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

Radiomics analyses for outcome prediction in patients with locally advanced rectal cancer and glioblastoma multiforme using multimodal imaging data

Dissertationsschrift

zur Erlangung des akademischen Grades
Doctor of Philosophy (Ph. D.)
vorgelegt
der Medizinischen Fakultät Carl Gustav Carus
der Technischen Universität Dresden

von

M. Sc. Iram Shahzadi

aus Chakwal, Punjab, Pakistan

Dresden 2023

1. Gutachter: Prof. Dr. Steffen Löck

2. Gutachter: Prof. Dr. Daniela Thorwarth

Tag der mündlichen Prüfung: (Verteidigungstermin)

gez.: ....

Vorsitzender der Promotionskommission

Abstract

Personalized treatment strategies for oncological patient management can improve outcomes of patient populations with heterogeneous treatment response. The implementation of such a concept requires the identification of biomarkers that can precisely predict treatment outcome. In the context of this thesis, we develop and validate biomarkers from multimodal imaging data for the outcome prediction after treatment in patients with locally advanced rectal cancer (LARC) and in patients with newly diagnosed glioblastoma multiforme (GBM), using conventional feature-based radiomics and deep-learning (DL) based radiomics. For LARC patients, we identify promising radiomics signatures combining computed tomography (CT) and T2-weighted (T2-w) magnetic resonance imaging (MRI) with clinical parameters to predict tumour response to neoadjuvant chemoradiotherapy (nCRT). Further, the analyses of externally available radiomics models for LARC reveal a lack of reproducibility and the need for standardization of the radiomics process. For patients with GBM, we use postoperative [^{11}C] methionine positron emission tomography (MET-PET) and gadolinium-enhanced T1-w MRI for the detection of the residual tumour status and to prognosticate time-to-recurrence (TTR) and overall survival (OS). We show that DL models built on MET-PET have an improved diagnostic and prognostic value as compared to MRI.

Kurzzusammenfassung

Personalisierte Therapiekonzepte bieten die Möglichkeit das Behandlungsergebnis von Krebspatienten mit einem derzeit heterogenen Ansprechen des Tumors zu verbessern. Die Umsetzung eines solchen Konzepts erfordert die Identifizierung von Biomarkern, die das Behandlungsergebnis präzise vorhersagen können. Im Rahmen dieser Arbeit werden Biomarker aus multimodalen Bildgebungsdaten entwickelt und validiert um das Outcome von Patienten mit lokal fortgeschrittenem Rektumkarzinom (LARC) und neu diagnostiziertem Glioblastoma multiforme (GBM) unter Verwendung von konventionellen Radiomics-Analysen und Deep-Learning (DL) basierten Radiomics-Verfahren vorherzusagen. Für LARC-Patienten werden Radiomics-Signaturen basierend auf der Computertomographie (CT) und der T2-gewichteten (T2-w) Magnetresonanztomographie (MRT) gemeinsam mit klinischen Parametern entwickelt um das Tumoransprechen auf eine neoadjuvante Chemoradiotherapie (nCRT) vorherzusagen. Weitere Analysen von extern verfügbaren Radiomics-Modellen für LARC zeigen einen Mangel an Reproduzierbarkeit und die Notwendigkeit einer Standardisierung des Radiomics-Prozesses. Für Patienten mit GBM werden Modelle zur Vorhersage des Resektionsstatus, der lokalen Tumorkontrolle (TTR) sowie des Gesamtüberlebens (OS) entwickelt, basierend auf Bildgebung der postoperativen [^{11}C]-Methionin Positronenemissionstomographie (MET-PET) und T1-w MRT. DL-Modelle, die MET-PET-Daten verwenden, zeigen dabei eine bessere diagnostische und prognostische Güte als MRT-basierte Modelle.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADC	apparent diffusion coefficient
AJCC	American Joint Committee on Cancer
AUC	area under the ROC curve
BCE	binary-cross entropy
BN	batch normalization
CC	Colon cancer
cCR	clinical complete response
CI	confidence interval
C-index	concordance index
CNN	convolutional neural networks
CNS	central nervous system
CPHM	Cox proportional hazard models
CRC	Colorectal cancer
CRT	chemoradiotherapy
CT	computed tomography
CTV	clinical target volume
CV	cross-validation
DKTK-ROG	German Consortium for Translational Cancer Research - Radiation Oncology Group
DL	deep learning
DM	distant metastases
DNA	deoxyribonucleic acid
DNN	deep neural network
DWI	diffusion weighted imaging
ECOG	Eastern Cooperative Oncology Group
EN	elastic-net
FBN	fixed bin number
FBS	fixed bin size
FC	fully connected
FDG	fluorodeoxyglucose
FET	f-fluoro-ethyl-l-tyrosine

FFDM	freedom from distant metastases
FN	false negative
FO	first-order
FP	false positive
FPR	false positive rate
GBM	Glioblastoma multiforme
GLCM	grey level co-occurence matrix
GLDZM	grey level distance zone matrix
GLM	generalized linear model
GLRLM	grey level run length matrix
GLSZM	grey level size zone matrix
GND test	Greenwood Nam D'Agostino test
GTV	gross tumor olume
HL test	Hosmer-Lemeshow goodness of fit test
HU	hounsfield unit
IBSI	Imaging Biomarker and Standardization Initiative
ICC	intraclass correlation coefficient
IDH	isocitrate dehydrogenase
KPS	Karnofsky performance status
LARC	locally advanced rectal cancer
LC	local control
LoG	Laplacian of Gaussian
MET	methionine
MFO	morphological and first order
MGMT	O6-methylguanine–DNA methyltransferase
MI	mutual information
MIM	mutual information maximization
MIRP	Medical Image Radiomcis Processor
ML	machine learning
MLEM	maximum likelihood expectation maximization
MR	magnetic resonance
MRI	magnetic resonance imaging
mRMR	minimal redundancy maximum Relevance

nCRT	neoadjuvant chemoradiotherapy
NGLDM	neighbouring grey level dependence matrix
NGTDM	neighbouring grey tone difference matrix
NMR	nuclear magnetic resonance
OS	overall survival
OSEM	ordered subset expectation maximization
PACS	Picture Archiving and Communication System
pCR	pathological complete response
PET	positron emission tomography
PFS	progression free survival
pNR	pathological non-responder
RC	Rectal cancer
RCT	chemoradiotherapy
RF	random forest
ROC	reciever operating curve
ROI	reigon-of-interest
RSF	random survival forest
RT	radiotherapy
SMBO	sequential model-based optimisation
SOT	second-order texture
SUV	standardized uptake values
TE	echo time
TME	total mesorectal excision
TN	true negative
TNR	true negative rate
TP	true positive
TPR	true positive rate
TR	relaxation time
TRG	tumour regression grade
TTR	time-to-recurrence
UDWT	undecimated discrete wavelet transform
UR	univariate regression
WHO	World Health Organization

1 Introduction

In recent years, the morbidity and mortality of cancer is continuously increasing due to global ageing population and environmental factors such as pollution, increased consumption of Western pattern diet, alcohol, and tobacco (Anand et al., 2008; Soerjomataram & Bray, 2021; Chhikara & Parang, 2023). The main goals of modern oncology are to prevent or diagnose cancer at an early stage and improve the quality and length of life after cancer through multimodal and personalized treatment. There are currently three main treatment methods for cancer management that are often used in combination: (i) surgery, (ii) chemotherapy, and (iii) radiation therapy. The present approaches to cancer treatment are largely empirical and often referred as "one size fits all" (Duffy & Crown, 2008). In general, evidence for treatment options is derived from clinical trials based on a patient population with similar diagnosis and staging. As a result, some patients with aggressive disease may be undertreated, and some with indolent disease may be overtreated. In addition, for those patients who receive treatment, only a proportion derives clinical benefit, whereas adverse side effects are common. In contrast, personalization of treatment offers the potential to improve treatment outcomes by considering the tumour characteristics of patients or subgroups individually. The implementation of this approach requires the identification of prognostic biomarkers, that can reliably stratify the patients into subgroups with similar tumour characteristics and treatment response, illustrated in Figure 1.1.

Currently, the main method of treatment personalization for cancer patients is the analysis of molecular profiles from tumour biopsies. Because a tumour may show marked subclonal heterogeneity, molecular profiles from a single biopsy may be insufficient for yielding robust biomarkers for treatment personalization. Multiple biopsies may be required to capture the heterogeneity present in the tumour and yield biomarkers to develop accurate prognostic models. Since acquiring tissue biopsies is invasive and not always possible, there is a need to develop non-invasive biomarkers for cancer treatment personalization that can incorporate whole-tumour information. Medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), X-ray, and ultrasound are widely used in clinical practice to aid decision-making in cancer patients and can capture three-dimensional whole-tumour information. Medical imaging has thus the potential to overcome challenges posed by molecular biomarkers. For example, studies have shown that simple imaging features such as tumour volume determined on multimodal imaging data, i.e. MRI, CT or PET, were able to predict treatment response (Partridge et al., 2005; Hutchings et al., 2006; Meneghetti et al., 2021).

Recently, radiomics has attracted considerable interest in the field of radiology and clinical oncology. Radiomics analyses perform a non-invasive, quantitative characterization of medical imaging to identify image biomarkers (Figure 1.2). Such analyses employ modern machine learning (ML) algorithms for the evaluation of cancer diagnosis or the prediction of treatment outcomes

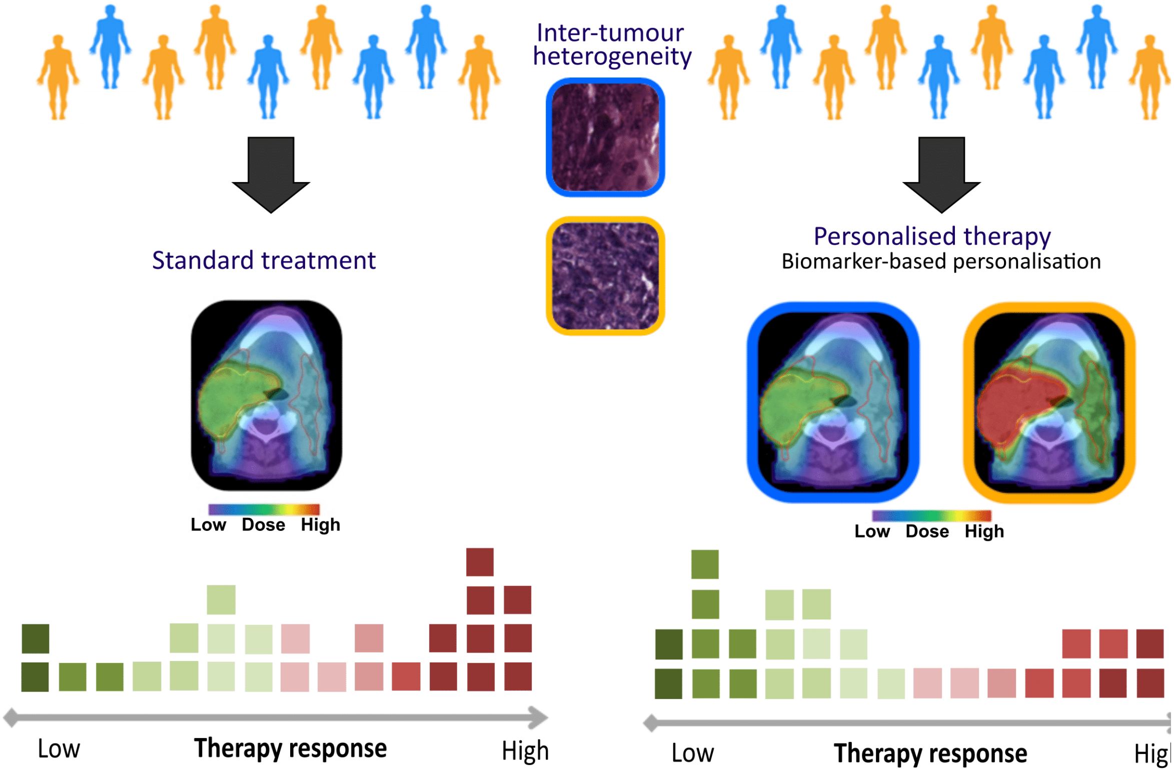


Figure 1.1: Schematic representation to individualize radiotherapy based on inter-tumour heterogeneity. In standard treatment, a heterogeneous patient population is treated with the same dose, which may lead to under or overtreatment of tumour regions. Biomarker-based patient stratification can be used to personalize treatment, e.g., by dose adaptation in clinical trials.

in different cancer types, incorporating different imaging modalities (Aerts, 2016). Within the field of cancer diagnosis, studies employing magnetic resonance (MR) have, for instance, tackled the differentiation of high grade and low grade gliomas (Cho et al., 2018), between malignant and benign prostate tumours (Wibmer et al., 2015) and positive and negative O6-methylguanine–DNA methyltransferase (MGMT) methylation status in glioblastoma (Kong et al., 2016; Korfiatis et al., 2016). Conventional radiomics approaches work with handcrafted features to capture tumour characteristics. However, manual feature engineering bears the risk of not capturing the full information content present in the data. Alternatively, deep learning (DL) based radiomics addresses this issue by automatically learning features from the data when training them on suitable tasks and could therefore improve performance for the prediction of radiotherapy outcomes compared to traditional statistical and ML models (Figure 1.3).

Rectal cancer (RC) is a distinct type of Colon cancer (CC) with high rate of local and regional recurrence (Lee, 1995). The standard treatment for locally advanced rectal cancer (LARC) is preoperative, neoadjuvant chemoradiotherapy (nCRT) followed by total mesorectal excision and postoperative adjuvant chemotherapy. This standard of care was established from various landmark clinical trials (Fisher et al., 1988; Franke et al., 2021). It has been shown that pre-operative

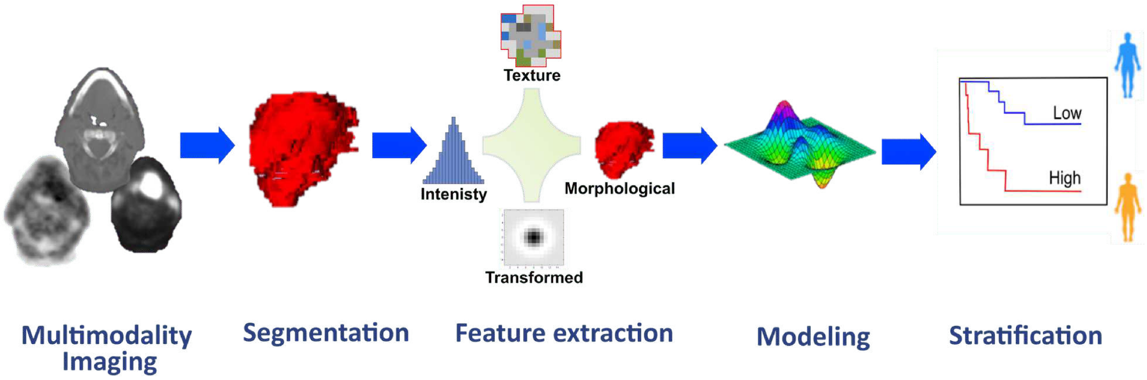


Figure 1.2: Schematic representation of the conventional radiomics workflow. The main steps are (1) image acquisition, (2) identification and delineation of the region or reigon-of-interest (ROI), (3) extraction of features with different levels of complexity, (4) feature modelling using classical statistics or machine learning approaches, and (5) final predictions based on modelling step.

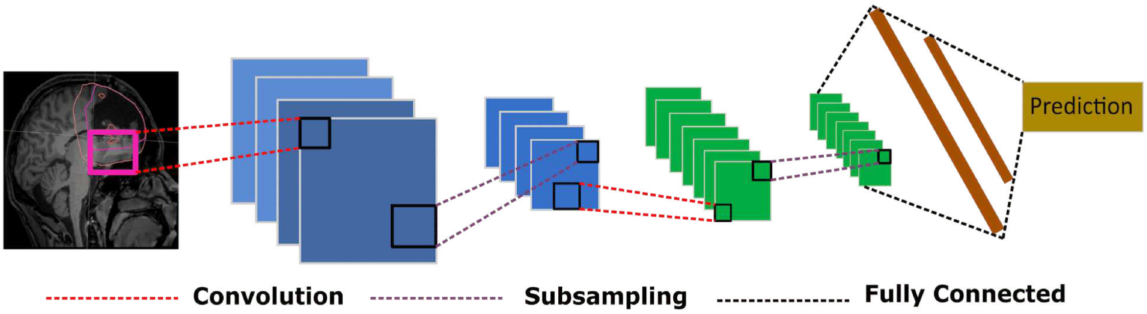


Figure 1.3: Schematic representation of the end-to-end feature extraction and prediction with convolutional neural networks (CNN). CNN can analyse patterns in imaging data by performing multiple steps of convolution and subsampling. Finally, the representation of images stored in fully connected layers are mapped to the target output.

radiotherapy helps to achieve a marked reduction in both acute and long-term treatment-related toxicity in rectal cancer (Häfner & Debus, 2016). Moreover, nCRT provides high local control and is able to reduce the local recurrence rate to about 10% (Sauer et al., 2004). The standard of operative management in rectal cancer is total mesorectal excision (TME) that has been shown to improve local control (LC) and overall survival (OS) in RC patients (Kapiteijn & van de Velde, 2000). However, the treatment process of LARC is associated with the risk of toxicities and long-term disabilities in patients. RC is commonly surrounded by other radiosensitive structures in the pelvic region, such as bone marrow, and future chemoradiotherapy interventions for local recurrences are impeded by accrued tissue damage from earlier treatment (Newman et al., 2016). Another concern is that while the current standard treatment of LARC has significantly reduced the rate of local recurrence, more than 25% of patients still succumb to RC due to distant metastases. After nCRT 15-20% of patients with LARC show pathological complete response.

There is increased interest in the adoption of organ preservation and low morbidity surgeries such as local excision in partial responders and watch-and-wait strategy for pathological complete response (pCR) (Dossa et al., 2017). Retrospective studies comparing survival outcomes between pCR's treated with wait-and-watch strategy and those who had surgical resection have shown 94% 5-year disease-free survival and 81-85% 5-year OS (Renehan et al., 2016). A study conducted by the Mercury group (Group et al., 2007) also suggests selective radiotherapy rather than reflexive radiotherapy in patients with good prognosis, i.e. without suspicious lymph nodes or extramural vascular invasion.

Tumour response to treatment depends on several factors including the type of chemotherapy administered, radiation dose, and surgical practice; however, also the biology of a tumour plays a most important role in describing treatment response. Several studies have analysed molecular data, such as gene expression, mutations, and single nucleotide polymorphisms as potential biomarkers of response to nCRT in LARC (Rimkus et al., 2008; Boige et al., 2010; Duldulao et al., 2013). Treatment outcome prediction and patient prognosis through radiomics analyses is another widely explored topic in literature, with more than 300 studies alone having been published for treatment response and long-term outcomes prediction in LARC (Staal et al., 2021). For outcome prognosis in LARC, MRI-based radiomic models have been widely developed for tumour response and freedom from distant metastases (FFDM) after nCRT (Nie et al., 2016; Dinapoli et al., 2018; Horvat et al., 2018; Jeon et al., 2019; Antunes et al., 2020). Some studies have considered radiomic features extracted from CT imaging (Chee et al., 2017; Bibault et al., 2018) or a combination of CT and MRI features (Li et al., 2020b; Zhang et al., 2020).

Glioblastoma multiforme (GBM) is the most common type of primary brain tumours that belongs to the heterogeneous and invasive tumours of the central nervous system arising from glial cells (Bailey, 1985; Ferguson & Lesniak, 2005). These tumours are highly invasive. They typically arise in the deep white matter and proceed to infiltrate grey matter and other structures (Rees et al., 1996). The standard treatment in glioblastoma is surgical resection followed by chemoradiotherapy (CRT) (Stupp et al., 2005). Despite intense multimodal treatment, patients with GBM have poor prognosis with 5-year OS rate of only 9.8% (Stupp et al., 2009). In order to prolong patient survival, GBM resection aims at maximal safe resection. However, for many patients, aggressive resection is not possible as GBM may reside in critical regions (Sanai & Berger, 2009; Sanai et al., 2011; Bloch et al., 2012). Thus, the residual tumour is left behind, leading to tumour recurrence and resulting in poor prognosis. Additionally, regardless of the extent of resection, infiltrative growth patterns may also lead to tumour recurrence and eventually to patient death. Thus, controlling and/or targeting residual tumours and infiltrative tumour cells may improve patient survival. Furthermore, the assessment of patient prognosis in GBM prior to the start of treatment may help to identify subgroups of patients that would benefit from escalated radiotherapy doses.

Researchers have discovered molecular, genetic and histopathological biomarkers pertaining to long-term outcomes in GBM patients, with the most widely accepted prognostic biomarkers for GBM being age and MGMT promoter methylation status. The effectiveness of treatment is lower in elderly patients due to reduced innate immunity and higher postoperative complication rates, thus resulting in poorer prognosis (Laigle-Donadey & Delattre, 2006). MGMT is a deoxyribonucleic acid (DNA)-repair gene whose high level of activity in cancer cells causes resistance to chemotherapy (Wen & Kesari, 2008). Suppression of the MGMT gene through methylation blunts DNA repair processes and sensitizes tumour cells to radiation, which leads to improved tumour control and improved patient survival (Laigle-Donadey & Delattre, 2006). Many studies have performed gene-expression profiling to identify genes whose expression can predict patient survival in GBM (Rich et al., 2005; Yamanaka et al., 2006; Candido et al., 2019). Similarly, many studies also predicted the chemotherapeutic response and survival of patients with GBM using MRI radiomics features (Cui et al., 2016; Kickingereder et al., 2016c; Li et al., 2022). Although the results of these analyses are encouraging, important processes of radiomics such as assessing feature robustness, were not always considered, and external validation is rarely performed. Despite the potential of radiomics for the individualization of cancer therapy, further research and developments are required, e.g., regarding the choice of suitable machine learning algorithms for risk modelling.

