0.69 and ρ=0.55, respectively) and with $\mathrm{TBR_{peak}}$ (ρ=0.63, ρ=0.68 and ρ=0.56, respectively). The morphological feature ($\mathrm{F2_M}$) showed a weak Spearman correlation with both hypoxia FMISO-PET imaging features (both ρ=-0.13).



(a) Pre-treatment

(b) Week 2

Figure 5.7.: Correlation plots of the developed signatures for the representative models for the prediction of loco-regional tumour control and the ^{18}F-fluoromisonidazole positron emission tomography (FMISO-PET) parameters $HV_{1.6}$ and TBR_{peak} (Zips et al., 2012; Löck et al., 2017). The correlations between the selected features within the signatures and the FMISO-PET parameters were determined using Spearman rank correlation coefficient. Abbreviations: $\overline{F}$ cluster feature consisting of several features represented by the mean value as a new meta-feature, F_S first order statistical feature, F_M morphological feature, F_T texture feature.

5.4. Summary and discussion

The aim of this study was to evaluate and compare the prognostic value of radiomics risk models using CT images obtained during treatment to models based on pre-treatment CT images.

For LRC, newly developed risk models trained on in-treatment scans (CT_{W2}) on average achieved a significantly higher prognostic performance and led to improved patient risk stratifications in comparison to pre-treatment CT scans ($CT_{W0\text{-}FDG}$). The improved performance was also observed for the published Aerts' signature. Models based on in-treatment imaging showed a higher prognostic value than tumour volume, which performed similar to the pre-treatment models. These results indicate that CT imaging during treatment contains additional prognostic information.

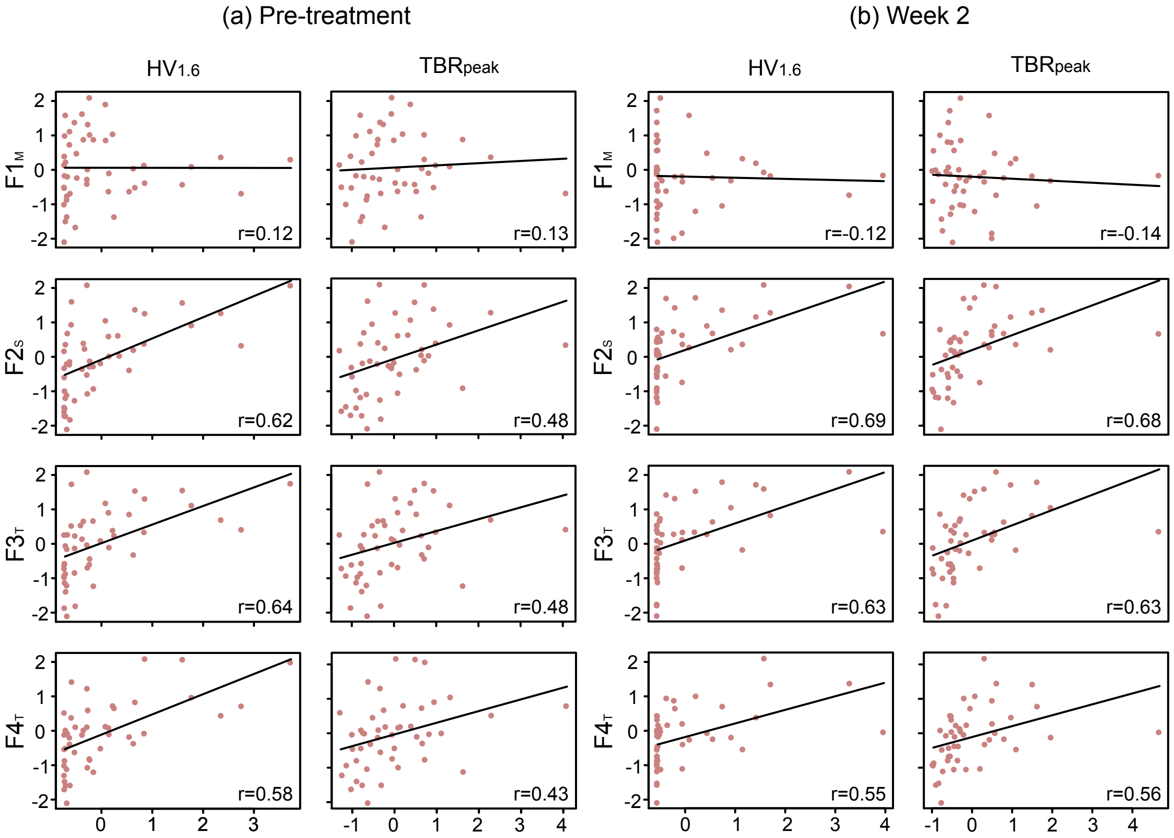


Figure 5.8.: Correlation plots of the Aerts' signature for the prediction of loco-regional tumour control and the ^{18}F-fluoromisonidazole positron emission tomography (FMISO-PET) parameters HV$_{1.6}$ and TBR$_{\text{peak}}$ (Zips et al., 2012; Löck et al., 2017). The correlations between the selected features within the signatures and the FMISO-PET parameters were determined using Spearman rank correlation coefficient. Abbreviations: F$_{\text{S}}$ first order statistical feature, F$_{\text{M}}$ morphological feature, F$_{\text{T}}$ texture feature.

Image features may be modified due to changes in tumour biology during the course of treatment. Such biological changes comprise RCT-induced re-oxygenation and shrinkage of the tumour which have been associated with treatment response (Stadler et al., 1998; Yaromina et al., 2011; Wiedenmann et al., 2015; Linge et al., 2016c). For instance, Zips *et al.* (Zips et al., 2012) showed in a prospective study that the hypoxic volume and the tumour to background ratio obtained from FMISO-PET images have a strong association with LRC. Furthermore, they observed an improved prognostic performance after week one and two of RCT in comparison to the pre-treatment images, which was validated recently (Löck et al., 2017). The developed CT-based radiomics signatures contained mostly texture and morphological features. The correlation of these features were assessed with tumour hypoxia, measured by the FMISO-PET imaging parameters hypoxic volume and the tumour to background ratio (Zips et al., 2012; Löck et al., 2017). Most texture-based features from the newly developed signatures and from Aerts' signature were moderately correlated with one or more hypoxia markers. This correlation slightly increased in the second week of treatment, where FMISO-PET showed improved prognostic value. This may be one explanation for the similarly improved performance of the CT-based radiomics models and gives an indication that changing tumour hypoxia may be observable in macroscopic CT imaging. However, the links between imaging features and tumour hypoxia, as well as other biological tumour mechanisms, should be studied in closer detail, as these are generally poorly understood.

The combined feature set, consisting of pre-treatment, week two and delta CT imaging features, likewise led to an improved model performance compared to models only based on pre-treatment CT-based features. However, the selected features were predominantly extracted from the second week CT scans or were delta features, which underlines the importance of the in-treatment data. Moreover, signatures containing only features from the second week CT scans showed a higher performance (e.g., MIM feature selection method) for predicting LRC than signatures including delta features (e.g., Spearman feature selection). This is in line with the study by Fave *et al.* (Fave et al., 2017), which showed that delta radiomics features are changing during treatment but provide limited additional prognostic information compared to baseline imaging.

For OS, similar results for in-treatment and pre-treatment imaging and a generally lower performance compared to LRC were observed. The overall decreased prognostic performance compared to LRC models may be due to the fact that the cause of death was not necessarily cancer-related, which causes the OS endpoint to be comparatively noisy. Furthermore, the combined feature set led to a lower performance in the validation cohort than the in-treatment feature set. One explanation for the reduced performance could be the selected features. In the developed signature, the two delta features were discordantly expressed between the exploratory and the validation cohort (appendix D.4). This may negatively effect the risk prediction. Interestingly, OS models based on Aerts' signature achieved the highest accuracy using pre-treatment CT scans, not in-treatment scans. This result is reasonable, since this signature was developed for the prediction of OS using pre-treatment CT scans.

Radiomics feature stability analysis was performed prior to model development, as several acquisition parameters differed between the pre-treatment and the in-treatment CT scans, e.g., the mean CT exposure settings. A lower CT exposure leads to increased image noise, which in turn affects imaging features. Feature stability was measured by the Spearman rank correlation coefficient to consider non-linear correlation effects. Moreover, the stability is also influenced by other factors, such as uncertainties in image registration or dissimilarities due to the GTV transfer. To further enhance the robustness, feature stability information, e.g., from test re-test or multiple tumour delineations datasets, may be included in future (Kim et al., 2016; Zhao et al., 2016). Furthermore, initiatives such as the Quantitative Imaging Network (QIN) of the National Institute of Health may help to establish open and standardised protocols for image acquisition, reconstruction and analysis (Buckler et al., 2011a; Buckler et al., 2011b; Clarke et al., 2014).

A limitation of this study is the relatively low number of patients and the small number of events for both endpoints. At the participating institutions, CT scans during treatment are generally not acquired as part of the clinical routine, except, e.g., for treatment re-planning. Therefore, only data from clinical imaging trials were available for this analysis. Further validation of the achieved findings is planned through retrospective analysis of additional data sets from other centres and data recorded in an on-going prospective clinical study. Due to the limited number of data samples additional cross validation experiments on the combined cohorts were performed to compare the average performance of several feature selection methods and machine learning algorithms between the time points. Both analyses showed a prognostic advantage of in-treatment imaging and limit the probability of false positive results. In addition, the analyses excluding twelve patients with differing CT-acquisition parameters in the exploratory cohort (see table 5.2) were performed, leading to similar results.

An alternative to in-treatment CT may be CBCT, which is routinely acquired in many centres during RCT for quality assurance, such as treatment position verification. However, the applicability of CBCT for radiomics risk modelling requires further investigation (Van Timmeren et al., 2016). For instance, the image quality of CBCT is low in comparison to conventional CT imaging, and the limited field of view may not be large enough to cover large tumours. These limitations may negatively affect the accuracy of radiomics risk models, but their influence may be somewhat mitigated by improved image reconstruction algorithms (van Timmeren et al., 2017b).

Radiomics risk models may not be limited to predicting survival-based endpoints, but may be used to predict the risk for occurrence of late radiation-induced side effects as well. For instance, radiomics models were recently built to predict the occurrence of xerostomia and sticky saliva for HNSCC patients (Pilz et al., 2017; van Dijk et al., 2017a; van Dijk et al., 2017b). Incorporating imaging during treatment to predict late side effects may lead to a higher prognostic accuracy, e.g., by capturing RCT-induced reactions of the normal tissue.

The present study showed that the incorporation and consideration of CT imaging acquired during treatment may be a promising way to improve radiomics risk models. This was demonstrated by newly developed radiomics models as well as by the established Aerts' signature for the endpoint LRC. Both models showed an improved validation performance compared to the tumour volume. Moreover, the investigated time point (second week of treatment) is suitable to make an early treatment adaptation in patients not responding to radio(chemo)therapy.

6. Tumour phenotype characterisation using tumour sub-volumes

6.1. Motivation

Characterisation of the tumour phenotype using imaging data is commonly based on radiomics features which were computed using the entire gross tumour volume ($\text{GTV}_{\text{entire}}$). Such an approach assumes that the individual tumour appearance is homogeneous, or heterogeneous but uniformly distributed over the entire tumour volume. However, tumours are biologically complex and exhibit substantial spatial variation, e.g., in gene expression and in macroscopic structure (Wu et al., 2016b). One reason for such spatial variations may be, e.g., necrosis which mainly appears in the tumour core and high cell proliferation mostly occurs at the tumour periphery. Some regional tumour variations, e.g., necrosis or contrast enhanced vascularisation are even apparent in imaging data (Gatenby et al., 2013; O'Connor et al., 2015). Furthermore, different regions within an individual tumour may differ in radiosensitivity, which depends on the tumour micro-environment, the distribution of cancer stem cells and localised genetic or molecular alterations (Schütze et al., 2007; Schütze et al., 2014). As a consequent, such spatial variations may effect the performance of image-based risk models.

The analysis of specific tumour sub-volumes revealed an improved prognostic performance of radiomics models. For instance, Grove *et al.* (Grove et al., 2015) showed that the expressions of 2D radiomics features computed on the rim of the tumour differed from those calculated on the tumour core. Furthermore, the ratio of tumour rim and core features led to an improved prediction of OS in NSCLC patients. Wu *et al.* (Wu et al., 2016b) identified clinically relevant tumour sub-volumes to characterise the regional heterogeneity of tumours in breast cancer patients based on dynamic contrast enhanced MRI. The resulting risk models based on the identified sub-volumes showed also an improved outcome prediction compared to models based on the $\text{GTV}_{\text{entire}}$. In another study, Wu *et al.* (Wu et al., 2016a) identified different tumour sub-volumes using CT and FDG-PET imaging of lung cancer patients. It was shown that spatially distinct sub-volumes are linked to higher risk of recurrence compared to the volume of the $\text{GTV}_{\text{entire}}$, resulting in an improved model prediction of OS.

Aside from such initial findings, in most of the previously described studies, only single clinical parameters or radiomics features (e.g., tumour volume) were investigated. Therefore, the investigation of the potential of radiomics risk models based on tumour sub-volumes are still sparse and has not been independently validated. In particular, for patients with HNSCC.

Based on the $\text{GTV}_{\text{entire}}$ two sub-volumes were defined using CT images of HNSCC patients: an outer tumour rim and the complementary tumour core. Subsequently, radiomics features were extracted from each sub-volume separately. For the prediction of LRC risk models were developed and applied on an external validation cohort. Parts of this work were presented at the international conference of the European Society for Radiotherapy and Oncology (Leger et al., 2018).

6.2. Patient cohort and experimental design

6.2.1. Characteristics of patient cohorts

A retrospective multi-centre cohort consisting of 302 patients with histologically confirmed loco regionally advanced HNSCC was used. All patients received primary RCT and underwent a non-contrast-enhanced CT scan for treatment-planning purpose. The multi-centre cohort was divided into an exploratory and a validation cohort by an approximate ratio of 2:1. In the exploratory cohort 152 of the 207 patients were treated in one of the six partner sites of the DKTK-ROG between 2005 and 2011 (Linge et al., 2016b). The remaining 55 patients were treated at the UKD between 1999 and 2006. The validation cohort consisted of 95 patients from which 50 patients received their treatment within a prospective clinical trial (NCT00180180) at the UKD between 2006 and 2012 (Zips et al., 2012; Löck et al., 2017). The remaining 45 patients were treated at the UKD or the RCDF between 2005 and 2009 as well as at the University Hospital Tübingen between 2008 and 2013. Patient characteristics for the exploratory and validation cohorts are summarised in table 6.1.

Radiomics risk models were developed to predict the primary clinical endpoint LRC. The event time for LRC was calculated from the first day of radio-chemotherapy to the date of the event or censoring. The number of events for LRC was 85 for the exploratory and 31 for the validation cohort, respectively. The median follow-up time was 15.8 months (range: 1.2–127.9 months) for the exploratory and 19.3 months (range: 1.3–75.2 months) for the validation cohort. Furthermore, the two-year LRC rate was 58.0% for the exploratory and 63.0% for the validation cohort (log-rank test: p=0.18). The corresponding Kaplan-Meier curves are shown in figure 6.1.

6.2.2. Experimental design

Tumour sub-volume definition and feature computation

The $\text{GTV}_{\text{entire}}$ of the primary tumour was manually delineated on each planning CT scan by a radiation oncologist. Subsequently, the voxel spacing was resampled using trilinear image interpolation to an isotropic voxel size of $1.0\times1.0\times1.0$ mm^3 to correct for differences in voxel

Table 6.1.: Patient characteristics of the exploratory and the validation cohort.

Clinical variable	Exploratory cohort	Validation cohort	*p*-value
Number of patients	207	95	-
Gender			
male	176	71	0.71^2
female	31	10	
missing	0	14	
Age in years			
median	58.5	54	0.007^3
range	39 - 80	37 - 74	-
TN staging			
T stage 1 / 2 / 3 / 4 / missing	2 / 24 / 51 / 130 / 0	3 / 10 / 34 / 47 / 1	0.08^1
N stage 0 / 1 / 2 / 3 / missing	24 / 8 / 160 / 15 / 0	10 / 8 / 58 / 9 / 10	0.18^1
UICC stage 2010			
I / II / III / IV / missing	0 / 0 / 13 / 136 / 58	1 / 2 / 9 / 69 / 14	0.10^1
Tumour volume in cm^3			
median	29.1	39.4	0.18^3
range	4.3 - 322.2	2.7 - 239.0	-
Prescribed total dose in Gy			
median	72	72	$<0.001^3$
range	68 - 77	69 - 77	-
HPV16 DNA			
negative / positive / missing	159 / 26 / 22	50 / 7 / 38	0.7^1
Number of events			
LRC	85 (41 %)	31 (33 %)	-
Follow up time of patients alive in months			
median	52.6	52.7	-
range	4.2 - 131.9	7.8 - 107.2	-

Abbreviations: T, clinical tumour stage; N, clinical nodal stage; UICC, Union internationale contre le cancer Gy, Gray; HPV, human papillomavirus; DNA, deoxyribonucleic acid; LRC, loco-regional tumour control

^{1} χ^2 test; ^{2} exact Fisher test; ^{3} Wilcoxon-Mann-Whitney test

spacings and slice thicknesses between the cohorts (Aerts et al., 2014; Shafiq-Ul-Hassan et al., 2017).

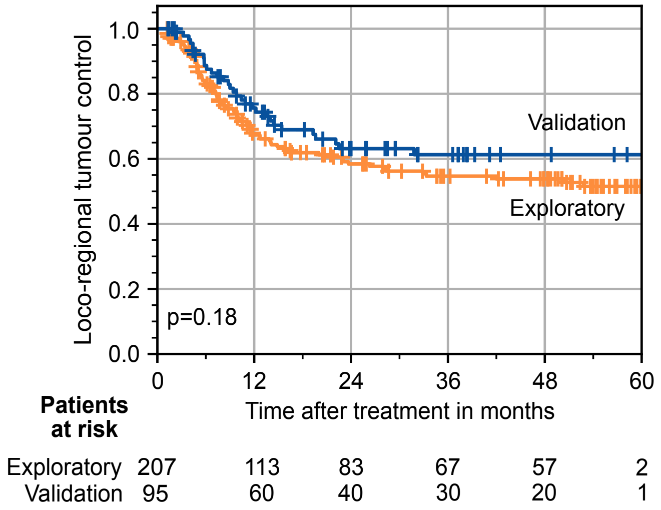
The analysis was divided into two subsequent steps which are shown in figure 6.2. Based on the delineated $\text{GTV}_{\text{entire}}$ two distinct sub-volumes were generated. The outer contour of the $\text{GTV}_{\text{entire}}$ was cropped by different widths (3, 5, 10 mm) to define the rim of the tumour (GTV_{rim}). The corresponding remaining sub-volumes were defined as tumour core (GTV_{core}). The minimum core volume was restricted to 40% of the entire tumour volume to avoid disappearance of the core sub-volume in small tumours. Furthermore, the best performing tumour rim sub-volume was selected and extended ($\text{GTV}_{\text{rim+ext}}$) into surrounding tissue with different distances (1, 2, 3, 5 mm) to assess the prognostic performance outside of the tumour delineation.

Nine additional images were created by applying spatial filtering to the base image to emphasise image characteristics such as edges and blobs. Eight additional images were created by applying a stationary coiflet-1 wavelet high-/low-pass filter along each of the three spatial dimensions (section 2.3.1). One further image was created by applying a Laplacian of Gaussian (LoG) filter consisting of five different filter kernel widths (1.0, 2.0, 3.0, 5.0, 6.0 mm, (section 2.3.1). Subsequently, the GTV mask was re-segmented to include only soft tissue voxels between -150 and 180 HU, thereby removing voxels containing air or bone, which may affect feature expression. Features were implemented in compliance with the Image Biomarker Standardisation Initiative (Zwanenburg et al., 2016). A total of 1538 features were computed and extracted from each sub-volume. 18 statistical, 38 histogram-based and 95 texture features were calculated on the base image and the nine transformed images. Moreover 28 morphological features were determined on the base image.

Radiomics risk modelling

Radiomics risk models were developed using the RMF, which consists of five steps: (I) feature pre-processing, (II) feature selection, (III) hyper-parameter optimisation, (IV) model development and (V) model validation (section 2.3.2). The risk models were generated as previously described in section 5.2.2. Briefly, after feature normalisation and clustering, feature selection was performed multiple times using 1000 bootstrap samples of the exploratory cohort. Subsequently, model training was conducted on 1000 bootstrap samples of the exploratory cohort, using the highest ranked features as well as the optimised hyper-parameter set. Finally, an ensemble prediction was made by averaging the predicted risk scores of each model for both the exploratory and the independent validation cohort separately.

Combinations of five feature selection methods and six learning algorithms were used for model development to reduce the risk of incidental findings based on the recommendation in chapter 4. The following feature selection methods were used: Spearman correlation, MIM,



Patients at risk	0	12	24	36	48	60
Exploratory	207	113	83	67	57	2
Validation	95	60	40	30	20	1

Figure 6.1.: Kaplan-Meier curves for loco-regional tumour control of the exploratory and the validation cohort.

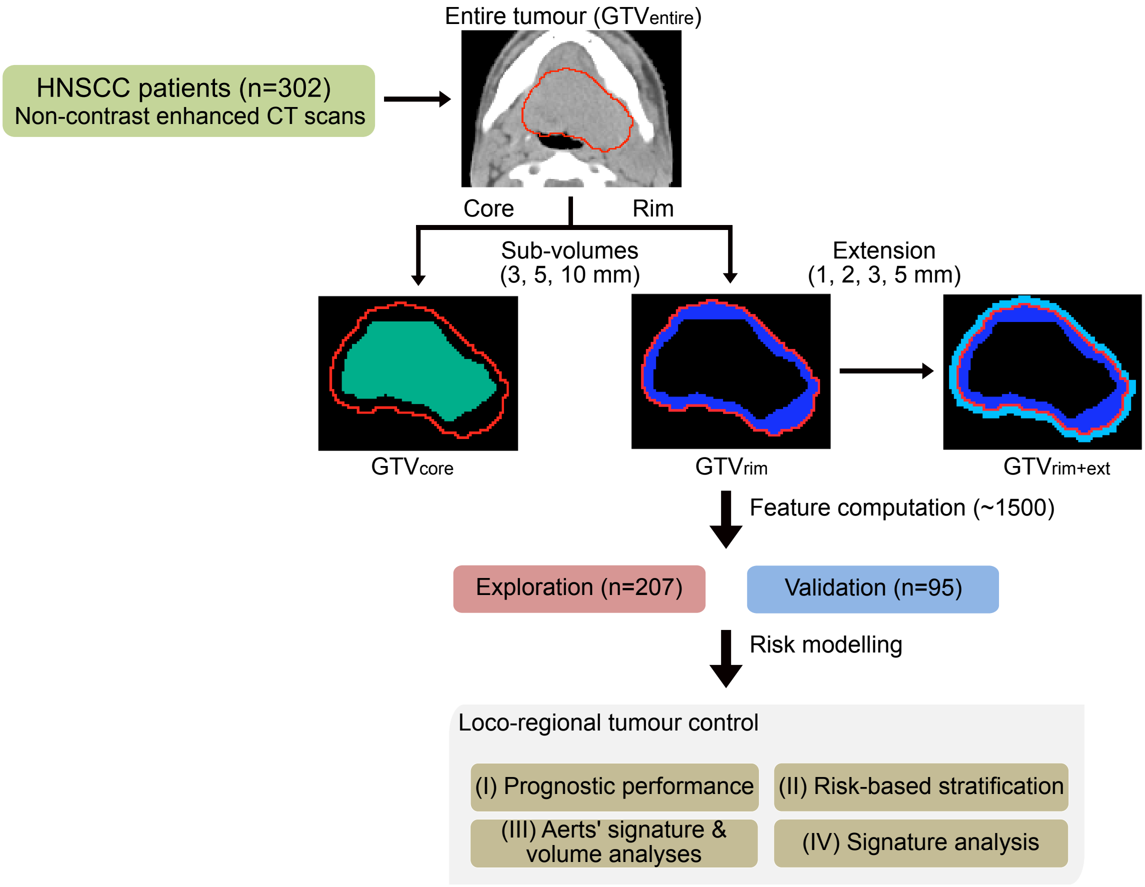

Entire tumour (GTVentire)
HNSCC patients (n=302)
Non-contrast enhanced CT scans
Core
Rim
Sub-volumes
(3, 5, 10 mm)
Extension
(1, 2, 3, 5 mm)
GTVcore
GTVrim
GTVrim+ext
Feature computation (~1500)
Exploration (n=207)
Validation (n=95)
Risk modelling
Loco-regional tumour control
(I) Prognostic performance
(II) Risk-based stratification
(III) Aerts' signature & volume analyses
(IV) Signature analysis

Figure 6.2.: Experimental design. A cohort of 302 patients with loco regionally advanced head and neck squamous cell carcinoma (HNSCC) was used to generate different sub-volumes based on the delineated entire tumour. The entire cohort was split into an exploratory and an external validation cohort for risk modelling. Prognostic model performance and patient risk group stratification were assessed on the validation cohort. Prognostic models were also trained using the radiomics signature obtained by Aerts *et al.* (Aerts et al., 2014) and the tumour volumes. Selected features within the developed signatures were analysed in terms of their univariate association with loco-regional tumour control using the entire cohort.

MIFS, MRMR and RF-VI. The six learning algorithms comprised: Cox, BT-Cox, BGLM-Cox, RSF and MSR-RF as well as the full-parametric BT-Weibull model.

Performance assessments

Four analyses were conducted to assess the radiomics models, as depicted in figure 6.2.

(I) The prognostic performance of the radiomics models was assessed on the exploratory and the validation cohort using the C-Index. Risk models were developed based on $\mathrm{GTV}_{\mathrm{entire}}$, $\mathrm{GTV}_{\mathrm{3mm\text{-}rim}}$, $\mathrm{GTV}_{\mathrm{5mm\text{-}rim}}$ and $\mathrm{GTV}_{\mathrm{10mm\text{-}rim}}$ as well as the corresponding core volumes $\mathrm{GTV}_{\mathrm{3mm\text{-}core}}$, $\mathrm{GTV}_{\mathrm{5mm\text{-}core}}$ and $\mathrm{GTV}_{\mathrm{10mm\text{-}core}}$. The median C-indices over all combination of feature selection methods and machine learning algorithms were determined based on the exploratory and the validation cohort for each tumour sub-volume to avoid incidental findings.

Subsequently, representative model combinations for each sub-volumes were selected, consisting of one feature section method and one learning algorithm. In particular, those model combinations were selected which showed the highest median performance of one feature selection method over all machine learning algorithms and vice-versa on the exploratory cohort. For the further analyses the bets performing representative models based on the sub-volume of (a) the $\text{GTV}_{\text{entire}}$, (b) the tumour rim, (c) the corresponding core and (d) the extended rim were investigated in more detail. The differences between the prognostic performance of the representative rim- (a) and the corresponding core-based model (b) was compared using a non-parametric analytical (NPA) approach based on the C-Index (Kang et al., 2015). The resulting *p*-values<0.05 were considered as statistically significant.