The aim of this thesis is to develop and validate imaging biomarkers using radiomics for patients with LARC and GBM. The presented results help to facilitate image-based treatment decisions and to gain a deeper understanding of radiomics for treatment personalization.

This thesis is structured as follows: Chapter 2 provides an overview of GBM and LARC followed by and the essential background information about imaging modalities. Afterward, statistical methods used for conventional and DL-based radiomics for diagnostic and prognostic modelling are introduced.

Traditional radiomics analyses commonly apply imaging features of different complexity for the prediction of the endpoint of interest. However, the prognostic value of each feature class is generally unclear. In this thesis, we developed and independently validated signatures for outcome prediction after nCRT in patients with LARC based on imaging datasets by analysing different feature classes (Chapter 3).

Over the past years, over 300 radiomics studies have been published on LARC indicating promising results for treatment outcome prediction. Because of the complexity, heterogeneity and increasing volume of that literature, it is challenging to interpret the results. Furthermore, many radiomics models lack independent external validation that is decisive for their clinical application. Therefore, in Chapter 4 we aim to validate previously published radiomics signatures for LARC on our multicentre cohort.

For patients with GBM, gross total resection cannot always be achieved due to infiltrative growth patterns and thus residual tumour cells persist after surgery. Patients with residual tumour may be candidates for escalated radiotherapy doses to eradicate residual disease. Imaging-based diagnosis of residual tumour is a complex evaluation process, and automatic methods may be helpful to support the clinical decision. In Chapter 5, we compared the diagnostic performance of both, conventional and DL-based radiomics, for automatic detection of residual tumour on MET-PET and T1c-w MRI data.

Patients with GBM may benefit from the development of accurate prognostic biomarkers, e.g. by selection for escalated radiotherapy doses after surgery. Therefore, in Chapter 6, we developed and independently validated conventional and DL-based radiomics biomarkers for the prognosis of time-to-recurrence (TTR) and OS.

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

2 Theoretical background

2.1 Introduction to locally advanced rectal cancer and glioblastoma multiforme

In this section, we discuss both considered tumour entities, LARC and GBM, including epidemiological and pathological features and conventional treatment.

2.1.1 Locally advanced rectal cancer

Colorectal cancer (CRC) is the third most commonly diagnosed cancer that accounts for 10% of global cancer incidence (Xi & Xu, 2021) is also the third leading cause of cancer-related deaths in both genders, however, higher incidence rate is found in males compared to females. CRC is regarded to be comprising CC and RC as both are developed in the colon, however, studies have shown that CC and RC are two distinct cancer types with differences in molecular carcinogenesis, pathology and multimodal treatment (Wei et al., 2004; Li et al., 2007; Paschke et al., 2018). RC is defined as a tumour whose margin is 16 cm or less from anal verge (Sobin, 2009). The patients with RC present with severe disabling symptoms, including rectal bleeding and change in bowel habit. These symptoms are sometimes attributed to local bowel conditions such as haemorrhoids, and due to delay in seeking proper medical treatment it results in expansion of disease to advanced stage (McCarthy et al., 2012). The causes of RC include alcohol abuse, smoking, western diet including consumption of processed food. Other factors that show association with RC include history of CRC and age with the majority of incidences occurring in patients with age above 50. Current screening of RC is commonly performed with colonoscopy or flexible sigmoidoscopy and once suspicion of extended disease is found, diagnostic CT and MRI is conducted for staging of the disease. Approximately 5% to 10% of patients with RC present with LARC. LARC comprises rectal tumours within 6-12 cm of the anal verge, reaching to and beyond mesorectal fascia. LARC is defined as T3 or T4 primary tumours or nodal metastases (T3-4 and/or N+) (Edge & Compton, 2010; Amin et al., 2017). The current management of LARC includes nCRT (40-50.4 Gy plus augmented 5-flurouracil (FU) alone or combined with oxaliplatin) followed by TME and adjuvant therapy with 5-FU combined with oxaliplatin (Sauer et al., 2004). This treatment regime has shown to improve the LC and OS of patients, however recurrent diseases and distant metastases (DM) still pose a major problem after treatment.

2.1.2 Glioblastoma multiforme

Gliomas are the most common cancer of central nervous system (CNS) and represent almost 80% of all CNS and primary brain tumours (Ostrom et al., 2022). Gliomas originate from glial cells that surround the nerve cells and are classified into three types based on the type of glial cells involved, thus including tumours of astrocytomas, ependymomas, and oligodendrogliomas. World Health Organization (WHO) classifies these tumours into malignancy grade ranging from I to IV, with GBM being most malignant astrocytic glioma (WHO grade IV) with poor prognosis (median survival of 15 months) (Stupp et al., 2005). GBM can occur at any age. However, it is most common in adults with more than 80% of patients diagnosed at the age above 55, and males have 1.6 times higher incident rate than women (Ostrom et al., 2022). Studies have suggested that immune surveillance mechanisms stimulated by the protective effect to infections and allergies result in the development of gliomas (Fisher et al., 2007; Ohgaki, 2009; Barnholtz-Sloan et al., 2018), however, no carcinogenetic causes of gliomas and more specifically of GBM has been identified. In addition, rare genetic syndromes are also held responsible for the development of approximately 10% of gliomas (Fisher et al., 2007; Bondy et al., 2008). GBM tumour mass is characterized with high level of regional heterogeneity and poor delineation. The lesions often occur in the subcortical of white matter of cerebral hemispheres, thus occupying a massive region of the brain lobe and invading rapidly into surrounding brain tissue. In high-resolution contrast-enhanced MRI scans, the infiltrating tumour cells can be seen dispersed within normal brain tissues around the necrotic region of GBM (Louis et al., 2007). The symptoms of GBM arise late after the origin of the disease and are depicted in patients as neurological symptoms including memory loss, personality changes and seizures, and raised intracranial pressure with tension headaches, nausea, and vomiting (Wen & Kesari, 2008). The standard treatment of GBM consists of surgery followed by radiotherapy up to a dose of 60 Gy with concomitant and adjuvant chemotherapy. Temozolomide is the main chemotherapeutic agent used for the treatment of GBM. Commonly, total resection of the tumour is not possible in GBM due to infiltrative growth patterns. Therefore, debulking is performed to reduce the tumour size. The tumour remnants are then irradiated by including a 2-3 cm margin around the resection. Even after multimodal treatment, tumour recurrence happens in almost all patients.

2.2 Basic physical principles of imaging modalities

Imaging is an integral component of present-day radiation oncology practice within the clinics. Radiological imaging has substantially improved our understanding of cancer biology, diagnosis, staging, and prognosis. The most used imaging modalities in radiation oncology include CT, MRI and PET. In this section, we explain basic principles of these imaging modalities.

2.2.1 Computed tomography

Computed tomography (CT) is a common imaging modality with many uses, including quick diagnosis of injuries, diseases and radiation treatment planning. In radiation oncology, CT is particularly useful as it provides three-dimensional (3D) information of anatomical structures that

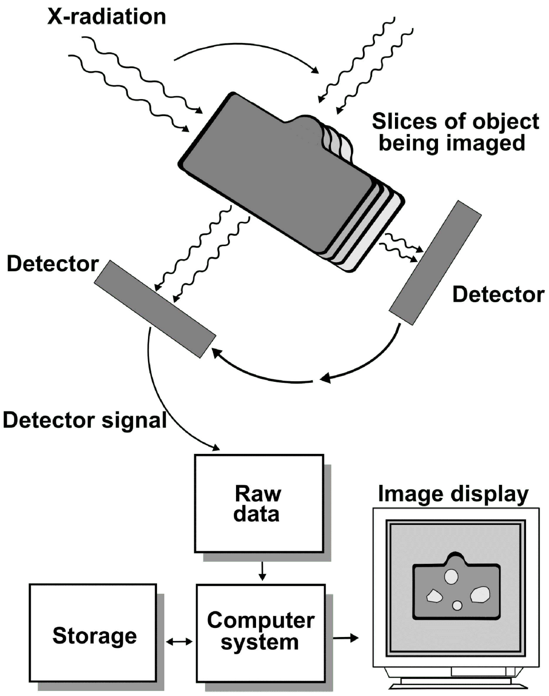


Figure 2.1: An overview of CT data acquisition and image reconstruction from projections. An X-ray beam passes through each slice of the object from different locations, by rotating the source and detector. The X-ray beam is partially attenuated by the object, and the remaining photons can be detected in the detector array. The detector array converts incident photons to an electric signal, which is processed and stored by a computer system. This system then reconstructs a CT image of the object. Figure reproduced from (Seeram, 2015) with permission from Elsevier.

allows better knowledge of target volumes and position of organs at risk. Figure 2.1 provides an illustration of CT image acquisition and reconstruction. In CT, data acquisition is done by passing an X-ray beam through a patient. Part of the beam is attenuated by the patient tissue, and a detector array measures the strength of the incident beam. In order to acquire the full volumetric attenuation profile of tissue, the X-ray tube and detector rotate around the patient at least 180 degrees. The relationship between transmitted and detected intensities of the X-ray beam for a non-homogenous material are given by the Beer-Lambert Law (Bouguer, 1729) as follows:

$$I(x, y) = I_0 \exp\left(-\int_l \mu(x, y)\, ds\right) \tag{2.1}$$

where $I(x, y)$ and I_0 are the intensities of the transmitted and the incident beam, and $\mu(x, y)$ is the attenuation coefficient of matter at position (x, y), a 2D function to rebuild. Usually, $l = l(\theta, \tau)$ *where* $(\theta, \tau) \in [0, 2\pi) \times R$. The total attenuation p in projection space at position r and angle θ is given by:

$$p_\theta(r) = \ln\left(\frac{I}{I_0}\right) = -\int_l \mu(x, y) \tag{2.2}$$

In order to explicitly calculate a value of the integral, a simple rotation matrix is used to transform x and y to point r in projection space. The above equation can be written as:

$$p_\theta(r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x \cos\theta + y \sin\theta - r)\, dx\, dy \tag{2.3}$$

where the Dirac function δ ensures that only relevant rays contribute to projection $p_\theta(r)$. Equation (2.3) is also known as the Radon transform (Radon, 1986). The CT image reconstruction is the process of computing the function $f(x, y)$ from projections $p_\theta(r)$, a process known as back-projection (Herman, 2009). This is achieved by computing the inverse Radon transform as follows:

$$(x, y) = \int_0^\pi p(s, \theta)|_{s=x\cos\theta+y\sin\theta}\, d\theta \tag{2.4}$$

CT imaging relies on the principle that the fraction of X-rays absorbed or scattered by an object depends on its material composition, which is quantified through the attenuation coefficient. In the typical energy range of an X-ray beam for CT, the attenuation coefficient is mainly defined by the electron density of the substance. Therefore, denser substances and substances containing elements with many electrons will have higher attenuation coefficients. In practice, hounsfield unit (HU) (Hounsfield, 1980), attenuation coefficients normalized to the attenuation coefficient of water are used, because it is a well-suited scale for the examination of objects with high concentration of water, e.g. humans. The HU unit of a tissue is defined as:

$$H_{tissue} = \frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \times 1000 \tag{2.5}$$

The typical clinical range of a CT scan, between air and bone, is approximately -1024 HU to 3071 HU. In this work, the primary target are soft tissues which represent a small portion of this range, i.e. -150 to 180 HU.

2.2.2 Magnetic resonance imaging

Magnetic resonance imaging (MRI) can provide high contrast and detailed assessment of morphology, and can also be combined with functional and metabolic evaluation. It is generally a safe imaging modality, with the possibility of repeating the scanning process without any side effects due to the lack of exposure to ionizing radiation or iodinated contrast agents. MRI scans are used to diagnose a variety of conditions, from torn ligaments to tumours.

MRI works on the principles of nuclear magnetic resonance (NMR), i.e. under the effect of a resonating electromagnetic radiofrequency the nucleus of an atom is perturbed or excited and responds by producing the radiofrequency signal (Bloch, 1946). MRI uses the natural magnetic property of hydrogen, which as a part of lipids or water constitutes 70%-90% of most tissues in human body. Protons in the nucleus of every hydrogen atom have a nuclear spin and are thus sensitive to the presence of an external magnetic field. Typically, due to thermal movement, the

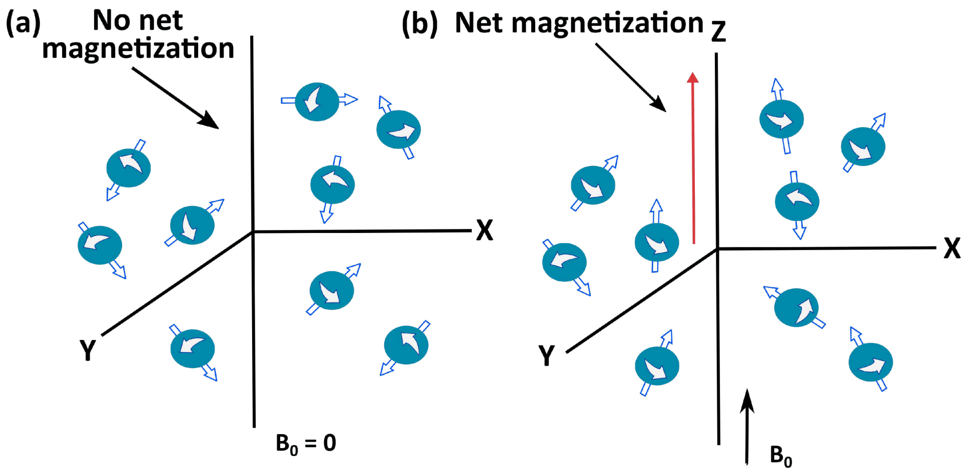


Figure 2.2: An illustration of molecular spin in MRI. (a) With no magnetic field, the spins are oriented randomly, producing no net magnetization. (b) The spins orient with an applied magnetic field B_0, producing a net magnetic field M_0 aligned along B_0. Figure reproduced from (Balaban & Peters, 2019) with permission from Elsevier.

magnetic moments of protons are randomly orientated, resulting in a non-observable magnetization (Figure 2.2(a)). When these protons come under the influence of a strong enough magnetic field, they undergo resonance, precessing at an angular frequency 0 proportional to the applied external magnetic field as described by the Larmor equation:

$$\omega_0 = \gamma B_0 \tag{2.6}$$

where B_0 is the applied static magnetic field and is a constant called gyro-magnetic ratio that depends on type of nuclei being excited. The proton population will then tend to show net alignment along the applied external magnetic field, producing their own magnetic field, known as net magnetization.

$$M_0 = \chi B_0 \tag{2.7}$$

where χ is the magnetic susceptibility. In normal or equilibrium configuration, M_0 is aligned along the magnetic field, B_0 which is conventionally shown as z direction (Figure 2.2(a)). M_0 can be decomposed into measurable longitudinal (M_z) and transversal (M_{xy}) components, and under equilibrium $M_z = M_0$ and $M_{xy} = 0$. By applying an additional radiofrequency pulse known as excitation pulse with associated magnetic field B_1, we can tilt M_0 away from B_0 The magnitude of this effect is referred as flip angle. Radiofrequency pulses are applied repeatedly to encode spatial information using magnetic gradients in the x, y and z dimension to acquire volumetric data.

In MR images, different tissues show different intensities, which is described as image contrast. MRI can generate a wide variety of contrasts by using different techniques, which are known as

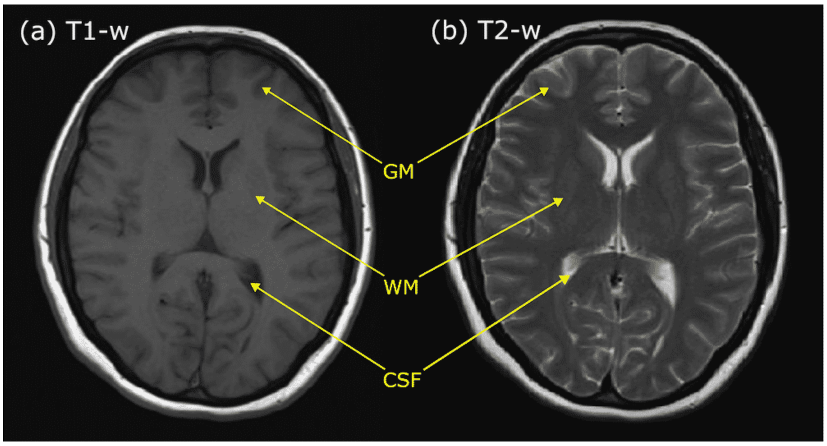


Figure 2.3: An axial slice of the brain measured using T1-weighted (T1-w) and T2-weighted (T2-w) MRI sequences. (a) In T1-w images, fluid appears dark, fatty tissues are bright and wet tissues are mid-grey. (b) In T2-w images, fluids appear bright, while wet and fatty tissues are mid-grey. The tissue types indicated are white matter (WM), grey matter (GM) and cerebrospinal fluid.

pulse sequences, and by controlling the time of these sequences. There are different sequences, however, they all have radiofrequency pulses and gradient pulses with carefully controlled timing values called relaxation time (TR) and echo time (TE). The time between each repetition of radiofrequency pulse is referred to as TR and body tissues can be characterized by two relaxation times: i) T1, and ii) T2. Following the application of the excitation pulse, the time required for the z component of M_0 to return to 63% of its original value is called T1 relaxation. The value of the net magnetization in the external field aligned axis then follow an exponential growth:

$$M_z(\tau) = M_0 \left(1 - e^{\frac{-\tau}{T1}}\right) \tag{2.8}$$

where is the time following the radiofrequency pulse and T1 is a time constant describing the rate of growth. Similarly, the time required for the transversal magnetization component M_{xy} to return to 37% of the original value is known as T2 relaxation. This return of magnetization also follows an exponential decay

$$M_{xy} = M_0 e^{\frac{-\tau}{T2}} \tag{2.9}$$

The strength of the signal produced by the above magnetization components is proportional to the number of protons contained within the voxel weighted by the T1 and T2 relaxation times for the tissues within the voxel. T1-weighted (T1-w) images show better contrast with clear boundaries between different tissues: a fluid shows a very dark contrast, fatty tissues are very bright and water-based are mid-grey as shown in Figure 2.3(a). T2-weighted (T2-w) images are also called 'pathology scans'. In T2-w images, fluids show higher contrast while water based and fat tissues are mid-grey as shown in Figure 2.3(b).

2.2.3 Positron emission tomography

Positron emission tomography (PET) is a nuclear imaging technique that is heavily used in the field of clinical oncology for detecting tumours and tumour metastases, and for the clinical diagnosis of diffuse brain diseases and injuries. PET imaging reveals functional or physiological activities such as blood flow, metabolism and absorption in certain tissues or organs. PET imaging can visualize biological processes via injecting radiotracers in the patient. Radiotracers are designed by connecting positron-emitting radioactive isotopes to molecules with different function in the human body (Becquerel, 1896). Such radiotracers are for example fluorodeoxyglucose (18-F FDG) and [^{11}C] methionine (MET). The PET images are typically acquired contemporaneously with MRI or with CT to produce fused images for more precise expert interpretation. Radionuclides with an excess of protons, decay by positron emission (β^{+} decay) alongside an outgoing neutrino. For example, ^{11}C decays by positron emission:

$$^{11}C \rightarrow {}^{11}B + e^{+} + v \tag{2.10}$$

The positron loses its energy almost immediately when it combines with electrons present in tissues and forms positronium that last for only 10^{-10} seconds before the process of annihilation occurs. In the annihilation process, the mass of the positron and electron is converted into electromagnetic energy. 1.022 MeV of energy is released in the form of two high-energy photons, each carrying 511 keV of energy, as shown in Figure 2.4. These photons have a high probability of escaping the body for external detection. In a typical PET scan, overall 106 to 109 events (decays) are detected externally by detectors surrounding the imaging object that are designed to convert the outgoing high-energy photons into an electrical signal. The line joining the detected location that passes directly through the annihilation point, as shown in Figure 2.4, is referred to as the coincidence line, or line-of-response. Numerous coincidence line form a dataset (also known as line integral data) is then reconstructed utilizing iterative reconstruction algorithms, e.g. the ordered subset expectation maximization (OSEM) algorithm, which is an optimized version of maximum likelihood expectation maximization (MLEM) (Qi & Leahy, 2006).

Once the PET scan is performed, the images defining the radioactivity concentration map of the human body are converted into a quantitative measure known as the standardized uptake values (SUV). In order to account for injection and body weight variability, SUV is commonly defined as follows:

$$SUV = \frac{\text{radioactivity concentration}}{\text{injected dose}} \times \text{body weight} \tag{2.11}$$

2.3 Conventional radiomics

Conventional radiomics aims to characterize the tumour phenotype by extracting quantitative features from medical imaging data. The most extensively used imaging modalities for radiomics

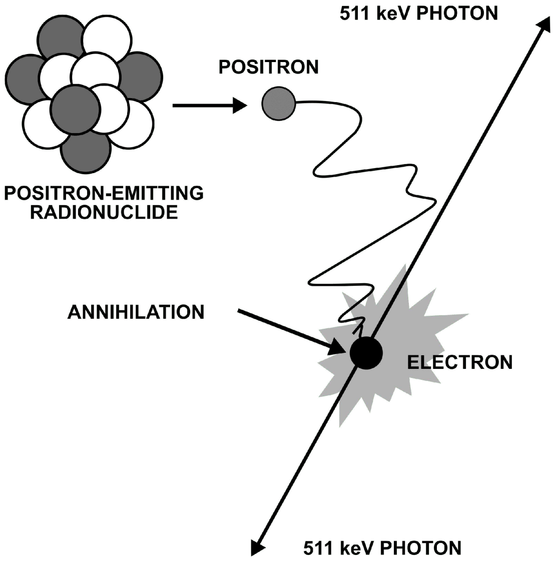


Figure 2.4: An illustration of the process of annihilation in PET. The annihilation process results in releasing 511 keV annihilation photons. Figure reproduced from (Dahlbom & Cherry, 2006) with permission from Springer Nature.

features extraction include CT, MRI, and PET. The essential steps of the radiomics workflow are: (i) image acquisition, reconstruction and segmentation, (ii) image pre-processing and feature extraction, and (iii) feature modelling. Radiomics features are extracted from multimodality medical imaging data using mathematical algorithms. After feature extraction, radiomic analyses employ classical statistics and modern machine learning algorithms to identify features that are correlated to tumour characteristics and are relevant for clinical outcome (Kumar et al., 2012). In the following section, we provide a detailed description of the main steps of the radiomics workflow used within this thesis.

2.3.1 Region-of-interest segmentation

Feature computation in radiomics requires a region-of-interest ROI i.e. segmented region relevant for a study from which features are to be extracted. In the context of this thesis, the clinical target volume (CTV) and the gross tumor olume (GTV) used for radiotherapy treatment planning form the ROI. In clinical settings, the treatment planning ROIs are delineated in a slice-by-slice manner by experienced radiation oncologists.

Medical imaging data comprises a series of stacked two-dimensional (2D) images to accurately represent the volumetric (3D) nature of human anatomy. Given a volume $V(x, y, z)$, the delineation of the ROI leads to the creation of a mask $M(x, y, z)$. Each voxel location (x, y, z) in the ROI is represented as:

$$M(x, y, z) = \begin{cases} 1 & \text{if V(x,y,z) in ROI} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

For radiotherapy treatment planning, the contours for targeted regions and organs at risk are stored. If multimodality imaging is done for the same tumour entity, image registration is used to transfer contours from one imaging modality to another without the need for re-segmentation. This step is carried by rigid image registration, where two sets of imaging volumes from the same patient are aligned anatomically and contours are propagated (Woods et al., 1993; Hajnal et al., 1995). Current clinical treatment planning systems such as Raystation (RaySearch Laboratories, Stockholm) incorporate image registration algorithms with a high degree of flexibility for managing multimodality data.

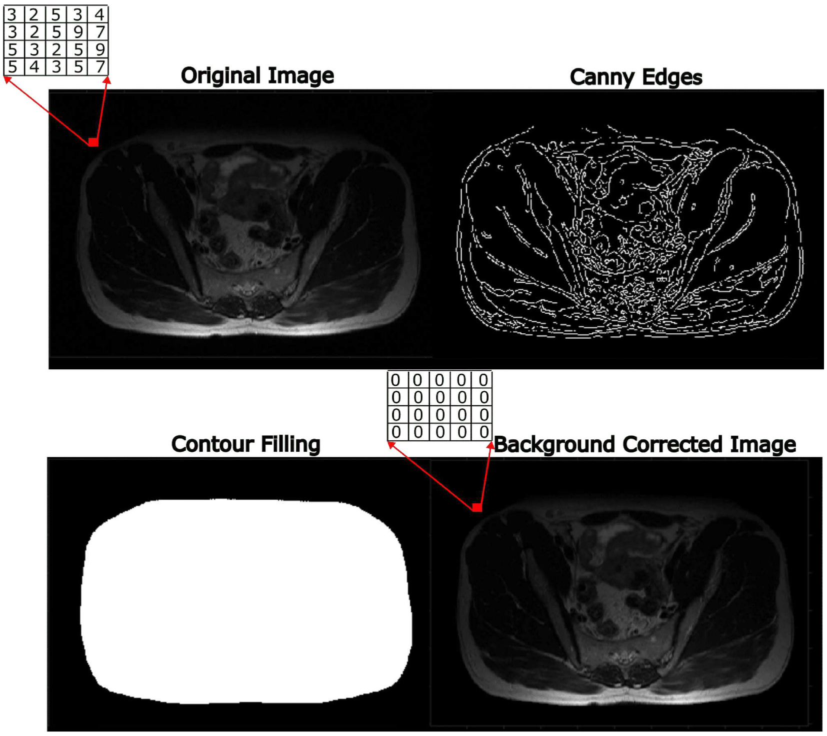
2.3.2 Image pre-processing

Texture-based image analyses are highly dependent on image spatial resolution and noise. Therefore, prior to feature extraction, spatial pre-processing of imaging data must be performed to improve image quality and thus facilitate modelling tasks. Below, we described pre-processing steps that are used within this thesis.

Background suppression

MR images often suffer from background phase variation, which arise due to magnetic field inhomogeneity and result in nonuniformity across and along the boundaries of the scanned organ. These regions contain image pixels (referred to as background pixels in this thesis) around the scanned organ that may decrease the performance of any image processing algorithm. A mask that separates the scanned organ is necessary to turn background pixels to zero. In this thesis, we corrected all MR images for background phase variation. This was achieved by creating a mask of the soft tissue region in the image using the Canny Edge detection algorithm (Canny, 1986) and multiplying the true image with the mask, setting all the background phase variations to zero.

In order to reduce the false detection of edges, the Canny Edge detection algorithm first applies Gaussian filtering for smoothing. In the next step, it applies intensity gradients to compute horizontal, vertical, and diagonal edges. Ideally, edges computed should be thin, which is done by non-maximum suppression. After non-maximum suppression, thresholding is applied to remove weak edges that are irrelevant to desired strong edges that are sufficient to create image boundaries (Canny, 1986). Once strong edges are detected by Canny edge detection, an image mask is created by contour filling (pixels inside contour=1, pixels outside contour=0) and finally



3	2	5	3	4
3	2	5	9	7
5	3	2	5	9
5	4	3	5	7

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Figure 2.5: An illustration of background supression in MRI. The original MR image contains background phase variations, thus pixels in background are > 0. Canny edge detection algorithm detects the edges, which are then contour filled to create a binary image mask. The mask image is finally multiplied with the original image to set background pixels to zero.

this mask is multiplied with the original image to turn the background pixels to zero. An example of the Canny Edge detection algorithm applied to T2-w rectal MR images with background variations is shown in Figure 2.5.

Bias field correction

After background suppression, the next step in the image pre-processing pipeline involves bias field correction of MR images. MR images are often corrupted by biased fields that cause gradual variations in the intensity of the same tissue at different locations across the image, due to difference in coil sensitivities. This spatial smoothing can be represented as a low-frequency signal:

$$S(x, y) = I(x, y)B(x, y) + \eta(x, y) \tag{2.13}$$

where S is the corrupted image, I is the original bias-free image, B is the bias field, and η is additive Gaussian noise. A widely used approach for bias field correction is the N4 bias correction method (Tustison et al., 2010). This method iteratively calculates the smooth multiplicative field by maximizing the high frequency components of the image intensity distribution. In each iteration, the corrected image is computed by using the results of a previous iteration as a prior. In this work, we used the N4 bias correction method before the feature computation step for MRI data.

Image interpolation

Following N4 bias correction, image interpolation is performed. Medical images are often acquired with different voxel size due to different scanner settings across different centres and patients. Therefore, in order to increase reproducibility in radiomics, interpolation of images to an isotropic voxel size is imperative. The dimensions of these isotropic voxels can be decided based by analysing the entire patient cohort. Given an imaging volume $V(x, y, z)$ with voxel size $0.971 \times 0.97 \times 1$ mm^3, the suitable isotropic voxel resampling would be $V_1(x, y, z) = 1 \times 1 \times 1$ mm^3. Consequently, the corresponding ROI mask $R(x, y, z)$ could be resampled to $R_1(x, y, z) = 1 \times 1 \times 1$ mm^3. Three common types of algorithms used for interpolation are: (i) nearest neighbour, (ii) linear, and (iii) cubic interpolation. In this work we used cubic interpolation for resampling imaging volumes while nearest neighbour interpolation was used for resampling the ROI mask.

Image normalization

MR images can be normalized to improve comparability. A relatively simple normalization method is z-score normalization:

$$z = \frac{V_1(x, y, z) - \mu}{\sigma} \tag{2.14}$$

where and are the mean and standard deviation of all voxels within an image, respectively. This type of normalization is sufficient when MR images are acquired with high spatial resolution, however more sophisticated methods such as histogram equalization or normalization by p^{th} percentile can be used when image intensities show more variability. Normalization by p^{th} percentile value of $V_1(x, y, z)$ gives new voxel intensity P:

$$P = \frac{V_1(x, y, z) - minV_1(x, y, z)}{p^{th}percentile} \tag{2.15}$$

In this thesis, we used normalization by the 95^{th} percentile for T2-w MRI in rectal cancer radiomics analysis, i.e. 5% of the highest image intensities were ignored due to the occurrence of outliers, while z-score normalization was applied to T1c-w MRI data for glioblastoma.

For CT, we typically re-segment the ROI mask, so that only soft tissue voxels are used for feature extraction. An intensity range between -150 to 180 HU is reasonable to remove air and bone voxels from the CT scan.

2.3.3 Radiomics feature extraction

Radiomics aims to characterize the tumour phenotype using quantitative imaging features computed and extracted from medical imaging. After feature computation, the resulting features are used to develop either prognostic or diagnostic models. In this section, we describe the main steps and different classes of feature extraction in detail. Theory in this section is adapted from the Imaging Biomarker and Standardization Initiative (IBSI) (Depeursinge et al., 2020; Zwanenburg et al., 2020).

ROI extraction and Image discretization

Before feature calculation, the ROI is separated from the surrounding voxels. This is done using the image ROI mask to keep voxels contained within the delineated ROI, and surrounding voxels are replaced with *NaN*. Given an interpolated image volume $V_1(x, y, z)$ and corresponding ROI mask $R_1(x, y, z)$, the ROI imaging volume is given as follows:

$$V_R(x, y, z) = \begin{cases} \mathrm{V}_1(x, y, z) & \text{if } R_1(x, y, z) = 1 \\ \mathrm{NaN} & \text{otherwise} \end{cases} \tag{2.16}$$

Once the ROI is extracted, for histogram and texture features it is essential to bin or discretize the intensity range within the ROI to reduce computation time and suppress image noise. Given a column vector v_R of image intensity values in the ROI, the discretization process maps v_R to b number of bins in range $[1, N_g]$, where $g = 1, 2, \ldots, N$ are the number of discretized grey-level values. The two common methods of discretization are fixed number and fixed bin size.

Fixed bin number discretization

As the name suggests, the fixed bin number (FBN) method discretizes the v_R intensity values to a fixed number of N_g bins as follows:

$$v_d = \begin{cases} \lfloor N_g \frac{v_R - v_{R,min}}{v_{R,max} - v_{R,min}} \rfloor + 1 & \text{if } v_R < v_{R,max} \\ \mathrm{N}_g & \text{if } v_R = v_{R,max} \end{cases} \tag{2.17}$$

where $v_{R,max}$ and $v_{R,min}$ are the lowest and highest intensity values in the ROI, respectively, and the bin width is given by $\frac{v_{R,max} - v_{R,min}}{N_g}$. Thus, in fixed bin number method, the voxel intensity in v_R is corrected by the minimum intensity value in v_R divided by the bin width. The resulting value is then rounded to the nearest integer. This type of discretization introduces a normalization effect, which may be beneficial for imaging data with arbitrary intensity values, such as MRI.

Fixed bin size discretization

In fixed bin size (FBS) discretization, a new bin with width w_b is assigned to every intensity value in v_R.The bin width w_b starts at the minimum value $v_{R,min}$, which is the user defined lower-bound of the re-segmentation range. The method is recommended when re-segmentation of intensity values is required, e.g. for CT and PET. However, if intensity values are arbitrary as in MRI, the use of fixed bin size discretization is not recommended. Fixed bin size discretization is defined as follows:

$$V_d = \lfloor \frac{v_R - v_{R,min}}{w_b} \rfloor + 1 \tag{2.18}$$

Please refer to the IBSI reference manual for a fully details list of discretization methods (Zwanenburg et al., 2020)

Feature classes

After image pre-processing, radiomics features can be extracted from the defined ROI, e.g., CTV, and GTV delineated on images. Radiomics features can be broadly categorized into three feature classes i.e. (i) morphological or shape-based features, (ii) first-order (FO) statistical and histogram-based features, and (iii) second-order texture (SOT) features. The detailed definition used to calculate different feature classes can be found in the IBSI reference manual (Zwanenburg et al., 2020) and are extracted and calculated according to those guidelines.

Morphological features

Morphological features based on the geometrical aspects of the ROI are calculated from the morphological mask. The common examples of morphological features include volume, surface area, and compactness, related to the size of the ROI. For instance, the ROI volume V can be approximated by counting voxels as follows:

$$V = \sum_{k=1}^{N_v} V_k \tag{2.19}$$

where N_v is a total number of voxels in the morphological mask and V_k is the volume of a voxel. Another feature of the morphological class is the compactness that measures how compact, or sphere-like, the ROI is. It is defined by:

$$C = \frac{V}{\sqrt{\pi A^3}} \tag{2.20}$$

where A is the total surface area of the morphological mask. Please refer to the IBSI reference manual for a fully detailed list of morphological features (Zwanenburg et al., 2020).

First-order features

The FO features can be broadly classified into local intensity features, intensity-based statistical features, intensity histogram features and intensity volume histogram features. All these features provide information about intensity distributions in the ROI. The local intensity features are computed by defining a neighbourhood around the centre voxel, where the centre voxel must be located within, ROI and the neighbourhood can be defined anywhere around the ROI. Local and global intensity peaks are two examples of local intensity features.

Intensity-based statistical features describe how continuous intensity values are distributed within ROI. An example of an intensity-based statistical feature is the mean intensity:

$$M = \frac{1}{N_v} \sum_{i=1}^{N_v} v_{R,i} \tag{2.21}$$

where v_R is the set of intensities in N_v voxels.

Intensity histogram features describe how discretized pixel intensity values are distributed within ROI. The intensity histogram features are obtained by discretizing the intensity distribution v_R into bins represented by, v_d as explained previously. The mean of the discretized intensities is calculated as follows:

$$M = \frac{1}{N_v} \sum_{i=1}^{N_v} v_{d,i} \tag{2.22}$$

Intensity volume histogram features describe the relationship between an intensity value i and the volume fraction v that contains the intensity value, equal or greater than i. Please refer to the IBSI reference manual for a fully detailed list of first-order features (Zwanenburg et al., 2020).

Second-order features

The SOT features characterize spatial patterns within the ROI of imaging volumes. Texture features describe image heterogeneity by analysing spatial groupings of discretized intensities, and are thus capable of expressing separation or clustering between different parts of tissue in the ROI. Texture features are computed from texture matrices derived from either 2D images (slice) or 3D volumes. In this thesis, six distinct types of texture matrices were extracted from the 3D ROI: the grey level co-occurence matrix (GLCM), grey level run length matrix (GLRLM), grey level size zone matrix (GLSZM), grey level distance zone matrix (GLDZM), neighbouring grey tone difference matrix (NGTDM), and neighbouring grey level dependence matrix (NGLDM).

For instance, the GLCM is one of most classical second-order statistical methods for texture analysis introduced by Haralick et al. (Haralick et al., 1973). In simple words, GLCM is a tabulation of how often different combinations of gray levels (i, j) co-occur distributed within the considered ROI. Therefore, it is also referred as co-occurrence distribution matrix. The two main parameters of GLCM are direction (θ) and distance (d). The distance metric used for computing texture features is the Chebyshev distance. Pixel pairs are then analysed in a specified direction and their frequency is recorded in a square matrix. The computed matrix is then normalized to convert frequencies into probabilities. A GLCM element, denoted by pixel pair $p(1,2)$ and direction 0°, will correspond to the number of pixel pairs that were found in an image with gray-level values 1 and 2 in the horizontal direction (Figure 2.6). Generally, there are 8 and 26 connected neighbourhoods in 2D images and 3D volumes, respectively. Therefore, it is possible to compute multiple GLCM matrices for a single image; one for each distance and direction pair, which can be aggregated using either averaging or merging (Further details on aggregation methods are available in IBSI reference manual). One example of a second-order statistics calculated from GLCM is entropy that measures the randomness of co-occurring grey levels within the image, with higher values indicating higher heterogeneity. It is calculated as follows:

$$E = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} p(i,j)\log p(i,j) \tag{2.23}$$

where N_g are distinct grey-levels in the discretized intensity mask and $p(i,j)$ is the $(i,j)_{th}$ entry in normalized GLCM.

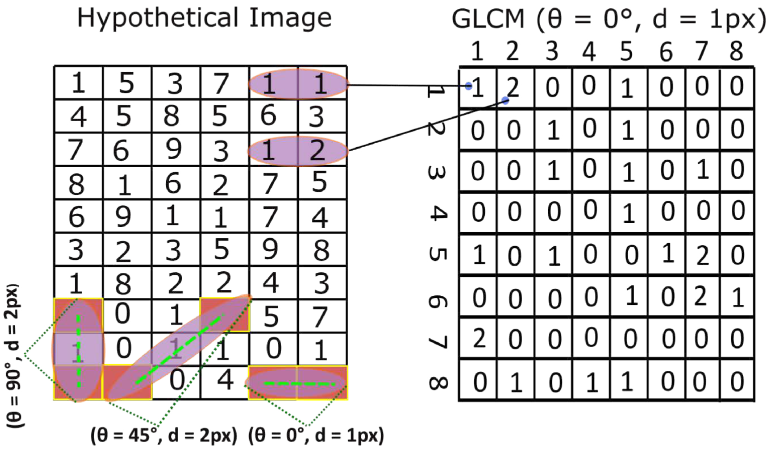
The GLRLM texture matrix was proposed by Galloway (Galloway, 1975), It captures the information about runs of grey-level values in a particular direction. Longer runs indicate finer texture, while the relatively shorter runs characterize coarser textures in the observed ROI. Similar to GLCM, GLRLM are also calculated for all possible (d,θ) pairs. Assuming that $p(i,j) = \frac{r_{i,j}}{N_s}$, where $r_{i,j}$ is the number of times there is a run of length j having a grey level i, $N_s = \sum_{i=0}^{N_g-1}\sum_{j=0}^{N_r-1} r_{i,j}$ is the total number of runs in the ROI, where N_r is the maximum number of runs along a direction θ.

Run entropy, a feature that can be derived from the GLRLM can be computed as follows:

$$E = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_r-1} p(i,j)\log p(i,j) \tag{2.24}$$

Discretized intensities or voxel values can also be identical, that result in formation of linked groups or zones. GLSZM counts the number of such zones of linked voxels (Thibault et al., 2013). In order to capture the relationship between grey level values and location, GLDZM can be computed that counts the number of zones of linked voxels that have specific grey level value and same distance to the ROI edge.

Finally, there are neighbourhood based matrices that either expresses the sum of grey level differences of central pixel/voxel from their neighbouring pixels/voxels (NGTDM) (Amadasun & King, 1989) or how neighbouring pixels/voxels are distributed around central pixel/voxel (NGLDM) (Sun & Wee, 1983) in the discretized ROI. Please refer to the IBSI reference manual for a fully detailed list of second-order features (Zwanenburg et al., 2020).



1	5	3	7	1	1
4	5	8	5	6	3
7	6	9	3	1	2
8	1	6	2	7	5
6	9	1	1	7	4
3	2	3	5	9	8
1	8	2	2	4	3
	0	1		5	7
1	0	1	1	0	1
		0	4		

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	1	0	1	0	1	0
4	0	0	0	0	1	0	0	0
5	1	0	1	0	0	1	2	0
6	0	0	0	0	1	0	2	1
7	2	0	0	0	0	0	0	0
8	0	1	0	1	1	0	0	0

Figure 2.6: An illustration of two important parameters, pixels pair distance d and direction θ, for the calculation of the grey level co-occurrence matrix (GLCM). A hypothetical image and its corresponding GLCM matrix, assuming a horizontal direction of analysis and a 1-pixel (px) distance, are shown here.

Transformed features

After image pre-processing, radiomics features can be extracted from the defined ROI e.g., CTV, and GTV delineated on base images or transformed images. Transformed images are obtained by applying so-called convolutional filters on the base image after image interpolation. It is worth mentioning that morphological features are extracted from the base image only and not from transformed images, as geometrical aspects of the ROI remain unaffected by image transformation. The two main types of transformation that are widely used in the literature are undecimated discrete wavelet transform (UDWT), and Laplacian of Gaussian (LoG) transforms. In this thesis, we will focus on features extracted from baseline and LoG transformed medical images. Below, we discuss the basic definition of feature classes. The theory is adapted from IBSI reference manual (Depeursinge et al., 2020).

Laplacian of Gaussian

LoG is a band-pass circularly symmetric convolutional filter that computes the second spatial derivate of the image $I(x, y)$ after applying Gaussian smoothing filter. In medical imaging, noise shows itself as a high frequency signal. The source of noise in CT is electronic noise of detection system or quantum noise from X-ray photons while in MRI, the noise is in images is primarily the thermal noise added due to resistance of receiver coils. Therefore, a smoothing operation is required that helps to remove the noise and large variations of signal that can be detected in image slices. The 2D LoG convolution filter function with Gaussian standard deviation is given as follows:

$$LoG(x, y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2 + y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.25}$$

where x and y are coordinates of pixels surrounding the central pixel on which the equation is used to calculate the convolution. The value of is particularly important in emphasizing image details. A small value of emphasizes fine image details, whereas larger values highlight coarser image details. The LoG filter profile depends on the 1D radius $||r||$ that corresponds to the radial second-order derivative of a general D-dimensional Gaussian filter. Hence, for a D-dimensional LoG filter, the above equation can be written as

$$LoG(r) = -\frac{1}{\sigma^2}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^D\left(D - \frac{||r||^2}{\sigma^2}\right)e^{-\frac{||k||^2}{2\sigma^2}} \tag{2.26}$$

The application of the LoG filter on images enhances the short-range differences in pixel intensities and reveals the texture appearance according to the value. In this work, 3D LoG filters were used to transform 3D volumetric data.

2.3.4 Radiomics modelling pipeline

After image pre-processing and feature extraction, the next steps within the computation pipeline of radiomics comprises feature selection and machine learning based predictive modelling to predict the outcome of interest. In the following section, modelling steps are described more in detail.