(II) Risk-based patient stratification into groups of low and high risk of loco-regional recurrence was performed for each tumour sub-volume and for all model combinations. The results for the selected models (a)-(d) are shown more in detail. Patients were stratified based on the median risk cut-off value ($\text{median}_{\text{risk}}$) and the cut-off value using the bootstrapped method ($\text{boot}_{\text{risk}}$, section 2.4) determined on the exploratory cohort. The resulting cut-off values were directly applied to the validation cohort. Survival curves were estimated using the Kaplan-Meier method and the stratification was compared using log-rank tests. Log-rank test *p*-values<0.05 were considered to be statistically significant.

(III) Assessment of the established Aerts' radiomics signature and the tumour volume. For further validation of the results and to reduce the risk of incidental findings an externally developed radiomics signature by Aerts *et al.* Aerts2014b was assessed. For the sub-volumes (a)-(d) risk models were trained and validated using the Aerts' signature consisting of four imaging features as previously described in section 4.2.2. The prognostic performance of the tumour volume determined on the sub-volumes (a)-(d) were assessed reflecting the clinical importance of this parameter. Subsequently, for the each sub-volume a representative model based on the Aerts' signature and the tumour volume were selected for further analyses.

(IV) The developed signatures were analysed in detail for the models trained on the different selected tumour sub-volumes (a)-(d). Features included in the signatures and their expression values are depicted as heatmaps for the exploratory and the validation cohort to represent the level of expressions and to show possible differences between the tumour sub-volumes. For this purpose, all patients were sorted according to their predicted risk and to their risk group stratification. To quantify the overall importance of the identified features, univariate prognostic power of the individual radiomics features included in the signatures were measured by the Cox model on the entire patient cohort.

6.3. Results of tumour sub-volumes evaluation

(I) Prognostic performance

The median volume fraction of the defined tumour rim sub-volumes was 47% (range: 26%–60%) for the $\text{GTV}_{\text{3mm-rim}}$ and 53% (range: 32%–60%) for the $\text{GTV}_{\text{5mm-rim}}$ and $\text{GTV}_{\text{10mm-rim}}$ sub-volumes, respectively, compared to the volume of the $\text{GTV}_{\text{entire}}$. The median performance on the exploratory cohort was similar between the $\text{GTV}_{\text{entire}}$ and the rim-based models (C-Index: 0.75–0.78). For the validation cohort, models based on the $\text{GTV}_{\text{entire}}$ achieved a median prognostic performance of 0.64±0.03 (median±SD). The models based on the tumour rim sub-volumes showed similar median performance on the validation cohort (C-Index: $\text{GTV}_{\text{3mm-rim}}$: 0.64±0.03, $\text{GTV}_{\text{5mm-rim}}$: 0.64±0.04 and $\text{GTV}_{\text{10mm-rim}}$: 0.64±0.02, respectively). The core-based risk models revealed lower prognostic performance on the validation cohort (C-Index: $\text{GTV}_{\text{3mm-core}}$: 0.60±0.05, $\text{GTV}_{\text{5mm-core}}$: 0.60±0.02 and $\text{GTV}_{\text{10mm-core}}$: 0.60±0.03, respectively).

The C-Indices of the representative models for each tumour sub-volume are shown in table 6.2. Among all $\text{GTV}_{\text{entire}}$-based risk models, the BT-Cox algorithm in combination with the MRMR feature selection method was selected for further analysis, as it showed the best performance in the exploratory cohort (C-Index: 0.81, 95% confidence interval [0.76–0.85]). On the validation cohort, this model achieved a C-Index of 0.68 ([0.60–0.77]). The BT-Cox–MIM model trained on the $\text{GTV}_{\text{5mm-rim}}$ achieved the highest prognostic performance compared to all other rim-based models on the exploratory cohort (C-Index: 0.92, [0.90–0.95]). This representative model attained a high performance on the validation cohort (C-Index: 0.68, [0.60–0.77]), which was similar to the $\text{GTV}_{\text{entire}}$-based model. The corresponding $\text{GTV}_{\text{5mm-core}}$-based model (BT-Cox–Spearman) showed a significantly lower prognostic performance on the validation cohort (C-Index: 0.57, [0.47–0.77]) compared to the $\text{GTV}_{\text{5mm-rim}}$ model (NPA-test: p=0.047). Figure 6.3 shows the prognostic performance based on the $\text{GTV}_{\text{entire}}$, $\text{GTV}_{\text{5mm-rim}}$ and $\text{GTV}_{\text{5mm-core}}$ of the feature selection methods and learning algorithms on the exploratory and the validation cohorts. The resulting C-Indices for the $\text{GTV}_{\text{3mm-rim}}$ and the $\text{GTV}_{\text{10mm-rim}}$ based models on the exploratory and on the validation cohort for the considered feature selection methods and learning algorithms are depicted in appendix E.1.

The $\text{GTV}_{\text{5mm-rim}}$ sub-volume, which achieved the highest prognostic performance among all rim-based models was subsequently extended by different widths beyond the originally delineated tumour into surrounding tissue. The tumour extension $\text{GTV}_{\text{5mm-rim+2mm}}$ (figure 6.3) and $\text{GTV}_{\text{5mm-rim+3mm}}$ showed the highest median performance on the validation cohort (both C-Indices: 0.64±0.04,). The other remaining extensions achieved a slightly reduced median performance in validation (appendix E.2). The C-Index of the representative models trained on the different tumour extensions are shown in table 6.2. The models (MSR-RF–

Table 6.2.: Concordance indices (C-Index) of the representative model combinations consisting of a feature selection (FS) and a learning algorithm (ML) using the entire gross tumour volume (GTV_{entire}) and the different tumour rim and core as well as the extended rim sub-volumes for the exploratory and the validation cohort. Furthermore, the *p*-values of the log-rank tests of the Kaplan-Meier analyses using the $median_{risk}$ and the $boot_{risk}$ cut-off calculation methods are shown.

Tumour sub-volume	C-Index		*p*-value	
Combination (ML–FS)	Exploratory	Validation	$median_{risk}$	$boot_{risk}$
GTV_{entire}				
BT-Cox–MRMR	0.81	0.68	0.005	0.001
$GTV_{3mm\text{-}rim}$				
MSR-RF–MRMR	0.86	0.65	0.02	0.001
$GTV_{3mm\text{-}core}$				
MSR-RF–MRMR	0.84	0.55	0.72	0.63
$GTV_{5mm\text{-}rim}$				
BT-Cox–MIM	0.92	0.68	<0.001	<0.001
$GTV_{5mm\text{-}core}$				
BT-Cox–Spearman	0.93	0.57	0.40	0.40
$GTV_{10mm\text{-}rim}$				
BT-Cox–MRMR	0.88	0.62	0.16	0.16
$GTV_{10mm\text{-}core}$				
BT-Cox–Spearman	0.87	0.60	0.30	0.30
Tumour extension based on $GTV_{5mm\text{-}rim}$				
$GTV_{5mm\text{-}rim+1mm}$				
BT-Cox–MIM	0.80	0.63	0.08	0.02
$GTV_{5mm\text{-}rim+2mm}$				
MSR-RF–MRMR	0.86	0.70	0.01	0.01
$GTV_{5mm\text{-}rim+3mm}$				
MSR-RF–MRMR	0.83	0.69	0.07	0.07
$GTV_{5mm\text{-}rim+5mm}$				
BT-Weibull–MRMR	0.81	0.64	0.06	0.06

MRMR) trained on the $GTV_{5mm\text{-}rim+2mm}$ and on the $GTV_{5mm\text{-}rim+3mm}$ achieved the highest performance in the exploratory cohort and a high performance in validation (C-Index: 0.70, [0.63–0.79] and 0.69, [0.60–0.79], respectively).

(II) Risk-based patient stratification

Patients were stratified into low and high risk groups based on the risk predicted by the radiomics risk models within the exploratory cohort. Table 6.2 shows the *p*-values of the log-rank test for LRC for all representative models on the validation cohort using the $median_{risk}$ and the $boot_{risk}$ cut-off calculation methods. Kaplan-Meier analyses using both calculation methods for the GTV_{entire}, $GTV_{5mm\text{-}rim}$, $GTV_{5mm\text{-}core}$ and $GTV_{5mm\text{-}rim+2mm}$ and the remaining sub-volumes are summarised in appendix E.3 and appendix E.4, respectively.

The BT-Cox model trained on GTV_{entire} was able to stratify patients into low and high risk groups with a significant difference in LRC using the $median_{risk}$ and the $boot_{risk}$ cut-off values (p=0.005 and p=0.001, respectively). An improved stratification could be achieved by

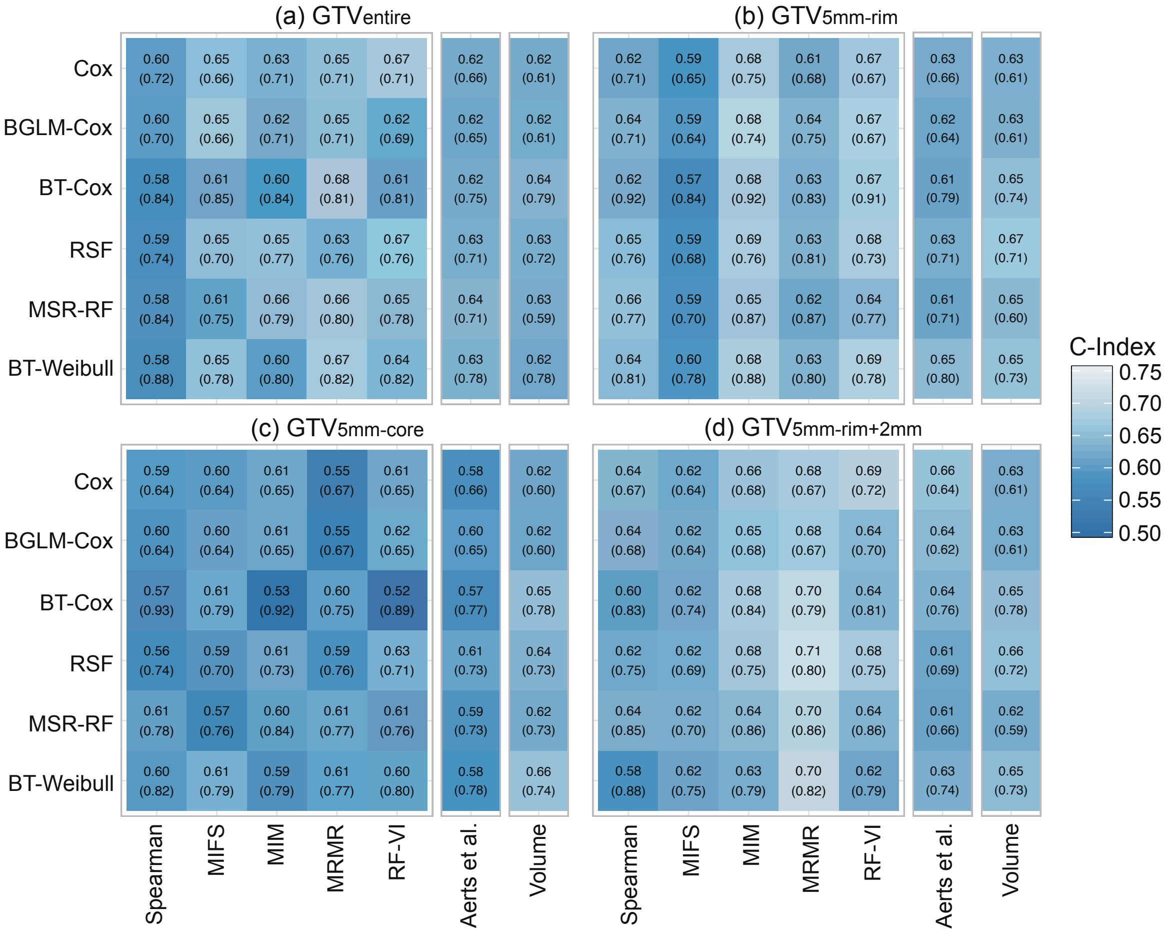
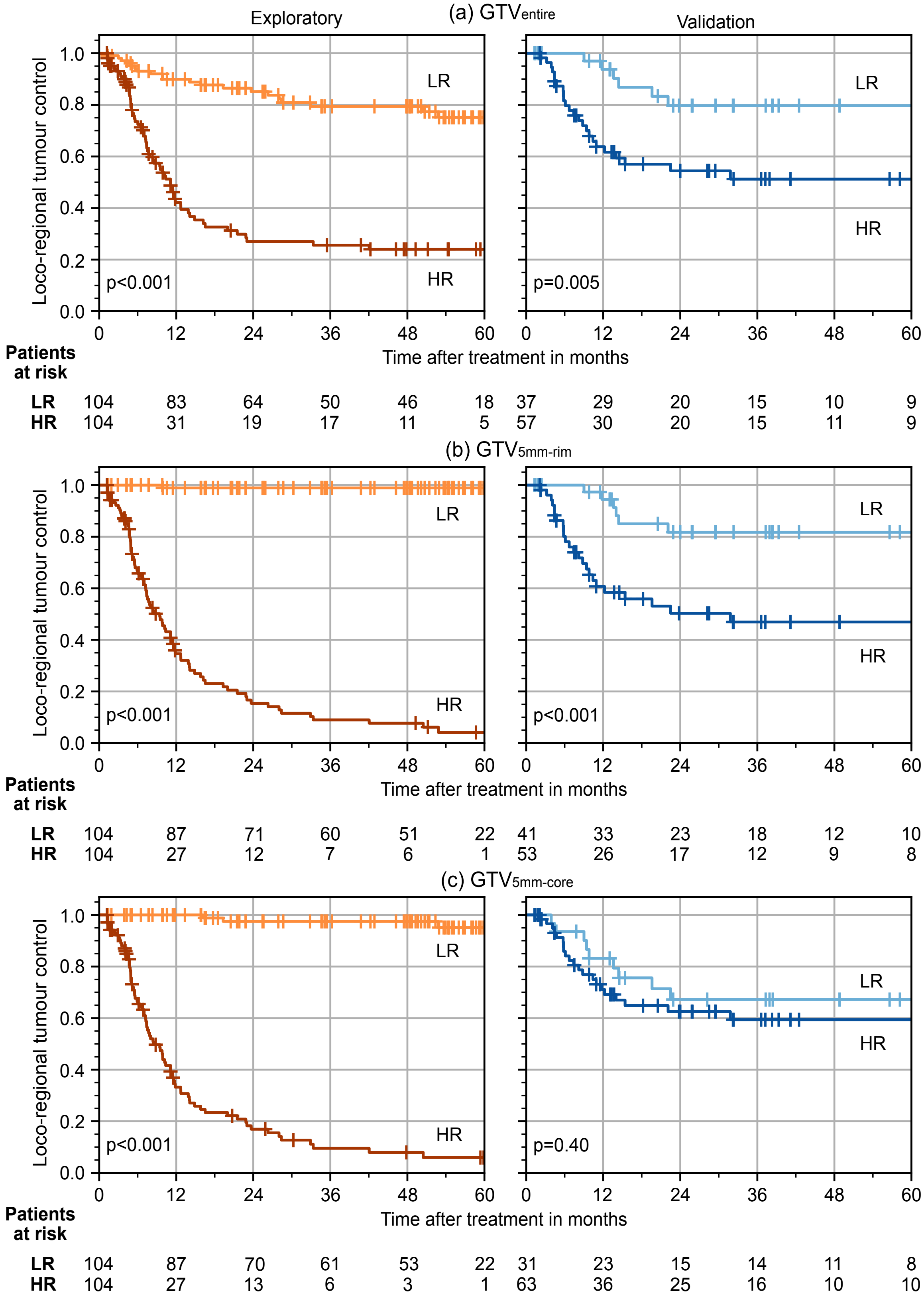


Figure 6.3.: Concordance indices for the risk models based on the entire gross tumour volume (GTV_{entire}), $GTV_{5mm\text{-}rim}$, $GTV_{5mm\text{-}core}$ and $GTV_{5mm\text{-}rim+2mm}$ sub-volumes on the exploratory (in parentheses) and the validation cohort for different feature selection methods (columns) and learning algorithms (rows). Model performance using the Aerts *et al.* (Aerts et al., 2014) signature and the volume parameter are depicted.

the $GTV_{5mm\text{-}rim}$-based model using both cut-off calculation methods (p<0.001). Stratification based on the predicted risk for the $GTV_{5mm\text{-}core}$ model did not lead to significant differences in LRC between both groups ($median_{risk}$: p=0.40 and $boot_{risk}$: p=0.40). Figure 6.4 shows the Kaplan-Meier curves using the $median_{risk}$ cut-off values for the representative models based on GTV_{entire}, $GTV_{5mm\text{-}rim}$ and $GTV_{5mm\text{-}core}$, respectively, for the exploratory and the validation cohort.

The p-values of the representative models based on extended tumour rim for the validation cohort are summarised in table 6.2. The selected model, which was based on the $GTV_{5mm\text{-}rim+2mm}$ was able to stratify the patients into low and high risk groups with a significant difference in LRC using both cut-off values (p=0.016). The Kaplan-Meier curves for this model on the exploratory and the validation cohort are shown in figure 6.5. The resulting



(a) GTV$_{\text{entire}}$

Patients at risk												
LR	104	83	64	50	46	18	37	29	20	15	10	9
HR	104	31	19	17	11	5	57	30	20	15	11	9

(b) GTV$_{\text{5mm-rim}}$

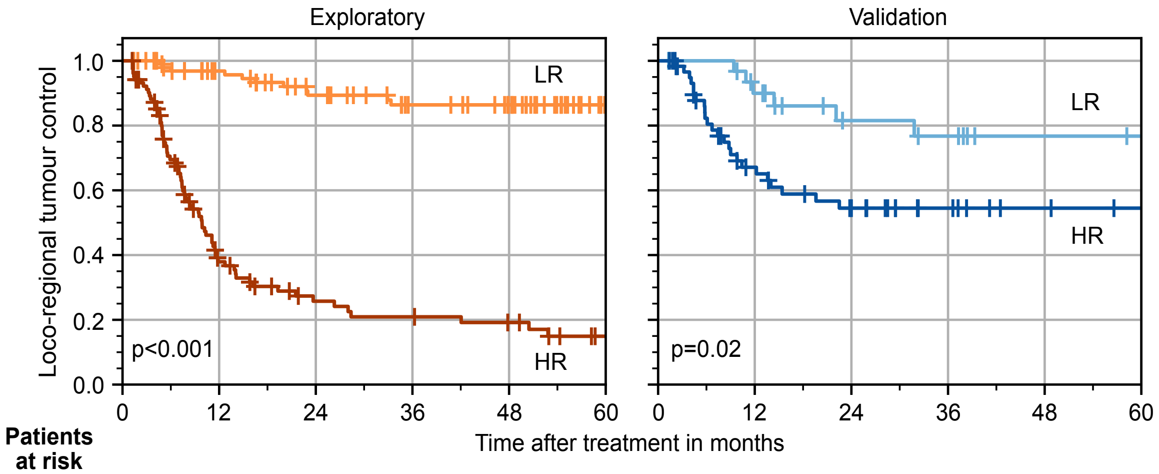
Patients at risk												
LR	104	87	71	60	51	22	41	33	23	18	12	10
HR	104	27	12	7	6	1	53	26	17	12	9	8

(c) GTV$_{\text{5mm-core}}$

Patients at risk												
LR	104	87	70	61	53	22	31	23	15	14	11	8
HR	104	27	13	6	3	1	63	36	25	16	10	10

Figure 6.4.: Kaplan-Meier curves for the prediction of loco-regional tumour control (LRC) of the representative models based on the GTV$_{\text{entire}}$, the GTV$_{\text{5mm-rim}}$ and the GTV$_{\text{5mm-core}}$ volumes for patients in the exploratory (left) and in the validation cohort (right). Patients were stratified into low (LR) and high (HR) risk groups based on the median risk of LRC determined on the exploratory cohort.

(d) $\mathrm{GTV}_{\text{5mm-rim+2mm}}$



Patients at risk	0	12	24	36	48	60	0	12	24	36	48	60
LR	104	81	67	54	47	20	32	26	17	15	11	10
HR	104	31	16	13	10	3	62	33	23	15	10	8

Figure 6.5.: Kaplan-Meier curves for the prediction of loco-regional tumour control of the representative model (table 6.2) based on the $\mathrm{GTV}_{\text{5mm-rim+2mm}}$ for patients in the exploratory (left) and in the validation cohort (right). Patients were stratified into low (LR) and high (HR) risk groups of recurrence based on the median risk of loco-regional recurrence determined on the exploratory cohort.

p-values for all tumour extensions and for all feature selection methods as well as machine learning algorithms are depicted in appendix E.5.

(III) Assessment of the Aerts' signature and tumour volume

In general, the median performance of the radiomics models using Aerts' signature determined on the sub-volumes (a)-(d) were reduced compared to the newly developed radiomics models on the exploratory (C-Index: 0.71–0.73) and the validation cohort (C-Index: 0.59–0.63).

The BT-Weibull model based on Aerts' signature determined on $\mathrm{GTV}_{\text{entire}}$ showed the highest performance on the exploratory cohort. It attained a C-Index of 0.63 (95% confidence interval, [0.54–0.73]) on the validation cohort. The BT-Weibull model based on $\mathrm{GTV}_{\text{5mm-rim}}$ achieved a slightly improved performance on the validation cohort (C-Index: 0.65, [0.56–0.73]), whereas the BT-Weibull model trained on the corresponding tumour core revealed a significantly reduced performance in validation (C-Index: 0.58, [0.48–0.69], NPA-test: p=0.049). Furthermore, the representative models based on the Aerts' signature for the extended $\mathrm{GTV}_{\text{5mm-rim+2mm}}$ (BT-Cox) showed also a similar performance (C-Index: 0.64, [0.54–0.74]), whereas the $\mathrm{GTV}_{\text{5mm-rim+3mm}}$ (MSR-RF) tumour extension showed an improved prognostic performances on the validation cohort (C-Index: 0.68 [0.59–0.78]) compared to the $\mathrm{GTV}_{\text{entire}}$-based model trained on the Aerts' signature.

The representative models trained on the tumour volume determined on the $\text{GTV}_{\text{entire}}$ (BT–Cox), $\text{GTV}_{\text{5mm-rim}}$ (BT–Cox), $\text{GTV}_{\text{5mm-core}}$ (BT–Cox) and $\text{GTV}_{\text{5mm-rim+2mm}}$ (MSR-RF) sub-volumes showed a good performance on the validation cohort (C-Indices: 0.63, 0.65, 0.65 and 0.62), respectively. The resulting C-Indices for the learning algorithms based on the Aerts' signature and the tumour volume parameter are summarised in figure 6.3, appendix E.1 and appendix E.5.

The representative $\text{GTV}_{\text{5mm-rim}}$-based model trained with the Aerts' signature could stratified the patients into low and high risk groups with significant differences on LRC using the bootstrapped cut-off calculation method (p=0.028). The remaining models did not lead to significant differences on LRC between both risk groups. For the representative tumour volume-based models (b) and (c) patients could be stratified with significant differences in LRC between both risk groups using both cut-off calculation methods (p<0.05). The remaining models based on the tumour volume did not lead to significant differences in LRC between both risk groups. The resulting p-values of the log-rank tests are depicted in appendix E.3, appendix E.4 and appendix E.5.

(IV) Signature analysis

Radiomics signatures were investigated for the representative models based on (a) $\text{GTV}_{\text{entire}}$, (b) $\text{GTV}_{\text{5mm-rim}}$, (c) $\text{GTV}_{\text{5mm-core}}$ and extended (d) $\text{GTV}_{\text{5mm-rim+2mm}}$. Figure 6.6 shows the feature expressions of the newly developed signatures for each patient in a heatmap. Feature names of the selected features within the signatures are explained in appendix E.C.

The signature of the $\text{GTV}_{\text{entire}}$ model consists of two first-order statistical radiomics features computed on the wavelet transformed images. For instance, feature F1_{S} is based on the intensity-volume histogram and describes the differences between the volume fractions at two different intensity fractions (El Naqa et al., 2009). The two radiomics features in the signature showed a significant association to LRC based on the univariate Cox model using the entire patient cohort (F1_{S}: p<0.001 and F2_{S}: p<0.001).

The $\text{GTV}_{\text{5mm-rim}}$-based model was trained on a signature that contains ten radiomics features which were mainly texture-based features extracted from the original base images. For instance, feature F5_{T} is based on the GLDZM, capturing the relation between location and grey level to measure the intra-tumour heterogeneity. All features within the signature were significantly prognostic for LRC in univariate analyses, except of feature F2_{S}.

The developed signature for the model based on the corresponding $\text{GTV}_{\text{5mm-core}}$ sub-volume also consisted of ten radiomics features which were either first-order statistical or texture-based features extracted from different wavelet transformed images. Only features F2_{S}, $\overline{\text{F4}}_{\text{S}}$ and $\overline{\text{F8}}_{\text{T}}$ were significantly associated with LRC using univariate Cox analyses.

The extended $\text{GTV}_{\text{5mm-rim+2mm}}$-based model was trained on a radiomics signature with seven features, which comprised mainly morphological and texture features. The texture-

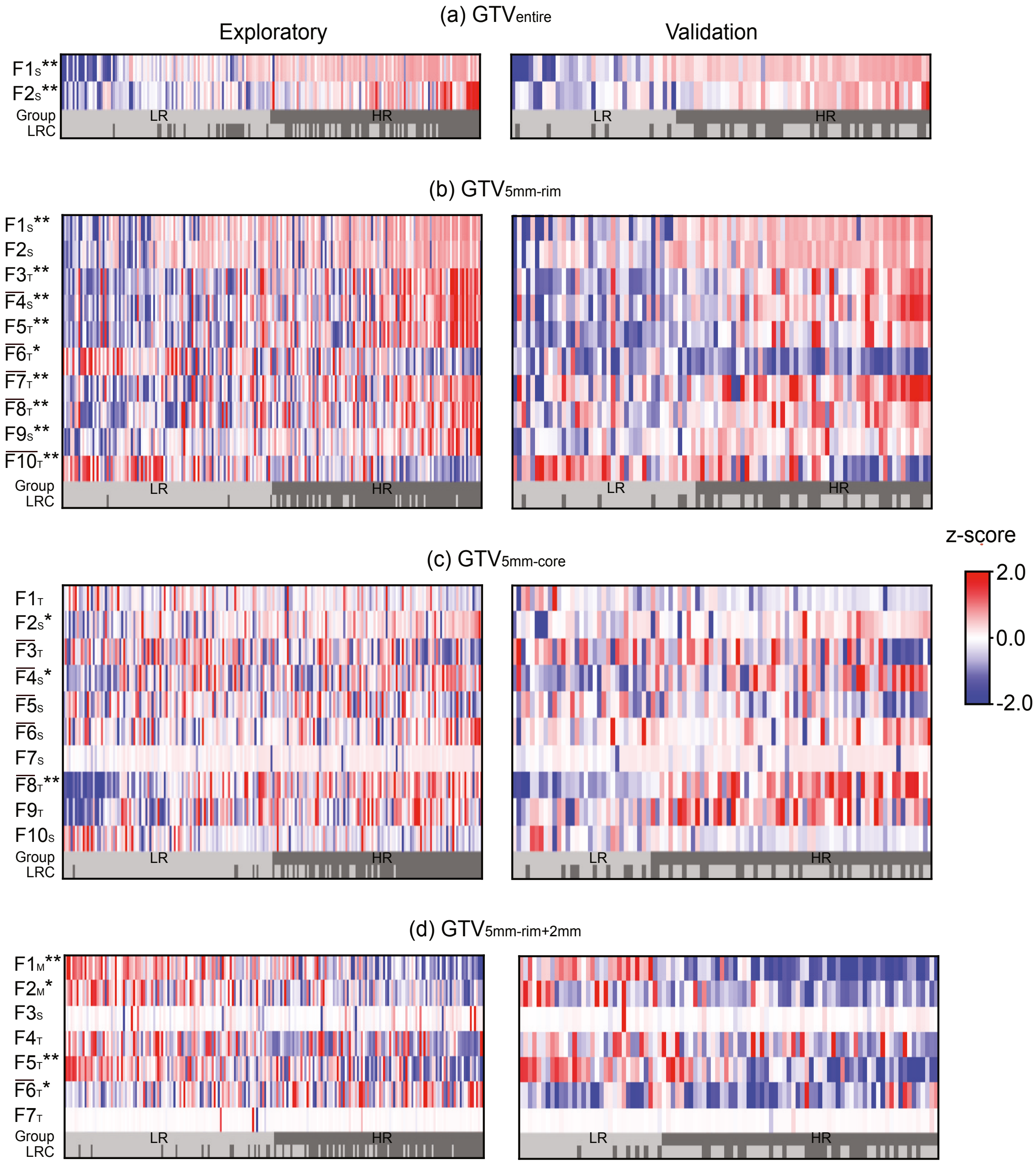


Figure 6.6.: Heatmaps showing different expression patterns of the radiomics features within the developed signatures for the representative models based on the (a) $\text{GTV}_{\text{entire}}$, (b) $\text{GTV}_{\text{5mm-rim}}$, (c) $\text{GTV}_{\text{5mm-core}}$ and (d) $\text{GTV}_{\text{5mm-rim+2mm}}$. Feature expression values are sorted according to the predicted risk and the risk group based on the determined $\text{median}_{\text{risk}}$ cut-off values. Loco-regional tumour control (LRC) during follow-up (yes, light; no, dark) and features with a significant association with LRC are shown (**p*<0.05 and ***p*<0.001). A detailed description of the feature abbreviations can be found in appendix E.C. Abbreviations: $\bar{F}$ cluster feature consisting of several features represented by the mean value as a new meta-feature, F_S first order statistical feature, F_M morphological feature and F_T texture feature.

based features were commonly extracted from wavelet transformed images. For example, feature $F2_M$ quantified the compactness of the $GTV_{5mm\text{-}rim+2mm}$ volume relative to that of a sphere. The univariate Cox regression revealed that features $F1_M$, $F2_M$, $F5_T$ and $\overline{F6}_T$ were significantly associated with LRC.

6.4. Summary and discussion

Tumours may contain biologically complex structures and exhibit substantially spatial variation, e.g., necrosis may appear in the core of the tumour and high cell proliferation may occur along the tumour periphery. The main objective of this study was to investigate and compare different sub-volumes based on the tumour rim and the core to identify those region which contains the relevant prognostic information using macroscopic CT imaging for patients with HNSCC.

Radiomics risk models based on tumour rim sub-volumes were superior for the prediction of LRC and the stratification of patients into low and high risk groups than models based on the corresponding core. This may indicate that the tumour rim is biologically more diverse and important treatment-related processes occur primarily in the rim. Moreover, it indicates that these processes are observable in macroscopic CT imaging. The strong variability of the radiomics feature expressions especially of patients with an event may be explain the reduced performance of the tumour-core based models. Furthermore, the heatmap showed that the expression values for several features within the signature were nearly zero for many patients of the representative tumour core-based model. Usually feature values around zero are not informative and thereby contain limited prognostic information for the radiomics risk models. These results are in-line with previously published data (Gatenby et al., 2013; O'Connor et al., 2015; Dou et al., 2017). For example, Grove *et al.* (Grove et al., 2015) showed that tumour-rim based radiomics features led to a higher expression compared to features extracted from corresponding tumour-core sub-volumes in NSCLC patients. The performance differences between the tumour rim and the corresponding core sub-volumes were also observed for the published Aerts' signature. In particular, the tumour-rim based model ($GTV_{5mm\text{-}rim}$) showed a significantly improved C-Index compared to the model based on the corresponding core ($GTV_{5mm\text{-}core}$). The tumour volume of the rim (b) and the core (c) sub-volume showed a good C-Index which demonstrated the relevance of this parameter.

Defining the precise extent of the macroscopic tumour is difficult, especially using CT imaging without contrast enhancement. Slight extensions of the delineated tumour volume into normal tissue did not reduce the performance of the radiomics risk models, which indicates that these regions may also contain prognostic information. In addition, the slightly extensions of the tumour may be useful for assessing feature stability using small tumour extensions, simulating different tumour delineations of different observers.

This study is based on the assumption that necrotic regions may appear in the tumour core due to inadequate vascular supply and that proliferating cancer cells are mainly occur in the tumour periphery (Vaupel et al., 1989). However, it is challenging to distinct between the proliferating and the necrotic part of the tumour by means of a static definition as used in this study. Furthermore, such a simple approach does not take into account complex spatial variations in tumours, e.g., non-uniform necrotic regions which may appear in different parts of the tumour. The identification and incorporation of tumour specific regional variations by more sophisticated image analysis techniques may help to overcome this gap. For instance, differential information from multi-modal imaging data such as PET-CT may be used. Moreover, super-voxel algorithms can be applied to group voxels into super-voxel segments based on their grey value, e.g., using the FDG uptake value (Kanungo et al., 2002). Subsequently, the resulting super-voxel segments can be further merged to generate tumour sub-volumes, e.g., by hierarchical or fuzzy c-means clustering algorithms across the entire patient cohort. Wu *et al.* (Wu et al., 2016a) proposed such a two-stage clustering process, for the identification and determination of sub-volumes based on CT imaging combined with FDG-PET scans in lung cancer patients. However, due to missing functional imaging, it is was not possible to use such imaging data in this thesis. Nevertheless, an adaptation of the proposed sub-volume generation process by a more advanced image analysis in combination with incorporating complementary imaging information may offer the potential to enhance radiomics risk models in the future.

The identification of tumour sub-volumes which enable the potential to improve the performance of conventional radiomics risk models may also be interesting for risk models based on deep learning algorithms. Deep learning algorithms, in particular convolutional neural networks (CNN) consist of a sequence of convolutional and sub-sampling operations to learn complex feature representation directly from the imaging data (Greenspan et al., 2016). Consequently these results in a large number of model parameters, which may lead to model over-fitting, especially in the case of few training data. Therefore, a reduction of the learning data by usage of using tumour sub-volumes instead of the entire tumour volume may help to develop generalisable and better performing deep learning models, e.g., by fewer convolutional operations.

In conclusion, the consideration and application of tumour sub-volumes are a promising techniques to improve the performance of radiomics risk models.

7. Summary and further perspectives

The personalisation of cancer treatment is one major objective in radiation oncology, e.g., to tailor the radiation dose individually to patients or to small subgroup. This approach requires the identification of biomarkers, which characterise the tumour phenotype and precisely predict therapy response. Next to molecular-based biomarkers, radiomics attempts to characterise the tumour using imaging data. Radiomics is based on the extraction and analysis of quantitative image features by machine learning algorithms to develop prognostic or predictive risk models.

This thesis was dedicated to establish and implement a radiomics workflow at the National Centre for Radiation Research in Oncology. Further, methodological developments are provided, which aims to enhance the prognostic performance of image-based risk models.

The precise prediction of therapy response requires different processing steps to evolve imaging-based risk models. Consequently, in this thesis two novel software-based frameworks were developed to compute a wide range of different quantitative features, which were extracted from medical imaging data and to build generalisable radiomics risk models. In particular, the feature computation framework provides the mathematical definitions of the radiomics features and different image pre-processing algorithms, e.g., to enhance the image quality prior to feature extraction. Hence, a novel data-driven physical correction model for the correction of intensity non-uniformity, which is a typical artefact in MRI data, was developed and integrated. The new pre-processing algorithm is motivated by the physical properties of a typical MRI coil system and provides smooth intensity corrections.

The developed risk modelling framework provides a wide range of different machine learning algorithms to identify informative features and to evolve automatic and unbiased predictive and prognostic risk models. In addition to the provided algorithms, the framework enables building risk models using data resampling strategies, e.g., bootstrapping or cross validation to select features and to build models, which are robust against influence of data perturbation.

The evaluation and identification of suitable feature selection methods and learning algorithms are integral components for the development of reliable clinical risk models. Within in this thesis, an extensive evaluation of twelve different feature selection methods and eleven machine learning algorithms for time-to-event survival data was realised. Moreover, radiomics risk models were developed and externally validated for the prediction of LRC and OS using CT scans of a multi-centre HNSCC patient cohort. Consequently, six different features selection methods and seven learning algorithms were identified and recommended for use in future radiomics studies.

In general, the improvement of radiomics risk models is an essential step to facilitate image-based risk models for treatment decisions in clinical care. Therefore, the potential of CT imaging during the course of treatment was investigated. For this purpose, pre-treatment CT images were compared with in-treatment CT imaging (after the second week of primary RCT) based on their prognostic value to predict LRC for patients with HNSCC. As a result, the risk model performance for the prediction of LRC could be significantly improved by more than 10% using in-treatment CT scans compared to pre-treatment CT imaging. Furthermore, an improvement of risk-based patient stratification was demonstrated. The incorporation of in-treatment CT imaging is a promising way to improve radiomics risk models. Moreover, the time-point of the in-treatment image acquisition still permits a clinical treatment adaptation.

Imaging-based risk models are usually built on the characteristics of the entire tumour. However, tumours are biologically complex and exhibit substantial spatial variation, e.g., expressing necrosis mainly in the core and high tumour cell proliferation at the periphery. Such spatial variations may influence the prognostic performance of the risk models. For that reasons, tumour sub-volumes were investigated, with the aim to gain a deeper understanding which part of the tumour contains more prognostic information and whether incorporation of this spatial diversity improves the model performance. In particular, risk models, which are based on the pre-defined tumour rim and the corresponding core, were developed and compared. The analyses demonstrated that the rim of the tumour contains more prognostic information, leading to higher model performance and better patient stratification compared to the core. Moreover this indicates that spatial variations within the tumour can be measured by macroscopic CT imaging.

The field of radiomics still offers many interesting challenges and open research questions for the future. Some of them are shortly discussed in the following paragraphs.

Radiomics risk modelling

Radiomics feature robustness and reproducibility are vital factors for the successful clinical application of radiomics risk models (Gillies et al., 2015). However, radiomics features may be influenced, e.g., by differences in patient positioning or the usage of different image acquisition parameters like image resolution or reconstruction algorithms. The resulting radiomics risk models based on non-robust features may lead to a poor model generalisability and to false outcome prediction. Therefore, feature stability analyses are recommended to identify radiomics features, which are non-robust against such influences and could be dismissed, consequently. These stability analyses are often performed using test-retest data, where two or more images of a patient or a phantom are acquired within a short time interval (Aerts et al., 2014; Mackin et al., 2015; Shafiq-Ul-Hassan et al., 2017). However, such data sets are scarcely available, e.g., due to effort of acquiring process. An additional problem is that feature robustness depends on the particular tumour phenotype and the used image modality (Leijenaar et al., 2013; van Timmeren et al., 2016). Consequently, these factors may hamper the application and translation of radiomics risk models into the clinical workflow. This leads to a need for increasing the robustness of radiomics features by new strategies. For instance, it is possible to incorporate radiomics features, which are computed and extracted from the original images as well as from images perturbed by translations or rotations. These data augmentation strategies are an effective technique to stabilise radiomics features, e.g., by averaging over the produced perturbed feature values to create new meta features. As an alternative, all generated perturbation values may be included as additional data samples into the modelling process. Moreover this approach enlarges the data distributions of the features and may enable the risk models to learn robust decision boundaries to reduce the risk of false model predictions.

A further important step for the acceptance of radiomics risk models is the linkage between the features derivated from imaging data and the underlying tumour biology (Segal et al., 2007; Grossmann et al., 2017). Aerts *et al.* (Aerts et al., 2014) reported significant associations between the radiomics features within the developed signature and the gene expression profiles of lung cancer patients. For instance, both texture based features in the signature were strongly correlated with cell-cycling pathways, indicating an increased proliferation in the more heterogeneous tumours (Aerts et al., 2014). However, only few studies systematically investigated the causal relationships between radiomics features and the underlying tumour biology (Panth et al., 2015; Incoronato et al., 2017). In particular, for patients with HNSCC further investigations are required to gain deeper understanding of potential causality. These analyses would allow to dismiss radiomics features which are not associated with known biological tumour mechanisms.

Deep learning based radiomics risk modelling

The predictive or prognostic performance of radiomics risk models may be further improved using the deep-learning approach. In particular CNN are able to learn feature representations directly from the imaging data instead of using engineered radiomics features. The application of CNN showed promising results in the medical imaging domain, e.g., for image segmentation or lesion detection (Greenspan et al., 2016; Pereira et al., 2016; Roth et al., 2016). However, only few studies investigated the potential of the deep learning approach for radiomics (Liu et al., 2016; Lao et al., 2017). For instance, Paul *et al.* (Paul et al., 2016) showed that deep learning features alone did not improve the model accuracy in comparison to traditional quantitative imaging features. Consequently, the adaptation of deep learning to radiomics risk modelling still requires substantial fundamental research. Deep learning is usually based on 2D images, which reduce the available information of the entire tumour and result in a loss of information concerning the relation between subsequent image slices.

Conceptually, 3D CNN are possible, however, no pre-trained networks are available. Hence, a large number of training data samples (i.e., $\gg 1000$) would be required, e.g., to avoid model overfitting, which is difficult to achieve. A further unresolved limitation of deep learning is the missing ability to handle (censored) continuous time-to-event survival data. Also, the linkage between the learned feature representations and tumour biology is challenging, e.g., due to the unclear meaning of these features. Therefore, new strategies and algorithms have to be implemented to increase the application of deep learning. Still, deep-learning is a promising approach for radiomics and the individualisation of cancer treatment in the future.

Combination of radiomics models with in-silico modelling

Radiomics risk models are usually based on imaging data acquired once at a fixed time point. For the precise characterisation and analysis of the tumour phenotype, it may not be sufficient to use the detailed tumour characteristics only based on information derived from one specific time point. Furthermore, several other challenge arise by the usage of imaging during treatment, e.g., late imaging time points (e.g., fourth week of treatment) may also contain additional informations but may be too late for consideration.

To overcome this gap, it is conceivable to combine conventional radiomics risk models by in-silico models. In particular, in-silico models enable to simulate the tumour behaviour and their micro-environment on a fine time scale, which allows a precise observation of the tumour dynamics. The gained knowledge from in-silico models may be used as additional features for conventional radiomics risk models to improve the prediction performance. Furthermore, the information derived from medical imaging data can be used to develop realistic in-silico models and the acquired imaging data during the treatment provides the ability to continuously update and adjust such models.

Ferranti *et al.* (Ferranti et al., 2017) generated and simulated networks analogous to biological systems and incorporated these data into machine learning approaches as additional knowledge about the underlying system. The usage of these data increased the prediction performance of the machine learning algorithms. Another common approach is the application of agent-based models, which are powerful simulations to investigate the interactions in complex systems (Bonabeau, 2002). An agent is the smallest unit in this model and can show different types of stochastic behaviour, by interaction with other agents. Although these models simplify many aspects of reality, they have been shown to be useful in cancer research, e.g., to study tumour growth processes or the mutational landscape of solid tumours (Waclaw et al., 2015; Poleszczuk and Enderling, 2016). Kather *et al.* (Kather et al., 2016) developed a multiagent-based model from quantitative histological and other microscopic data. This agent-based model simulated interactions between tumour cells, immune cells, and stroma and represented diverse spatial patterns observed in histological samples of human colorectal cancer. Subsequently, a Cox model was trained based on the number of stroma cells and additional clinical parameters to predict OS. Jalalimanesh *et al.* (Jalalimanesh et al., 2017) developed an agent based model to simulate different scenarios of radiotherapy with the aim to optimise the therapy of solid tumours. The agent-based approach considered the heterogeneity of tumour oxygen diffusion and effects of hypoxia on radiotherapy. The mathematical modelling and simulation of tumour behaviour based on imaging or other types of data may provide additional information about the complex biological system of a tumour. Incorporation of these data into conventional radiomics risk models may be a promising and effective way to improve outcome prediction.

In summary, the presented thesis established the developed radiomics workflow at National Centre for Radiation Research in Oncology. Furthermore, the observed advantage of in-treatment imaging and the consideration of spatial diversity into radiomics risk models are important contributions for improving their accuracy. Moreover, the results provide an substantial step towards personalisation of cancer treatment and may be applied in interventional clinical trials after prospective validation.

8. Zusammenfassung

Die Personalisierung der Krebsbehandlung ist ein wesentliches Ziel in der modernen Radioonkologie, indem z.B. die Strahlendosis für einen Patienten und dessen Tumor individuell angepasst wird. Die Implementierung eines solchen Ansatzes erfordert die Identifizierung von spezifischen Merkmalen, welche den Tumor charakterisieren, um das Therapieansprechen präzise vorhersagen zu können. Neben der molekularen Charakterisierung, hat sich radiomics als ein vielversprechendes Verfahren erwiesen, welches den Tumor anhand von Bilddaten beschreibt. Dazu werden quantitative Merkmale aus den medizinischen Bilddaten extrahiert, die anschließend mittels intelligenter Lernverfahren analysiert werden, um prognostische oder prädiktive Risikomodelle zu entwickeln.

Die vorliegende Arbeit diente der Etablierung und Implementierung von radiomics innerhalb des Nationalen Zentrums für Strahlenforschung in der Onkologie, sowie der methodischen Weiterentwicklung mit dem Ziel die Genauigkeit von bildbasierten Risikomodellen zu verbessern.

Die präzise Vorhersage des Therapieansprechens erfordert eine Vielzahl von Verarbeitungsschritten, um bildbasierte Risikomodelle zu entwickeln. Daher wurden in dieser Arbeit zwei neuartige software-basierte Systeme (frameworks) erarbeitet, welche die Extraktion der quantitativen Bildmerkmale sowie die Erstellung von Risikomodellen ermöglichen. Das framework zur Merkmalsberechnung stellt die mathematischen Definitionen der Bildmerkmale sowie verschiedene Algorithmen für die Bildvorverarbeitung zur Verfügung, um beispielsweise die Bildqualität zu verbessern. Des Weiteren wurde ein neues Verfahren zur Korrektur von Signalinhomogenitäten, die ein typisches Bildartefakt in Magnetresonanztomographie (MRT)-Bilddaten sind. entwickelt und in das framework integriert. Das vorgeschlagene Korrekturmodell basiert auf den physikalischen Verhalten eines typischen MRT-Spulensystems und verhindert somit zu starke Intensitätskorrekturen. Das zweite framework umfasst eine Vielzahl von unterschiedlichen intelligenten Lernverfahren für die Erkennung von relevanten Merkmalen sowie zur automatischen Entwicklung von Vorhersagemodellen. Zudem ermöglicht es zufällige Stichproben zu generieren z.B. mittels Kreuzvaliderungsverfahren, um Merkmale zu selektieren und Modelle zu generieren, die robust gegenüber dem Einfluss von Datenveränderungen sind. Die Evaluierung und Identifizierung geeigneter Methoden für die Merkmalsselektion sowie des maschinellen Lernverfahrens ist ein integraler Bestandteil für die Entwicklung von klinisch anwendbaren Risikomodellen. In einer umfassenden Analyse wurden zwölf Methoden zur Merkmalsselektion und elf maschinelle Lernverfahren für Überlebenszeitdaten untersucht. Die Risikomodelle wurden anhand von Röntgen-Computertomographie (CT) Bilddaten für die Vorhersage der loko-regionären Tumorkontrolle (LRC) und des Gesamtüberlebens (OS) von Patienten mit fortgeschrittenen

Kopf-Hals-Tumoren (HNSCC) entwickelt und bewertet. Insgesamt wurden sechs Methoden zur Merkmalsselektion und sieben Lernverfahren als geeignet vorgeschlagen, welche in zukünftigen radiomics-basierten Studien verwendet werden sollten. Die kontinuierliche Verbesserung der bildbasierten Risikomodelle ist eine Voraussetzung für die Einführung in den therapeutischen Entscheidungsprozess. Daher wurden in dieser Arbeit CT-Bildgebung die während der Therapie (in-treatment) erstellt wurde hinsichtlich ihrer prognostischen Aussagekraft für die Vorhersage von LRC und OS innerhalb einer HNSCC Patientenkohorte, untersucht. In dieser Untersuchung konnte gezeigt werden, dass die Genauigkeit der Risikomodelle für die Vorhersage von LRC mittels in-treatment CT-Bildgebung im Vergleich zur Bildgebung vor der Therapie um mehr als 10% verbessert werden konnte. Die Verwendung von in-treatment CT-Bildgebung stellt somit einen vielversprechenden Ansatz dar, um die Genauigkeit von bildbasierten Risikomodellen zu verbessern.

Radiomics-basierte Risikomodelle verwenden häufig Bildmerkmale, welche unter Verwendung des gesamten sichtbaren Tumors berechnet und extrahiert wurden. Tumore sind jedoch biologisch komplex und weisen häufig eine unterschiedliche räumlich Verteilung auf molekularer Ebene auf, z.B. nekrotische Areale im Tumorzentrum und eine starke Zellproliferation entlang der Peripherie. Solche räumlichen Variationen können die Genauigkeit der Risikomodelle beeinflussen. Daher wurden in dieser Arbeit verschiedene Subvolumina innerhalb des Tumors untersucht, um ein tieferes Verständnis über die prognostische Aussagekraft der verschiedenen Tumorareale zu erlangen. Des Weiteren wurde analysiert, ob die Einbeziehung dieser Diversität zu einer Verbesserung der Genauigkeit von Risikomodellen führen kann. Die Analyse zeigte, dass der Tumorrand die wesentlichen prognostischen Informationen, im Vergleich zum Tumorzentrum, enthält und das diese biologischen Variationen innerhalb des Tumors mit Hilfe von makroskopischer CT-Bildgebung gemessen werden kann.

In der vorliegenden Arbeit wurde radiomics am Nationalen Zentrum für Strahlenforschung in der Onkologie erfolgreich etablierte und eingeführt. Die entwickelten frameworks ermöglichen somit die Durchführung von radiomics-basierten Analysen. Zudem leisten die vorgeschlagen methodischen Weiterentwicklungen, wie die Verwendung von CT-Bildgebung während der Therapie und die Berücksichtigung der räumlichen Diversität, einen wichtigen Beitrag zur Steigerung der Genauigkeit von bildbasierten Risikomodellen. Die erreichten Verbesserungen, stellen darüber hinaus einen wichtigen Schritt zur Individualisierung der Krebsbehandlung von Patienten dar und könnten somit nach einer prospektiven Validierung in interventionellen klinischen Studien angewendet werden.

Bibliography

Aerts H. 2016. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncology 2:1636–1642.

Aerts H, Velazquez E, Leijenaar R, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen M, Leemans C, Dekker A, Quackenbush J, Gillies R, and Lambin P. 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 5:4006.

Amadasun M and King R. 1989. Textural features corresponding to textural properties. IEEE Transactions on Systems, Man and Cybernetics 19:1264–1273.

Axel L, Costantini J, and Listerud J. 1987. Intensity correction in surface-coil MR imaging. American Journal of Roentgenology 148:418–420.

Battiti R. 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5:537–550.

Baumann M, Krause M, Overgaard J, Debus J, Bentzen S, Daartz J, Richter C, Zips D, and Bortfeld T. 2016. Radiation oncology in the era of precision medicine. Nature Reviews Cancer 16:234–249.

Belliveau J, Kennedy D, McKinstry R, Buchbinder B, Weisskoff R, Cohen M, Vevea J, Brady T, and Rosen B. 1991. Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254:716–719.

Bellman R. 1961. Adaptive control processes: a guided tour. Princeton University Press.

Bellon E, Haacke E, Coleman P, Sacco D, Steiger D, and Gangarosa R. 1986. MR artifacts: A review. American Journal of Roentgenology 147:1271–1281.

Bentzen J, Toustrup K, Eriksen J, Primdahl H, Andersen L, and Overgaard J. 2015. Locally advanced head and neck cancer treated with accelerated radiotherapy, the hypoxic modifier nimorazole and weekly cisplatin. Results from the DAHANCA 18 phase II study. Acta Oncologica 54:1001–1007.

Bergstra J and Bengio Y. 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13:281–305.

Bezdek J, Ehrlich R, and Full W. 1984. FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences 10:191–203.

Bloch D. 1946. Nuclear induction. PhysicalRview 70:460.

Bonabeau E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences 99:7280–7287.

Bouguer P. 1922. Essai d'Optique sur la Gradation de la Lumiere. Journal of the Röntgen Society 18:93–93.

Breiman L. 2001. Random forests. Machine Learning 45:5–32.

Brown H and Prescott R. 2014. Applied mixed models in medicine. John Wiley & Sons.

Buckler A, Bresolin L, Dunnick M, and Daniel C. 2011a. A Collaborative Enterprise for Multi-Stakeholder Participation in the Advancement of Quantitative Imaging. Radiology 258:906–914.

Buckler A, Bresolin L, Dunnick N, Sullivan D, and the others. 2011b. Quantitative Imaging Test Approval and Biomarker Qualification: Interrelated but Distinct Activities. Radiology 259:875–884.

Burrus C, Gopinath R, Guo H, Odegard J, and Selesnick I. 1998. Introduction to wavelets and wavelet transforms: a primer. Vol. 1. Prentice Hall New Jersey.

Bushberg J. 2002. The essential physics of medical imaging. Lippincott Williams & Wilkins.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A, et al. 2016. Stan: A probabilistic programming language. Journal of Statistical Software 20:1–37.

Carr H and Purcell E. 1954. Effects of diffusion on free precession in nuclear magnetic resonance experiments. Physical Review 94:630.

Chang H, Huang W, Wu C, Huang S, Guan C, Sekar S, Bhakoo KK, and Duan Y. 2017. A new variational method for bias correction and its applications to rodent brain extraction. IEEE Transactions on Medical Imaging 36:721–733.

Chicklore S, Goh V, Siddique M, Roy A, Marsden P, and Cook G. 2013. Quantifying tumour heterogeneity in 18 F-FDG PET/CT imaging by texture analysis. European Journal of Nuclear Medicine and Molecular Imaging 40:133–140.