Robustness

The previously mentioned features from different feature classes are sensitive to changes in the experimental conditions used for image acquisition and must be assessed for repeatability and reproducibility. ‘Repeatability’ refers to features that remain the same when imaging is performed multiple times in the same subject. ‘Reproducibility’ refers to features that remain the same when imaging is performed using different software applications, image acquisition settings or imaging performed at different clinics. Studies have shown that features that lack repeatability and reproducibility (non-robust features) deteriorate the performance of machine learning models and increase the risk of false predictions when applied to new data (Traverso et al., 2018). Therefore, careful filtering of non-robust features is very important for reducing the risk of poor generalizability of radiomics models. Conventionally, feature robustness is assessed by computing features from test-retest imaging (Tixier et al., 2012; Leijenaar et al., 2013; Balagurunathan et al., 2014a; van Velden et al., 2016). In test-retest imaging, the imaging process is repeated after a few hours or days, acquiring similar but not identical images. Thus, the features extracted are also not identical and have reduced intra-class correlation. This helps to identify non-robust features that change significantly due to small variations in imaging conditions.

However, acquiring images under test-retest condition for every radiomics study is not clinically feasible. Therefore, in order to mimic the perturbations induced by difference in imaging conditions, various handcrafted perturbations, e.g., rotation, translation, noise, volume adaptation, and contour randomization, are applied to acquire single images as proposed by Zwanenburg et al. (Zwanenburg et al., 2019a). Rotation perturbs the image and corresponding segmentation mask by rotating them at specified angles $\theta \in [-\theta^{\circ}, \theta^{\circ}]$ along the temporal axis (*z*-axis). When perturbing the image by translation, both mage and mask are shifted by a specified fraction η of isotropic voxel spacing along *x*, *y* and *z*. The translation fraction is permuted over the different directions, thus generating a new image for each permutation. Noise is an additive perturbation drawn from a normal distribution with mean 0 and standard deviation equal to the noise present in the images. Volume adaptation expands and/or shrinks the ROI by a specified fraction. Contour randomization is based on linear iterative clustering (Scharstein & Pal, 2007) and perturbs the mask by randomly selecting supervoxels based on the overlap with the original mask. Noise and contour randomization can be repeated multiple times, and each repetition generates a new image. Similarly, each rotation angle and volume adaptation lead to the creation of a new image and ROI mask.

The pipeline for introducing image perturbation is incorporated in the radiomics feature extraction software MIRP (Zwanenburg et al., 2019a; Zwanenburg et al., 2019b). Feature computation is then executed on base images resulting in baseline or unperturbed features, and features extracted with combinations of perturbations. Robustness of each perturbed feature is then evaluated on a scale of 0-1 using the intraclass correlation coefficient (ICC) (1,1)(Shrout & Fleiss, 1979). The ICC value of 1 indicates that the feature is fully repeatable between perturbations, while lower values indicate increasing variance between different perturbations and thus lower repeatability. Normally, ICC values are computed with a 95% confidence interval (CI). Within this thesis, features with the lower boundary of the 95% CI of the ICC below 0.80 are considered as not reproducible. After the filtering step, the baseline version of the reproducible features is used for modelling.

Clustering

Considering different feature classes as mentioned previously, it becomes obvious that hundreds of radiomics features can be extracted from a single image. Features may show high degrees of correlation with each other and this increases the risk of model overfitting resulting in unfit model coefficients and false predictions or type I error (Chalkidou et al., 2015). Therefore, after performing a preliminary analysis for selecting robust features, the feature space can be further reduced by correlation and redundancy analysis (Balagurunathan et al., 2014b; Parmar et al., 2015). Cluster analysis refers to the discovery of distinct and non-overlapping subpopulations within a large population (Jain & Dubes, 1988). Cluster analysis for the radiomics features aims to create groups (clusters) of similar features with high redundancy. There are different types of clustering algorithms such as k-means clustering (McQueen, 1967; Hartigan, Wong, et al., 1979), agglomerative hierarchical clustering (Hastie et al., 2001), and model-based probabilistic clustering (Titterington et al., 1985; Banfield & Raftery, 1993; Cheeseman, Stutz, et al., 1996). However, in this work, we focused on agglomerative hierarchical clustering due to its intuitive appeal and its data visualization properties. Hierarchical clustering starts with as many clusters as there are features. The number of clusters is reduced by one at each step by combining two clusters based on some optimization criteria. The most used criterion for merging clusters is dissimilarity between two clusters. In single linkage (smallest dissimilarity) the distance between two clusters is represented by the minimum distance between all possible pairs of features in clusters. In average linkage (average dissimilarity) the distance used is the average of all pairs of features and in complete linkage (maximum dissimilarity) the distance is the maximum between all possible pairs of features in two clusters. Several distance metrics can be used, such as Euclidean distance or correlation dissimilarity. In this work, average linkage in combination with Spearman correlation (ρ) was used for optimization and $\rho \leq 0.8$ was defined as a threshold for placing features into the same cluster after hierarchical clustering. The feature with the highest mutual information with the endpoint was selected as the representative for each cluster and used for further modelling.

Feature transformation and normalization

After selecting non-redundant features for downstream modelling of some endpoint of interest, we attempt to improve feature distributions. Imaging features might vary in their distribution due to skewness. In order to map features to a more central, normal, distribution, a simple log-transformation, more complex Box-Cox (Box & Cox, 1964) or Yeo-Johnson transformations (Yeo & Johnson, 2000) from a parametric family of transformations can be used. Within this thesis, features were transformed using the Yeo-Johnson transformation, which is defined as:

$$\psi(\lambda, y) = \begin{cases} ((\text{y+1})^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y + 1) & \text{if } \lambda = 0, y \geq 0 \\ [(\text{-y+1})^{2-\lambda} - 1] & \text{if } \lambda \neq 2, y < 0 \\ \text{-}\log(-y + 1) & \text{if } \lambda = 0, y < 0 \end{cases} \tag{2.27}$$

where y is a value of a feature that can be negative and is a transformation parameter which optimized via maximum likelihood estimator. Yeo-Johnson transformation is also known as a power transformation, as the variable y is raised to a particular power. Figure 2.7 shows the distribution of an example feature before and after applying the Yeo-Johnson transformation. After applying the power transformation, a widely used strategy to is to standardize feature values to have a zero-mean and unit-variance. This can be achieved by applying z-score normalization, defined by

$$y' = \frac{y - \mu(y)}{\sigma(y)} \tag{2.28}$$

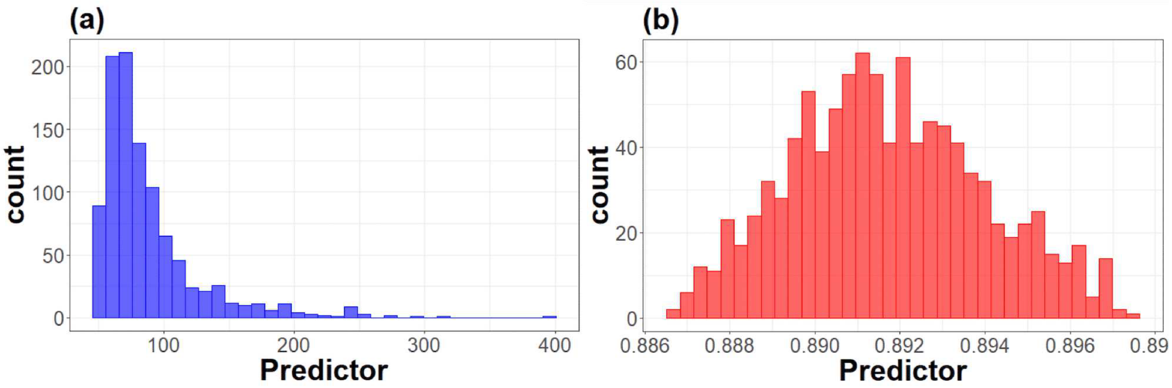


Figure 2.7: An example of a skewed feature before (a) and after (b) applying the Yeo-Johnson transformation. Code and data taken from (Kuhn & Johnson, 2019).

where $\mu(y)$ and $\sigma(y)$ are the mean and standard deviation of the variable y.

2.4 Prognostic and diagnostic modelling in oncology

Clinical decision support tools can be divided into diagnostic and prognostic tools. Diagnostic prediction models may assess the probability of disease when it is already present, while prognostic models may evaluate the risk of a particular health state (e.g., tumour recurrence) occurring in the future. In the field of machine learning, dedicated algorithms are available for both types of modelling. In this thesis, we used survival analysis paired with regression-based machine learning models to predict prognostic endpoints, while for patient diagnosis we used binary classification models. In this section, we describe some basics of prognostic analysis.

2.4.1 Basics of survival analysis and Cox regression

The primary focus of survival modelling is to build a statistical model for prognosticating survival or the hazard of an event for an individual patient that is part of a patient population. One of the challenges in creating statistical models for time-to-event data or survival endpoints is that the period of observation can be censored for some individuals i. In clinical studies, censoring (more specifically right-censoring) occurs when the event is not observed, while the patient is being monitored. The target attributes for survival analysis also comprise the event indicator function δ_i, which is 1 when an event is observed and 0 for right censored observations. Given a set of features or covariates x_i, the survival time or hazard function can be predicted with survival analysis (Cox & Oakes, 1984).

The survival time distribution is characterized by a survival function with a probability density function $f(t)$ and a cumulative distribution function $F(t)$

$$S(t) = P[T > t] = 1 - F(t) = \int_t^\infty f(x)\,dx \tag{2.29}$$

where t ranges from 0 to infinity. $S(t)$ is the probability that an individual is still alive at time t. Theoretically, the survival function can be portrayed as a smooth curve, as shown in Figure 2.8(a). The survival function is non-increasing with $S(0) = 1$, as no event can occur prior to $t = 0$. For a limited dataset, the survival curve is estimated as a step function as shown in Figure 2.8(b) through use of the Kaplan-Meier estimator. $S(t)$ is accompanied by a hazard function, which is the instantaneous rate of event and defined as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr[t < T \leq t + \Delta t | T > t]}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.30}$$

where - $f(t)$ is the derivative of $S(t)$, which suggests writing Equation 2.29 as

$$\lambda(t) = -\frac{d}{dt} \log S(t) \tag{2.31}$$

Integrating Equation 2.31 from limits 0 to t with the boundary condition $S(0) = 1$, we can obtain the probability of surviving to duration t as a function of hazard at all durations up to t:

$$S(t) = exp(-\int_0^t \lambda(x)\, dx) \tag{2.32}$$

$$\Lambda t = \int_0^t \lambda(x)\, dx \tag{2.33}$$

where Λt is called cumulative hazard function, and thus it holds that

$$S(t) = e^{-\Lambda(t)} \tag{2.34}$$

One method of studying the effect of covariates in survival models is the Cox proportional hazard models (CPHM).CPHM assumes that a unit increase in a covariate has a multiplicative effect on the hazard function, i.e. the relative hazard function $e^{\beta^T x}$ (β being a set of unknown regression parameters) is multiplicative with the baseline hazard $\lambda_0(t)$

$$\lambda(t|x) = \lambda_0(t) e^{\beta^T x} \tag{2.35}$$

where $\lambda(t|x)$ is the hazard function of a person with covariates x. The CPHM is a semi-parametric model because the baseline hazard $\lambda_0(t)$ remains unspecified, while the parameter estimate in the CPHM is carried out by maximizing the Cox's partial likelihood of the coefficients.

$$L(\omega) = \prod_{T_i uncensored} \frac{e^{\beta^T x_i}}{\sum_{T_j \geq T_i} e^{\beta^T x_j}} \tag{2.36}$$

The above equation provides the probability that one event occurred for the i_{th} individual at time T_i with respect to the remaining individuals with $T_j \geq T_i$ still at risk of an event.

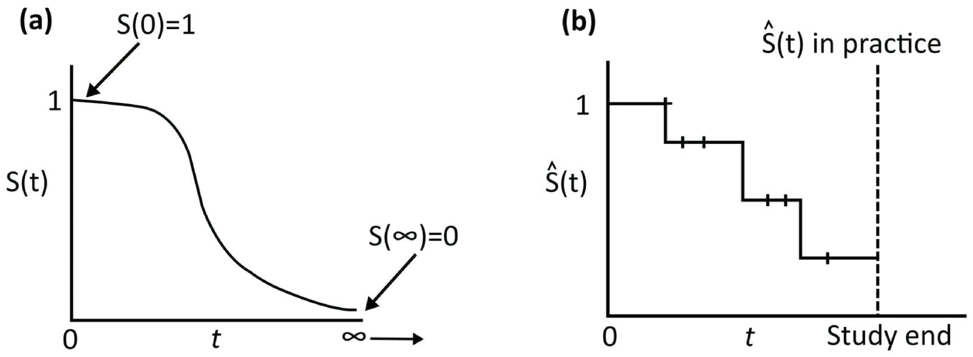


Figure 2.8: An illustration a of survival function. (a) Theoretically, the survival function is given as a decreasing smooth curve. (b) In practice, we usually obtain estimated survival curves that are step functions rather than smooth curves (Kaplan-Meier estimate).

2.4.2 Logistic regression

One of the commonly used algorithms for diagnostic analysis with binary outcome labels is logistic regression (Cox, 1958). Logistic regression essentially estimates the probabilities of outcome by determining the relationship between features and outcome. For the binary classification problem, conditional probabilities of outcome labels y_i (where i=1,2,3,. . . , n) are computed, given features $X = x_{ij}$ (where j=1,2,3,. . . , k represents available features). $P(y|X)$ is approximated by a sigmoid function applied to a linear combination of features as follows:

$$P(y|x) = \frac{1}{1+e^{-z}} \text{ with } z = \beta^T X = \sum_{j=1}^{k} \beta_j x_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (2.37)$$

The parameters β_j of the logistic function can be estimated using maximum likelihood estimation that searches for the parameters that best fit the joint probability of features.

2.5 Performance evaluation

In this section, two types of performance metrics are explained. The first metric is the concordance index (C-index) that is used to evaluate the performance of survival models. The second type of metrics contain the confusion matrix and the reciever operating curve (ROC) curve that are used to evaluate the performance of binary classification models.

2.5.1 Concordance index

The C-index (Harrell et al., 2001)is a generalization of the area under the ROC curve (AUC) for time-to-event survival models, while taking into account the censored data. The survival time of two subjects (i and j) can be assessed only if (1) both of them had an event or (2) if the censored survival time (T_i) of one subject is greater than the uncensored survival time (T_j) of the other subject. Thus, the C-index can be interpreted as the fraction of all possible subject pairings whose predicted and actual survival times (or hazards) are orders in the same way. It can be written as:

$$C = \frac{1}{|N|} \sum_{T_i uncensored} \sum_{T_j > T_i} 1_{f(x_i) < f(x_j)} \tag{2.38}$$

where $|N|$ is the number of all pairs, f is the model applied for prediction and $f(x_i)$ is the time predicted for i_{th} the subject. The indicator function in above equation $1_{a<b} = 1$ if $a < b$ and 0 otherwise. The value of the C-index ranges between 0 and 1: a value of 0.5 indicates that the predicted values do not reflect the correct order of events according to their observed survival, whereas the values 0.0 and 1.0 indicate that the predicted values reflect the correct order of events according to their observed survival, with the value of 0.0 being applicable to predicted values that have an inverse relationship to survival.

2.5.2 Confusion matrix

In a binary classification problem, the given data point can either be positive with an assigned label of 1, or negative with label 0. Hence, there are four possible combinations of predicted and true label: If a positive label was observed, and it is classified as positive by a classifier, it is referred as true positive (TP), whereas if it is classified as negative by a classifier, it is referred as false negative (FN). Similarly, if a negative was observed, and it is also classified as negative by a classifier it is referred as true negative (TN), and if it is classified as positive by a classifier it is referred to as false positive (FP). These combinations can be arranged as a confusion matrix, as shown in Figure 2.9(a). The confusion metric can be used to compute a number of different metrics. However, an important aspect that should be considered before computing any evaluation metric is class imbalance. Class imbalance refers to the skewed distribution of class labels, with the most frequent class referred to as majority class. Some metrics, such as accuracy, may produce misleading values when class imbalance is present. Below we will discuss the metrics that were used in this work and the effect of class imbalance on each of them.

2.5.3 Sensitivity

Sensitivity is the ratio of correctly classified positive samples to the total number of actual positive samples and hence it is also referred as true positive rate (TPR):

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.39}$$

When positive samples are disproportionally more frequent, a simple naive classifier that always predicts the majority class yields a high sensitivity value. This indicates that sensitivity may be optimistically biased under class imbalance.

2.5.4 Specificity

Specificity is the ratio of correctly classified negative samples to the total number of actual negative samples, hence it is also referred as true negative rate (TNR):

$$Specificity = \frac{TN}{TN + FP} \tag{2.40}$$

When negative samples are disproportionally higher, generally the classifier would be biased towards the negative class and thus specificity would be high.

2.5.5 Accuracy

Accuracy is defined as the ratio of correctly classified samples (TP, TN) to the total number of samples:

$$Accuracy = \frac{TP + TN}{FP + TP + TN + FN} \tag{2.41}$$

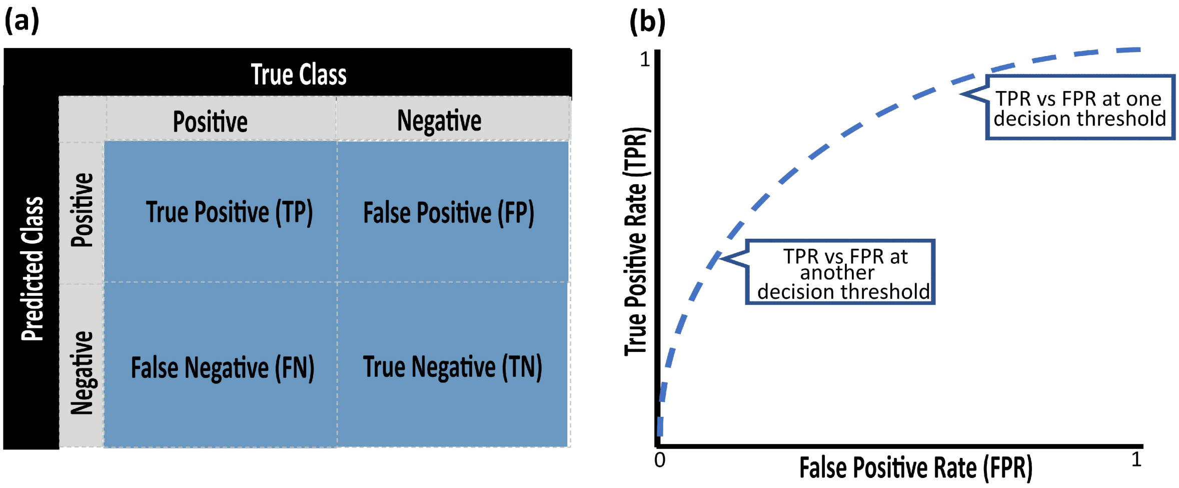
Accuracy treats all error types 8 (FP, FN) as equal. However, equal is not always preferred, in particular when the samples are imbalance in terms of positive and negative class distribution.

2.5.6 Area under the ROC curve

Finally, another metric summarizing the performance of a binary classifier is the AUC. After fitting a classifier for a binary classification problem on training data, the model is applied on test data and predictions are made in the form of probabilities or scores, which are then transformed into normalized probabilities. Each of these probabilities is then used as a threshold to convert predicted probabilities into class labels. An ROC curve is a plot of TPR against false positive rate (FPR) (1-specificity) for all possible threshold values. As shown in Figure 2.9(b) reducing the threshold value decreases both TPR and FPR. AUC measures the 2D area underneath the ROC curve, thus providing an aggregate performance measure across all possible thresholds. It is measured on the scale of 0 to 1 where a value of 0.5 indicates that the predicted values do not reflect the correct labels, whereas the values 0.0 and 1.0 indicate that the predicted values reflect the correct labels, with the value of 0.0 being applicable to predicted values that have an inverse relationship to labels. AUC measures the quality of a model's prediction irrespective of selected thresholds, therefore it is generally a desirable evaluation metric in presence of imbalanced data.

2.6 Machine learning and deep learning

In this section, we provide an overview of feature selection methods together with machine learning algorithm and basic principles and terminologies of DL used in this thesis.


(a)
(b)
1
True Positive Rate (TPR)
TPR vs FPR at one decision threshold
TPR vs FPR at another decision threshold
0
False Positive Rate (FPR)
1

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 2.9: An illustration of (a) Confusion matrix, (b) receiver operating curve (ROC) at different threshold values.

2.6.1 Feature selection methods

In radiomics analysis, it is essential to use only the most important features for model building to reduce overfitting of models and for improved model interpretability. This can be achieved with the help of feature selection methods, which assess the importance of individual features.

Mutual information maximization

The mutual information maximization (MIM) method estimates the correlation between class labels y and features x_i using a linear approximation (Gelfand & I A glom, 1959). The mutual information (MI) is defined as

$$I(x_i, y) = -\frac{1}{2}\ln(1 - \rho(x_i, y)^2) \tag{2.42}$$

where I is the mutual information and is a correlation coefficient between x_i and y. For a binary classification problem, the correlation between a continuous feature x_i and binary labels y is computed using either Spearman's (Spearman, 1910) or Pearson's correlation (Rodgers & Nicewander, 1988). In case of time-to-event data, correlation is based on the C-index such that $\rho(x_i, y) = 2(\text{C-index} - 0.5)$.

Minimum redundancy maximum relevance

The minimal redundancy maximum Relevance (mRMR) method seeks to identify the subset S of features that have maximum predictive power when put together but also minimum intercorrelation (Peng et al., 2005). This is achieved by selecting the feature that maximizes the score (f) using greedy forward selection based on MI (I) as follows:

$$f(x_i, y) = argmax \left(I(x_i, y) - \frac{1}{|S|} \sum_{S_j \in S} I(x_i, S_j) \right) \tag{2.43}$$

where S is the number of already selected features.

Univariate and multivariate regression

For each feature (univariate) or subset of features (multivariate), logistic regression and Cox regression models are built for binary classification problems and for time-to-event endpoints, respectively. These models are then assessed on a holdout test set, and ranked according to AUC values for logistic regression and C-index for Cox regression models.