Clarke L, Nordstrom R, Zhang H, Tandon P, Zhang Y, Redmond G, Farahani K, Kelloff G, Henderson L, Shankar L, Deye J, Capala J, and Jacobs P. 2014. The Quantitative Imaging Network: NCI's Historical Perspective and Planned Goals. Translational Oncology 7:1–4.

Cocosco C, Kollokian V, Kwan R, Pike G, and Evans A. 1997. Brainweb: Online interface to a 3D MRI simulated brain database. In: *NeuroImage*.

Cohen M, DuBois R, and Zeineh M. 2000. Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging. Human Brain Mapping 10:204–211.

Collewet G, Davenel A, Toussaint C, and Akoka S. 2002. Correction of intensity nonuniformity in spin-echo T 1-weighted images. Magnetic Resonance Imaging 20:365–373.

Condon B, Patterson J, Wyper D, Jenkins A, and Hadley D. 1987. Image non-uniformity in magnetic resonance imaging: its magnitude and methods for its correction. The British Journal of Radiology 60:83–87.

Coroller T, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee S, Mak R, and Aerts H. 2016. Radiomic phenotype features predict pathological response in non-small cell lung cancer. Radiotherapy and Oncology 119:480–486.

Coroller T, Grossmann P, Hou Y, Velazquez E, Leijenaar R, Hermann G, Lambin P, Haibe-Kains B, Mak R, and Aerts H. 2015. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiotherapy and Oncology 114:345–350.

Cox D. 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) 34:187–220.

Cubes M. 1987. A high resolution 3d surface construction algorithm/william e. Lorensen, Harvey E Cline–SIG ?87.

Cunliffe A, Armato S, Castillo R, Pham N, Guerrero T, and Al-Hallaq H. 2015. Lung texture in serial thoracic computed tomography scans: Correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. International Journal of Radiation Oncology Biology Physics 91:1048–1056.

Dasarathy B and Holder E. 1991. Image characterizations based on joint gray level-run length distributions. Pattern Recognition Letters 12:497–502.

Dawant B, Zijdenbos A, and Margolin R. 1993. Correction of intensity variations in MR images for computer-aided tissue classification. IEEE Transactions on Medical Imaging 12:770–781.

Demidenko E. 2013. Mixed models: theory and applications with R. John Wiley & Sons.

Depeursinge A and Fageot J. 2017. Biomedical Texture Operators and Aggregation Functions. Academic Press, p. 55.

Dietterich T. 2000. Ensemble methods in machine learning. In: *International Workshop on Multiple Classifier Systems*. Springer, pp. 1–15.

Dietz A, Vanselow B, Rudat C, Conradt X, Weidauer H, Kallinowski F, and Dollner R. 2003. Prognostic impact of reoxygenation in advanced cancer of the head and neck during the initial course of chemoradiation or radiotherapy alone. Head Neck 25:50–58.

Dou T, Aerts H, Coroller T, and Mak R. 2017. Radiomic-Based Phenotyping of Tumor Core and Rim to Predict Survival in Nonsmall Cell Lung Cancer. International Journal of Radiation Oncology Biology Physics 99:S84.

Dupuy A and Simon R. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute 99:147–57.

Dzyubachyk O, van der Geest R, Staring M, Börnert P, Reijnierse M, Bloem J, and Lelieveldt B. 2013. Joint Intensity Inhomogeneity Correction for Whole-Body MR Data. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, pp. 106–113.

El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R, et al. 2009. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognition 42:1162–1171.

El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R, Thorstad W, and Deasy J. 2009. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognition 42:1162–1171.

Eustace A, Mani N, Span P, Irlam J, Taylor J, Betts G, Denley H, Miller C, Homer J, Rojas A, et al. 2013. A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder cancer. Clinical Cancer Research 19:4879–4888.

Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones A, Stingo F, Liao Z, Mohan R, and Court L. 2017. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific Reports 7:588.

Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 2012. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic Resonance Imaging 30:1323–1341.

Ferranti D, Krane D, and Craft D. 2017. The value of prior knowledge in machine learning of complex network systems. Bioinformatics 33:3610–3618.

Friedman J, Hastie T, and Tibshirani R. 2001. The Elements of Statistical Learning. Vol. 1. Springer series in statistics.

Galloway M. 1975. Texture analysis using gray level run lengths. Computer Graphics and Image Processing 4:172–179.

Ganeshan B, Abaleke S, Young R, Chatwin C, and Miles K. 2010. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 10:137.

Gatenby R, Grove O, and Gillies R. 2013. Quantitative imaging in cancer evolution and ecology. Radiology 269:8–15.

Gatta G, Botta L, Sánchez M, Anderson L, Pierannunzio D, Licitra L, Hackl M, Zielonke N, Oberaigner W, Van Eycken E, et al. 2015. Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: The EUROCARE-5 population-based study. European Journal of Cancer 51:2130–2143.

Gelfand I and Iaglom A. 1959. Calculation of the Amount of Information about a Random Function Contained in Another Such Function. American Mathematical Society Translations. American Mathematical Society.

George M, Kalaivani S, and Sudhakar M. 2017. A non-iterative multi-scale approach for intensity inhomogeneity correction in MRI. Magnetic Resonance Imaging 42:43–59.

Gerlinger M, Rowan A, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine 366:883–892.

Gillies R, Kinahan P, and Hricak H. 2015. Radiomics: images are more than pictures, they are data. Radiology 278:563–577.

Greenspan H, van Ginneken B, and Summers R. 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35:1153–1159.

Grossmann P, Stringfield O, El-Hachem N, Bui M, Velazquez E, Parmar C, Leijenaar R, Haibe-Kains B, Lambin P, Gillies R, et al. 2017. Defining the biological basis of radiomic phenotypes in lung cancer. eLife 6.

Grove O, Berglund A, Schabath M, Aerts H, Dekker A, Wang H, Rios Velazquez E, Lambin P, Gu Y, Balagurunathan Y, Eikman E, Gatenby R, Eschrich S, and Gillies R. 2015. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. PLoS One 10:1–14.

Guyon I and Elisseeff A. 2003. An Introduction to Variable and Feature Selection. 3:1157–1182.

Haacke E, Mittal S, Wu Z, Neelavalli J, and Cheng Y. 2009. Susceptibility-weighted imaging: technical aspects and clinical applications, part 1. American Journal of Neuroradiology 30:19–30.

Haralick R, Shanmugan K, and Dinstein I. 1973. Textural features for image classification.

Harrell F, Lee K, and Mark D. 1996a. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Statistics in Medicine 15:361–387.

Harrell F, Lee K, and Mark D. 1996b. Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Tutorials in Biostatistics, Statistical Methods in Clinical Studies 1:223–249.

Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest C, Groheux D, Hindié E, Martineau A, Pradier O, Hustinx R, et al. 2015. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. Journal of Nuclear Medicine 56:38–44.

Hatt M, Tixier F, Pierce L, Kinahan P, Le Rest C, and Visvikis D. 2016. Characterization of PET/CT images using texture analysis: the past, the present... any future? European Journal of Nuclear Medicine and Molecular Imaging.

Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby R, Balagurunathan Y, et al. 2016. Predicting malignant nodules from screening CT scans. Journal of Thoracic Oncology 11:2120–2128.

Hentschel M, Appold S, Schreiber A, Abolmaali N, Abramyuk A, Dörr W, Kotzerke J, Baumann M, and Zöphel K. 2011. Early FDG PET at 10 or 20 Gy under chemoradiotherapy is prognostic for locoregional control and overall survival in patients with head and neck cancer. European Journal of Nuclear Medicine and Molecular Imaging 38:1203–1211.

Herrick R, Hayman L, Taber K, Diaz-Marchan P, and Kuo M. 1997. Artifacts and pitfalls in MR imaging of the orbit: a clinical review. RadioGraphics 17:707–24.

Holschneider M, Kronland-Martinet R, Morlet J, and Tchamitchian P. 1990. A real-time algorithm for signal analysis with the help of the wavelet transform. In: *Wavelets*. Springer, pp. 286–297.

Hothorn T and Lausen B. 2003. On the exact distribution of maximally selected rank statistics. Computational Statistics and Data Analysis 43:121–137.

Hutchings M, Loft A, Hansen M, Pedersen L, Buhl T, Jurlander J, Buus S, Keiding S, D'Amore F, Boesen A, et al. 2006. FDG-PET after two cycles of chemotherapy predicts treatment failure and progression-free survival in Hodgkin lymphoma. Blood 107:52–59.

Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi A, Mirabelli P, Monti S, and Salvatore M. 2017. Radiogenomic analysis of oncological data: a technical survey. International Journal of Molecular Sciences 18:805.

Ishwaran H and Kogalur U. 2007. Random survival forests for R. New Functions for Multivariate Analysis:25.

Ishwaran H, Kogalur U, Blackstone E, and Lauer M. 2008. Random survival forests. The Annals of Applied Statistics:841–860.

Ishwaran H, Kogalur U, Chen X, and Minn A. 2011. Random survival forests for high-dimensional data. Statistical Analysis and Data Mining: The ASA Data Science Journal 4:115–132.

Ishwaran H, Kogalur U, Gorodeski E, Minn A, and Lauer M. 2010. High-dimensional variable selection for survival data. Journal of the American Statistical Association 105:205–217.

Ivanovska T, Laqua R, Wang L, Schenk A, Yoon J, Hegenscheid K, Völzke H, and Liebscher V. 2016. An efficient level set method for simultaneous intensity inhomogeneity correction and segmentation of MR images. Computerized Medical Imaging and Graphics 48:9–20.

Jackson DF and Hawkes DJ. 1981. X-ray attenuation coefficients of elements and mixtures. Physics Reports 70:169–233.

Jain A and Dubes R. 1988. Algorithms for Clustering Data. Prentice-Hall, Inc.

Jain R, Poisson L, Gutman D, Scarpace L, Hwang S, Holder C, Wintermark M, Rao A, Colen R, Kirby J, et al. 2014. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. Radiology 272:484–493.

Jalalimanesh A, Haghighi H, Ahmadi A, and Soltani M. 2017. Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. Mathematics and Computers in Simulation 133:235–248.

Janitza S, Celik E, and Boulesteix A. 2015. A computationally fast variable importance test for random forests for high-dimensional data. Advances in Data Analysis and Classification:1–31.

Jemal A, Bray F, Center M, Ferlay J, Ward E, and Forman D. 2011. Global cancer statistics. CA: A Cancer Journal for Clinicians 61:69–90.

Kachelriess M, Knaup M, and Kalender W. 2004. Extended parallel backprojection for standard three-dimensional and phase-correlated four-dimensional axial and spiral cone-beam CT with arbitrary pitch, arbitrary cone-angle, and 100% dose usage. Medical Physics 31:1623–1641.

Kang L, Chen W, Petrick N, and Gallas B. 2015. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. Statistics in Medicine 34:685–703.

Kanungo T, Mount D, Netanyahu N, Piatko C, Silverman R, and Wu A. 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:881–892.

Kaplan EL and Meier P. 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481.

Karaboga D. 2005. An idea based on Honey Bee Swarm for Numerical Optimization. Technical Report:10.

Karaboga D and Basturk B. 2007. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. Journal of Global Optimization 39:459–471.

Karaboga D and Basturk B. 2008. On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing 8:687–697.

Kather J, Weis C, Bianconi F, Melchers S, Schad L, Gaiser T, Marx A, and Zöllner F. 2016. Multi-class texture analysis in colorectal cancer histology. Scientific Reports 6:27988.

Kickingereder P, Burth S, Wick A, Götz M, Eidel O, Schlemmer H, Maier-Hein K, Wick W, Bendszus M, Radbruch A, et al. 2016. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. Radiology 280:880–889.

Kim H, Park C, Lee M, Park S, Song Y, Lee J, Hwang E, and Goo J. 2016. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. PLoS One 11:1–11.

King A, Chow K, Yu K, Mo FK, Yeung D, Yuan J, Bhatia K, Vlantis A, and Ahuja A. 2013. Head and neck squamous cell carcinoma: diagnostic performance of diffusion-weighted MR imaging for the prediction of treatment response. Radiology 266:531–8.

Krupa K and Bekiesińska-Figatowska M. 2015. Artifacts in magnetic resonance imaging. Polish Journal of Radiology 80:93–106.

Kumar V, Gu Y, Basu S, Berglund A, Eschrich S, Schabath M, Forster K, Aerts H, Dekker A, Fenstermacher D, et al. 2012. Radiomics: the process and the challenges. Magnetic Resonance Imaging 30:1234–1248.

Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout R, Granton P, Zegers C, Gillies R, Boellard R, Dekker A, et al. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer 48:441–446.

Langfelder P and Horvath S. 2012. Fast R functions for robust correlations and hierarchical clustering. Journal of Statistical Software 46.

Lao J, Chen Y, Li Z, Li Q, Zhang J, Liu J, and Zhai G. 2017. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Scientific Reports 7:10353.

Leger S, Bandurska-Luque A, Pilz K, Zöphel K, Baumann M, Troost EGC, Löck S, and Richter C. 2016. OC-0262: Comparison of machine-learning methods for predictive radiomic models in locally advanced HNSCC. Radiotherapy and Oncology 119:121–122.

Leger S, Löck S, Hietschold V, Haase R, Böhme H, and Abolmaali N. 2014. Reproducible and Accurate Automatic Correction of Intensity Non-Uniformity in MRI Data. In: *Joint Conference of the SSRMP, DGMP, ÖGMP 2014*, pp. 45–47.

Leger S, Löck S, Hietschold V, Haase R, Böhme H, and Abolmaali N. 2015. Automatic Intensity Non-Uniformity Correction in MRI Data. In: *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*.

Leger S, Löck S, Hietschold V, Haase R, Böhme H, and Abolmaali N. 2017a. Physical correction model for automatic correction of intensity non-uniformity in magnetic resonance imaging. Physics and Imaging in Radiation Oncology 4:32–38.

Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, Kotzerke J, Schreiber A, Tinhofer I, Budach C, Sak A, Stuschke M, Balermpas P, Rödel C, Ganswindt U, Belka C, Pigorsch S, Combs S, Mönnich D, Zips D, Krause M, Baumann M, Richter C, Troost E, and Löck S. 2018. Identification of tumour sub-volumes for improved radiomic risk modelling in locally advanced HNSCC. Radiotherapy and Oncology.

Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, Kotzerke J, Schreiber A, Tinhofer I, Budach C, Sak A, Stuschke M, Balermpas P, Rödel C, Ganswindt U, Belka C, Pigorsch S, Combs S, Mönnich D, Zips D, Krause M, Baumann M, Troost E, Löck S, and Richter C. 2017b. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Scientific Reports 7:13206.

Leijenaar R, Carvalho S, Velazquez E, Van Elmpt W, Parmar C, Hoekstra O, Hoekstra C, Boellaard R, Dekker A, Gillies R, et al. 2013. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncologica 52:1391–1397.

Lewiner T, Lopes H, Vieira A, and Tavares G. 2003. Efficient implementation of marching cubes' cases with topological guarantees. Journal of Graphics Tools 8:1–15.

Li Q, Bai H, Chen Y, Sun Q, Liu L, Zhou S, Wang G, Liang C, and Li Z. 2017. A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. Scientific Reports 7:14331.

Liney G, Turnbull L, and Knowles A. 1998. A simple method for the correction of endorectal surface coil inhomogeneity in prostate imaging. Journal of Magnetic Resonance Imaging 8:994–997.

Linge A, Löck S, Krenn C, Appold S, Lohaus F, Nowak A, Gudziol V, Baretton G, Buchholz F, Baumann M, et al. 2016a. Independent validation of the prognostic value of cancer stem cell marker expression and hypoxia-induced gene expression for patients with locally advanced HNSCC after postoperative radiotherapy. Clinical and Translational Radiation Oncology 1:19–26.

Linge A, Lohaus F, Löck S, Nowak A, Gudziol V, Valentini C, von Neubeck C, Jütz M, Tinhofer I, Budach V, et al. 2016b. HPV status, cancer stem cell marker expression, hypoxia gene signatures and tumour volume identify good prognosis subgroups in patients with HNSCC after primary radiochemotherapy: A multicentre retrospective study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). Radiotherapy and Oncology 121:364–373.

Linge A, Löck S, Gudziol V, Nowak A, Lohaus F, von Neubeck C, Jütz M, Abdollahi A, Debus J, Tinhofer I, et al. 2016c. Low cancer stem cell marker expression and low hypoxia identify good prognosis subgroups in HPV (-) HNSCC after postoperative radiochemotherapy: a multicenter study of the DKTK-ROG. Clinical Cancer Research 22:2639–2649.

Liu R, Hall L, Goldgof D, Zhou M, Gatenby R, and Ahmed K. 2016. Exploring deep features from brain tumor magnetic resonance images via transfer learning. In: *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, pp. 235–242.

Ljungkvist A, Bussink J, Kaanders J, and van der Kogel A. 2007. Dynamics of tumor hypoxia measured with bioreductive hypoxic cell markers. Radiation Research 167:127–145.

Löck S, Perrin R, Seidlitz A, and Bandurska-luque A. 2017. Residual tumour hypoxia in head-and-neck cancer patients undergoing primary radiochemotherapy , final results of a prospective trial on repeat FMISO-PET imaging. Radiotherapy and Oncology 124:533–540.

Lohaus F, Linge A, Tinhofer I, Budach V, Gkika E, Stuschke M, Balermpas P, Rödel C, Avlar M, Grosu A, et al. 2014. HPV16 DNA status is a strong prognosticator of loco-regional control after postoperative radiochemotherapy of locally advanced oropharyngeal carcinoma: results from a multicentre explorative study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG). Radiotherapy and Oncology 113:317–323.

Low R. 2007. Abdominal MRI advances in the detection of liver tumours and characterisation. The Lancet Oncology 8:525–535.

Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones A, and Court L. 2015. Measuring CT scanner variability of radiomics features. Investigative Radiology 50:757.

Mantel N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50:163–170.

Marr D and Hildreth E. 1980. Theory of edge detection. Proceedings of the Royal Society of London B: Biological Sciences 207:187–217.

Milles J, Zhu Y, Gimenez G, Guttmann C, and Magnin I. 2007. MRI intensity nonuniformity correction using simultaneously spatial and gray-level histogram information. Computerized Medical Imaging and Graphics 31:81–90.

Milletari F, Navab N, and Ahmadi S. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *Fourth International Conference on 3D Vision*, pp. 565–571.

Muthupillai R, Lomas D, Rossman P, Greenleaf JF, Manduca A, and Ehman RL. 1995. Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. Science 269:1854–1857.

Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder C, Gutman D, Jain R, Colen R, Rubin D, Zinn P, et al. 2015. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. Journal of Neuroradiology 42:212–221.

Nomden C, de Leeuw A, Van Limbergen E, De Brabandere M, Nulens A, Nout R, Laman M, Ketelaars M, Lutgens L, Reniers B, et al. 2013. Multicentre treatment planning study of MRI-guided brachytherapy for cervical cancer: comparison between tandem-ovoid applicator users. Radiotherapy and Oncology 107:82–87.

O'Connor J, Rose C, Waterton J, Carano R, Parker G, and Jackson A. 2015. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clinical Cancer Research 21:249–257.

Ogawa S, Lee T, Nayak A, and Glynn P. 1990. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magnetic Resonance in Medicine 14:68–78.

Otsu N. 1975. A threshold selection method from gray-level histograms. Automatica 11:23–27.

Overgaard J, Hansen H, Overgaard M, Bastholt L, Berthelsen A, Specht L, Lindeløv B, and Jørgensen K. 1998. A randomized double-blind phase III study of nimorazole as a hypoxic radiosensitizer of primary radiotherapy in supraglottic larynx and pharynx carcinoma. Results of the Danish Head and Neck Cancer Study (DAHANCA) Protocol 5-85. Radiotherapy and Oncology 46:135–146.

Overgaard J, Hansen H, Specht L, Overgaard M, Grau C, Andersen E, Bentzen J, Bastholt L, Hansen O, Johansen J, et al. 2003. Five compared with six fractions per week of conventional radiotherapy of squamous-cell carcinoma of head and neck: DAHANCA 6&7 randomised controlled trial. The Lancet 362:933–940.

Panth K, Leijenaar R, Carvalho S, Lieuwes N, Yaromina A, Dubois L, and Lambin P. 2015. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiotherapy and Oncology 116:462–466.

Parmar C, Grossmann P, Bussink J, Lambin P, and Aerts H. 2015a. Machine learning methods for quantitative radiomic biomarkers. Scientific Reports 5:13087.

Parmar C, Grossmann P, Rietveld D, Rietbergen M, Lambin P, and Aerts H. 2015b. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Frontiers in Oncology 5:272.

Parmar C, Leijenaar R, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, Rietbergen M, Haibe-Kains B, Lambin P, and Aerts H. 2015c. Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer. Scientific Reports 5:1–10.

Partridge S, Gibbs J, Lu Y, Esserman L, Tripathy D, Wolverton D, Rugo H, Hwang E, Ewing C, and Hylton N. 2005. MRI measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival. American Journal of Roentgenology 184:1774–1781.

Paul R, Hawkins S, Balagurunathan Y, Schabath M, Gillies R, Hall L, and Goldgof D. 2016. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. Tomography: A Journal for Imaging Research 2:388.

Pencina M and D’Agostino R. 2004. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. Statistics in Medicine 23:2109–2123.

Peng H, Long F, and Ding C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:1226–1238.

Pereira S, Pinto A, Alves V, and Silva C. 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Transactions on Medical Imaging 35:1240–1251.

Pilz K, Leger S, Zwanenburg A, Richter C, Krause M, Baumann M, Löck S, and Troost E. 2017. EP-1065: Prediction of Dysphagia and Xerostomia based on CT imaging features of HNSCC Patients. Radiotherapy and Oncology 123:S585–S586.

Poleszczuk J and Enderling H. 2016. Cancer stem cell plasticity as tumor growth promoter and catalyst of population collapse. Stem Cells International 2016.

Purcell E, Torrey H, and Pound R. 1946. Resonance absorption by nuclear magnetic moments in a solid. Physical Review 69:37.

RJ L and Nicewander W. 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42:59–66.

Roth H, Lu L, Liu J, Yao J, Seff A, Cherry K, Kim L, and Summers R. 2016. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Transactions on Medical Imaging 35:1170–1181.

Rutherford R, Pullan B, and Isherwood I. 1976. Measurement of effective atomic number and electron density using an EMI scanner. Neuroradiology 11:15–21.

Schmidt S, Linge A, Zwanenburg A, Leger S, Lohaus F, Krenn C, Appold S, Gudziol V, Nowak A, von Neubeck C, et al. 2018. Development and validation of a gene signature for patients with head and neck squamous cell carcinomas treated by postoperative radio (chemo) therapy. Clinical Cancer Research.

Schütze C, Bergmann R, Brüchner K, Mosch B, Yaromina A, Zips D, Hessel F, Krause M, Thames H, Kotzerke J, et al. 2014. Effect of [18F] FMISO stratified dose-escalation on local control in FaDu hSCC in nude mice. Radiotherapy and Oncology 111:81–87.

Schütze C, Bergmann R, Yaromina A, Hessel F, Kotzerke J, Steinbach J, Baumann M, and Beuthien-Baumann B. 2007. Effect of increase of radiation dose on local control relates to pre-treatment FDG uptake in FaDu tumours in nude mice. Radiotherapy and Oncology 83:311–315.

Segal E, Sirlin C, Ooi C, Adler A, Gollub J, Chen X, Chan B, Matcuk G, Barry C, Chang H, et al. 2007. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature Biotechnology 25:675.

Shafiq-Ul-Hassan MZ GG, Latifi K, Ullah G, Hunt D, Balagurunathan Y, Abdalah M, Matthew B, Goldgof D, Mackin D, Court L, James R, and Moros E. 2017. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Medical Physics 44:1050–1062.

Shannon C. 1948. A mathematical theory of communication. Bell System Technical Journal 27:379–423.

Shrout P and Fleiss J. 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86:420.

Simmons A, Tofts P, Barker G, and Arridge S. 1994. Sources of intensity nonuniformity in spin echo images at 1.5 T. Magnetic Resonance in Medicine 32:121–128.

Simon N, Friedman J, Hastie T, and Tibshirani R. 2011. Regularization paths for Cox?s proportional hazards model via coordinate descent. Journal of Statistical Software 39:1.

Sled J, Zijdenbos P, and Evans C. 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17:87–97.

Song J, Liu Z, Zhong W, Huang Y, Ma Z, Dong D, Liang C, and Tian J. 2016. Non-small cell lung cancer: quantitative phenotypic analysis of CT images as a potential marker of prognosis. Scientific Reports 6:1–9.

Spearman C. 1910. Correlation calculated from faulty data. British Journal of Psychology 3:271–295.

Stadler P, Feldmann H, Creighton C, Kau R, and Molls M. 1998. Changes in tumor oxygenation during combined treatment with split-course radiotherapy and chemotherapy in patients with head and neck cancer. Radiotherapy and Oncology 48:157–64.

Stejskal E and Tanner J. 1965. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. The Journal of Chemical Physics 42:288–292.

Stollnitz E, DeRose A, and Salesin D. 1995. Wavelets for computer graphics: a primer. 1. IEEE Computer Graphics and Applications 15:76–84.

Sun C and Wee W. 1983. Neighboring gray level dependence matrix for texture classification. Computer Vision, Graphics and Image Processing 23:341–352.

Therneau T and Grambsch P. 2000. Modeling Survival Data: Extending the Cox Model. New York.

Thibault G, Angulo J, and Meyer F. 2014. Advanced statistical matrices for texture characterization: application to cell classification. IEEE Transactions on Biomedical Engineering 61:630–637.

Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J, and Mari J. 2009. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. Pattern Recognition and Information Processing.

Toustrup K, Sørensen B, Nordsmark M, Busk M, Wiuf C, Alsner J, and Overgaard J. 2011. Development of a hypoxia gene expression classifier with predictive impact for hypoxic modification of radiotherapy in head and neck cancer. Cancer Research 71:5923–5931.

Tustison N, Cook P, and Gee J. 2011. N4ITK: Improved N3 Bias Correction. IEEE Transactions on Medical Imaging 29:1310–1320.

Usmanij E, de Geus-Oei L, Troost E, Peters-Bax L, van der Heijden E, Kaanders J, Oyen W, Schuurbiers O, and Bussink J. 2013. 18F-FDG PET Early Response Evaluation of Locally Advanced Non–Small Cell Lung Cancer Treated with Concomitant Chemoradiotherapy. Journal of Nuclear Medicine 54:1528–1534.

Vallières M, Freeman C, Skamene S, and El Naqa I. 2015. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Physics in Medicine and Biology 60:5471–96.

Vallières M, Kay-Rivest E, Perrin L, Liem X, Furstoss C, Aerts H, Khaouam N, Nguyen-Tan P, Wang C, Sultanem K, et al. 2017. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports 7:10117.

Van de Moortele P, Auerbach E, Olman C, Yacoub E, Uurbil K, and Moeller S. 2009. T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. NeuroImage 46:432–446.

Van Essen D, Smith S, Barch D, Behrens T, Yacoub E, Ugurbil K, Consortium WMH, et al. 2013. The WU-Minn Human Connectome Project: An Overview David. NeuroImage 70:646–656.

Van Leemput K, Maes F, Vandermeulen D, and Suetens P. 1999. Automated model-based bias field correction of MR images of the brain. IEEE Transactions on Medical Imaging 18:885–896.

Van Timmeren J, Leijenaar R, Van Elmpt W, and Lambin P. 2016. Are planning CT radiomics and cone-beam CT radiomics interchangeable? Radiotherapy and Oncology:446–447.

Van Dijk L, Brouwer C, van der Schaaf A, Burgerhof J, Beukinga R, Langendijk J, Sijtsema N, and Steenbakkers R. 2017a. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. Radiotherapy and Oncology 122:185–191.

Van Dijk L, Noordzij W, Brouwer C, Boellaard R, Burgerhof J, Langendijk J, Sijtsema N, and Steenbakkers R. 2017b. 18F-FDG PET image biomarkers improve prediction of late radiation-induced xerostomia. Radiotherapy and Oncology 120:89–95.

Van Putten L. 1968. Tumour reoxygenation during fractionated radiotherapy; studies with a transplantable mouse osteosarcoma. European Journal of Cancer 4:172–82.

Van Timmeren J, Leijenaar R, van Elmpt W, Reymen B, and Lambin P. 2017a. Feature selection methodology for longitudinal cone-beam CT radiomics. Acta Oncologica 56:1537–1543.

Van Timmeren J, Leijenaar R, van Elmpt W, Reymen B, Oberije C, Monshouwer R, Bussink J, Brink C, Hansen O, and Lambin P. 2017b. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and Oncology 123:363–369.

Van Timmeren J, Leijenaar R, van Elmpt W, Wang J, Zhang Z, Dekker A, and Lambin P. 2016. Test-retest data for radiomics feature stability analysis: generalizable or study specific. Tomography: A Journal for Imaging Research 2:361–365.

Vaupel P, Kallinowski F, and Okunieff P. 1989. Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: a review. Cancer Research 49:6449–6465.

Vovk U, Pernuš F, and Likar B. 2006. Intensity inhomogeneity correction of multispectral MR images. NeuroImage 32:54–61.

Vovk U, Pernuš F, and Likar B. 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Transactions on Medical Imaging 26:405–421.

Waclaw B, Bozic I, Pittman M, Hruban R, Vogelstein B, and Nowak M. 2015. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature 525:261.

Wald R, Khoshgoftaar T, Dittman D, Awada W, and Napolitano A. 2012. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In: *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*. IEEE, pp. 377–384.

Wells W, Grimson W, Kikinis R, and Jolesz F. 1996. Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging 15:429–442.

Wibmer A, Hricak H, Gondo T, Matsumoto K, Veeraraghavan H, Fehr D, Zheng J, Goldman D, Moskowitz C, Fine S, et al. 2015. Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. European Radiology 25:2840–2850.

Wicks D, Barker G, and Tofts P. 1993. Correction of intensity nonuniformity in MR images of any orientation. Magnetic Resonance Imaging 11:183–196.

Wiedenmann N, Bucher S, Hentschel M, Mix M, Vach W, Bittner M, Nestle U, Pfeiffer J, Weber W, and Grosu A. 2015. Serial [18F]-fluoromisonidazole PET during radiochemotherapy for locally advanced head and neck cancer and its correlation with outcome. Radiotherapy and Oncology 117:113–117.

Wright M, Dankowski T, and Ziegler A. 2016. Random forests for survival analysis using maximally selected rank statistics. arXiv preprint arXiv:160503391.

Wu J, Gensheimer M, Dong X, Rubin D, Napel S, Diehn M, Loo B, and Li R. 2016a. Robust Intratumor Partitioning to Identify High-Risk Subregions in Lung Cancer: A Pilot Study. International Journal of Radiation Oncology Biology Physics 95:1504–1512.

Wu J, Gong G, Cui Y, and Li R. 2016b. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. Journal of Magnetic Resonance Imaging 44:1107–1115.

Yamamoto S, Maki D, Korn R, and Kuo M. 2012. Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape. American Journal of Roentgenology 199:654–663.

Yamaner S, Güllüoğlu M, Kapran Y, and Özel S. 2012. Can diffusion-weighted MRI determine complete responders after neoadjuvant chemoradiation for locally advanced rectal cancer? Diagnostic and Interventional Radiology 18:574–581.

Yaromina A, Kroeber T, Meinzer A, Boeke S, Thames H, Baumann M, and Zips D. 2011. Exploratory study of the prognostic value of microenvironmental parameters during fractionated irradiation in human squamous cell carcinoma xenografts. International Journal of Radiation Oncology Biology Physics 80:1205–1213.

Yip S and Aerts H. 2016. Applications and limitations of radiomics. Physics in Medicine and Biology 61:R150.

Zhao B, Tan Y, Tsai W, Qi J, Xie C, Lu L, and Schwartz L. 2016. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Scientific Reports 6:23428.

Zhou L, Zhu Y, Bergot C, Laval-Jeantet A, Bousson V, Laredo J, and Laval-Jeantet M. 2001. A method of radio-frequency inhomogeneity correction for brain tissue segmentation in MRI. Computerized Medical Imaging and Graphics 25:379–389.

Zhuge Y, Udupa J, Liu J, Saha P, and Iwanage T. 2002. Scale-based method for correcting background intensity variation in acquired images. In: *Medical Imaging*. International Society for Optics and Photonics, pp. 1103–1111.

Zips D, Zöphel K, Abolmaali N, Perrin R, Abramyuk A, Haase R, Appold S, Steinbach J, Kotzerke J, and Baumann M. 2012. Exploratory prospective trial of hypoxia-specific PET imaging during radiochemotherapy in patients with locally advanced head-and-neck cancer. Radiotherapy and Oncology 105:21–28.

Zschaeck S, Löck S, Leger S, Haase R, Bandurska-Luque A, Appold S, Kotzerke J, Zips D, Richter C, Gudziol V, et al. 2017. FDG uptake in normal tissues assessed by PET during treatment has prognostic value for treatment results in head and neck squamous cell carcinomas undergoing radiochemotherapy. Radiotherapy and Oncology 122:437–444.

Zwanenburg A, Leger S, Vallières M, Löck S, et al. 2016. Image biomarker standardisation initiative. arXiv preprint arXiv:161207003.

Appendix

A. Hyper-parameters for the feature selection methods

Table A.1.: Definition of the hyper-parameters for the feature selection methods. These parameters were kept fixed and were not optimised during hyper-parameter optimisation.

Feature selection method	Hyper-parameter name	Parameter value(s)
MRMR	topFeatures	100
	RelativeImportanceThreshold	0
MIFS	topFeatures	100
	RelativeImportanceThreshold	0.05
uni-Cox	nIterations	10
	nFolds	2
multi-Cox	nIterations	20
	nFolds	3
	nTopFeatures	20
	modelSize	3
	α	0.1
PVI-RF	topFeatures	100
	nTree	1000
	mTry	10
	nodeSize	45
	splitrule	maxstat
	α	0.5
	minprop	0.5
MSR-RF	topFeatures	100
	nTree	1000
	mTry	10
	nodeSize	20
	splitrule	maxstat
	α	0.5

Feature selection method	Hyper-parameter name	Parameter value(s)
	minprop	0.5
RF-VI	topFeatures	20
	nTree	1000
	K	2
	nRepetition	50
	nSteps	2
	nSplits	1
	splitrule	logrank
	nodeSize	45
	mTry	500
	nVariables	10
RF-MD	topFeatures	20
	nTree	2000
	K	5
	nRepetition	50
	nSteps	5
	nSplits	1
	splitrule	logrank
	nodeSize	20
	mTry	100
	nVariables	100
RF-VH	topFeatures	100
	nTree	1000
	K	2
	nRepetition	50
	nSteps	2
	nSplits	1
	splitrule	logrank
	nodeSize	20
	mTry	500
	nVariables	10

B. Hyper-parameters for the machine learning algorithms

The following paragraph defined the different hyper-parameters of the individual machine learning algorithms. The hyper-parameter optimisation aims to find an optimal configuration for the machine learning algorithms to adjust the specific algorithms to the prediction task, which may improve the prognostic accuracy and reduce the influence of model over-fitting. This is particularly important for the more complex models (e.g., BT-Weibull, BT-Cox and RSF), as the choice of their hyper-parameters influences how well they can learn the underlying data distribution. One challenge of hyper-parameter optimisation arises with the high number of different model parameters, which requires computational resources to optimise them. To limit these resources, hyper-parameter ranges for each algorithm were manually defined with the aim to reduce the possible parameter space. The parameter space was defined based on prior knowledge, e.g., the maximum signature size was derived by the number of events, i.e., 10 events per predictor variable as well as identifying those settings that led to balanced performance in the internal cross-validation of the exploratory cohort. In future, a further time reduction could be achieved by replacing the exhaustive grid search optimisation by a random search strategy (Bergstra and Bengio, 2012).

Table B.1.: Definition of the hyper-parameters for the for the machine learning algorithms, which were used during the hyper-parameter optimisation.

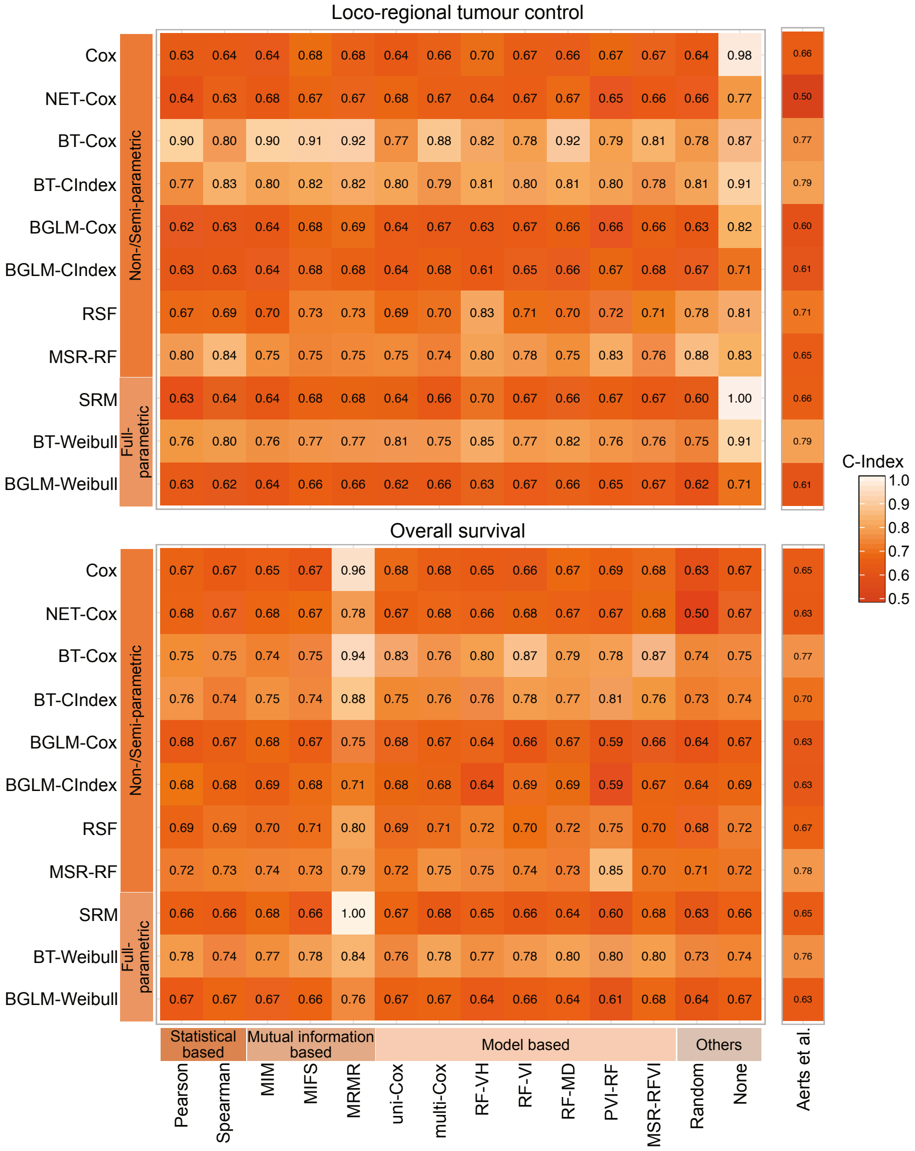
Machine learning algorithm	Hyper-parameter name	Parameter value(s)
Cox model		
	Signature size	2, 3, 4, 5, 7, 10
NET-Cox model		
	Signature size	2, 3, 4, 5, 7, 10
	α	0–1.0, step size 0.2
	ω	lambda.min[1], lambda.1se[2]
BT based models		
	Signature size	2,3,4,5,7,10
	α	0.001, 0.01, 0.05
	ω	lambda.min[1], lambda.1se[2]
	mStop	200
RSF model		
	Signature size	2, 3, 4, 5, 7, 10
	ntree	2000
	mtry	100
	$node-Size$	25–50, step size 1
	maxDepth	10, 15
	nSplit	1, 2, 100
	splitRule	logrank, logrankscore
MSR-RF model		
	Signature size	2, 3, 4, 5, 7, 10
	ntree	2000
	mtry	100
	$node-Size$	25–50, step size 1
	minprop	0.1
	α	0.1, 0.5
	splitRule	C, maxstat
SRM model		
	Signature size	2, 3, 4, 5, 7, 10
	Distribution	weibull, gaussian, exponential

[1] minimum mean cross-validated error

[2] Error within one standard error of the minimum

C. Comparison of feature selection methods and machine learning algorithms for time-to-event survival models

C.A. Prognostic performances on the exploratory cohort



Loco-regional tumour control

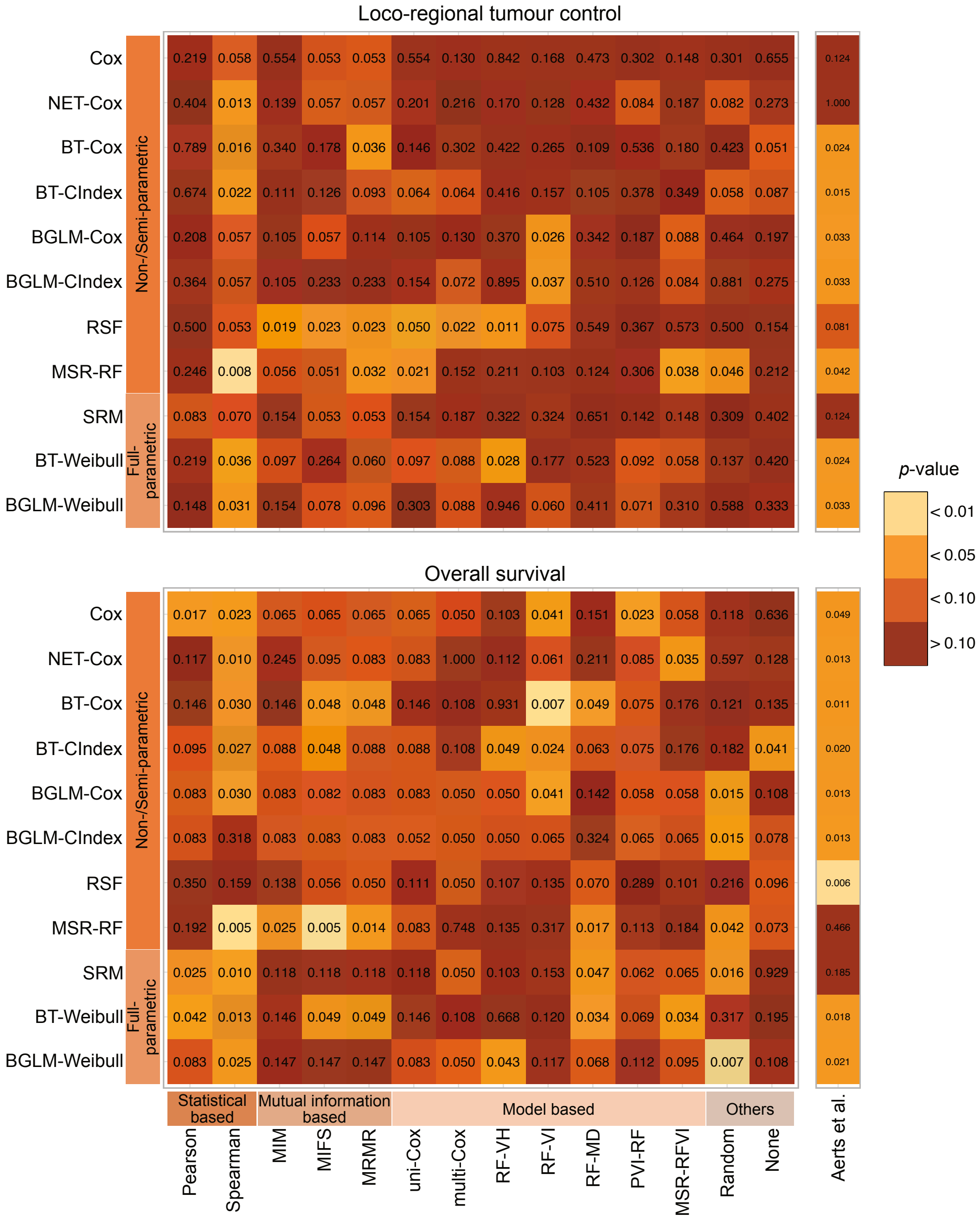
		Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Non-/Semi-parametric	Cox	0.63	0.64	0.64	0.68	0.68	0.64	0.66	0.70	0.67	0.66	0.67	0.67	0.64	0.98	0.66
	NET-Cox	0.64	0.63	0.68	0.67	0.67	0.68	0.67	0.64	0.67	0.67	0.65	0.66	0.66	0.77	0.50
	BT-Cox	0.90	0.80	0.90	0.91	0.92	0.77	0.88	0.82	0.78	0.92	0.79	0.81	0.78	0.87	0.77
	BT-CIndex	0.77	0.83	0.80	0.82	0.82	0.80	0.79	0.81	0.80	0.81	0.80	0.78	0.81	0.91	0.79
	BGLM-Cox	0.62	0.63	0.64	0.68	0.69	0.64	0.67	0.63	0.67	0.66	0.66	0.66	0.63	0.82	0.60
	BGLM-CIndex	0.63	0.63	0.64	0.68	0.68	0.64	0.68	0.61	0.65	0.66	0.67	0.68	0.67	0.71	0.61
	RSF	0.67	0.69	0.70	0.73	0.73	0.69	0.70	0.83	0.71	0.70	0.72	0.71	0.78	0.81	0.71
	MSR-RF	0.80	0.84	0.75	0.75	0.75	0.75	0.74	0.80	0.78	0.75	0.83	0.76	0.88	0.83	0.65
Full-parametric	SRM	0.63	0.64	0.64	0.68	0.68	0.64	0.66	0.70	0.67	0.66	0.67	0.67	0.60	1.00	0.66
	BT-Weibull	0.76	0.80	0.76	0.77	0.77	0.81	0.75	0.85	0.77	0.82	0.76	0.76	0.75	0.91	0.79
	BGLM-Weibull	0.63	0.62	0.64	0.66	0.66	0.62	0.66	0.63	0.67	0.66	0.65	0.67	0.62	0.71	0.61

Overall survival

		Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Non-/Semi-parametric	Cox	0.67	0.67	0.65	0.67	0.96	0.68	0.68	0.65	0.66	0.67	0.69	0.68	0.63	0.67	0.65
	NET-Cox	0.68	0.67	0.68	0.67	0.78	0.67	0.68	0.66	0.68	0.67	0.67	0.68	0.50	0.67	0.63
	BT-Cox	0.75	0.75	0.74	0.75	0.94	0.83	0.76	0.80	0.87	0.79	0.78	0.87	0.74	0.75	0.77
	BT-CIndex	0.76	0.74	0.75	0.74	0.88	0.75	0.76	0.76	0.78	0.77	0.81	0.76	0.73	0.74	0.70
	BGLM-Cox	0.68	0.67	0.68	0.67	0.75	0.68	0.67	0.64	0.66	0.67	0.59	0.66	0.64	0.67	0.63
	BGLM-CIndex	0.68	0.68	0.69	0.68	0.71	0.68	0.68	0.64	0.69	0.69	0.59	0.67	0.64	0.69	0.63
	RSF	0.69	0.69	0.70	0.71	0.80	0.69	0.71	0.72	0.70	0.72	0.75	0.70	0.68	0.72	0.67
	MSR-RF	0.72	0.73	0.74	0.73	0.79	0.72	0.75	0.75	0.74	0.73	0.85	0.70	0.71	0.72	0.78
Full-parametric	SRM	0.66	0.66	0.68	0.66	1.00	0.67	0.68	0.65	0.66	0.64	0.60	0.68	0.63	0.66	0.65
	BT-Weibull	0.78	0.74	0.77	0.78	0.84	0.76	0.78	0.77	0.78	0.80	0.80	0.80	0.73	0.74	0.76
	BGLM-Weibull	0.67	0.67	0.67	0.66	0.76	0.67	0.67	0.64	0.66	0.64	0.61	0.68	0.64	0.67	0.63

Figure C.1.: Concordance indices (C-Index) to predict loco-regional tumour (top) control and overall survival (bottom) depending on the feature selection methods (columns) and learning algorithms (rows) for the exploratory cohort. Performance of the Aerts *et al.* (Aerts et al., 2014) signature is depicted.

C.B. Risk-based patient stratification



Loco-regional tumour control

		Statistical based		Mutual information based			Model based							Others		
		Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Non-/Semi-parametric	Cox	0.219	0.058	0.554	0.053	0.053	0.554	0.130	0.842	0.168	0.473	0.302	0.148	0.301	0.655	0.124
	NET-Cox	0.404	0.013	0.139	0.057	0.057	0.201	0.216	0.170	0.128	0.432	0.084	0.187	0.082	0.273	1.000
	BT-Cox	0.789	0.016	0.340	0.178	0.036	0.146	0.302	0.422	0.265	0.109	0.536	0.180	0.423	0.051	0.024
	BT-CIndex	0.674	0.022	0.111	0.126	0.093	0.064	0.064	0.416	0.157	0.105	0.378	0.349	0.058	0.087	0.015
	BGLM-Cox	0.208	0.057	0.105	0.057	0.114	0.105	0.130	0.370	0.026	0.342	0.187	0.088	0.464	0.197	0.033
	BGLM-CIndex	0.364	0.057	0.105	0.233	0.233	0.154	0.072	0.895	0.037	0.510	0.126	0.084	0.881	0.275	0.033
	RSF	0.500	0.053	0.019	0.023	0.023	0.050	0.022	0.011	0.075	0.549	0.367	0.573	0.500	0.154	0.081
	MSR-RF	0.246	0.008	0.056	0.051	0.032	0.021	0.152	0.211	0.103	0.124	0.306	0.038	0.046	0.212	0.042
Full-parametric	SRM	0.083	0.070	0.154	0.053	0.053	0.154	0.187	0.322	0.324	0.651	0.142	0.148	0.309	0.402	0.124
	BT-Weibull	0.219	0.036	0.097	0.264	0.060	0.097	0.088	0.028	0.177	0.523	0.092	0.058	0.137	0.420	0.024
	BGLM-Weibull	0.148	0.031	0.154	0.078	0.096	0.303	0.088	0.946	0.060	0.411	0.071	0.310	0.588	0.333	0.033

Overall survival

		Statistical based		Mutual information based			Model based							Others		
		Pearson	Spearman	MIM	MIFS	MRMR	uni-Cox	multi-Cox	RF-VH	RF-VI	RF-MD	PVI-RF	MSR-RFVI	Random	None	Aerts et al.
Non-/Semi-parametric	Cox	0.017	0.023	0.065	0.065	0.065	0.065	0.050	0.103	0.041	0.151	0.023	0.058	0.118	0.636	0.049
	NET-Cox	0.117	0.010	0.245	0.095	0.083	0.083	1.000	0.112	0.061	0.211	0.085	0.035	0.597	0.128	0.013
	BT-Cox	0.146	0.030	0.146	0.048	0.048	0.146	0.108	0.931	0.007	0.049	0.075	0.176	0.121	0.135	0.011
	BT-CIndex	0.095	0.027	0.088	0.048	0.088	0.088	0.108	0.049	0.024	0.063	0.075	0.176	0.182	0.041	0.020
	BGLM-Cox	0.083	0.030	0.083	0.082	0.083	0.083	0.050	0.050	0.041	0.142	0.058	0.058	0.015	0.108	0.013
	BGLM-CIndex	0.083	0.318	0.083	0.083	0.083	0.052	0.050	0.050	0.065	0.324	0.065	0.065	0.015	0.078	0.013
	RSF	0.350	0.159	0.138	0.056	0.050	0.111	0.050	0.107	0.135	0.070	0.289	0.101	0.216	0.096	0.006
	MSR-RF	0.192	0.005	0.025	0.005	0.014	0.083	0.748	0.135	0.317	0.017	0.113	0.184	0.042	0.073	0.466
Full-parametric	SRM	0.025	0.010	0.118	0.118	0.118	0.118	0.050	0.103	0.153	0.047	0.062	0.065	0.016	0.929	0.185
	BT-Weibull	0.042	0.013	0.146	0.049	0.049	0.146	0.108	0.668	0.120	0.034	0.069	0.034	0.317	0.195	0.018
	BGLM-Weibull	0.083	0.025	0.147	0.147	0.147	0.083	0.050	0.043	0.117	0.068	0.112	0.095	0.007	0.108	0.021

Figure C.2.: Resulting *p*-values of the log-rank tests for loco-regional tumour control (top) and overall survival (bottom) for the considered feature selection methods (columns) and learning algorithms (rows) as well as the signature by Aerts *et al.* (Aerts et al., 2014). The cut-off values used for stratification were based on the median predicted risk value determined on the exploratory cohort. Cut-off values were applied to the validation cohort unchanged.