Elastic net

The standard algorithms for binary and time-to-event outcome are logistic regression and Cox regression, respectively, that assume a linear relationship between features x and output y. An extension to these algorithms involves adding penalties or regularization to the loss function, which encourages simpler models that have smaller coefficient values. The elastic-net (EN) algorithm is a type of regularized model that combines two types of penalties, i.e. L1 and L2 penalty functions (Zou & Hastie, 2005). L1 regularization adds a penalty equal to the absolute value of the magnitude of the coefficients of the model. A regression model that uses L1 regularization is called Lasso regression. L2 regularization minimizes the size of coefficients but no coefficient is turned exactly to zero, thus preventing coefficients from getting removed from the model. The final estimates from the EN method with L1 and L2 regularization for regression (linear or survival) model is defined as

$$E = Loss + \lambda(\alpha \sum_{j=1}^{k} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{m} \beta_j^2) \tag{2.44}$$

where $\beta_j = \beta_1, \beta_2, ..., \beta_k$ are the coefficients of the model and λ denotes the amount of shrinkage. When $\lambda = 0$ no feature is removed and when $\lambda = \infty$, all features are removed. penalty term that decides how much we want to penalize the model. When $\alpha = 1$ the loss function reduces to L1 regularization and when $\alpha = 0$ the loss function reduces to L2 regularization.

2.6.2 Machine learning algorithms

Machine learning algorithms are used to find the mapping between input features and an outcome of interest. In this section, we describe the more advanced machine learning algorithms used in this thesis.

Random forest

The random forest (RF) (Breiman, 2001) is a machine learning algorithm that leverages the power of multiple decision trees to map input features to outcome labels. Decision trees (Magee, 1964) are developed incrementally by breaking down the dataset into smaller subsets, resulting in a tree with leaf nodes and decision nodes. Each decision node has two or more branches, while each leaf node represents the final decision of classification or regression for the subset of samples contained within this node. Important features are on average selected early in the tree-forming process, depending on hyperparameters of the algorithm like the number of features assessed for each node. One of the widely used methods for selecting features on each node is called the Gini index (G). In case of a binary classification problem, it is calculated as follows:

$$G = 1 - P^2(y_i = 0) - P^2(y_i = 1) \tag{2.45}$$

where P is the probability of a data point being classified for a distinct class. The Gini index measures how effectively a predictor splits the mixed classes into two or more groups. However, decision trees tend to perform worse if complex interactions are present in the data and small variations in the data can generate very different looking trees, resulting in high variance. RFs tend to eliminate some of these limitations. RF methods reduce the variance of individual decision trees by constructing M different decision trees, each with its own predictions for outcome. The final predictions are then ensembled and one of the commonly used ensembling method for getting final model predictions is majority voting. This concept can be represented by the following function

$$f(x_i) = \begin{cases} 1 & \text{if } \frac{1}{M}\sum_{m=1}^{M} f_N > \frac{1}{2} \\ 0 & \textit{otherwise} \end{cases} \tag{2.46}$$

where M is the number of trees, f_N is a tree with n training examples and x is the feature vector. In case of binary classification, f_N is a measurable function of the training data and feature vector x_i, that is used to estimate label y. Moreover, feature subsets are randomly created when growing RF, and the splitting algorithm searches for the best feature in this subset. This results in a wider diversity of features being used, which may help create a better model.

For time-to-event problems, a random survival forest (RSF) is used (Ishwaran et al., 2008). The main difference between RF for classification and RSF for estimating survival endpoints is that the feature split on root nodes of the decision tree is created by maximizing the log-rank test statistic instead of computing the Gini index.

Boosted gradient linear and tree models

Gradient boosting is a type of ensemble method that combines output from multiple weak learners with the aim of creating an ensemble model that is better than any of the underlying weak learners. New learners are crated to correct the residual errors in the predictions from previous learners. The gradient in gradient boosting refers to the gradient of the loss function, which is the target value for each new learner to predict. If we let $F(X, \theta)$ be a function that maps input features x to output labels y, then the boosting procedure fits an additive model of the form:

$$F(X, \theta) = \sum_{m=1}^{M} \beta_m f_m(x, \theta_m) \tag{2.47}$$

where $f_m(x, \theta_m)$ is the individual learner, θ_m is a set of parameters and β_m is the coefficient of the m_{th} model. Multiple learners are created in a consecutive manner, each of them minimizes the loss function over training data by considering the output of previous learners. Thus, boosting is a forward stage-wise additive technique. In case of gradient boosting the optimization method is the following minimization using the loss function L:

$$\widehat{f_m}(X) = \underbrace{min \sum_{i=1}^{n} L(y_i, \sum_{m=1}^{M} \beta_m f_m(x, \theta_m))}_{\{\beta_m, \theta_m\}} \tag{2.48}$$

One of the most popular methods of gradient boosting is Extreme Gradient Boosting (XGB) (Chen & Guestrin, 2016) that still minimizes the same loss function, however, it performs second order Taylor expansion to gain analytical tractability. In this thesis, we used XGB linear and XGB tree models for binary classification and time-to-event outcome data.

2.6.3 Deep learning

In this thesis, CNN are used for the diagnostic and prognostic studies on GBM. Therefore, in this section, we provide a summary of key concepts in DL. For a more detailed and in-depth discussion on deep neural network (DNN) and CNN, please read the referred literature to (LeCun et al., 2015; Weidman, 2019).

Neural network

Neural networks are inspired by networks of neurons in the human brain. Therefore, each computational unit in a neural network is also referred to as a neuron. Multiple neurons are organized in layers to form a neural network (Figure 2.10). A typical (shallow) neural network consists of an input layer that takes in the data, one hidden layer that perform transformations on the data and an output layer that maps these transformations to the results. To get meaningful predictions, neural networks can be trained to recognize the patterns in the data. Each neuron in a neural network takes one or more inputs and computes an output as follows:

$$ouput = h(w^T x + b) \tag{2.49}$$

where x represents an input vector, w is the weight vector, b is the bias and h is a non-linear activation function. During computation, every input value in x is multiplied with a weight in w and finally offset is provided to data by adding bias b. The linear combination thus formed is transformed via a non-linear activation function h to produce the final output. Common choices for non-linear activation function are the sigmoid function (Bishop & Nasrabadi, 2006):

$$h(x) = \frac{1}{1 + e^{-x}} \tag{2.50}$$

the hyperbolic tangent function:

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.51}$$

and the rectified linear unit (ReLU):

$$h(x) = max(0, x) \tag{2.52}$$

A single hidden layer neural network with a linear combination of N individual neurons can approximate any continuous function $\hat{f}(x)$ as follows:

$$\hat{f}(x) = \sum_{i=0}^{N-1} v_i h(w_i^T x + b_i) \tag{2.53}$$

where v_i is the combination weights between neurons. A DNN can be composed by combining more hidden layers that are connected to each other. Due to the multi-layer architecture, deep neural networks are capable of representing features in a hierarchical fashion, i.e. layers close to the input extract lower-level or minor features from the input data, while layers close to the output extract higher-level or more detailed information. The ability to represent features as a non-linear combination of lower level features sets them apart from conventional machine learning methods (Bengio et al., 2009). A neural network with multiple layers can be defined as:

$$\hat{f}(x;\theta) = (f_m, \ldots, f_1)(x) = h^m(h^{m-1}(\ldots(h^2(h^1(w_1^T x + b_1) + b_2) + \ldots) + b_{m-1}) + b_m) \tag{2.54}$$

where θ defines parameter set with weights and biases and these parameters are learned during the training process of the neural network. The aim of parameter learning is to reduce the error $E(\theta)$ between ground truth and predicted values, however, this cannot be achieved by finding an analytical solution of the equation $\nabla E(\theta) = 0$. This is because the optimization problem is non-convex with multiple local minima beside one global minimum. Therefore, iterative numerical

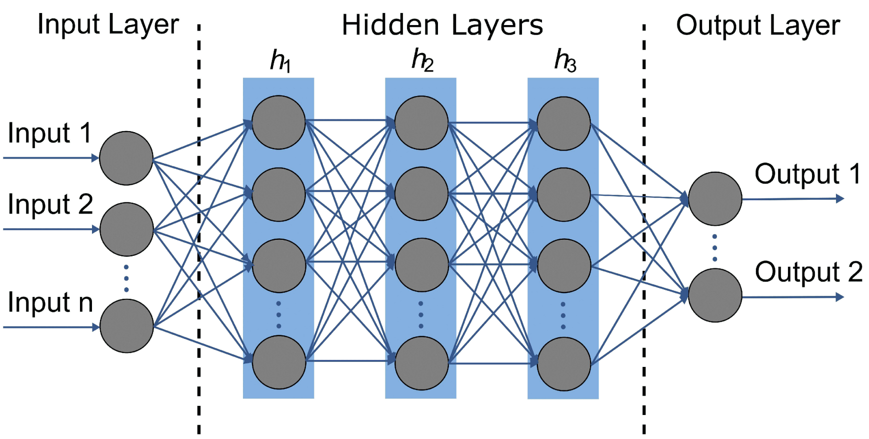


Figure 2.10: An illustration of a basic neural network architecture with input, output and three hidden layers. Each layer is composed of connected computational units called neurons.

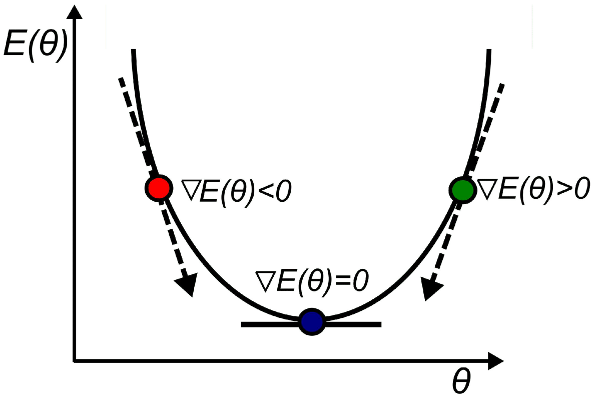


Figure 2.11: Basic concept of the gradient descent algorithm. Two different random initial values for parameters are indicated by different colours. The parameters are updated in the direction of the greatest rate of decrease of the error function, i.e. towards the minimum. When the gradient of the loss function is < 0 the red parameter is moved to the right and when the gradient of the loss function is > 0 the green parameter is moved to the left.

procedures such as gradient descent algorithms are commonly used to find the set of hyperparameters that leads to a sufficiently good approximation:

$$\theta^{\tau+1} = \theta^{(\tau)} - \eta \nabla E(\theta^{\tau}) \tag{2.55}$$

where η is the learning rate, τ is the iteration index, and $\nabla E(\theta)$ points in the direction of the greatest rate of increase of the error function. Initially, a random parameter-vector θ° is selected. The steps of changing the parameters of the neural network are performed until a global minimum or a solution close to it is reached. In particular, a single iteration steps involves a small step in the direction of the negative gradient. Figure 2.11 illustrates the basic idea of gradient descent algorithm.

Convolutional neural network

A CNN is a type of deep neural network that is specifically designed for image classification and detection tasks. CNNs are capable of extracting spatial relationships from nD (where n=1,2,3,4) imaging data by using the mechanism of local receptive fields, image subsampling and weight sharing. Typically, a CNN has convolution layers, activation or nonlinearity layers, pooling or subsampling layers, fully connected layers and finally an output layer.

Convolution layer: In a CNN an input is a tensor of shape:

$$N \times H \times W \times C \tag{2.56}$$

where N is the number of input samples also known as batch size, H and W are height and width of inputs, respectively, while C is the number of input channels (one channel for greyscale images and three channels for RGB images). Convolutional layers convolve the input with connected weights or learnable filter kernels. Normally, for 2D imaging data such as natural images, 2D filters are used to extract spatial features and kernel slides along 2 dimensions of the image (single slice for medical imaging) as shown in Figure 2.12(a). In 3D-CNNs, the kernel slides in 3 dimensions, as illustrated in Figure 2.12(b). Medical images such as CT and MRI are normally acquired in a 3D fashion and information in each slice is related to the slice next to it, thus providing a complete view of internal organs. Therefore, 2D CNNs inherently fail to leverage context from the adjacent slice if used for 3D medical imaging data. 3D-CNNs address this issue by using 3D filters in order to extract features from a volumetric patch of a scan. The ability to extract interslice context can lead to improved predictions. However, this improvement in performance comes at the cost of a large number of parameters and computational resources required to train a 3D-CNN (Singh et al., 2020).

Non-linearity layer: The convolutional layer is followed by a non-linearity layer that consists of an activation function. The purpose of activation functions is mainly to add non-linearity to the network, which otherwise would constitute a linear model. The most commonly use activation function is ReLU.

Pooling layer: The feature map of the convolution layer is downsampled in the pooling layer. The basic role of the pooling layer is to reduce the spatial size of the feature map and thereby reduce the number of parameters and computational burden of the network. Normally, a pixel window of size 2 × 2 is selected on a feature map and maximum (max pooling) or average (average pooling) of these pixels is considered as new pixel value for the feature map. Thus, if we select a pixel window of size 2 × 2 with a stride (number of pixels shift over the input matrix) on a feature map is of size 6 × 6 (36 pixels), then the output pooled feature map will be of size 3 × 3 (9 pixels).

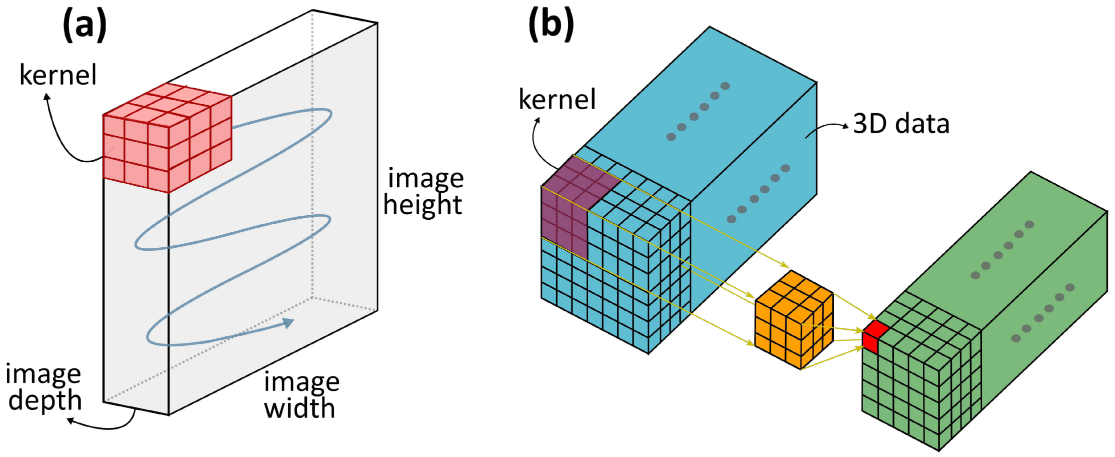


Figure 2.12: An illustration of (a) 2D convolutional kernel that slides over each channel (image depth) of the image in a 2D manner, while in (b) a 3D convolutional kernel is applied also in slice dimension of stacked data. Each pixel value in the output feature map (shown in red) is the result of a 3D convolution.

Fully connected Layer: The fully connected (FC) layer has full connection to the previous layer. A convolutional neural network often has two to five FC layers, where the first FC layer converts the incoming 2D features into a 1D feature vector. The next FC layers use the feature vector and apply weights until the final FC layer that predicts the final labels. Parameters of the CNN are trained and computed using backpropagation of the gradient descent algorithm.

In a typical CNN, convolutional layers, activation function and pooling layers are stacked together, followed by fully connected layers as shown previously in Figure 1.3. The precise organization of different layers and their connections form the network architecture. For image classification, well-known CNN architectures include AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), Inception (Szegedy et al., 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017). Multiple variants of these architectures are further created by adapting the number of layers in the CNN. The baseline dataset used for evaluating the performance of different CNN architectures is ImageNet. This dataset consists of fifteen million natural images which are labelled with approximately twenty-two thousand classes. Below we describe the basic architecture of VGGNet, ResNet, and DenseNet.

VGGNet: This architecture was described in the 2014 paper titled “Very Deep Convolutional Networks for Large-Scale Image Recognition” (Simonyan & Zisserman, 2014) and achieved top results in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). It has a simple, uniform structure of serially ordered convolutional and pooling layers. Each convolutional layer uses small filters (e.g. 2 × 2 pixels) followed by a max pooling layer. The definition and repetition of serially ordered convolution and pooling layers are referred as VGG-blocks. A 16-layer VGGNet architecture (VGG16) is shown in Figure 2.13. VGGNet is easy to understand and implement, and it is also effective at extracting features from images.

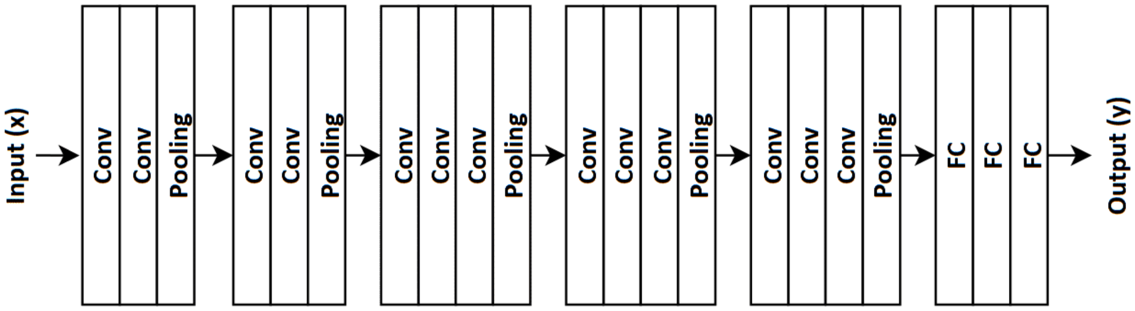


Figure 2.13: An illustration of VGGNet. Conv: convolution layer, FC: fully connected layer.

ResNet: This architecture for CNN was proposed in the paper “Deep Residual Learning for Image Recognition,” which achieved success on the 2015 version of the ImageNet ILSVRC challenge (He et al., 2016). First contribution of ResNet paper shows that if you just keep stacking convolutional layers on the top of activations, such as in VGGNet, performance starts getting worse as the network grows deep. A key innovation in the ResNet was the residual module or residual block (Figure 2.14). A residual block can be realized by feedforward neural networks with shortcut connections. The shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked convolution layers. Thus, the output from each layer is not only fed into the layer next to it, but also added to layers that are 2–3 hops away. With the help of shortcut connections training of convolutional layer in some residual blocks of ResNet can be skipped and thus different part of network are trained at different rates. This helps to produce improved accuracy with increased depth of network. A 18-layer ResNet architecture (ResNet18) with repeated residual blocks is shown in Figure 2.14

DenseNet: This architecture was published in the paper ’Densely Connected Convolutional Networks’ (Huang et al., 2017). DenseNet achieved higher accuracy than previous published models (such as ResNet) on ImageNet dataset. The network is composed of dense blocks. Within each dense block convolution layers are densely connected together i.e. the output feature maps from each layer are fed into the input of all subsequent layers. A dense block is composed of repeated stack of batch normalization (BN) layer followed ReLU activation and 3×3 convolution layer as shown in Figure 2.15. The dense connectivity ensures the maximum flow of information because each layer has direct access to the original input signal as well as gradients from the loss function, thus making it easy to train the model. DenseNet has an additional layer called transition layer that assures the concatenation of feature maps that each layer receives from the previous layer. Figure 2.15 shows the basic architecture of DenseNet.

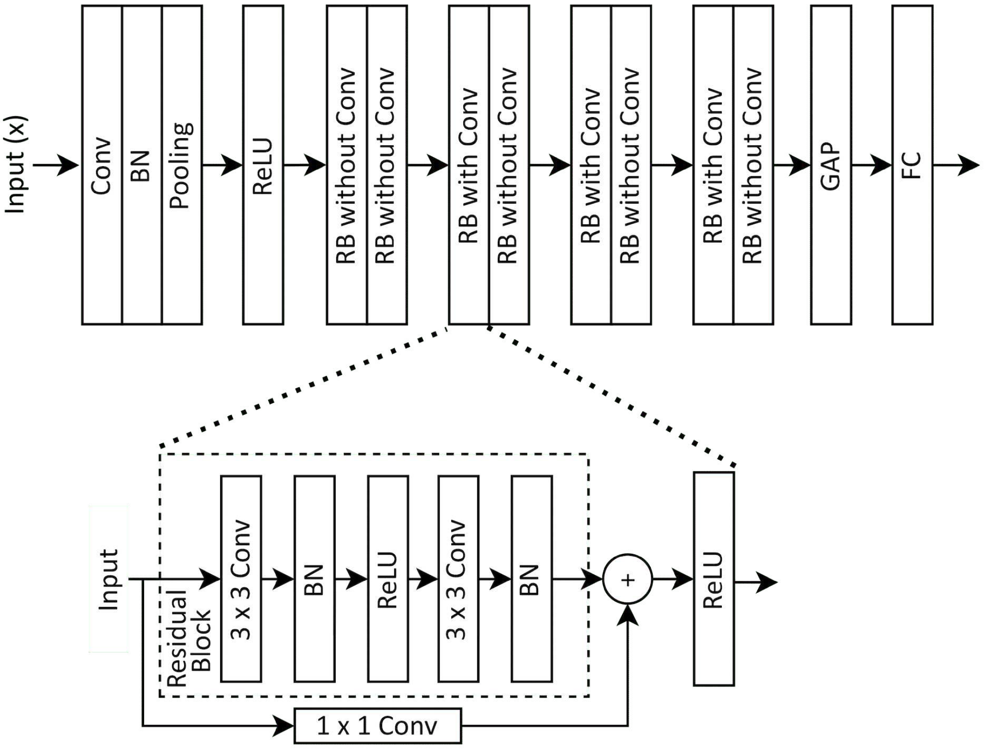


Figure 2.14: An illustration of 18 layer ResNet and its building block, i.e. residual block with 1 × 1 convolution. Residual blocks with and without 1 × 1 convolution are stacked together to form the ResNet architecture. BN: batch normalization, RB: residual block, Conv: convolution layer, GAP: global average pooling layer, FC: fully connected layer.