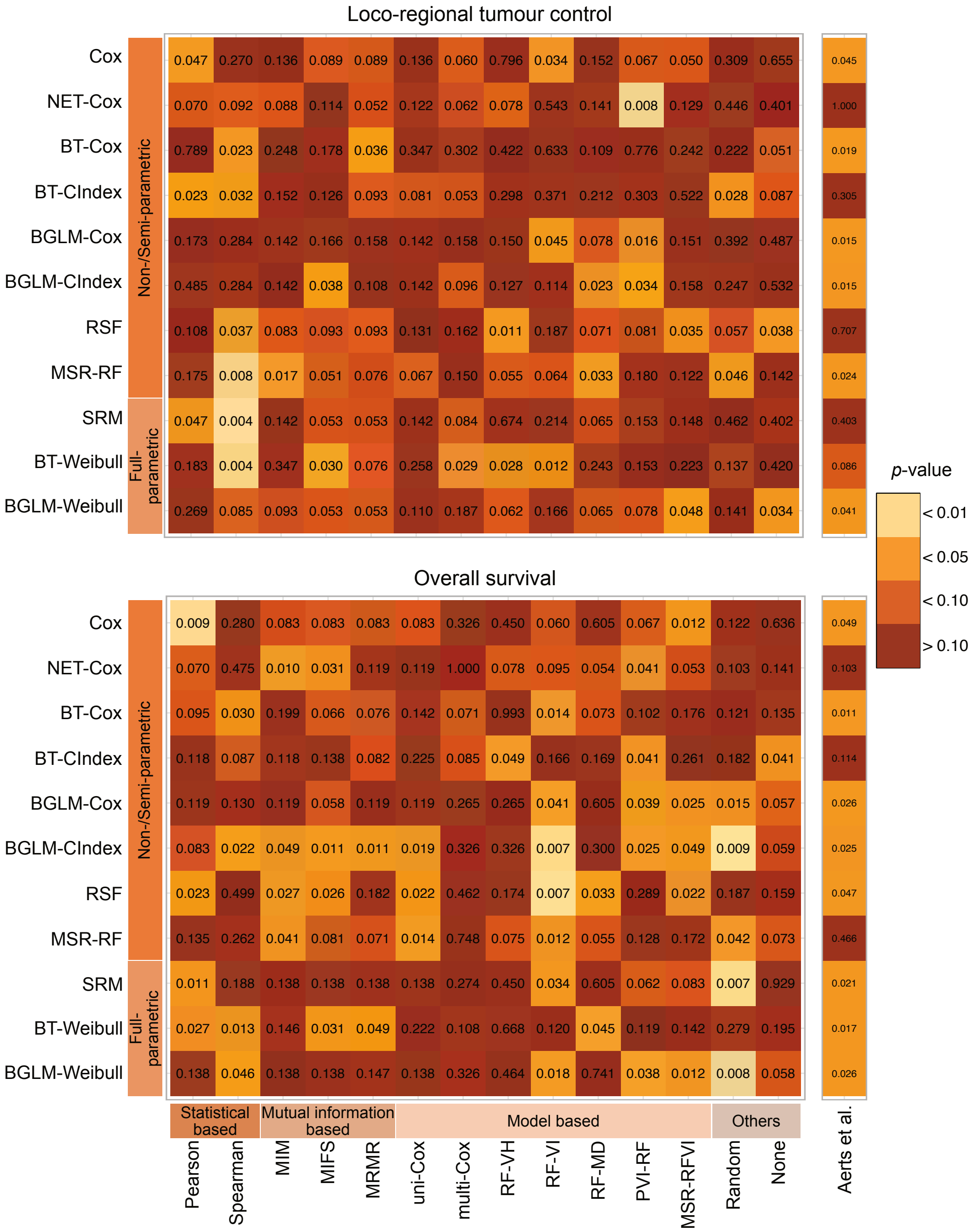


Figure C.3.: Resulting *p*-values of the log-rank tests for loco-regional tumour control (top) and overall survival (bottom) for the considered feature selection methods (columns) and learning algorithms (rows) as well as the signature by Aerts *et al.* (Aerts et al., 2014). The cut-off values used for stratification were selected by 1000 bootstrap samples based on the exploratory cohort. The fraction of significant stratification results was calculated for each cut-off, leading to the optimal value which has the largest power. Cut-off values calculated on the exploratory cohort were applied to the validation cohort unchanged.

D. Characterisation of tumour phenotype using computed tomograpy imaging during treatment

D.A. Multi-level model

The MLM consists of three levels (L) representing the different levels of effect. The top level L_0 describes the effect of the imaging time point as well as the effects of the feature selection method i and learning algorithm j:

$$\begin{aligned} y_{i,j} &= \alpha_{\text{method},i,j} + \beta_{\text{time}} x_{time} + \varepsilon_{\text{time}} \\ \varepsilon_{\text{time}} &\sim \mathcal{N}(0, \sigma^2_{\text{time}}), \end{aligned}$$

where y$_{i,j}$ defines the C-Index of the bootstrap sample. The term $\alpha_{\text{method},i,j}$ is an offset term, modelled separately in levels L_1 and L_2 and x$_{\text{time}}$ is a contrast variable, which is 0 for pre-treatment imaging and 1 for imaging in the second week. β_{time} is the effect of second week imaging compared to pre-treatment imaging, and has a weakly informative prior $\mathcal{N}(0, 1)$ and is limited to the range $[-1, 1]$. The error term $\varepsilon_{\text{time}}$ is modelled with a normal distribution with mean 0 and standard deviation $\sigma_{\text{time}} \in [0, 1]$. The level L_1 models the effect of learning algorithm (learner) j dependent on feature selection (fs) method i:

$$\begin{aligned} \alpha_{\text{method},ij} &= \alpha_{fs,i} + \beta_{\text{learner},j} + \varepsilon_{\text{learner},j} \\ \varepsilon_{\text{learner},j} &\sim \mathcal{N}(0, \sigma^2_{\text{learner},j}), \end{aligned}$$

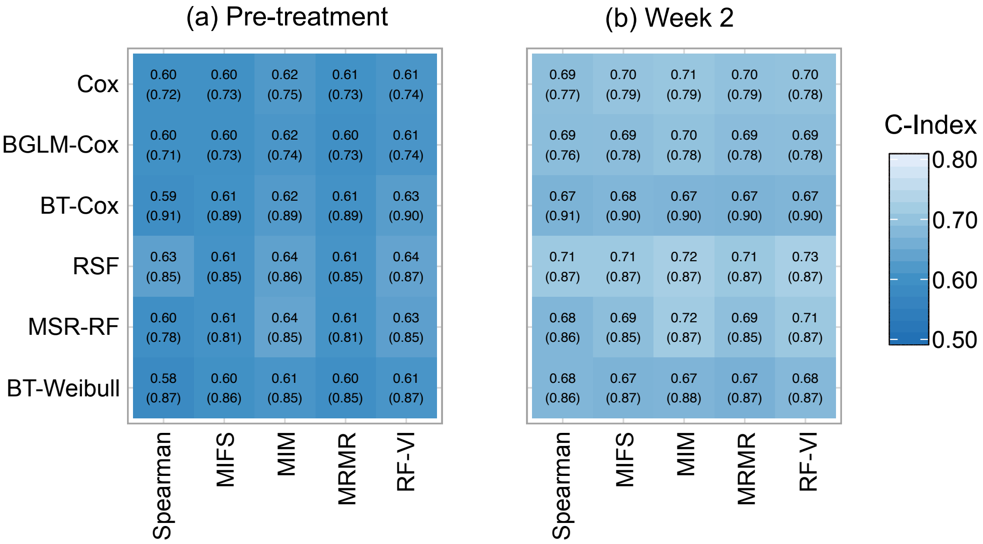
where $\alpha_{\text{fs},i}$ is an offset modelled separately in level L_2. $\beta_{\text{learner},j}$ defines the effect of the learning algorithm j. It has a weakly informative prior $\mathcal{N}(0, 1)$ and is limited to the range $[-1, 1]$. The error term $\varepsilon_{learner,j}$ is modelled with a normal distribution with mean 0 standard deviation $\sigma_{\text{learner},j}$, with $\sigma_{\text{learner},j} \in [0, 1]$. The level L_2 models the effect of feature selection method i:

$$\begin{aligned} \alpha_{\text{fs},i} &= \beta_{\text{fs},i} + \varepsilon_{\text{fs},i} \\ \varepsilon_{\text{fs},i} &\sim \mathcal{N}(0, \sigma^2_{\text{fs},i}), \end{aligned} \tag{D.1}$$

where $\beta_{\text{fs},i}$. $\beta_{\text{fs},i}$ has a weakly informative prior $\mathcal{N}(0.5, 1)$ and is limited to the range $[0, 1]$. The error term $\varepsilon_{\text{fs},i}$ is modelled with a normal distribution with mean 0 and standard deviation $\sigma_{\text{fs},i}$, with $\sigma_{\text{fs},i} \in [0, 1]$.

D.B. Prognostic performance

The results of the internal cross-validation experiments using the combined cohorts for the primary endpoint LRC are shown in figure D.1. The models were trained using a three-fold cross validation scheme with 33 repetitions. At each fold, feature selection and model training were performed 20 times. The C-Index is shown based on (a) pre-treatment and (b) week two CT scans for the internal validation cohorts and the internal training cohorts (brackets). On average the validation C-Index increased by 0.09 in week two ($\mathrm{CT}_{\mathrm{W0\text{-}FDG}}$: 0.61±0.01 and $\mathrm{CT}_{\mathrm{W2}}$: 0.70±0.05, respectively) (MLM: p=0.16).



(a) Pre-treatment

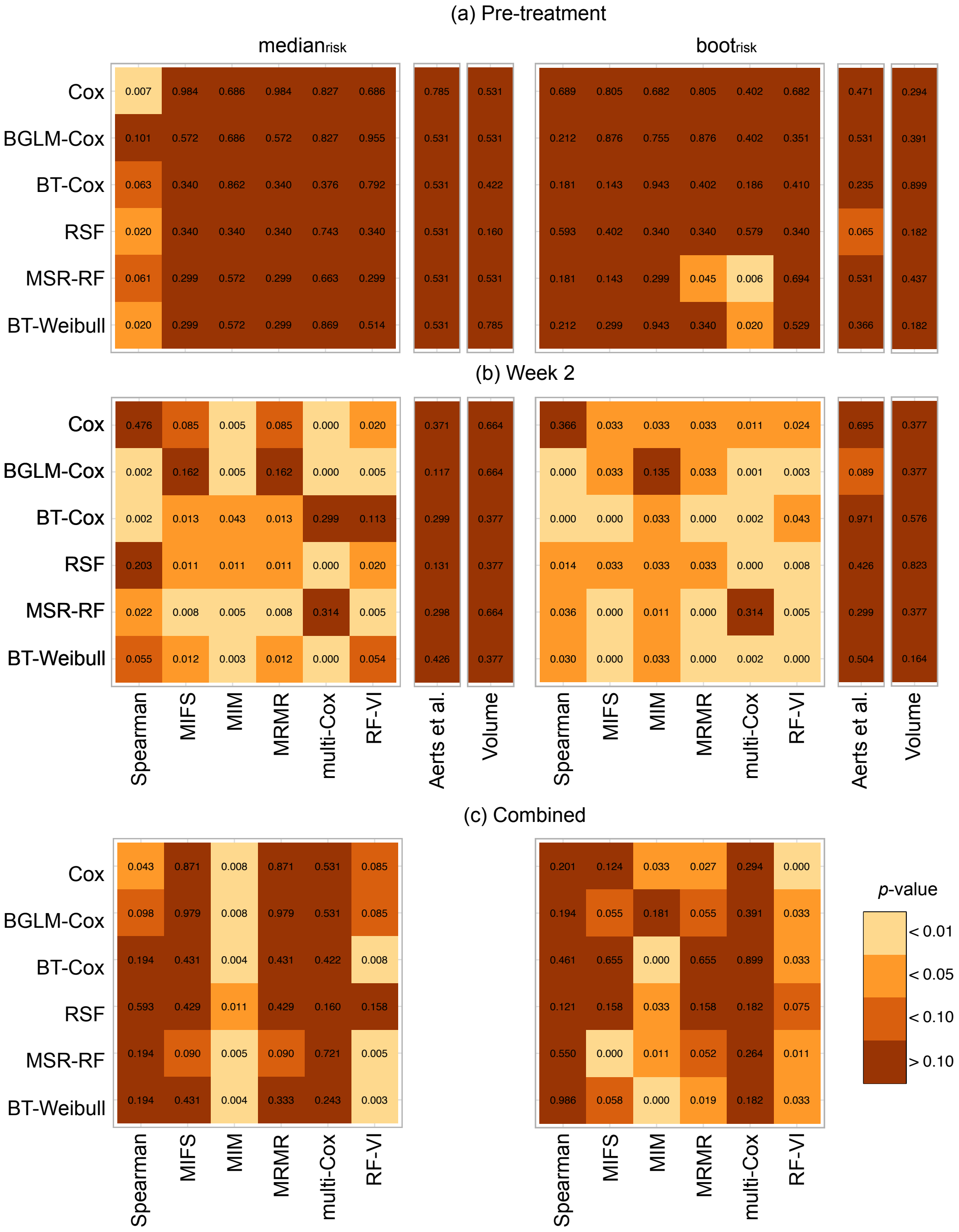
	Spearman	MIFS	MIM	MRMR	RF-VI
Cox	0.60 (0.72)	0.60 (0.73)	0.62 (0.75)	0.61 (0.73)	0.61 (0.74)
BGLM-Cox	0.60 (0.71)	0.60 (0.73)	0.62 (0.74)	0.60 (0.73)	0.61 (0.74)
BT-Cox	0.59 (0.91)	0.61 (0.89)	0.62 (0.89)	0.61 (0.89)	0.63 (0.90)
RSF	0.63 (0.85)	0.61 (0.85)	0.64 (0.86)	0.61 (0.85)	0.64 (0.87)
MSR-RF	0.60 (0.78)	0.61 (0.81)	0.64 (0.85)	0.61 (0.81)	0.63 (0.85)
BT-Weibull	0.58 (0.87)	0.60 (0.86)	0.61 (0.85)	0.60 (0.85)	0.61 (0.87)

(b) Week 2

	Spearman	MIFS	MIM	MRMR	RF-VI
Cox	0.69 (0.77)	0.70 (0.79)	0.71 (0.79)	0.70 (0.79)	0.70 (0.78)
BGLM-Cox	0.69 (0.76)	0.69 (0.78)	0.70 (0.78)	0.69 (0.78)	0.69 (0.78)
BT-Cox	0.67 (0.91)	0.68 (0.90)	0.67 (0.90)	0.67 (0.90)	0.67 (0.90)
RSF	0.71 (0.87)	0.71 (0.87)	0.72 (0.87)	0.71 (0.87)	0.73 (0.87)
MSR-RF	0.68 (0.86)	0.69 (0.85)	0.72 (0.87)	0.69 (0.85)	0.71 (0.87)
BT-Weibull	0.68 (0.86)	0.67 (0.87)	0.67 (0.88)	0.67 (0.87)	0.68 (0.87)

Figure D.1.: Concordance indices (C-Index) are shown for the feature selection methods (columns) and learning algorithms (rows) based on (a) pre-treatment and (b) week 2 CT scans for the internal validation cohorts and the internal training cohorts (in parentheses).

D.C. Risk-based patient stratification



(a) Pre-treatment

median$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.007	0.984	0.686	0.984	0.827	0.686	0.785	0.531
BGLM-Cox	0.101	0.572	0.686	0.572	0.827	0.955	0.531	0.531
BT-Cox	0.063	0.340	0.862	0.340	0.376	0.792	0.531	0.422
RSF	0.020	0.340	0.340	0.340	0.743	0.340	0.531	0.160
MSR-RF	0.061	0.299	0.572	0.299	0.663	0.299	0.531	0.531
BT-Weibull	0.020	0.299	0.572	0.299	0.869	0.514	0.531	0.785

boot$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.689	0.805	0.682	0.805	0.402	0.682	0.471	0.294
BGLM-Cox	0.212	0.876	0.755	0.876	0.402	0.351	0.531	0.391
BT-Cox	0.181	0.143	0.943	0.402	0.186	0.410	0.235	0.899
RSF	0.593	0.402	0.340	0.340	0.579	0.340	0.065	0.182
MSR-RF	0.181	0.143	0.299	0.045	0.006	0.694	0.531	0.437
BT-Weibull	0.212	0.299	0.943	0.340	0.020	0.529	0.366	0.182

(b) Week 2

median$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.476	0.085	0.005	0.085	0.000	0.020	0.371	0.664
BGLM-Cox	0.002	0.162	0.005	0.162	0.000	0.005	0.117	0.664
BT-Cox	0.002	0.013	0.043	0.013	0.299	0.113	0.299	0.377
RSF	0.203	0.011	0.011	0.011	0.000	0.020	0.131	0.377
MSR-RF	0.022	0.008	0.005	0.008	0.314	0.005	0.298	0.664
BT-Weibull	0.055	0.012	0.003	0.012	0.000	0.054	0.426	0.377

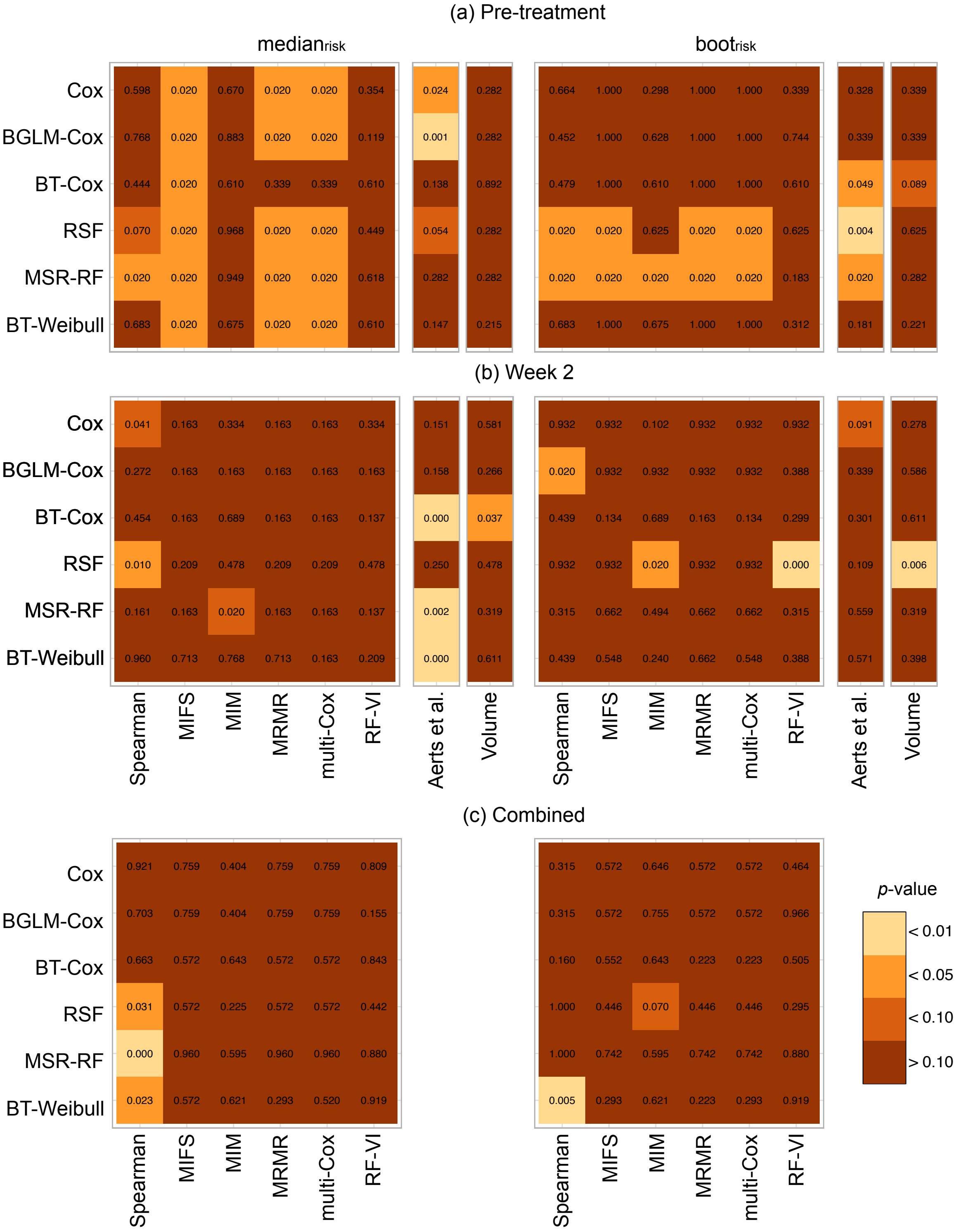
boot$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.366	0.033	0.033	0.033	0.011	0.024	0.695	0.377
BGLM-Cox	0.000	0.033	0.135	0.033	0.001	0.003	0.089	0.377
BT-Cox	0.000	0.000	0.033	0.000	0.002	0.043	0.971	0.576
RSF	0.014	0.033	0.033	0.033	0.000	0.008	0.426	0.823
MSR-RF	0.036	0.000	0.011	0.000	0.314	0.005	0.299	0.377
BT-Weibull	0.030	0.000	0.033	0.000	0.002	0.000	0.504	0.164

(c) Combined

median$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI
Cox	0.043	0.871	0.008	0.871	0.531	0.085
BGLM-Cox	0.098	0.979	0.008	0.979	0.531	0.085
BT-Cox	0.194	0.431	0.004	0.431	0.422	0.008
RSF	0.593	0.429	0.011	0.429	0.160	0.158
MSR-RF	0.194	0.090	0.005	0.090	0.721	0.005
BT-Weibull	0.194	0.431	0.004	0.333	0.243	0.003

boot$_{risk}$	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI
Cox	0.201	0.124	0.033	0.027	0.294	0.000
BGLM-Cox	0.194	0.055	0.181	0.055	0.391	0.033
BT-Cox	0.461	0.655	0.000	0.655	0.899	0.033
RSF	0.121	0.158	0.033	0.158	0.182	0.075
MSR-RF	0.550	0.000	0.011	0.052	0.264	0.011
BT-Weibull	0.986	0.058	0.000	0.019	0.182	0.033

Figure D.2.: Resulting *p*-values of the log-rank tests for loco-regional tumour control for the considered feature selection methods (columns) and learning algorithms (rows) based on (left) median (median$_{\text{risk}}$) and (right) bootstrapped (boot$_{\text{risk}}$) cut-off values using the predicted risk value. Furthermore, *p*-values of the log-rank tests for the signature by Aerts *et al.* (Aerts et al., 2014) and tumour volume based on the pre-treatment (CT$_{\text{W0-FDG}}$) and in-treatment (CT$_{\text{W2}}$) images are shown. The cut-off values used for stratification were determined on the exploratory cohort and applied to the validation cohort unchanged.



(a) Pre-treatment

	median$_{risk}$								boot$_{risk}$							
	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.598	0.020	0.670	0.020	0.020	0.354	0.024	0.282	0.664	1.000	0.298	1.000	1.000	0.339	0.328	0.339
BGLM-Cox	0.768	0.020	0.883	0.020	0.020	0.119	0.001	0.282	0.452	1.000	0.628	1.000	1.000	0.744	0.339	0.339
BT-Cox	0.444	0.020	0.610	0.339	0.339	0.610	0.138	0.892	0.479	1.000	0.610	1.000	1.000	0.610	0.049	0.089
RSF	0.070	0.020	0.968	0.020	0.020	0.449	0.054	0.282	0.020	0.020	0.625	0.020	0.020	0.625	0.004	0.625
MSR-RF	0.020	0.020	0.949	0.020	0.020	0.618	0.282	0.282	0.020	0.020	0.020	0.020	0.020	0.183	0.020	0.282
BT-Weibull	0.683	0.020	0.675	0.020	0.020	0.610	0.147	0.215	0.683	1.000	0.675	1.000	1.000	0.312	0.181	0.221

(b) Week 2

	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Aerts et al.	Volume
Cox	0.041	0.163	0.334	0.163	0.163	0.334	0.151	0.581	0.932	0.932	0.102	0.932	0.932	0.932	0.091	0.278
BGLM-Cox	0.272	0.163	0.163	0.163	0.163	0.163	0.158	0.266	0.020	0.932	0.932	0.932	0.932	0.388	0.339	0.586
BT-Cox	0.454	0.163	0.689	0.163	0.163	0.137	0.000	0.037	0.439	0.134	0.689	0.163	0.134	0.299	0.301	0.611
RSF	0.010	0.209	0.478	0.209	0.209	0.478	0.250	0.478	0.932	0.932	0.020	0.932	0.932	0.000	0.109	0.006
MSR-RF	0.161	0.163	0.020	0.163	0.163	0.137	0.002	0.319	0.315	0.662	0.494	0.662	0.662	0.315	0.559	0.319
BT-Weibull	0.960	0.713	0.768	0.713	0.163	0.209	0.000	0.611	0.439	0.548	0.240	0.662	0.548	0.388	0.571	0.398

(c) Combined

	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI	Spearman	MIFS	MIM	MRMR	multi-Cox	RF-VI
Cox	0.921	0.759	0.404	0.759	0.759	0.809	0.315	0.572	0.646	0.572	0.572	0.464
BGLM-Cox	0.703	0.759	0.404	0.759	0.759	0.155	0.315	0.572	0.755	0.572	0.572	0.966
BT-Cox	0.663	0.572	0.643	0.572	0.572	0.843	0.160	0.552	0.643	0.223	0.223	0.505
RSF	0.031	0.572	0.225	0.572	0.572	0.442	1.000	0.446	0.070	0.446	0.446	0.295
MSR-RF	0.000	0.960	0.595	0.960	0.960	0.880	1.000	0.742	0.595	0.742	0.742	0.880
BT-Weibull	0.023	0.572	0.621	0.293	0.520	0.919	0.005	0.293	0.621	0.223	0.293	0.919

Figure D.3.: Resulting *p*-values of the log-rank tests for overall survival for the considered feature selection methods (columns) and learning algorithms (rows) based on (left) median (median$_{\text{risk}}$) and (right) bootstrapped (boot$_{\text{risk}}$) cut-off values using the predicted risk value. Furthermore, *p*-values of the log-rank tests for the signature by Aerts *et al.* (Aerts et al., 2014) and tumour volume based on the pre-treatment (CT$_{\text{W0-FDG}}$) and in-treatment (CT$_{\text{W2}}$) scans are shown. The cut-off values used for stratification were determined on the exploratory cohort and applied to the validation cohort unchanged.

D.D. Developed radiomics signatures

Table D.1.: Radiomics signatures for loco-regional tumour control (LRC) based on Spearman feature selection and for overall survival (OS) based on mutual information maximisation (MIM) feature selection for different time points. The mathematical description and the abbreviations of features can be found in Zwanenburg *et al.* (Zwanenburg et al., 2016).

Endpoint	Time point	Feature name	Synonym	Image	Cluster
LRC	Pre-treatment	stat_mean	$F1_{\mathrm{S}}$	wav_coif1_lhl	Yes
		dzm_lde, dzm_zd_var	$\overline{F2}_{\mathrm{T}}$	wav_coif1_lhl	No
		ngl_dcnu, szm_lzlge	$F3_{\mathrm{T}}$	wav_coif1_lhh	No
		morph_pca_flatness	$F4_{\mathrm{M}}$	Base	Yes
	Week 2	cm_inv_diff, cm_inv_diff_mom, rlm_lre, rlm_r_perc, rlm_rl_var, rlm_rlnu_norm, rlm_sre	$\overline{F1}_{\mathrm{T}}$	wav_coif1_lll	Yes
		cm_info_corr2	$F2_{\mathrm{T}}$	wav_coif1_hll	No
		morph_pca_flatness	$F3_{\mathrm{M}}$	Base	No
		ngt_busyness, ngt_coarseness, rlm_glnu_3d	$\overline{F4}_{\mathrm{T}}$	wav_coif1_hlh	Yes
	Combined	ih_max_grad_delta, ih_min_grad_delta, ngl_glnu_delta	$\overline{\Delta F1}_{\mathrm{T}}$	wav_coif1_lhh	Yes
		dzm_zdnu_delta	$\Delta F2_{\mathrm{T}}$	wav_coif1_lhh	No
		rlm_glnu_delta	$\Delta F3_{\mathrm{T}}$	wav_coif1_lhh	No
		ngt_contrast_delta	$\Delta F4_{\mathrm{T}}$	wav_coif1_llh	No
		dzm_glnu_delta, szm_glnu_delta	$\overline{\Delta F5}_{\mathrm{T}}$	wav_coif1_llh	No
OS	Pre-treatment	cm_contrast, cm_diff_avg, cm_diff_entr, cm_diff_var, cm_dissimilarity, cm_inv_diff_mom_norm, cm_inv_diff_norm, dzm_z_perc, ngl_lde, ngl_ldhge, ngt_contrast, rlm_rlnu_norm, rlm_sre, szm_z_perc	$\overline{F1}_{\mathrm{T}}$	wav_coif1_lll	Yes
		morph_pca_min_axis	$F2_{\mathrm{M}}$	Base	No
		cm_inv_diff, cm_inv_diff_mom, rlm_lre, rlm_r_perc, rlm_rl_var	$\overline{F3}_{\mathrm{T}}$	wav_coif1_lll	No
		ngt_strength, rlm_rlnu	$\overline{F4}_{\mathrm{T}}$	wav_coif1_llh	Yes
		morph_area_dens_conv_hull	$F5_{\mathrm{M}}$	Base	Yes
	Week 2	dzm_glnu, szm_glnu, dzm_glnu, szm_glnu	$\overline{F1}_{\mathrm{T}}$	wav_coif1_lhh wav_coif1_lhl	Yes

Endpoint	Time point	Feature name	Synonym	Image	Cluster
		dzm_glnu, rlm_rlnu, szm_glnu	$\overline{F2}_{\mathrm{T}}$	LoG	Yes
		morph_area_mesh, morph_pca_least_axis, morph_vol_approx, morph_volume, ngl_dcnu, ngl_glnu, ngt_busyness, ngt_coarseness, ngt_strength, rlm_glnu, rlm_rlnu, szm_glnu, szm_lze, szm_lzhge, szm_lzlge, szm_zs_var, dzm_glnu, ih_max, ih_min	$\overline{F3}_{\mathrm{T}}$	Base	Yes
		ih_max, ih_min, ngl_dcnu, ngl_glnu, ngt_busyness, ngt_coarseness, ngt_strength, rlm_glnu, szm_lze, szm_lzhge, szm_lzlge, szm_zs_var		LoG	
		dzm_glnu, ih_max, ih_min, ngl_dcnu, ngl_glnu, rlm_glnu, szm_glnu		wav_coif1_hhh	
		ngl_dcnu, ngt_busyness, ngt_coarseness, rlm_glnu		wav_coif1_hhl	
		ih_max, ih_min, ngl_dcnu, ngl_glnu		wav_coif1_hlh	
		ngl_dcnu, ngt_busyness, ngt_coarseness, rlm_glnu		wav_coif1_hll	
		ih_max, ih_min, ngl_dcnu, ngl_glnu, szm_lze, szm_zs_var		wav_coif1_lhh	
		ih_max, ih_min, ngl_dcnu, ngl_glnu, szm_lze, szm_zs_var		wav_coif1_lhl	

Endpoint	Time point	Feature name	Synonym	Image	Cluster
		dzm_glnu, ih_max, ih_min, ngl_dcnu, ngl_glnu, ngt_coarseness, rlm_glnu, szm_glnu, szm_lze, szm_zs_var		wav_coif1_llh	
		ih_max, ih_min, ngl_dcnu, ngl_glnu, szm_lze, szm_zs_var		wav_coif1_lll	
		cm_inv_diff, cm_inv_diff_mom, rlm_lre, rlm_r_perc, rlm_rl_var, rlm_rlnu_norm, rlm_sre	$\overline{\Delta F4}_{\mathrm{T}}$	wav_coif1_lll	Yes
		morph_pca_min_axis	$F5_{\mathrm{M}}$	Base	No
		morph_area_dens_conv_hull	$F6_{\mathrm{M}}$	Base	No
	Combined	dzm_glnu_delta, szm_glnu_delta	$\overline{\Delta F1}_{\mathrm{T}}$	wav_coif1_lhl	Yes
		ngt_complexity_delta, rlm_lre_3d_delta, rlm_r_perc_delta, rlm_rl_var_delta	$\overline{\Delta F2}_{\mathrm{T}}$	LoG	Yes
		dzm_glnu_2W, szm_glnu_2W, dzm_glnu_2W, szm_glnu_2W	$\overline{F3}_{\mathrm{T}}$	wav_coif1_lhh	Yes
		dzm_glnu_delta, rlm_rlnu_delta, szm_glnu_delta	$\overline{F4}_{\mathrm{T}}$	LoG	Yes
		ih_max_grad_2W, ih_min_grad_2W, ngl_glnu_2W, szm_lze_2W, szm_lzhge_2W, szm_lzlge_2W, szm_zs_var_2W	$\overline{F5}_{\mathrm{T}}$	Base	Yes
		ih_max_grad_2W, ih_min_grad_2W, ngl_glnu_2W, szm_lze_2W, szm_lzhge_2W, szm_lzlge_2W, szm_zs_var_2W		wav_coif1_lll	

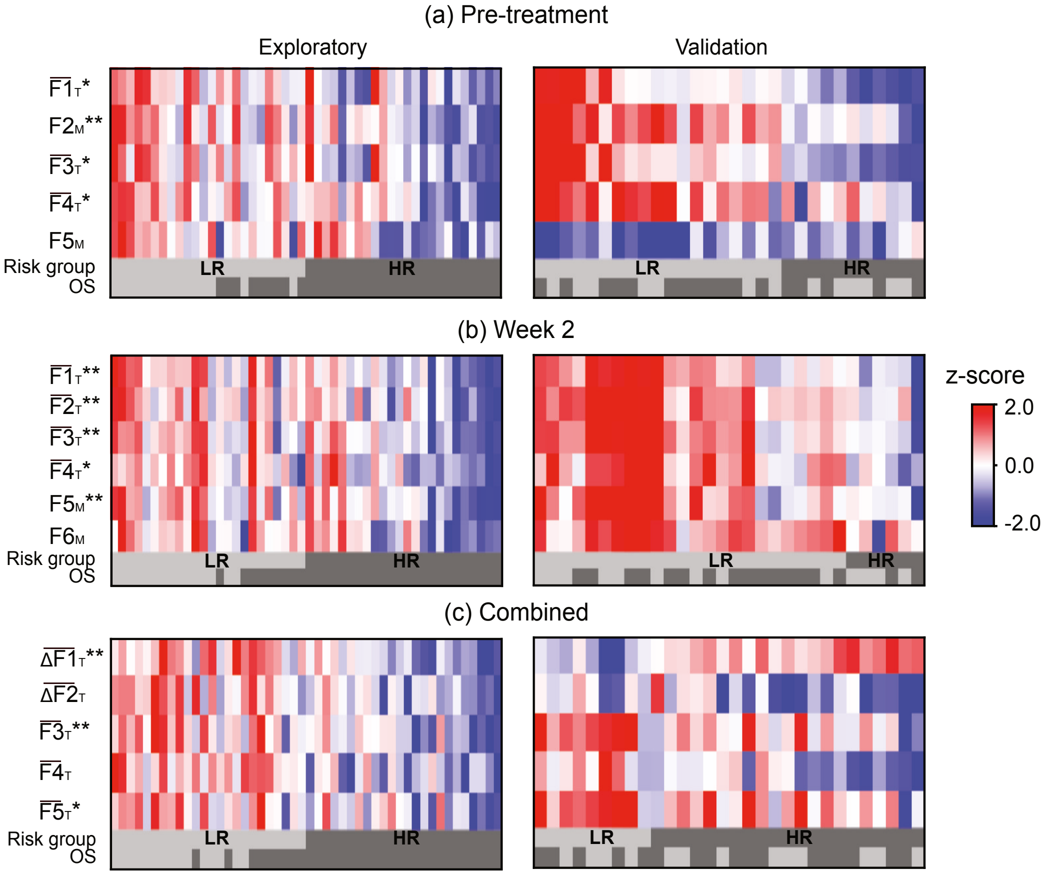
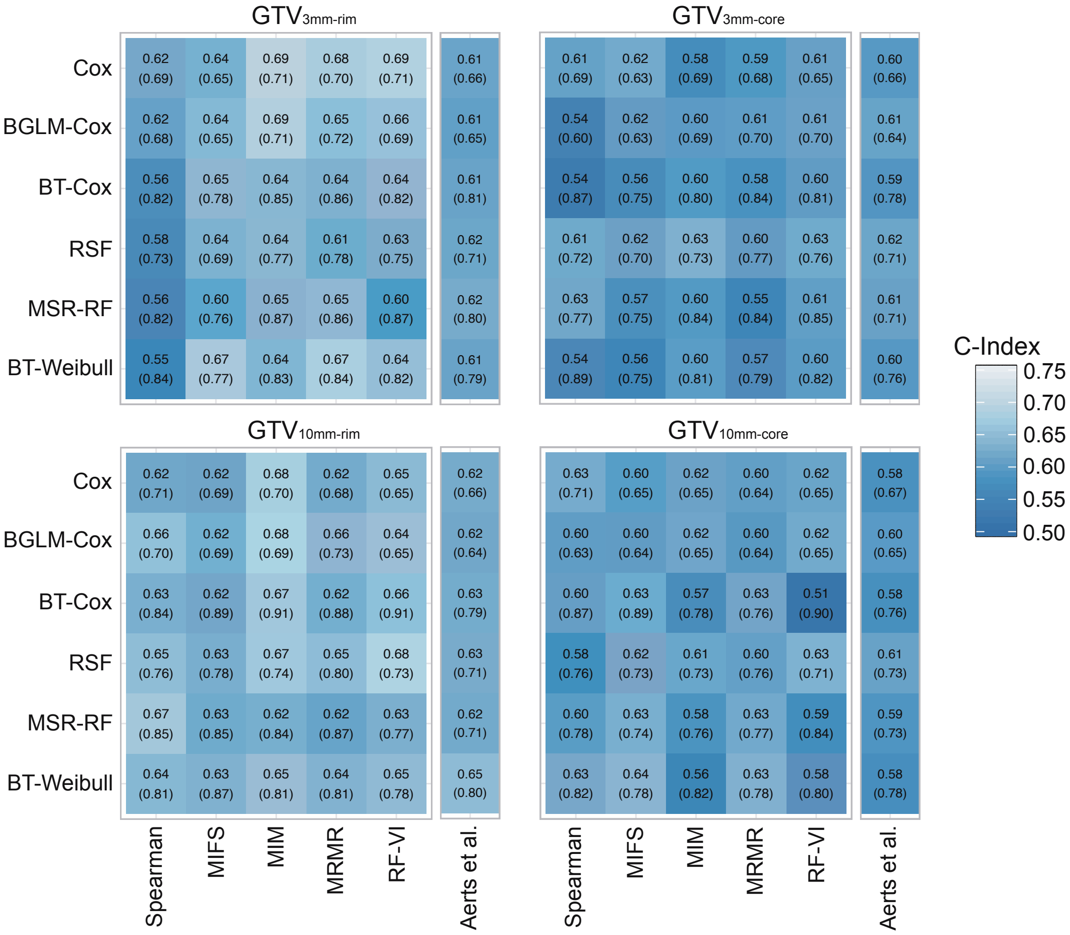


Figure D.4.: Feature expressions of developed signatures for the representative models, boosted tree Cox model in combination with the mutual information maximisation feature selection method, trained on the $CT_{W0\text{-}FDG}$, CT_{W2} and combined feature set. Overall survival (OS) during follow-up (yes, light; no, dark) and features with a significant correlation with OS are shown (*p<0.05 and **p<0.001). A detailed description of the feature abbreviations can be found in appendix D.1. Abbreviations: $\overline{F}$ cluster feature consisting of several features represented by the mean value as a new meta-feature, F_S first order statistical feature, F_M morphological feature, F_T texture feature, Δ F delta feature.

E. Tumour phenotype characterisation using tumour sub-volumes

E.A. Prognostic performance



GTV$_{3mm\text{-}rim}$

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.62 (0.69)	0.64 (0.65)	0.69 (0.71)	0.68 (0.70)	0.69 (0.71)	0.61 (0.66)
BGLM-Cox	0.62 (0.68)	0.64 (0.65)	0.69 (0.71)	0.65 (0.72)	0.66 (0.69)	0.61 (0.65)
BT-Cox	0.56 (0.82)	0.65 (0.78)	0.64 (0.85)	0.64 (0.86)	0.64 (0.82)	0.61 (0.81)
RSF	0.58 (0.73)	0.64 (0.69)	0.64 (0.77)	0.61 (0.78)	0.63 (0.75)	0.62 (0.71)
MSR-RF	0.56 (0.82)	0.60 (0.76)	0.65 (0.87)	0.65 (0.86)	0.60 (0.87)	0.62 (0.80)
BT-Weibull	0.55 (0.84)	0.67 (0.77)	0.64 (0.83)	0.67 (0.84)	0.64 (0.82)	0.61 (0.79)

GTV$_{3mm\text{-}core}$

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.61 (0.69)	0.62 (0.63)	0.58 (0.69)	0.59 (0.68)	0.61 (0.65)	0.60 (0.66)
BGLM-Cox	0.54 (0.60)	0.62 (0.63)	0.60 (0.69)	0.61 (0.70)	0.61 (0.70)	0.61 (0.64)
BT-Cox	0.54 (0.87)	0.56 (0.75)	0.60 (0.80)	0.58 (0.84)	0.60 (0.81)	0.59 (0.78)
RSF	0.61 (0.72)	0.62 (0.70)	0.63 (0.73)	0.60 (0.77)	0.63 (0.76)	0.62 (0.71)
MSR-RF	0.63 (0.77)	0.57 (0.75)	0.60 (0.84)	0.55 (0.84)	0.61 (0.85)	0.61 (0.71)
BT-Weibull	0.54 (0.89)	0.56 (0.75)	0.60 (0.81)	0.57 (0.79)	0.60 (0.82)	0.60 (0.76)

GTV$_{10mm\text{-}rim}$

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.62 (0.71)	0.62 (0.69)	0.68 (0.70)	0.62 (0.68)	0.65 (0.65)	0.62 (0.66)
BGLM-Cox	0.66 (0.70)	0.62 (0.69)	0.68 (0.69)	0.66 (0.73)	0.64 (0.65)	0.62 (0.64)
BT-Cox	0.63 (0.84)	0.62 (0.89)	0.67 (0.91)	0.62 (0.88)	0.66 (0.91)	0.63 (0.79)
RSF	0.65 (0.76)	0.63 (0.78)	0.67 (0.74)	0.65 (0.80)	0.68 (0.73)	0.63 (0.71)
MSR-RF	0.67 (0.85)	0.63 (0.85)	0.62 (0.84)	0.62 (0.87)	0.63 (0.77)	0.62 (0.71)
BT-Weibull	0.64 (0.81)	0.63 (0.87)	0.65 (0.81)	0.64 (0.81)	0.65 (0.78)	0.65 (0.80)

GTV$_{10mm\text{-}core}$

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.63 (0.71)	0.60 (0.65)	0.62 (0.65)	0.60 (0.64)	0.62 (0.65)	0.58 (0.67)
BGLM-Cox	0.60 (0.63)	0.60 (0.64)	0.62 (0.65)	0.60 (0.64)	0.62 (0.65)	0.60 (0.65)
BT-Cox	0.60 (0.87)	0.63 (0.89)	0.57 (0.78)	0.63 (0.76)	0.51 (0.90)	0.58 (0.76)
RSF	0.58 (0.76)	0.62 (0.73)	0.61 (0.73)	0.60 (0.76)	0.63 (0.71)	0.61 (0.73)
MSR-RF	0.60 (0.78)	0.63 (0.74)	0.58 (0.76)	0.63 (0.77)	0.59 (0.84)	0.59 (0.73)
BT-Weibull	0.63 (0.82)	0.64 (0.78)	0.56 (0.82)	0.63 (0.78)	0.58 (0.80)	0.58 (0.78)

Figure E.1.: Concordance indices for the feature selection methods (columns) and the learning algorithms (rows) trained on the GTV$_{3mm\text{-}rim}$ and the GTV$_{10mm\text{-}rim}$ as well as the corresponding tumour core sub-volumes on the exploratory (in parentheses) and the validation cohort. Furthermore, the performance of the models using the Aerts *et al.* (Aerts et al., 2014) signature are shown.

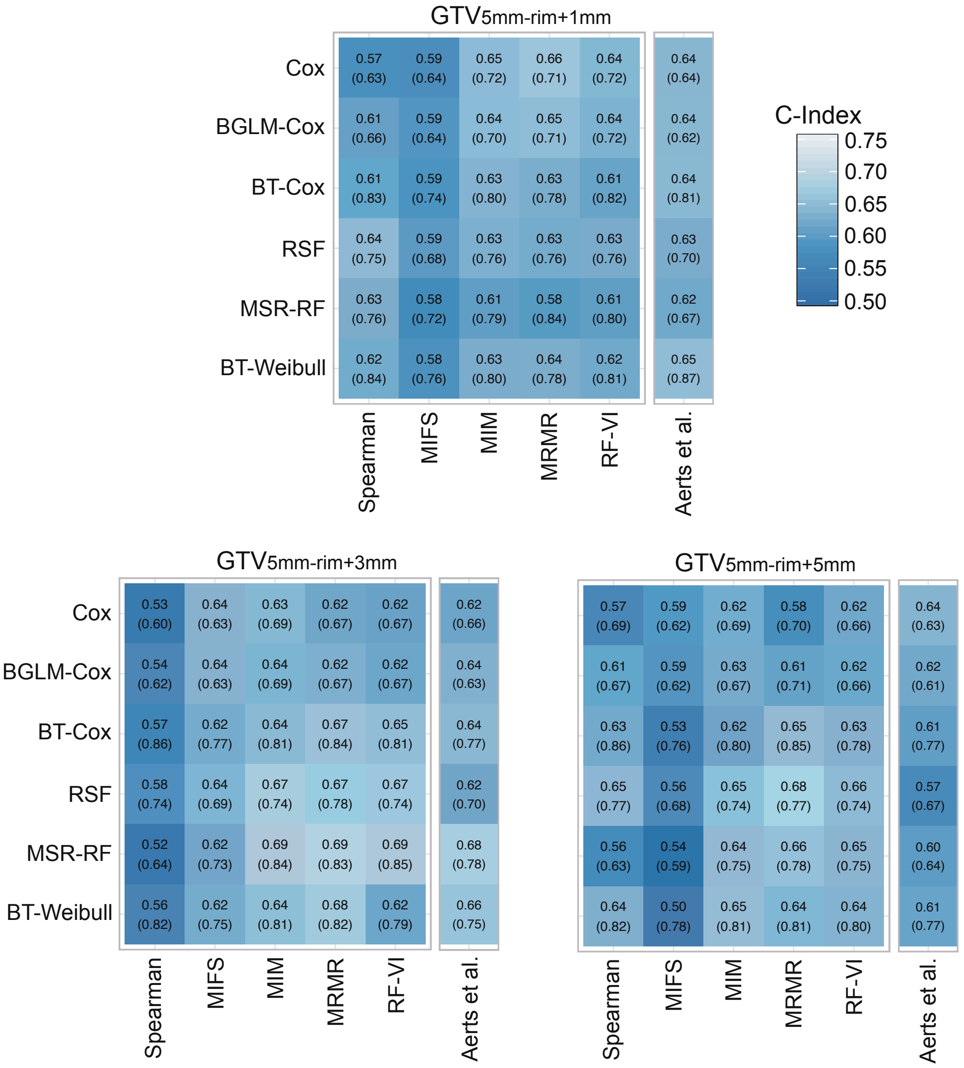
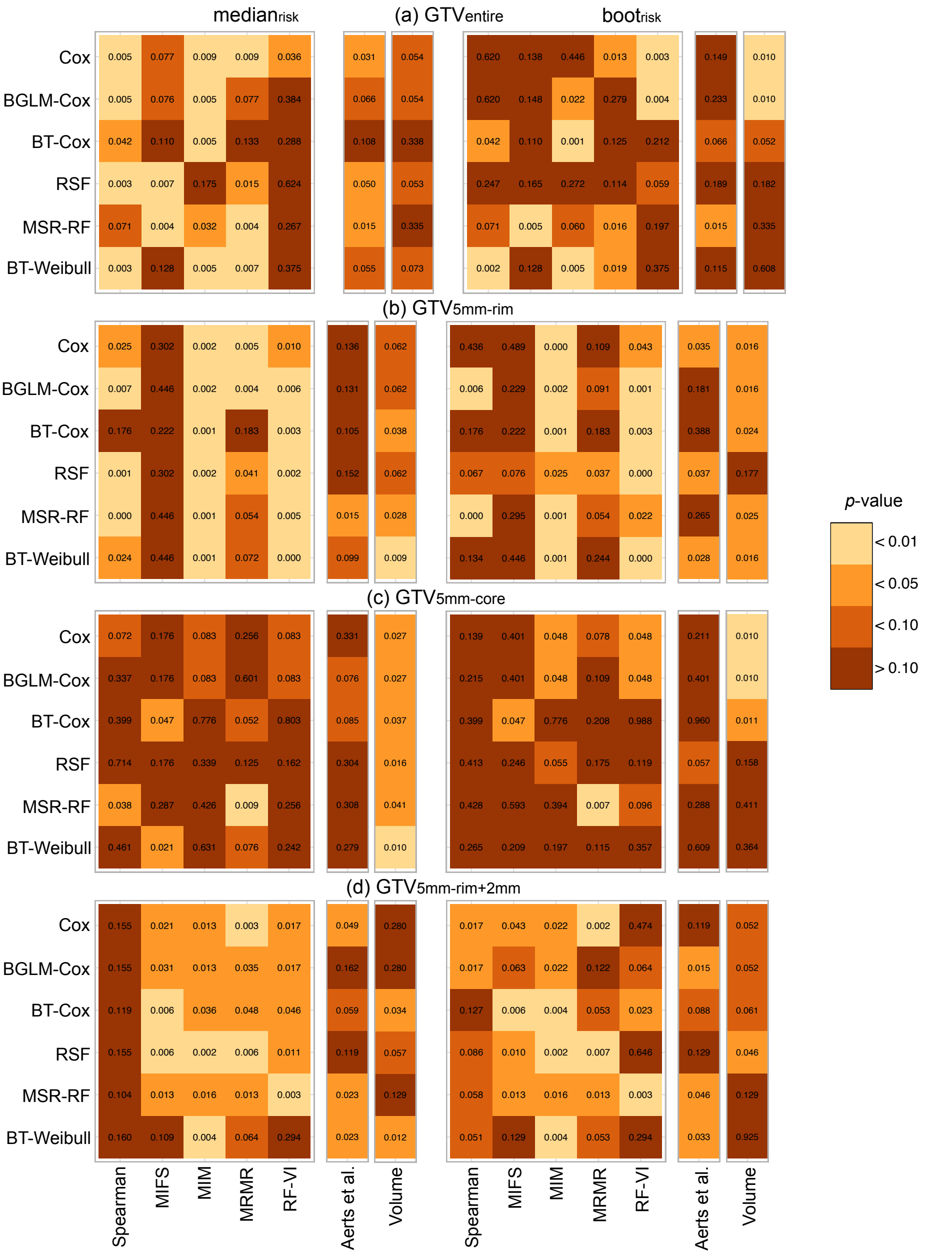


Figure E.2.: Concordance indices for the feature selection methods (columns) and the learning algorithms (rows) trained on the $\text{GTV}_{\text{5mm-rim+1mm}}$, the $\text{GTV}_{\text{5mm-rim+3mm}}$ and the $\text{GTV}_{\text{5mm-rim+5mm}}$ tumour sub-volumes on the exploratory (in parentheses) and the validation cohort. Furthermore, the performance of the models using the Aerts *et al.* (Aerts et al., 2014) signature are shown.