Data augmentation in deep learning

In order to build generalizable DL models, a large amount of data is required to avoid overfitting of highly parameterized models (LeCun et al., 2015). It is particularly challenging to build DL models in medical image analysis, where high-quality data is expensive and human-dependent for collection and data labeling. To deal with the problem of limited training data, synthetic training examples are created using data augmentation techniques that can help large-capacity learners to benefit from more representative training data. Data augmentation can increase robustness of DL a model by increasing its ability to correctly predict unseen examples that are noisy or slightly perturbed (Rozsa et al., 2016). Data augmentation can be broadly classified into the following categories: affine image transformations, elastic transformation, and pixel-level transformations. The most common technique of data augmentation is affine transformation that involves rotation, translation, cropping, flipping, and zooming of the image. However, such augmentations produce correlated images and thus provide limited improvement for deep neural network training and generalization to unseen datasets (Shin et al., 2018). Nevertheless, affine transformations are easy to implement for both 2D and 3D data and work with fewer number of hyperparameters. On the other hand, pixel-level transformations do not alter the geometry of the image but affect

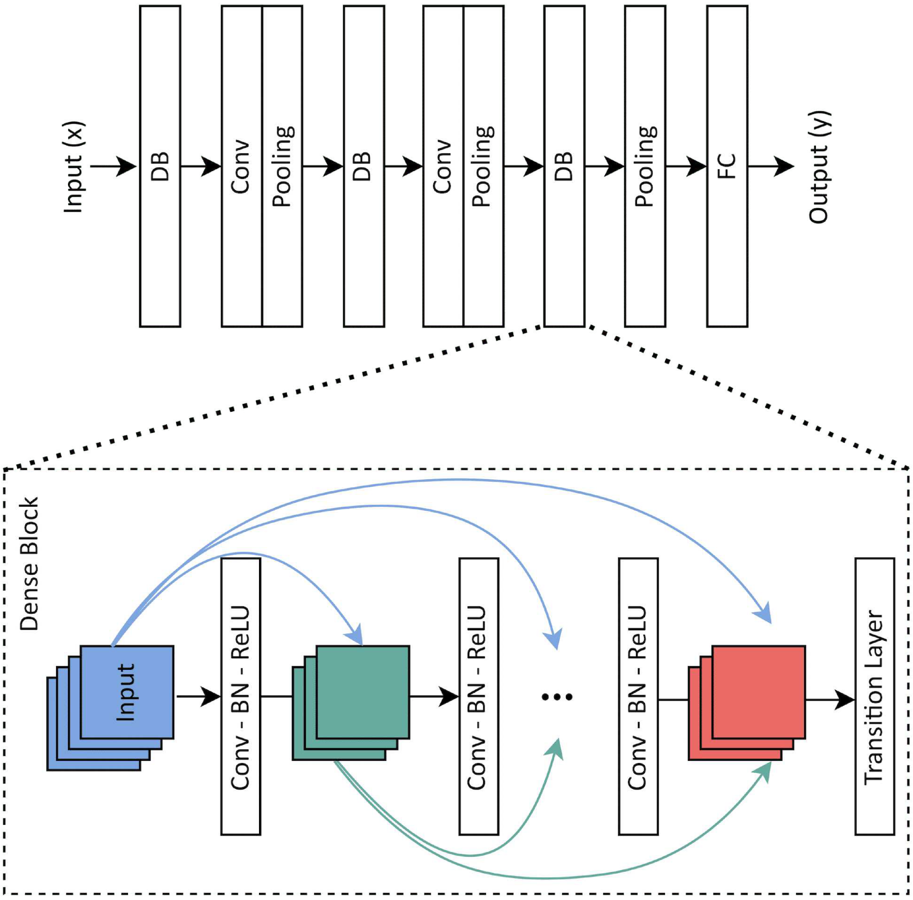


Figure 2.15: An illustration of DenseNet and its building block, i.e. dense block. In each dense block, output from each convolution layer is combined with output of all subsequent layers within the desne block. BN: batch normalization, Conv: convolution layer, DB: dense block, FC: fully connected layer.

the pixel intensity values. Pixel-level transformations are particularly useful for medical imaging data when images vary in pixel intensities because they are acquired with different scanners or image quality is affected due to patient motion. In pixel-level augmentation, pixels intensities are randomly perturbed with a given probability using random additive Gaussian noise and Gaussian blur. In addition, pixel-level transformations can modify pixel intensities by scaling or shifting of pixel-intensity values, e.g. modifying image contrast or brightness and applying gamma corrections.

2.6.4 Cox proportional hazard model in deep learning

When it comes to adaptation of neural networks for survival analysis, there exist two major approaches: (i) adaptation of the CPHM assumption (Ching et al., 2018; Katzman et al., 2018), and (ii) fully parametric DL survival models (Giunchiglia et al., 2018; Gensheimer & Narasimhan, 2019). Here, we focus on the CPHM adaptation approach. The estimation of parameters for training a DNN and CNN requires the optimization of a loss function. Faraggi et al. (Faraggi & Simon, 1995) first introduced the CPHM in a single hidden layer feed forward neural network by replacing the output of hidden node from the logistic function to the CPHM. Katzman et al. (Katzman et al., 2018) introduced DeepSurv, a feedforward deep neural network that estimates the log-risk function $\beta^T x$ as shown in Equation (38) parametrized by the weights of the neural network. Similarly, CNNs can be adapted to optimize the Cox proportional hazard likelihood for predicting survival, as shown by Starke et al. (Starke et al., 2020) and Meier et al. (Meier et al., 2020). The CPHM assigns hazards or risks of event to every observation, which can subsequently be used to assign observations to high and low risk groups. In case of CNNs, we first restrict the output of the network by using a single neuron with *tanh* activation. In CNNs, the layer before the output layer defines a feature vector x extracted from the input image volumes which is then multiplied by a weight vector w to give a scalar risk value $\beta^T x$. In the CNN forward pass, n feature vectors are extracted from the input patch of images $X = \{x_i\}$ where i=1,2,3. . . , n. In case of the CPHM for CNN, the loss is given by the following negative log partial likelihood

$$l(X, w) = -\sum_{i;\delta_i=1} \left(\beta^T x_i - w \log \sum_{j:T_j \geq T_i} e^{\beta^T x_j} \right) \tag{2.57}$$

where $\beta^T x$ is the output of the CNN and i indicates a patch coming from an uncensored observation ($\delta_i = 1$) and T_i is the corresponding survival time. The inner sum is over all patches j from patients that have longer or equal survival time than T_i, $\beta^T x_i$ is the risk associated with patch i, while $e^{\beta^T x_j}$ is called the partial hazard. It is worth mentioning that the minimization of the negative of the Cox partial log-likelihood function is carried out over the batch of images for each epoch of data in the forward pass.

2.6.5 Binary cross entropy loss in deep learning

In case of diagnostic modelling for binary outcome, the most commonly used loss function is the binary-cross entropy (BCE) loss. With BCE loss, the CNN returns the probability of a data point to come from the positive class, i.e. $\pi_i = P = (y_i = 1|x_i)$. This loss function is equal to the negative log-likelihood of a Bernoulli distributed response variable with parameter n and has the form

$$L(y_i, \pi_i) = -y_i \ln \pi_i - (1 - y_i) \ln(1 - \pi_i) \tag{2.58}$$

BCE is the Kullback-Leibler divergence between the true label y_i and the predicted probability π_i. It assigns a cost to the misclassified sample based on how incorrect the prediction is, i.e. the cost is not the same for all incorrectly classified samples.

3 MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer

3.1 Motivation

The treatment of LARC has evolved substantially during the past decade. Along with improvements in preoperative staging and surgical techniques, the use of nCRT before surgery has helped to decrease LC rates (Sauer et al., 2004). With this focus on loco-regional neoadjuvant treatment options, LC occur less often than with upfront surgery so that now distant failures have become the primary cause of morbidity and mortality for patients with locally advanced tumours. Furthermore, the response to nCRT in LARC patients varies greatly, ranging from pCR with no viable remaining tumour cells to continuing illness (pathological non-responder (pNR)) (Thies & Langer, 2013). For patients who have achieved a clinical complete response (cCR) following nCRT, the implementation of organ-preserving and low-morbidity procedures (Chau et al., 2006), or watch-and-wait methods, is receiving more attention (Dossa et al., 2017).

Many recent studies have focused on personalized treatment strategies to improve outcomes of patient populations with heterogeneous treatment response in LARC. However, there is still an unmet need of validated biomarkers that enable precise identification of the patient population that can benefit from organ preserving techniques or adjuvant therapies for improving long-term outcome. Numerous studies have examined molecular information, including gene expressions, mutations, and single nucleotide polymorphisms as potential biomarkers of response to nCRT in LARC (Rimkus et al., 2008; Boige et al., 2010; Duldulao et al., 2013). The inclusion of non-invasive biomarkers from clinical imaging may further increase the robustness and accuracy of corresponding prognostic models.

Recently, radiomic analyses employing classical statistics and modern machine learning algorithms to identify biomarkers based on multimodality imaging have shown a great potential for treatment outcome prediction in different cancer entities (Parmar et al., 2015; Gillies et al., 2016; Song et al., 2020). Radiomics models are widely developed on features extracted from T2-w MRI to predict patient's response to nCRT and long-term outcomes including FFDM and overall survival in LARC (Caruso et al., 2018; Cusumano et al., 2018; Dinapoli et al., 2018; Antunes et al., 2020; Petkovska et al., 2020; Petresc et al., 2020), and multiparametric MRI (mpMRI) (De Cecco et al., 2016; Nie et al., 2016; Giannini et al., 2019; Zhou et al., 2019). Few studies have considered radiomic features extracted from CT imaging (Chee et al., 2017; Bibault et al., 2018), PET (Bang et al., 2016; Van Helden et al., 2018), or a combination of CT and MRI features (Li et al., 2020b). Although the outcomes of these analyses are positive, crucial factors such as determining feature robustness were not always considered, and external validation was rarely carried out.

One key challenge in radiomic-based prognostic modelling is the selection of features that correlate well to the endpoint of interest. Different classes of features are commonly extracted, either directly from the images or after applying different filters (Shahzadi et al., 2021). The different feature classes normally extracted include: (i) morphological features that describe the shape of the ROI, (ii) FO features that describe the voxel intensity distribution, (iii) SOT features that describe statistical interrelationships between neighbouring voxels, and (iv) higher order features, where (i)-(iii) are extracted after applying transformations on the base images (see Section 2.3 for more details). In several studies, morphological and first order (MFO) extracted from pre-treatment T2-w MRI (De Cecco et al., 2016; Chidambaram et al., 2017; Cusumano et al., 2018; Coppola et al., 2021) had a high association to treatment response in LARC. Other studies considered SOT features only (Caruso et al., 2018; Cheng et al., 2021; Delli Pizzi et al., 2021) or in combination with MFO and SOT features (Zhou et al., 2019; Antunes et al., 2020; Petkovska et al., 2020; Petresc et al., 2020). However, it is generally unclear which of these feature classes will have the highest prognostic impact. Further, the direct combination of radiomic features may bring redundant information to the final model that may reduce its performance in external validation.

Therefore, the main objective of this study is to identify and independently validate novel radiomic signatures for the prognosis of tumour response to nCRT and FFDM in patients with LARC using a multicentre retrospective cohort of the German Consortium for Translational Cancer Research - Radiation Oncology Group (DKTK-ROG). In particular, we investigated the prognostic value of different feature classes, and developed multimodal radiomics signatures combining pre-treatment CT and T2-w MRI with clinical characteristics. The work presented within this chapter has been published in an international journal (Shahzadi et al., 2021) and was presented at an international conference (Shahzadi et al., 2022b).

3.2 Materials and methods

3.2.1 Patient cohort

In this retrospective study, multicentre data consisting of 190 patients was collected within the DKTK-ROG from four partner sites and divided into training and validation data based on the site (122 and 68 patients, respectively). 94 out of 122 patients of the training data were treated at the University Hospital Dresden (UKD, Germany) between 2006 and 2014. The remaining 28 patients were treated at the Klinikum rechts der Isar Munich (MTU) between 2007 and 2013. In the validation data, 12 out of 68 patients were treated at the University Hospital Freiburg between 2008 and 2013, while the remaining 56 patients in validation data were treated at the University Hospital Frankfurt between 2007 and 2015.

Patients included in this study were diagnosed with histopathologically confirmed LARC and underwent nCRT followed by surgery. Additional inclusion criteria for our study were the availability of pre-treatment MRI, treatment planning CT with sufficient image quality, and endpoint

Table 3.1: Patient, tumour, and treatment characteristics for the LARC training and validation data.

		Training data (122)		Validation data (68)		
Variable		**Median**	**Range**	**Median**	**Range**	**p-value**
Age (years)		59.5	24-79	63.5	21-86	0.26
		Number	**%**	**Number**	**%**	
Gender	Male/female	79/43	65/35	48/20	71/29	0.51
cT	2/3/4/unknown	6/98/18/0	5/80/15/0	7/53/7/1	10/78/10/2	0.23
cN	0/1/2/3/unknown	7/112/2/1/0	6/92/2/1/0	8/54/1/4/1	11/79/2/6/2	0.06
Grading	0/1/2/3/unknown	10/5/71/36/0	8/4/58/30/0	4/3/53/5/3	6/4/78/8/4	0.001
UICC stage	1/2/3/4/unknown	0/7/115/0/0	0/6/94/0/0	1/7/52/3/5	2/10/77/4/7	<0.001
Localization (cm)	3-6/>6-12/>12-16	65/54/3/0	53/44/3/0	24/37/6/1	35/54/9/2	0.02
RT dose (Gy)	50.4/45	95/27	78/22	66/2	97/3	<0.001
Chemotherapy regimen	5FU/5FU+OX/CAP/CAP+other	97/10/7/8	80/8/6/7	59/7/2/0	87/10/3/0	0.13
Response (TRG)	0/1/2/3/4	0/23/61/24/14	0/19/50/20/11	3/14/30/10/11	4/21/44/15/16	0.13
Distant metastases	No/yes	103/19	84/16	52/16	76/24	0.25

Abbreviations: cT=clinical T stage; cN=clinical N stage; RT=radiation therapy;
TRG= tumour regression grade; CAP= capecitabine; OX=oxaliplatine; FU=fluorouracil.

information. Ethical approval for the multicentre retrospective analyses was obtained from the Ethics Committee at the Technische Universität Dresden, Germany (BO-EK-385082020). The details of the patient characteristics for training and validation cohorts are summarized in Table 3.1.

The endpoints considered for evaluation were tumour response to nCRT and FFDM. Tumour response was determined by expert pathologists from the work-up of the surgical specimens. The response to nCRT depicts regressive changes in the tumour, and it is evaluated on a scale from 0-4 following Dworak et al. (Dworak et al., 1997), with 0 being no regression to 4 being complete regression. A detailed description of the tumour regression grade (TRG) is presented in Appendix Table A.1.

The survival endpoint FFDM was calculated from the first day of nCRT to the day of event or censoring. For patients with observed distant metastases, the event time was indicated by an event indicator variable of 1, whereas for patients without an observed event, the last follow-up time was used together with an event indicator variable of 0.

3.2.2 Experimental design

In this study, we develop and independently validate radiomic signatures for the prognosis of tumour response to nCRT and FFDM in patients with LARC using pre-treatment CT and T2-w MR imaging based on different radiomic feature classes. Figure 3.1 summarizes the design of this study. Imaging features were computed from the GTV individually on the treatment-planning

CT and pre-treatment T2-w MRI, including MFO, SOT, and intensity features of LoG transformed imaging. The features were filtered for stability under small image perturbations and clustered. In order to assess which image modality is more suited for the prediction of the endpoints and which feature class has the highest prognostic value, four radiomic models were developed on the training cohort individually for each imaging modality based on (i) MFO, (ii) SOT, (iii) LoG, and (iv) all features, i.e., the combination of MFO, SOT, and LoG features. In an additional analysis, the selected features from CT and T2-w MRI were combined for each of the cases (i) to (iv) to assess the benefit of multimodal radiomic models. The performance of each signature was then validated on the independent validation data using the AUC and the C-index for the prognosis of tumour response and FFDM, respectively. In the following paragraphs, the details of image processing and modelling are outlined.

3.2.3 Image acquisition

The training and validation imaging datasets were retrieved from the Picture Archiving and Communication System (PACS) in the respective centres. Staging T2-w MRI were acquired before nCRT with either a 1.5 T or a 3T scanner. Patients received a CT scan for treatment planning prior to radiotherapy. Appendix Table A.2 summarizes MR and CT image acquisition and reconstruction parameters for training and validation data. The GTV was delineated for each patient on T2-w transversal MR images by an experienced radiation oncologist and confirmed by a radiologist on both training and validation data. CT images were coregistered with MRI using rigid registration in RayStation 8B SP2 (RaySearch Laboratories, Stockholm, Sweden) and the GTV were transferred to the CT.

3.2.4 Image pre-processing, and feature extraction

Appendix Figure A.1 represents the process of image preprocessing as previously described (Shahzadi et al., 2022b). MRI images were first corrected for background phase variations that arise due to magnetic field inhomogeneities. This was achieved by creating a mask of the soft tissue region in the image using the Canny Edge detection algorithm (Canny, 1986) and multiplying the true image with the mask, setting all the background phase variations to zero. Bias field effect in MR images was minimized by using N4ITK bias correction method (Tustison et al., 2010). Image intensities were scaled using the 95th percentile of image intensities, i.e. 5% of the highest image intensities were ignored, representing potential outliers. Further image pre-processing and subsequent feature extraction was performed with Medical Image Radiomcis Processor (MIRP) Python toolkit (version 1.1.3) (Zwanenburg et al., 2019b) (see Section 2.3.2 for image processing details). In order to adjust the voxel spacing and slice thickness between the datasets, MR and CT image voxels were resampled to an isotropic voxel size of $1.0 \times 1.0 \times 1.0$ mm^3 using trilinear interpolation. In order to remove voxels containing air and bone, the GTV re-segmentation between -150 and 180 HU was performed on CT images to cover only soft tissue voxels. A set of

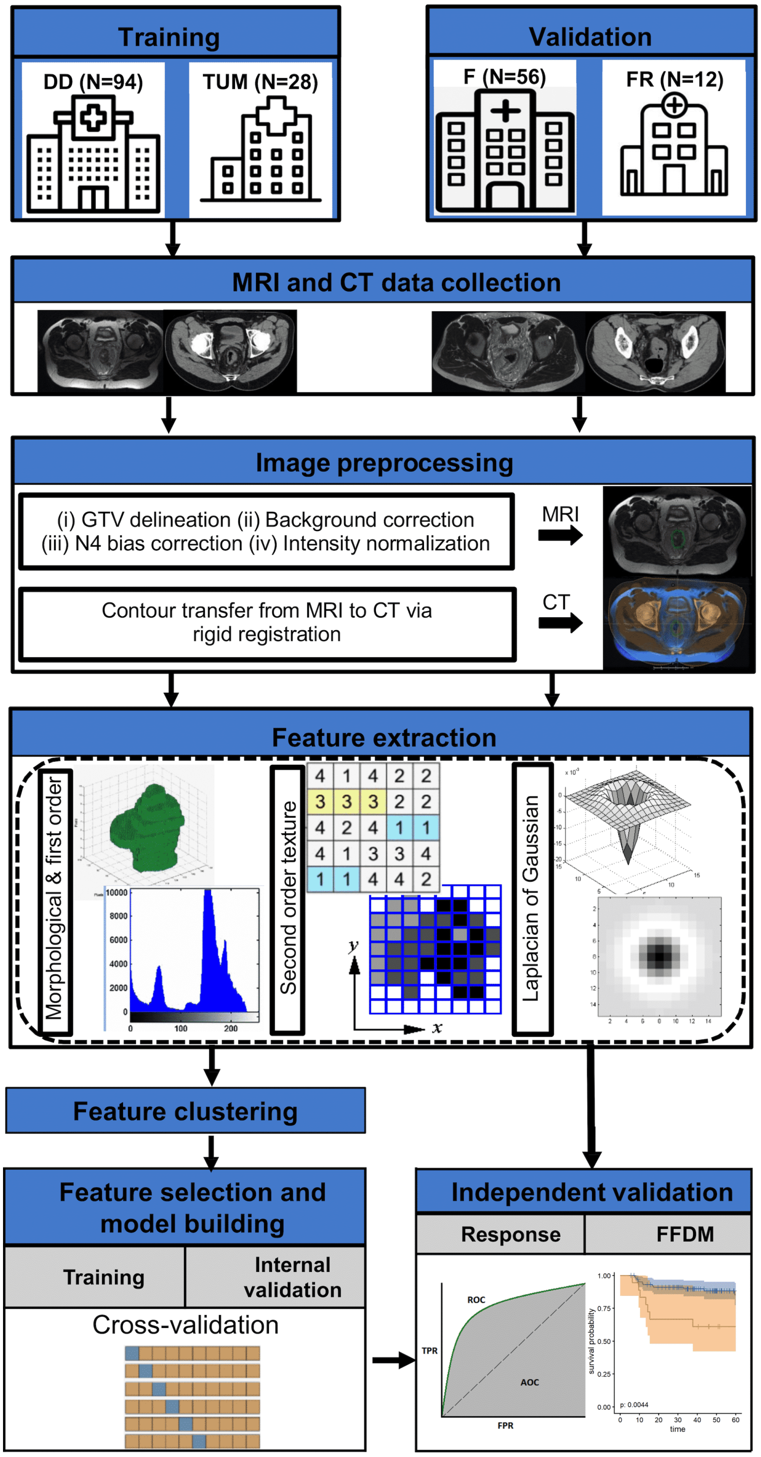


Figure 3.1: Design of the modelling study as presented in (Shahzadi et al., 2022b). Treatment planning computed tomography (CT) and pre-treatment T2-w MRI data were collected from four centres and divided into training and validation data. MRI data were pre-processed and gross tumour volume (GTV) was delineated, which was then transferred to CT images after rigid registration. Different feature classes were extracted from both modalities and signatures were developed on training data for tumour response prediction to nCRT and FFDM in a cross-validation cross-validation (CV) setting. These signatures were validated independently for both endpoints.