E.B. Risk-based patient stratification



(a) GTVentire

	medianrisk							bootrisk						
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume
Cox	0.005	0.077	0.009	0.009	0.036	0.031	0.054	0.620	0.138	0.446	0.013	0.003	0.149	0.010
BGLM-Cox	0.005	0.076	0.005	0.077	0.384	0.066	0.054	0.620	0.148	0.022	0.279	0.004	0.233	0.010
BT-Cox	0.042	0.110	0.005	0.133	0.288	0.108	0.338	0.042	0.110	0.001	0.125	0.212	0.066	0.052
RSF	0.003	0.007	0.175	0.015	0.624	0.050	0.053	0.247	0.165	0.272	0.114	0.059	0.189	0.182
MSR-RF	0.071	0.004	0.032	0.004	0.267	0.015	0.335	0.071	0.005	0.060	0.016	0.197	0.015	0.335
BT-Weibull	0.003	0.128	0.005	0.007	0.375	0.055	0.073	0.002	0.128	0.005	0.019	0.375	0.115	0.608

(b) GTV5mm-rim

	medianrisk							bootrisk						
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume
Cox	0.025	0.302	0.002	0.005	0.010	0.136	0.062	0.436	0.489	0.000	0.109	0.043	0.035	0.016
BGLM-Cox	0.007	0.446	0.002	0.004	0.006	0.131	0.062	0.006	0.229	0.002	0.091	0.001	0.181	0.016
BT-Cox	0.176	0.222	0.001	0.183	0.003	0.105	0.038	0.176	0.222	0.001	0.183	0.003	0.388	0.024
RSF	0.001	0.302	0.002	0.041	0.002	0.152	0.062	0.067	0.076	0.025	0.037	0.000	0.037	0.177
MSR-RF	0.000	0.446	0.001	0.054	0.005	0.015	0.028	0.000	0.295	0.001	0.054	0.022	0.265	0.025
BT-Weibull	0.024	0.446	0.001	0.072	0.000	0.099	0.009	0.134	0.446	0.001	0.244	0.000	0.028	0.016

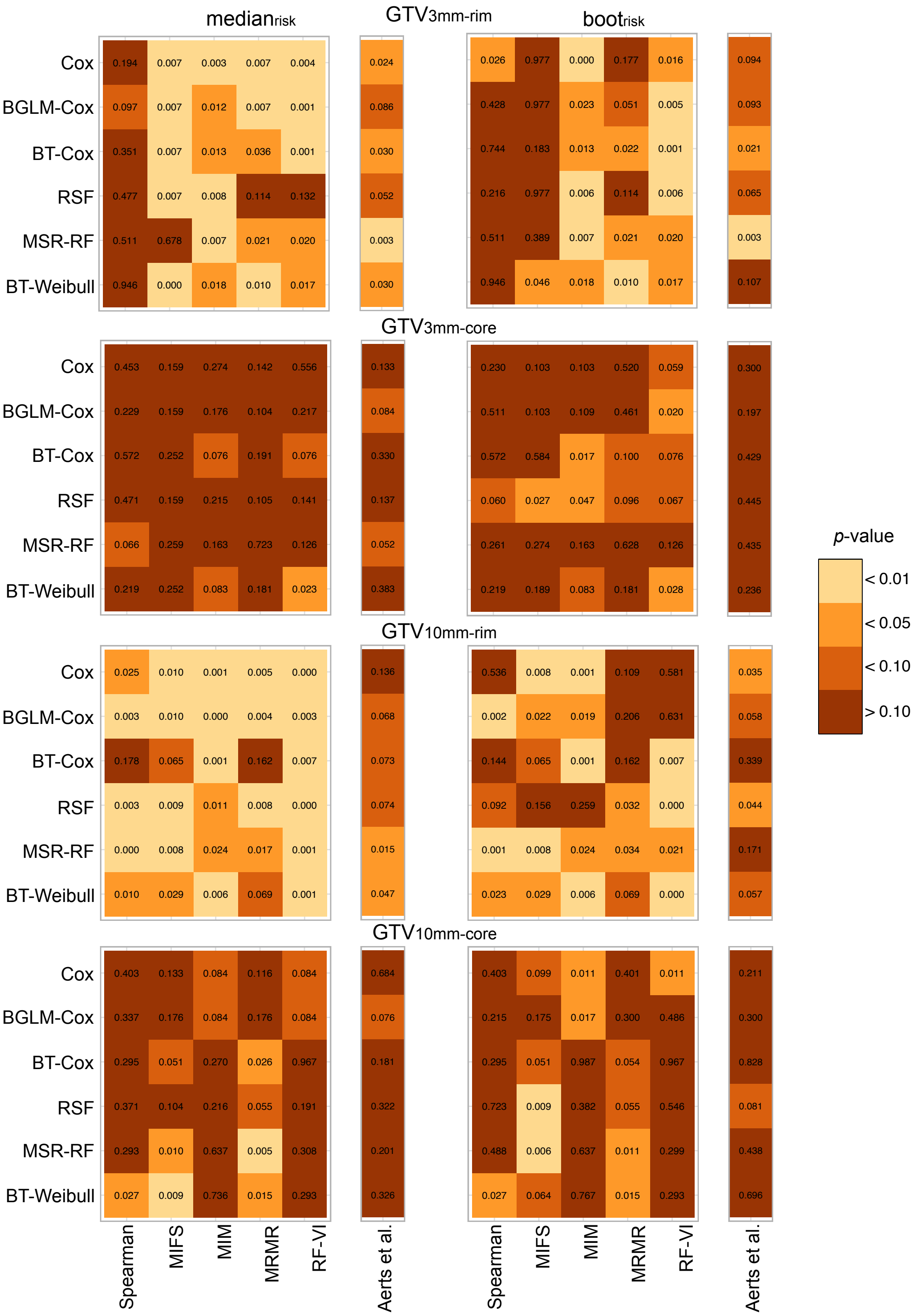
(c) GTV5mm-core

	medianrisk							bootrisk						
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume
Cox	0.072	0.176	0.083	0.256	0.083	0.331	0.027	0.139	0.401	0.048	0.078	0.048	0.211	0.010
BGLM-Cox	0.337	0.176	0.083	0.601	0.083	0.076	0.027	0.215	0.401	0.048	0.109	0.048	0.401	0.010
BT-Cox	0.399	0.047	0.776	0.052	0.803	0.085	0.037	0.399	0.047	0.776	0.208	0.988	0.960	0.011
RSF	0.714	0.176	0.339	0.125	0.162	0.304	0.016	0.413	0.246	0.055	0.175	0.119	0.057	0.158
MSR-RF	0.038	0.287	0.426	0.009	0.256	0.308	0.041	0.428	0.593	0.394	0.007	0.096	0.288	0.411
BT-Weibull	0.461	0.021	0.631	0.076	0.242	0.279	0.010	0.265	0.209	0.197	0.115	0.357	0.609	0.364

(d) GTV5mm-rim+2mm

	medianrisk							bootrisk						
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Volume
Cox	0.155	0.021	0.013	0.003	0.017	0.049	0.280	0.017	0.043	0.022	0.002	0.474	0.119	0.052
BGLM-Cox	0.155	0.031	0.013	0.035	0.017	0.162	0.280	0.017	0.063	0.022	0.122	0.064	0.015	0.052
BT-Cox	0.119	0.006	0.036	0.048	0.046	0.059	0.034	0.127	0.006	0.004	0.053	0.023	0.088	0.061
RSF	0.155	0.006	0.002	0.006	0.011	0.119	0.057	0.086	0.010	0.002	0.007	0.646	0.129	0.046
MSR-RF	0.104	0.013	0.016	0.013	0.003	0.023	0.129	0.058	0.013	0.016	0.013	0.003	0.046	0.129
BT-Weibull	0.160	0.109	0.004	0.064	0.294	0.023	0.012	0.051	0.129	0.004	0.053	0.294	0.033	0.925

Figure E.3.: Resulting *p*-values of the log-rank tests for loco-regional tumour control for the feature selection methods (columns) and the learning algorithms (rows) trained on the $\mathrm{GTV}_{\mathrm{entire}}$, the $\mathrm{GTV}_{\mathrm{5mm\text{-}rim}}$, the corresponding $\mathrm{GTV}_{\mathrm{5mm\text{-}core}}$ and the extended $\mathrm{GTV}_{\mathrm{5mm\text{-}rim+2mm}}$ sub-volumes based on median ($\mathrm{median}_{\mathrm{risk}}$) (left) and bootstrapped ($\mathrm{boot}_{\mathrm{risk}}$) cut-off values (right) using the predicted risk values. The cut-off values used for stratification were determined on the exploratory cohort and applied to the validation cohort unchanged. Furthermore, the results of the log-rank tests for the models using the Aerts' signature (Aerts et al., 2014) and the volume parameter are shown.



GTV3mm-rim

	median$_{risk}$						boot$_{risk}$					
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.194	0.007	0.003	0.007	0.004	0.024	0.026	0.977	0.000	0.177	0.016	0.094
BGLM-Cox	0.097	0.007	0.012	0.007	0.001	0.086	0.428	0.977	0.023	0.051	0.005	0.093
BT-Cox	0.351	0.007	0.013	0.036	0.001	0.030	0.744	0.183	0.013	0.022	0.001	0.021
RSF	0.477	0.007	0.008	0.114	0.132	0.052	0.216	0.977	0.006	0.114	0.006	0.065
MSR-RF	0.511	0.678	0.007	0.021	0.020	0.003	0.511	0.389	0.007	0.021	0.020	0.003
BT-Weibull	0.946	0.000	0.018	0.010	0.017	0.030	0.946	0.046	0.018	0.010	0.017	0.107

GTV3mm-core

	median$_{risk}$						boot$_{risk}$					
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.453	0.159	0.274	0.142	0.556	0.133	0.230	0.103	0.103	0.520	0.059	0.300
BGLM-Cox	0.229	0.159	0.176	0.104	0.217	0.084	0.511	0.103	0.109	0.461	0.020	0.197
BT-Cox	0.572	0.252	0.076	0.191	0.076	0.330	0.572	0.584	0.017	0.100	0.076	0.429
RSF	0.471	0.159	0.215	0.105	0.141	0.137	0.060	0.027	0.047	0.096	0.067	0.445
MSR-RF	0.066	0.259	0.163	0.723	0.126	0.052	0.261	0.274	0.163	0.628	0.126	0.435
BT-Weibull	0.219	0.252	0.083	0.181	0.023	0.383	0.219	0.189	0.083	0.181	0.028	0.236

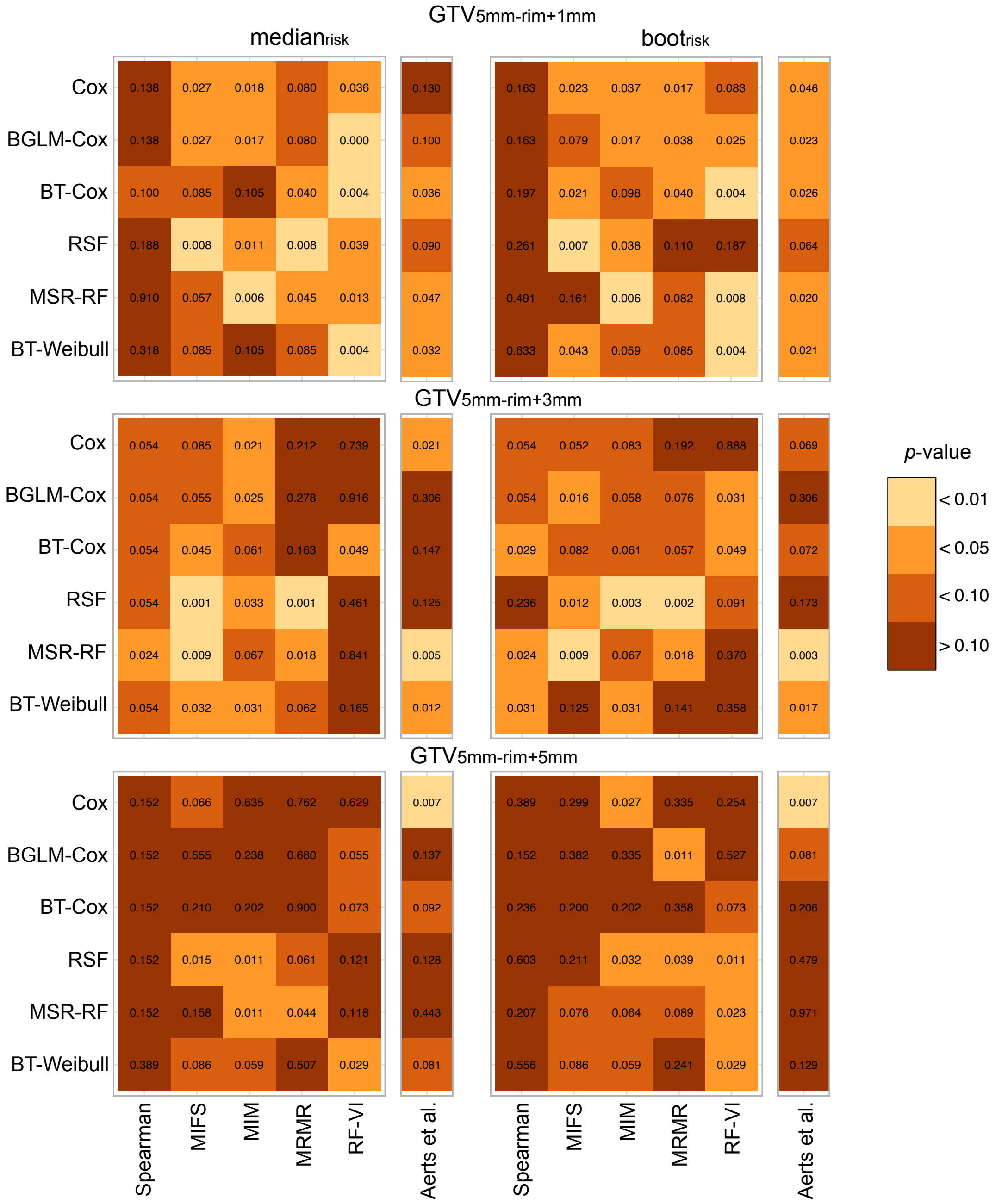
GTV10mm-rim

	median$_{risk}$						boot$_{risk}$					
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.025	0.010	0.001	0.005	0.000	0.136	0.536	0.008	0.001	0.109	0.581	0.035
BGLM-Cox	0.003	0.010	0.000	0.004	0.003	0.068	0.002	0.022	0.019	0.206	0.631	0.058
BT-Cox	0.178	0.065	0.001	0.162	0.007	0.073	0.144	0.065	0.001	0.162	0.007	0.339
RSF	0.003	0.009	0.011	0.008	0.000	0.074	0.092	0.156	0.259	0.032	0.000	0.044
MSR-RF	0.000	0.008	0.024	0.017	0.001	0.015	0.001	0.008	0.024	0.034	0.021	0.171
BT-Weibull	0.010	0.029	0.006	0.069	0.001	0.047	0.023	0.029	0.006	0.069	0.000	0.057

GTV10mm-core

	median$_{risk}$						boot$_{risk}$					
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.403	0.133	0.084	0.116	0.084	0.684	0.403	0.099	0.011	0.401	0.011	0.211
BGLM-Cox	0.337	0.176	0.084	0.176	0.084	0.076	0.215	0.175	0.017	0.300	0.486	0.300
BT-Cox	0.295	0.051	0.270	0.026	0.967	0.181	0.295	0.051	0.987	0.054	0.967	0.828
RSF	0.371	0.104	0.216	0.055	0.191	0.322	0.723	0.009	0.382	0.055	0.546	0.081
MSR-RF	0.293	0.010	0.637	0.005	0.308	0.201	0.488	0.006	0.637	0.011	0.299	0.438
BT-Weibull	0.027	0.009	0.736	0.015	0.293	0.326	0.027	0.064	0.767	0.015	0.293	0.696

Figure E.4.: Resulting *p*-values of the log-rank tests for loco-regional tumour control for the feature selection methods (columns) and the learning algorithms (rows) trained on the GTV$_{3mm\text{-}rim}$, the GTV$_{10mm\text{-}rim}$ and the corresponding core sub-volumes based on median (median$_{risk}$) (left) and bootstrapped (boot$_{risk}$) cut-off values (right) using the predicted risk values. The cut-off values used for stratification were determined on the exploratory cohort and applied to the validation cohort unchanged. Furthermore, the results of the log-rank tests for the models using the Aerts' signature (Aerts et al., 2014) are shown.



GTV5mm-rim+1mm

	median$_{risk}$						boot$_{risk}$					
	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.138	0.027	0.018	0.080	0.036	0.130	0.163	0.023	0.037	0.017	0.083	0.046
BGLM-Cox	0.138	0.027	0.017	0.080	0.000	0.100	0.163	0.079	0.017	0.038	0.025	0.023
BT-Cox	0.100	0.085	0.105	0.040	0.004	0.036	0.197	0.021	0.098	0.040	0.004	0.026
RSF	0.188	0.008	0.011	0.008	0.039	0.090	0.261	0.007	0.038	0.110	0.187	0.064
MSR-RF	0.910	0.057	0.006	0.045	0.013	0.047	0.491	0.161	0.006	0.082	0.008	0.020
BT-Weibull	0.318	0.085	0.105	0.085	0.004	0.032	0.633	0.043	0.059	0.085	0.004	0.021

GTV5mm-rim+3mm

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.054	0.085	0.021	0.212	0.739	0.021	0.054	0.052	0.083	0.192	0.888	0.069
BGLM-Cox	0.054	0.055	0.025	0.278	0.916	0.306	0.054	0.016	0.058	0.076	0.031	0.306
BT-Cox	0.054	0.045	0.061	0.163	0.049	0.147	0.029	0.082	0.061	0.057	0.049	0.072
RSF	0.054	0.001	0.033	0.001	0.461	0.125	0.236	0.012	0.003	0.002	0.091	0.173
MSR-RF	0.024	0.009	0.067	0.018	0.841	0.005	0.024	0.009	0.067	0.018	0.370	0.003
BT-Weibull	0.054	0.032	0.031	0.062	0.165	0.012	0.031	0.125	0.031	0.141	0.358	0.017

GTV5mm-rim+5mm

	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.	Spearman	MIFS	MIM	MRMR	RF-VI	Aerts et al.
Cox	0.152	0.066	0.635	0.762	0.629	0.007	0.389	0.299	0.027	0.335	0.254	0.007
BGLM-Cox	0.152	0.555	0.238	0.680	0.055	0.137	0.152	0.382	0.335	0.011	0.527	0.081
BT-Cox	0.152	0.210	0.202	0.900	0.073	0.092	0.236	0.200	0.202	0.358	0.073	0.206
RSF	0.152	0.015	0.011	0.061	0.121	0.128	0.603	0.211	0.032	0.039	0.011	0.479
MSR-RF	0.152	0.158	0.011	0.044	0.118	0.443	0.207	0.076	0.064	0.089	0.023	0.971
BT-Weibull	0.389	0.086	0.059	0.507	0.029	0.081	0.556	0.086	0.059	0.241	0.029	0.129

Figure E.5.: Resulting p-values of the log-rank tests for loco-regional tumour control for the feature selection methods (columns) and the learning algorithms (rows) trained on the different extended rim sub-volumes based on median (median$_{risk}$) (left) and bootstrapped (boot$_{risk}$) cut-off values (right) using the predicted risk values. The cut-off values used for stratification were determined on the exploratory cohort and applied to the validation cohort unchanged. Furthermore, the results of the log-rank tests for the models using the Aerts' signature (Aerts et al., 2014) are shown.

E.C. Developed signatures

Table E.1.: Radiomics signatures for predicting loco-regional tumour control for the representative models based on entire gross tumour volume, the selected rim and the core tumour as well as the selected extended tumour rim sub-volumes. The mathematical description and the abbreviations of features can be found in Zwanenburg *et al.* (Zwanenburg et al., 2016).

	Tumour sub-volume	Feature name	Synonym	Image	Cluster
$\text{GTV}_{\text{entire}}$	ivh_diff_v10_v90	$F1_S$	wav_coif1_llh	No	
	loc_peak_loc	$F2_S$	wav_coif1_hlh	No	
$\text{GTV}_{\text{5mm-rim}}$	ivh_diff_v10_v90	$F1_S$	wav_coif1_llh	No	
	ivh_v10	$F2_S$	wav_coif1_llh	No	
	dzm_ldhge	$F3_T$	Base	No	
	morph_integ_int, stat_energy, stat_energy	$F4_S$	Base, wav_coif1_lll	Yes	
	dzm_ldhge	$F5_T$	wav_coif1_lll	No	
	dzm_lgze, dzm_sdlge, szm_lgze, szm_szlge	$F6_T$	Base	Yes	
	szm_zs_entr	$F7_T$	Base, wav_coif1_lll	Yes	
	dzm_hgze, szm_hgze, szm_szhge	$F8_T$	Base	Yes	
	loc_peak_loc	$F9_S$	wav_coif1_hlh	No	
	dzm_lgze, szm_lgze, szm_szlge	$F10_T$	wav_coif1_hlh	No	
$\text{GTV}_{\text{5mm-core}}$	cm_clust_shade	$F1_T$	wav_coif1_lhh	No	
	ivh_i50	$F2_S$	wav_coif1_lhl	No	
	cm_auto_corr, cm_joint_avg, cm_sum_avg, dzm_hgze, ih_max_grad, ih_mean, ih_median, ih_min_grad, ih_mode, ih_p90, ivh_auc, ivh_v50, ngl_hgce, ngl_lgce rlm_hgre, rlm_lgre, rlm_srhge, szm_hgze, szm_szhge,	$F3_T$	wav_coif1_hhh	Yes	
	ih_skew, stat_skew	$F4_S$	wav_coif1_llh	Yes	
	ih_skew, stat_skew	$F5_S$	wav_coif1_lhh	Yes	
	ih_skew, stat_skew	$F6_S$	wav_coif1_hhh	Yes	
	ivh_diff_v25_v75	$F7_S$	wav_coif1_hhl	No	
	ngt_strength, rlm_rlnu	$F8_T$	wav_coif1_llh	Yes	
	dzm_zdnu	$F9_T$	wav_coif1_llh	No	

	Tumour sub-volume	Feature name	Synonym	Image	Cluster
	stat_median	F10$_{S}$	wav_coif1_hhh	No	
GTV$_{5mm\text{-}rim+2mm}$	morph_moran_i	F1$_{M}$	Base	No	
	morph_com	F2$_{M}$	Base	No	
	stat_cov	F3$_{S}$	wav_coif1_lhl	No	
	ngl_dcnu_norm	F4$_{T}$	wav_coif1_hhh	No	
	rlm_srlge	F5$_{T}$	wav_coif1_lhl	No	
	dzm_ldhge	F6$_{T}$	Base, wav_coif1_lll	Yes	
	stat_qcod	F7$_{S}$	wav_coif1_hhl	No	

Acknowledgment

First of all, I want to thank my advisor PD Dr. Steffen Löck for his valuable scientific exchange and his continuous support over the last years. Especially his expert advices for all theoretical and practical aspects of my thesis and his calm and thoughtful manner during discussions. It has been an honour to be his first PhD student. Furthermore, I thank Steffen, Prof. Dr. Hans-Joachim Böhme and Prof. Dr. David Craft for the acceptance to review my PhD thesis. I would like to thank my supervisor Dr. Christian Richter for the opportunity to continue my PhD thesis in his group and on the radiomics project at OncoRay as well as for his support and helpful discussions. I would also like to thank my colleague Dr. Alex Zwanenburg for the many meetings and open-minded scientific discussions and the proof-reading of my thesis. Only due to the outstanding teamwork, we could achieve all of our well-founded scientific results. Furthermore I would like to thank all colleagues of the group “High precision radiotherapy“ and “Modelling and Biostatistics in Radiation Oncology“ for the inspiring discussions and great group excursions. Especially, I thank Dr. Annika Jakobi and Dr. Kristin Stützer for their support and helpful advices at the beginning of the radiomics project. It is also a pleasure for me to thank Almut Dutz for proof-reading this thesis and the stimulating scientific and non-scientific discussions as well as her helpful advices during internal and external scientific presentations. A special thanks also to my friend and colleague Patrick Wohlfahrt for our daily meetings in the morning with a cup of coffee, inspired discussions and our sports and culinary activities. Moreover, I am grateful our interdisciplinary radiomics team for the constructive discussions and their clinical support. Especially, I thank Prof. Dr. Esther Troost and Karoline Pilz for their clinical input and for the many contouring tasks during the data collection. I would also like to thank Prof. Dr. Michael Baumann and Prof. Dr. Mechthild Krause for their support and the great scientific discussions as well as helpful advices during internal and external review processes of publications and abstracts. I am grateful for time spent with my “classmates“ Sonja Schellhammer and Johannes Müller for the scientific contributions and the barbecue activities. Finally, I would like to thank my family and all of my friends, for their support during this time.

Technische Universität Dresden
Medizinische Fakultät Carl Gustav Carus
Promotionsordnung vom 24.10.2014

Erklärungen zur Eröffnung des Promotionsverfahrens

1. Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

2. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten: PD Dr. Steffen Löck, Dr. Christian Richter, Dr. Alex Zwanenburg, Prof. Dr. Hans-Joachim Böhme und M.Sc. Almut Dutz.

3. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

4. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

5. Die Inhalte dieser Dissertation wurden in folgender Form veröffentlicht:
   - Leger S, Löck S, Hietschold V, Haase R, Böhme HJ, Abolmaali N Reproducible and Accurate Automatic Correction of Intensity Non-Uniformity in MRI Data. (2014) *Joint Conference of the SSRMP, DGMP, ÖGMP 2014*
   - Leger S, Löck S, Hietschold V, Haase R, Böhme HJ, Abolmaali N Automatic Intensity Non-Uniformity Correction in MRI Data. (2015) *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*
   - Leger S, Bandurska-Luque A, Pilz K, Zöphel K, Baumann M, Troost EGC, Löck S, Richter C OC-0262: Comparison of machine-learning methods for predictive radiomic models in locally advanced HNSCC (2016) *Radiotherapy and Oncology*

- Zwanenburg A, Leger S, Vallières M, Löck S and others Image biomarker standardisation initiative (2016) *arXiv*
- Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, Kotzerke J, Schreiber A, Tinhofer I, Budach C, Sak A, Stuschke M, Balermpas P, Rödel C, Ganswindt U, Belka C, Pigorsch S, Combs SE, Mönnich D, Zips D, Krause M, Baumann M, Troost EGC, Löck S, Richter C A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling (2017) *Scientific Reports*
- Pilz K, Leger S, Zwanenburg A, Richter C, Krause M, Baumann M, Löck S, Troost EGC EP-1065: Prediction of Dysphagia and Xerostomia based on CT imaging features of HNSCC Patients (2017) *Radiotherapy and Oncology*
- Leger S, Löck S, Hietschold V, Haase R, Böhme HJ, Abolmaali N: Physical correction model for automatic correction of intensity non-uniformity in magnetic resonance imaging (2017) *Physics and Imaging in Radiation Oncology*
- Zschaeck S, Löck S, Leger S, Haase R, Bandurska-Luque A, Appold S, Kotzerke J, Zips D, Richter C, Gudziol V and others FDG uptake in normal tissues assessed by PET during treatment has prognostic value for treatment results in head and neck squamous cell carcinomas undergoing radiochemotherapy (2017) *Radiotherapy and Oncology*
- Schmidt S, Linge A, Zwanenburg A, Leger S, Lohaus F, Krenn C, Appold S, Gudziol V, Nowak A, von Neubeck C and others Development and validation of a gene signature for patients with head and neck squamous cell carcinomas treated by postoperative radio (chemo) therapy (2018) *Clinical Cancer Research*
- Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, Kotzerke J, Schreiber A, Tinhofer I, Budach C, Sak A, Stuschke M, Balermpas P, Rödel C, Ganswindt U, Belka C, Pigorsch S, Combs SE., Mönnich D, Zips D, Krause M, Baumann M, Richter C, Troost EGC., Löck S Identification of tumour sub-volumes for improved radiomic risk modelling in locally advanced HNSCC (2018) *Radiotherapy and Oncology*
- Leger S, Zwanenburg A, Pilz K, Zschaeck S, Zöphel K, Kotzerke J, Schreiber A, Zips D, Krause M, Baumann M, Troost EGC., Richter C, Löck S Identification of tumour sub-volumes for improved radiomic risk modelling in locally advanced HNSCC (2018) *Radiotherapy and Oncology submitted*
- Zwanenburg A, Leger S, Vallières M, Löck S and others Standardized quantitative radiomics for high-throughput image-based phenotyping (2018) *Medical Image Analysis submitted*

6. Ich bestätige, dass es keine zurückliegenden erfolglosen Promotionsverfahren gab.

7. Ich bestätige, dass ich die Promotionsordnung der Medizinischen Fakultät Carl Gustav Carus der Technischen Universität Dresden anerkenne.

Dresden, 20.06.2018

Hiermit bestätige ich die Einhaltung der folgenden aktuellen gesetzlichen Vorgaben im Rahmen meiner Dissertation (Nicht angekreuzte Punkte sind für meine Dissertation nicht relevant.)

☐ das zustimmende Votum der Ethikkommission bei Klinischen Studien, epidemiologischen Untersuchungen mit Personenbezug oder Sachverhalten, die das Medizinproduktegesetz betreffen

☐ die Einhaltung der Bestimmungen des Tierschutzgesetzes

☐ die Einhaltung des Gentechnikgesetzes

☒ die Einhaltung von Datenschutzbestimmungen der Medizinischen Fakultät und des Universitätsklinikums Carl Gustav Carus.

Dresden, 20.06.2018