LoG filters with 5 different kernel widths (1 mm, 2 mm, 3 mm, 4 mm, 5 mm) was applied individually to the base MRI and CT images. The five response maps thus generated were averaged to create a single image.

Once image pre-processing was done, imaging features were computed from baseline and LoG transformed images. A set of 25 morphological and 57 first-order intensity-based features (MFO) was extracted from the 3D GTV on the treatment planning CT and on the pre-treatment T2-w MRI, respectively. In addition, 95 SOT features were calculated for every modality. Finally, the same 57 first-order intensity-based features were extracted from the GTV on the LoG transformed images. This resulted in a total of 234 features extracted from each imaging modality. SOT features were extracted from the 3D GTV based on GLCM, GLRLM, GLSZM, GLDZM, NGTDM, NGLDM. Image pre-processing and feature extraction in MIRP were implemented according to the recommendations of the IBSI (Zwanenburg et al., 2020). The definitions of the formulas used to calculate the features can be found in the IBSI reference manual. Image processing parameters used for feature extraction are summarized in Appendix Table A.3.

In order to obtain reproducible results, imaging features have to be stable under small image perturbations, as e.g. caused by differing acquisition parameters or positioning uncertainties (Zwanenburg et al., 2019a). We evaluated feature robustness by applying the following image augmentations based on the training data: adding Gaussian noise (mean:0, standard deviation:as present in the image), random volume changes of the GTV (0%, -15%, 15%), and translations (0.0, 0.25, and 0.75 mm) in all three spatial dimensions. All combinations of these perturbations were considered, leading to 81 perturbed images for each original dataset. The ICC was calculated with a 95% confidence interval, quantifying the similarity of feature values under different perturbations for every feature. Features with the lower boundary of the 95% confidence interval of the ICC below 0.8 were removed (see Robustness in Section 2.3.4 for more details).

Hierarchical clustering was used to mitigate the redundancy of features in MRI and CT individually, including (i) MFO features only, (ii) SOT features only, (iii) LoG features (statistical and intensity histogram) only, and (iv) all features, corresponding to the analyses based on the different feature classes. The Spearman correlation coefficient (ρ) was used as a similarity metric with average linkage as a criterion for merging two clusters; $\rho \geq$ 0.8 was defined for placing features into the same cluster. The feature with the highest mutual information with the endpoint was selected as the representative for each cluster (see Clustering in Section 2.3.4 for more details).

3.2.5 Radiomics modelling

In our analyses, we evaluated the prognostic performance of MRI and CT radiomic signatures for the prediction of tumour response to nCRT and FFDM. We evaluated 12 different radiomic models based on different (combinations) of feature classes and imaging modalities, as shown in Figure 3.2. First, four radiomic signatures were created individually for MRI and CT based on

(i) MFO, (ii) SOT, (iii) LOG, and (iv) all features. Once these signatures were developed for each dataset, four joint signatures were created by joining the respective MRI and CT signatures from (i) to (iv).

In order to create the eight single-modality signatures, a workflow containing four major processing steps was applied after feature clustering using an in-house end-to-end statistical learning software package: (i) feature pre-processing, (ii) feature-selection, (iii) model building with internal validation, and (iv) external validation. Steps (i)-(iii) were first performed using 33 repetitions of 3-fold stratified CV nested in the training dataset to identify an optimal signature, i.e. the steps were repeatedly performed on the internal training part and validated on the internal validation part of the cross-validation folds. After identifying the final signature, a final model was developed on the entire training data and validated on the independent validation data.

The following procedure was performed for each of the 99 CV runs: (i) Features were transformed using the Yeo-Johnson transformation to align their distribution to a normal distribution. Afterwards, features were z-transformed to mean zero and standard deviation one (see Feature transformation and normalization in Section 2.3.4 for more details). Both transformations were performed on the internal training part, and the resulting transformation parameters were applied unchanged to the features of the internal validation part. (ii) Four supervised feature-selection algorithms were considered: mRMR (Peng et al., 2005), MIM (Gelfand & I A glom, 1959), EN (Zou & Hastie, 2005), and univariate regression (Cox, 1958; Cox & Oakes, 1984). To avoid potential overfitting, only the five most relevant features were selected. (iii) The features selected by each of these methods were used to build prognostic models on the internal training part, which were validated on the internal validation part. Multivariable logistic regression was applied for the prognosis of tumour response and Cox regression for FFDM. Average model performance was assessed by the median cross validation AUC and C-index for tumour response and FFDM prognosis, respectively, for every feature selection method.

After a model has been created on CV folds of training data, the signature was created as follows. For each of the above-mentioned feature selection methods, the occurrence of every feature in the 99 modelling steps was counted and features were ranked according to their occurrences across the cross-validation folds. Features with occurrence $\geq$50% across each feature selection method were further considered. Finally, features that show repeated occurrences across at least 75% of the feature selection methods were selected and the cumulative occurrence of each feature was calculated as a sum of its occurrences. If a subset of these features showed a mutual Spearman correlation ρ > 0.5 on the entire training data, only the feature with the highest cumulative occurrence was considered. Below, we describe our feature selection scheme with an example.

Ranking scheme for feature selection: Here, we explain an example of feature selection for LoG features for tumour response prediction. The same technique applies to FFDM prediction as well. Appendix Table A.4 shows fourteen LoG MRI features and fifteen CT features

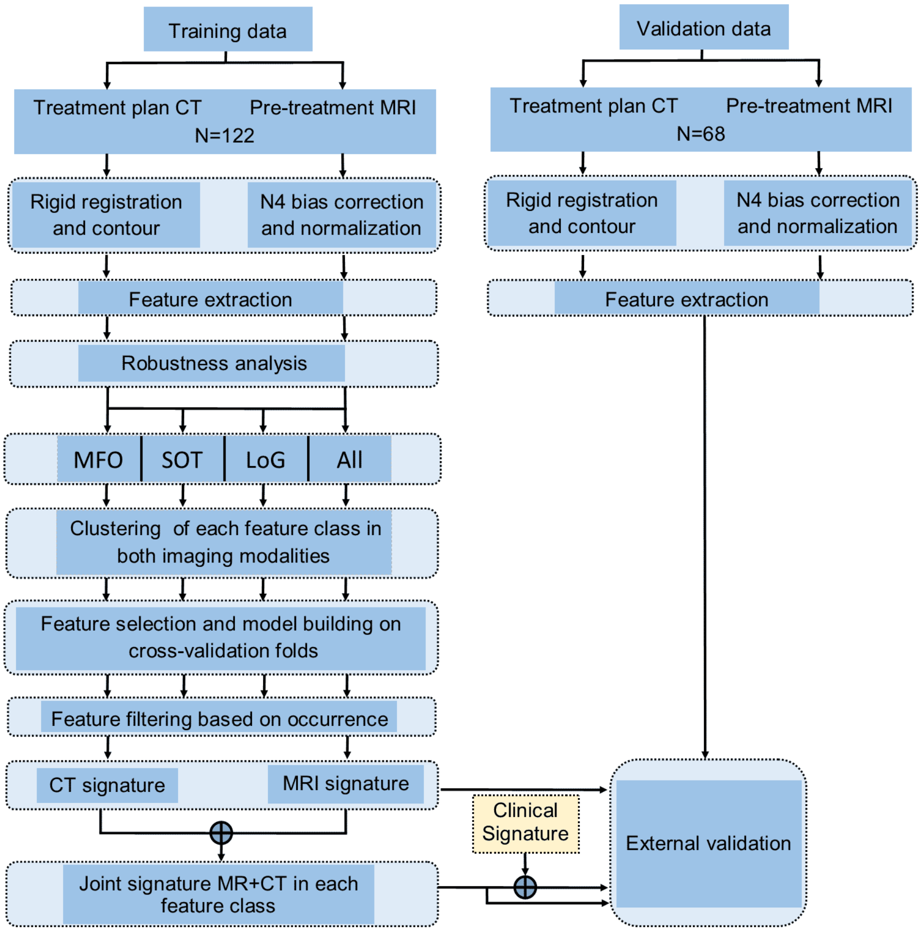

Training data
Treatment plan CT
Pre-treatment MRI
N=122
Rigid registration and contour
N4 bias correction and normalization
Feature extraction
Robustness analysis
MFO
SOT
LoG
All
Clustering of each feature class in both imaging modalities
Feature selection and model building on cross-validation folds
Feature filtering based on occurrence
CT signature
MRI signature
Joint signature MR+CT in each feature class
Validation data
Treatment plan CT
Pre-treatment MRI
N=68
Rigid registration and contour
N4 bias correction and normalization
Feature extraction
Clinical Signature
External validation

Figure 3.2: Modelling study workflow. After image pre-processing, radiomic features were extracted from pre-treatment T2-w magnetic resonance imaging (MRI) and treatment planning computed tomography (CT) and analysed for robustness. Features were separated into morphological and first-order (MFO), second-order texture (SOT), and LoG features. Also, features in each modality were analysed without any separation to feature classes, represented here by 'All'. Clustering was performed and four radiomic signatures were created individually for T2-w MRI and CT based on (i) MFO, (ii) SOT, (iii) LoG, and (iv) all features using a cross-validation approach and validated independently for both endpoints. Once these signatures were developed, four joint MRI and CT signatures in each feature category (i) to (iv) for both endpoints were validated with and without adding significant clinical features.

with the highest mutual information with tumour response selected after hierarchical clustering. These features were then used to build a prognostic model. Feature selection and model building with internal validation was first performed within 33 repetitions of 3-fold CV nested in the training dataset to identify an optimal signature. Four supervised feature-selection algorithms were considered: mRMR, MIM, EN, and univariate logistic regression (UR). To avoid potential overfitting, only the five most relevant features were selected in each cross-validation fold. These features were then used to build a multivariable logistic-regression model on the internal training part, and validated on the internal validation part. For each of the above-mentioned feature selection methods, the occurrence of every feature in the 99 modelling steps was counted and features were ranked according to their occurrences across the cross-validation folds. Table 3.2 shows features with $\geq$50% occurrence across each feature selection method that were further considered. Finally, features that showed repeated occurrences across at least 75% of the feature selection methods were selected (MR_log_ih_max_grad_fbn_n32 and MR_log_stat_min, CT_log_ih_max_grad_fbn_n32 and CT_log_stat_energy). MR_log_ih_max_grad_fbn_n32 showed the highest cumulative occurrence (i.e. the highest sum of occurrences across all feature selection methods) of 365, while MR_log_stat_min showed a cumulative occurrence of 251. Both features showed a Spearman correlation of <0.5 on the entire training data, thus forming the MR-based LoG radiomic signature. A model with this signature was then fitted on the entire training data and the trained model was applied to the external validation data. Similarly, for CT data, CT_log_ih_max_grad_fbn_n32 showed the highest cumulative occurrence of 363, while CT_log_stat_energy showed a cumulative occurrence of 277. The features were strongly correlated with a Spearman correlation >0.5. Therefore, only CT_log_ih_max_grad_fbn_n32 was considered for the final CT-based LoG radiomic signature. The final performance in internal cross validation was considered as the average of the cross-validation training AUC (CV training) and validation AUC (CV validation). The finally selected signature and the average AUC in internal training and external validation are shown in Appendix Table A.5. After feature selection was done, the resulting radiomic signature was then used to build prognostic models on the entire training data and (iv) the trained model was applied to the independent validation data. For creating the four joint signatures combining CT and MRI, the selected signatures in each feature class were pooled together and the same procedure as described in the last paragraphs was performed: clusters with ρ>0.5 were reduced to one feature, models were trained on entire training data and validated on external validation data. Finally, clinical features that were significantly associated to tumour response in univariable logistic regression or to FFDM in univariable Cox regression were added to the selected radiomic signature.

3.2.6 Statistical analysis

The following baseline clinical parameters were available: gender, age, tumour localization, UICC stage, grading, T stage, N stage, surgery type, chemotherapy type. Continuous variables of the clinical data were compared between training and validation cohort using the Mann-Whitney-U test, while categorical variables were compared by the χ^2 test. Model performance for the prediction of endpoints were evaluated by the AUC for tumour response to nCRT and by the C-index for FFDM prognosis. The estimated value and the 95% confidence interval of these metrics were computed using the bias-corrected bootstrap confidence interval method on 400 bootstraps of the data (Efron & Hastie, 2013). For creating a confusion matrix based on the final signature for tumour response prediction, an optimal cutoff was selected on the training data using Youden index and transferred to the validation data. For association with FFDM, patients were stratified into an optimally separated low and a high-risk group. The cutoff for risk group stratification was

Table 3.2: Median AUC for tumour response prognosis for LoG intensity features based on MRI and CT using CV of the training data. Features with an occurrence ≥50% are shown here. Features with a repeated occurrence across at least 75% (3 out of 4) of the feature selection methods are presented in bold. AUC: area under the curve, CV: cross-validation, CT: computed tomography, EN: Elastic net, UR: logistic regression, LoG: Laplacian of Gaussian, MRMR: minimum redundancy maximum relevance, MIM: mutual information maximization, MRI: magnetic resonance imaging.

Modality	Feature selection	CV training AUC	CV validation AUC	Features	Occurrence	Cumulative occurrence of selected features
MRI	**MRMR**	0.68	0.58	**log_ih_max_grad_fbn_n32**	73	**log_ih_max_grad_fbn_n32=365** **log_stat_min=251** **Remarks:** Both features occurred in at least 3 out of 4 (75%) feature selection methods. Both were weakly correlated so they were considered for the MRI_LoG signature.
	MIM	0.70	0.57	**log_ih_max_grad_fbn_n32**	98	
				log_stat_min	92	
				log_stat_max	60	
	EN	0.72	0.56	**log_ih_max_grad_fbn_n32**	98	
				log_stat_min	73	
				log_ivh_v25	50	
	UR	0.70	0.58	**log_ih_max_grad_fbn_n32**	96	
				log_stat_min	86	
				log_stat_max	56	
CT	**MRMR**	0.71	0.67	**log_ih_max_grad_fbn_n32**	70	**log_ih_max_grad_fbn_n32=363** **log_stat_energy=277** **Remarks:** Both features occurred in at least 3 out of 4 (75%) feature selection methods. Both were correlated with a Spearman correlation >0.5, therefore only log_ih_max_grad_fbn_n32 was considered for the CT_LoG signature.
	MIM	0.73	0.62	**log_ih_max_grad_fbn_n32**	98	
				log_stat_energy	97	
				log_stat_median	87	
				log_ivh_diff_v25_v75	61	
	EN	0.74	0.64	**log_ih_max_grad_fbn_n32**	99	
				log_stat_energy	86	
	UR	0.72	0.63	**log_ih_max_grad_fbn_n32**	96	
				log_stat_energy	94	
				log_stat_median	64	
				log_ivh_diff_v25_v75	51	

selected on the training data using maximally selected rank statistics (Hothorn & Lausen, 2003) and transferred to the validation data. The FFDM of stratified groups was assessed with Kaplan Meier curves compared with the log-rank test.

Calibration for the prediction of tumour response to nCRT and FFDM was evaluated via the Hosmer-Lemeshow goodness of fit test (HL test) (Hosmer & Lemesbow, 1980) and Greenwood Nam D'Agostino test (GND test) (Demler et al., 2015), respectively. Correlations between features were assessed by the Spearman correlation coefficient (ρ). All tests were two-sided with a significance level of 0.05. In order to evaluate the importance of individual features in the final signature, univariate fitting of a logistic regression (tumour response) or Cox regression (FFDM) models was performed and Wald-test p-values were computed. All analyses were performed in R version 4.0.3.

3.3 Results

Patient characteristics of the training and validation data are summarized and compared in Table 3.3. Patients in the training data had a higher tumour grading (p=0.001) and higher UICC stage (p<0.001). Patients of the validation data were treated with a higher dose (p<0.001). The endpoints tumour response and FFDM were similar for training and validation data (p=0.13 and p=0.25, respectively). In univariate analysis, a significant association was observed only between clinical T stage (cT) and tumour response (Appendix Table A.6).

For radiomics modelling, 234 radiomic features were extracted from the GTV in the T2-w MR and in the CT imaging dataset. Stability analysis reduced these to 208 features (MFO: 74, SOT: 82, LoG: 52) and 222 (MFO: 76, SOT: 95, LoG: 51) for MRI and CT, respectively. Clustering of correlated features further reduced the feature number to (i) $\mathrm{MRI}_{\mathrm{MFO}}$:24, $\mathrm{CT}_{\mathrm{MFO}}$:22; (ii) $\mathrm{MRI}_{\mathrm{SOT}}$:16, $\mathrm{CT}_{\mathrm{SOT}}$:19; (iii) $\mathrm{MRI}_{\mathrm{LoG}}$:14, $\mathrm{CT}_{\mathrm{LoG}}$:15; and (iv) $\mathrm{MRI}_{\mathrm{All}}$:39, $\mathrm{CT}_{\mathrm{All}}$:47.

Table 3.3 presents the results for the prognosis of tumour response, including the names of finally selected features. In internal cross validation, models based on CT data showed better prognostic performance than models based on MRI. Among feature classes, SOT features showed a high prognostic value (MRI: $\mathrm{AUC}_{\mathrm{SOT}}$=0.68, $\mathrm{AUC}_{\mathrm{MFO}}$=0.57, $\mathrm{AUC}_{\mathrm{LoG}}$=0.57, $\mathrm{AUC}_{\mathrm{All}}$=0.65; CT: $\mathrm{AUC}_{\mathrm{SOT}}$=0.70, $\mathrm{AUC}_{\mathrm{MFO}}$=0.65, $\mathrm{AUC}_{\mathrm{LoG}}$=0.64, $\mathrm{AUC}_{\mathrm{All}}$=0.67). This result, however, did not translate to the independent validation data, where SOT features performed poorly. Here, the overall best performance was achieved by LoG features for both imaging modalities (MRI: $\mathrm{AUC}_{\mathrm{LoG}}$=0.66, CT: $\mathrm{AUC}_{\mathrm{LoG}}$=0.61). Joint MRI+CT signatures performed almost similar to MRI only signatures in independent validation for all four models.

The clinical model containing only cT stage achieved training and validation AUCs of 0.60. Combining cT stage with the combined signature from MRI and CT achieved the best validation result with an AUC of 0.70. At a threshold of 0.248 this signature was able to accurately classify 16/21 responders and 20/47 non-responders (Appendix Figure A.2). Figure 1.3 shows ROC and the corresponding calibration plots for this signature on training and validation data. All features represented independent information (Appendix Figure A.3) and significantly contributed to the prediction in training ($p<0.05$), while only MR_log_stat_min was significant in validation (p=0.04). The MRI feature log_stat_min (IBSI:1GSF) represents the minimum intensity, while the CT feature log_ih_max_grad_fbn_n32 (IBSI:12CE) represents the gradient of the discretised histogram (32 bins) within the GTV on the LoG transformed image. Image-based interpretation of these features is presented in Figure 3.4. In the non-responder group, MR_log_stat_min showed relatively low values, which translates to the existence of bright voxels in the GTV on the original baseline T2-w MRI (Figure 3.4(b)). In comparison, responders showed no such high grey values (Figure 3.4(a)). Box plots of these features (Yeo-Johnson transformed and z-score normalized) in the two response groups are shown in Supplementary Appendix Figure A.4.

Table 3.4 presents the results for the prognosis of FFDM, including the names of finally selected features. Median follow-up time in training and validation data was 49.1 (5.7-111.8) months and 29.5 (1.2-94.1) months, respectively. Most of the metastases occurred until 24 months after treatment (training: 76%, validation: 56%). Until that time, 7 patients (training: 5, validation: 2) were lost to follow-up because of death, i.e. the competing risk of death was small. In internal cross validation, models based on MRI data showed a better prognostic performance than models based on CT. Among feature classes, LoG features showed a somewhat higher prognostic value (MRI: C-$\mathrm{index}_{\mathrm{LoG}}$=0.65, C-$\mathrm{index}_{\mathrm{MFO}}$=0.60, C-$\mathrm{index}_{\mathrm{SOT}}$=0.59, C-$\mathrm{index}_{\mathrm{All}}$=0.60, CT:

Table 3.3: Median AUC values for CV and external validation for tumour response prediction based on MRI, CT, joint MRI+CT, and imaging combined with clinical T stage. Values in parentheses represent the 95% confidence interval.

Modality	Feature level	CV training AUC	CV validation AUC	Signature	Final training AUC	External validation AUC
MRI	**All**	0.76	0.65	MR_dzm_zd_entr_3d_fbn_n32	0.72 (0.62-0.82)	0.34 (0.19-0.50)
	MFO	0.74	0.57	MR_morph_av MR_morph_geary_c	0.70 (0.60-0.79)	0.57 (0.39-0.73)
	SOT	0.75	0.68	MR_dzm_zd_entr_3d_fbn_n32	0.72 (0.62-0.81)	0.34 (0.10-0.50)
	LoG	0.70	0.57	MR_log_ih_max_grad_fbn_n32 MR_log_stat_min	0.67 (0.57-0.75)	0.66 (0.51-0.82)
CT	**All**	0.78	0.67	CT_dzm_zd_var_3d_fbn_n32 CT_cm_corr_d1_3d_v_mrg_fbn_n32	0.77 (0.69-0.84)	0.47 (0.34-0.63)
	MFO	0.77	0.65	CT_morph_av	0.72 (0.60-0.82)	0.52 (0.38-0.66)
	SOT	0.78	0.70	CT_dzm_zd_var_3d_fbn_n32 CT_cm_corr_d1_3d_v_mrg_fbn_n32	0.77 (0.59-0.80)	0.47 (0.36-0.66)
	LoG	0.73	0.64	CT_log_ih_max_grad_fbn_n32	0.70 (0.60-0.79)	0.61 (0.44-0.76)
Joint MRI +CT	**MRI_All + CT_All**	-	-	MR_dzm_zd_entr_3d_fbn_n32 CT_cm_corr_d1_3d_v_mrg_fbn_n32	0.76 (0.67-0.84)	0.38 (0.24-0.56)
	MRI_MFO + CT_MFO	-	-	MR_morph_geary_c CT_morph_av	0.74 (0.64-0.83)	0.57 (0.40-0.67)
	MRI_SOT + CT_SOT	-	-	MR_dzm_zd_entr_3d_fbn_n32 CT_cm_corr_d1_3d_v_mrg_fbn_n32	0.76 (0.67-0.84)	0.38 (0.24-0.56)
	MRI_LoG + CT_LoG	-	-	MR_log_stat_min CT_log_ih_max_grad_fbn_n32	0.71 (0.62-0.80)	0.66 (0.50-0.82)
Clinical+ MRI/CT	**No Radiomics**	-	-	cT	0.60 (0.53-0.66)	0.60 (0.50-0.70)
	MRI_LoG	-	-	cT MR_log_ih_max_grad_fbn_n32 MR_log_stat_min	0.69 (0.59-0.78)	0.69 (0.53-0.82)
	CT_LoG	-	-	cT CT_log_ih_max_grad_fbn_n32	0.72 (0.61-0.81)	0.66 (0.51-0.81)
	MRI_LoG + CT_LoG	-	-	cT MR_log_stat_min CT_log_ih_max_grad_fbn_n32	0.72 (0.62-0.80)	0.70 (0.54-0.84)

Abbreviations: AUC=area under a curve; cT=clinical T stage; CT=computed tomography; CV=cross-validation; LoG=Laplacian of Gaussian; MRI=magnetic resonance imaging; MFO=morphological and first order; SOT=second order texture.

C-index$_{\text{LoG}}$=0.52, C-index$_{\text{MFO}}$=0.47, C-index$_{\text{SOT}}$=0.51, C-index$_{\text{All}}$=0.46). In external validation, CT-based features showed a slightly higher performance compared to MRI. While both SOT and LoG features achieved similar prognostic value on MRI data (MRI: C-index$_{\text{SOT}}$=0.57, CT: C-index$_{\text{LoG}}$=0.57), the overall best prognostic performance in CT was achieved by SOT features (CT: C-index$_{\text{SOT}}$=0.69). No additional benefit was achieved by joining the MRI and CT signatures. Patient stratification into groups at low and high risk of distant metastases was performed based on the SOT models for each modality, i.e. for MRI, CT, and joint MRI+CT. While the CT and

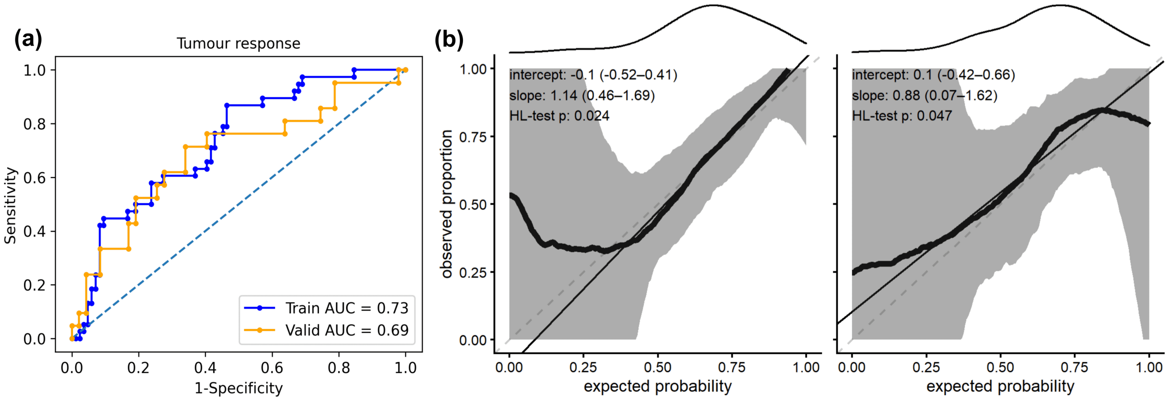


Figure 3.3: (a) Receiver operating characteristics (ROC) curves and (b) calibration plots for tumour response prognosis in training (left) and validation (right) resulting from best performing joint signature combining clinical T stage and Laplacian of Gaussian (LoG) features from T2-w MRI and CT. For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines) that follow the optimal expectation (dashed lines). Density of expected probabilities is shown above the calibration plot.

MRI+CT-based signatures achieved a significant patient stratification in independent validation (p<0.01), this was not the case for the MRI-based signature (p=0.68). Kaplan-Meier curves and corresponding calibration plots for the best performing CT signature are shown in Figure 3.5 and for the MRI and MRI+CT signatures in Appendix Figure A.5. The definition and interpretation of selected features are presented in Appendix Table A.7.

Final model parameters for the best performing signatures for the prognosis of tumour response and FFDM are presented in Appendix Table A.8

3.4 Summary and discussion

In this modelling study, radiomics signatures based on pre-treatment T2-w MRI and treatment planning CT imaging were developed and validated for the prediction of tumour response to nCRT and FFDM in patients with LARC. For both imaging modalities, the predictive performance of three feature classes i.e. MFO, SOT, LoG, and the combination of all features was independently validated. The best predictive performance for tumour response prediction on validation cohort was achieved by combining clinical T stage with LoG features from CT and MRI (AUC=0.70), while SOT features from CT showed the best performance for FFDM (C-index=0.69).

In comparison to other independently validated MRI-based multicentre radiomic studies for patients with LARC, our best performing signature for tumour response prediction (AUC=0.70) showed similar results to the study by Antunes et al. (Antunes et al., 2020) (AUC=0.71), but showed lower performance than results presented by Dinapoli et al. (Dinapoli et al., 2018) (AUC=0.75), and Cusumano et al. (Cusumano et al., 2018) (AUC=0.79) who also assessed tumour response to nCRT in LARC patients using T2-w MRI data. Dinapoli et al. (Dinapoli et

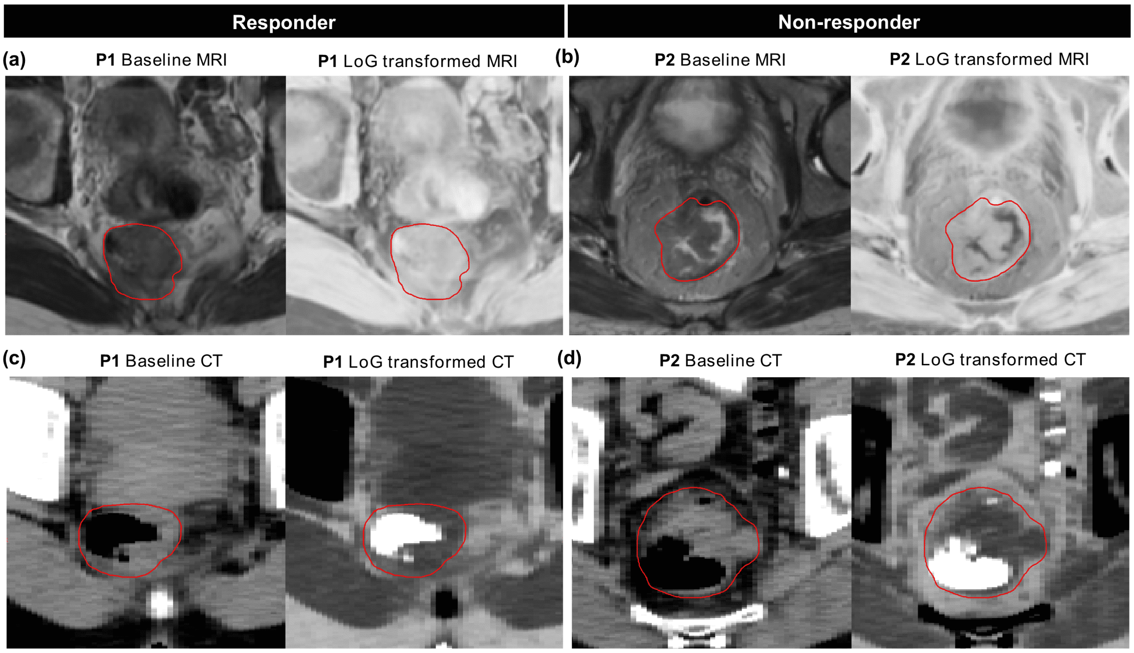


Figure 3.4: Representative images from MRI (a, b) and CT (c, d) with corresponding Laplacian of Gaussian (LoG) transformed images from two patients (P) in the two response groups, i.e. responder: P1 and non-responder: P2 on the training data. Red contours mark the gross tumour volume (GTV). P1 (responder: TRG=4) showed an overall homogenous appearance on the baseline MRI. On the contrary, P2 (non-responder: TRG=1) showed a more heterogeneous GTV with a low stat_min value on the LoG transformed MR image, which corresponds to some high pixel intensities on the baseline MRI. Similarly, a more homogenous GTV (excluding the air voxels) can be seen in P1 compared to P2 on the baseline and LoG transformed CT slices, possibly causing low gradients in the intensity histogram for the responder.

al., 2018) used first-order intensity histogram-based features, while the study by Cusumano et al. (Cusumano et al., 2018) additionally used fractal features in the final signature to build the model. Both studies also combined radiomics signature with clinical features (cT and cN).

Majority of studies presented for tumour response prediction in LARC are retrospective and single centric. These studies have shown promising results for tumour response prediction in LARC. De Cecco et al. (De Cecco et al., 2016) and Caruso et al. (Caruso et al., 2018) showed a significant association ($p<0.05$) of FO statistical and GLCM features, respectively, with tumour response to nCRT on small cohorts ($\leq$15 subjects). Ferrari et al. (Ferrari et al., 2019) showed that complete responders have higher GLCM energy and good responders have high expression of histogram features (AUC=0.87). Coppola et al. (Coppola et al., 2021) showed that heterogeneity of local skewness is associated to tumour response (AUC=0.90). Some studies also showed the association of SOT features with tumour response prediction. The studies by Delli Pizzi et al. (Delli Pizzi et al., 2021) and Petresc et al. (Petresc et al., 2020) showed an AUC of 0.79 and 0.80 in internal validation, respectively.

Compared to T2w-MRI, CT imaging is rarely used for diagnostic evaluation in LARC. Some studies have investigated the performance of CT imaging for tumour response prediction to nCRT

Table 3.4: Median C-index values for CV and external validation for FFDM prediction in MRI, CT, and joint MRI+CT. Values in parentheses represent the 95% confidence interval.

Modality	Feature level	CV training C-Index	CV validation C-index	Signature	Final training C-index	External validation C-index
MRI	**All**	0.79	0.60	MR_log_stat_median	0.69 (0.56-0.81)	0.54 (0.36-0.69)
	MFO	0.77	0.60	MR_stat_median	0.68 (0.54-0.82)	0.52 (0.34-0.68)
	SOT	0.75	0.59	MR_ ngl_dc_var_d1_a0_0_3d_fbn_n32 MR_ szm_sze_3d_fbn_n32 MR_cm_clust_prom_d1_3d_v_mrg_fbn_n32	0.70 (0.58-0.82)	0.57 (0.40-0.74)
	LoG	0.75	0.65	MR_log_stat_median MR_log_stat_iqr MR_log_ih_entropy_fbn_n32	0.69 (0.56- 0.82)	0.57 (0.39 - 0.73)
CT	**All**	0.74	0.46	No feature selected	-	-
	MFO	0.73	0.47	CT_morph_volume	0.62 (0.50 - 0.75)	0.58 (0.42 - 0.73)
	SOT	0.70	0.51	CT_szm_zsnu_3d_fbn_n32	0.64 (0.49- 0.80)	0.69 (0.51- 0.81)
	LoG	0.70	0.52	CT_log_stat_energy	0.65 (0.53 - 0.76)	0.63 (0.46 - 0.77)
Joint MRI+CT	**MRI_All + CT_All**	-	-	MR_log_stat_median	0.69 [0.56-0.81]	0.54 (0.36-0.69)
	MRI_MFO + CT_MFO	-	-	MR_stat_median CT_morph_volume	0.70 [0.55-0.81]	0.55 (0.37-0.70)
	MRI_SOT + CT_SOT	-	-	MR_ ngl_dc_var_d1_a0_0_3d_fbn_n32 MR_ szm_sze_3d_fbn_n32 MR_cm_clust_prom_d1_3d_v_mrg_fbn_n32 CT_szm_zsnu_3d_fbn_n32	0.73 (0.61-0.84)	0.62 (0.45-0.79)
	MRI_LoG + CT_LoG	-	-	MR_log_stat_median MR_log_stat_iqr MR_log_ih_entropy_fbn_n32 CT_log_stat_energy	0.72 (0.59-0.85)	0.59 (0.41-0.75)

Abbreviations: C-index=concordance-index; CT=computed tomography; CV=cross-validation; LoG=Laplacian of Gaussian; MRI=magnetic resonance imaging; MFO=morphological and first order; SOT=second order texture.

using patient populations treated with standard procedures, i.e. nCRT followed by TME (Rao et al., 2016; Chee et al., 2017; Bibault et al., 2018; Hamerla et al., 2019), or combined CT and MR imaging (Li et al., 2020b; Zhang et al., 2020). Based on radiomics features extracted from treatment plan CT, Bibault et al. (Bibault et al., 2018) developed a model for the prognosis of tumour response using DNN with an AUC of 0.72. Chee et al. (Chee et al., 2017) showed that FO features extracted from pre-treatment contrast enhanced CT were associated with tumour response prediction (responders showed low entropy, high uniformity, and low standard deviation). Some studies have also indicated poor performance of CT features for predicting tumour response in LARC. Exemplarily, Rao et al. (Rao et al., 2016) and Hamerla et al. (Hamerla et al., 2019) showed that CT features failed to predict tumour response. Regarding the combination of CT and MRI, Zhang et al. (Zhang et al., 2020) used MFO and SOT features extracted from pre-treatment CT and MRI and achieved an AUC of 0.87, while Li et al. (Li et al., 2020b) showed that contrast enhanced CT and multimodality MRI is able to achieve an AUC of 0.93. Although

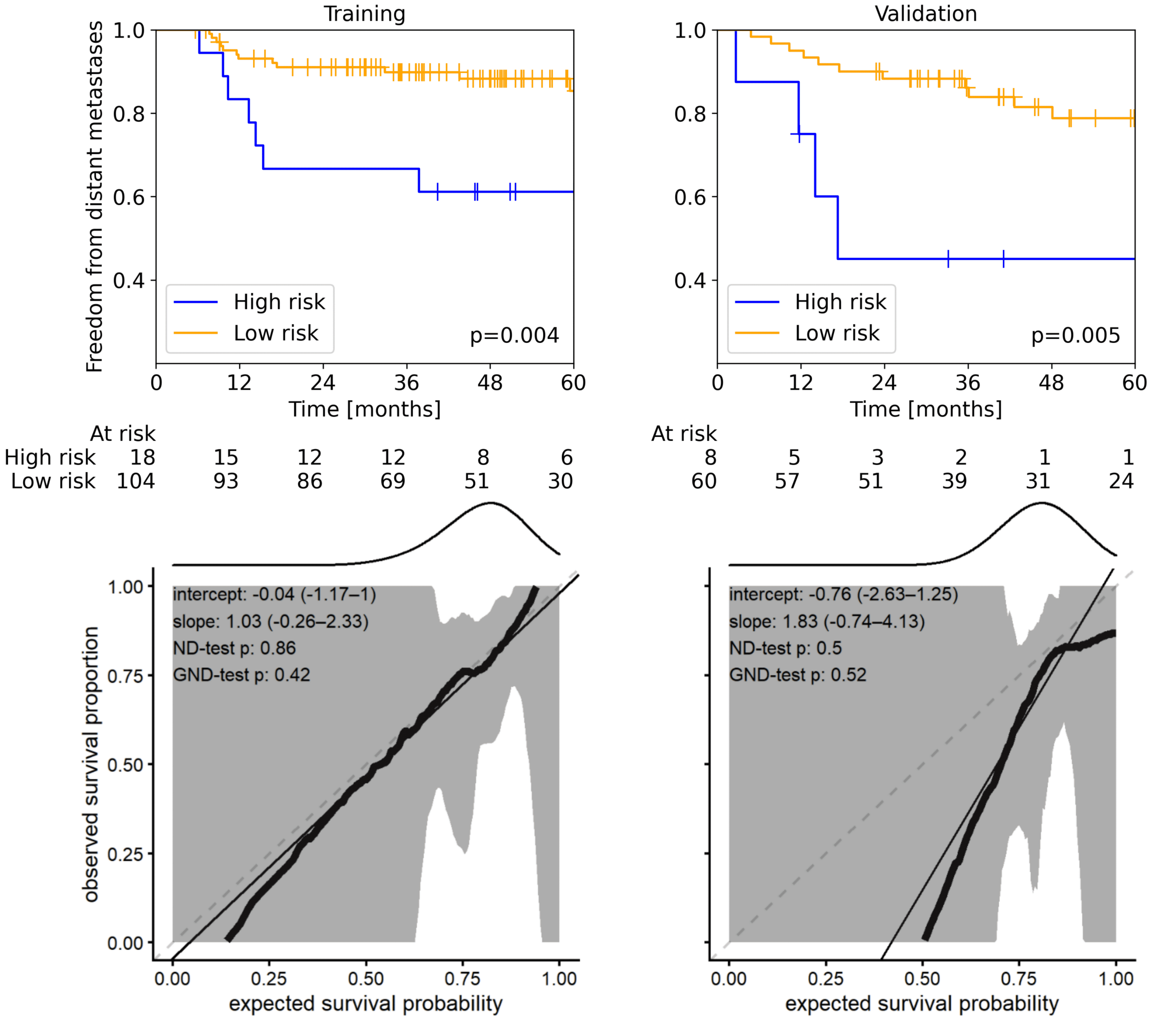


Figure 3.5: Kaplan-Meier (top) and calibration plots (bottom) on training (left) and validation (right) data for the prediction of FFDM using the best performing CT-based SOT feature, resulting in significant patient stratifications (p<0.01). For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines) that should follow the optimal expectation (dashed lines). Density of expected probabilities is shown above the calibration plot.

the results from most of these studies are promising, external validation is rarely performed in them.

The prognostic performance of models in LARC may be improved by including additional imaging modalities such as PET, other MRI sequences such as T1-w, T1c-w and diffusion weighted imaging (DWI), or images acquired at different time point during the course f treatment. For example, Jeon et al. (Jeon et al., 2019) built a predictive signature for treatment outcomes in LARC by using delta-radiomic features extracted from pre- and post-nCRT T2-w MRI. Their signature showed significant risk group stratification for FFDM (p<0.05). Chiloiro et al. (Chiloiro et al., 2020) also used delta radiomics to predict FFDM as binary outcome with an AUC of 0.78. Gianni et al. (Giannini et al., 2019) showed that radiomic signatures based on PET, T1-w MRI, and apparent diffusion coefficient (ADC) images had an increased performance for tumour response prediction (AUC=0.86) compared to PET only (AUC=0.84) and T1-w MRI only (AUC=0.72). To the best of our knowledge, no study was yet performed to predict FFDM combining pre-treatment MRI and treatment-planning CT for LARC

One of the major challenge in radiomics analyses is the selection of features, as numerous features of different complexity can be extracted and most often their number is larger than the study population. This can not only lead to substantial model overfitting, but also causes difficult feature interpretability. In our analyses, we observed that more complex SOT features showed a high predictive performance for tumour response prediction, while LoG transformed intensity features showed a high performance for the prognosis of FFDM in internal CV. However, the same was not true for external validation, and we observed the opposite behaviour, i.e. LoG transformed statistical, and intensity histogram features showed a high performance for the prediction of tumour response, while SOT features showed a somewhat higher performance for FFDM prediction. Also, it is noteworthy that the performance trend of feature classes in internal and external validation was similar for both modalities, i.e. similar feature classes were predictive for both CT and MRI.

In our analyses, features in final signature were somewhat interpretable on imaging level. Specifically, we discovered one MRI-based statistical feature, i.e. log_stat_min, which was predictive of tumour response to nCRT. This feature represents the minimum intensity on LoG transformed images, which is closely related to the maximum intensity (i.e. stat_max) on baseline images. The predictive performance of both features was analysed separately using univariate logistic regression. In training, stat_max was less predictive (AUC=0.57) than log_stat_min (AUC=0.64), while both features showed similar performance in validation with an AUC of 0.66. The high association of LoG transformed intensity features with the training data can be attributed to the fact that the LoG kernels help to reduce large variations in the signal, which can be detected within a single image slice (e.g. irregularities due to magnetic field, respiratory motion, or patient movement). Further, we interpret log_stat_min as a potential biomarker for tumour response prediction to nCRT based on the fact that a tumour normally is represented by low to intermediate signal intensity on T2-w MRI, excluding the intestinal lumen (Horvat et al., 2019; Jeon et al., 2019). The increased expression of log_stat_min in non-responders indicates the presence of high intensities within the GTV on baseline T2-w MRI, possibly indicating an aggressive or resistive tumour resulting in incomplete remission.

Limitations of this study are relatively low number of patients in the training and validation data and its retrospective nature. In addition, there is a class imbalance due to the smaller number of events for both endpoints, leading to wide confidence intervals in Tables 3.3 and 3.4 often including the value 0.5, i.e., the external validation results have a relatively large uncertainty. We aimed to mitigate the class imbalance problem by internal CV on the training data for feature selection. A 3-fold CV approach was used and repeated 33 times, ensuring that each fold contained sufficient events for training and validation and that the finally considered average model performance was sufficiently robust. Another common strategy used in machine learning to deal with the problem of imbalanced data is random undersampling of the majority class. We tested this procedure during stratified splitting of training data in internal cross-validation. We did not observe significant differences in feature selection for both endpoints, and therefore do not present the results from these experiments.

In conclusion, in the present modelling study, we developed and independently validated radiomic signatures for the prognosis of tumour response to nCRT and FFDM in patients based on T2-w MR and CT imaging. We studied feature classes of differing complexity and observed that a combination of LoG transformed intensity features from MRI and CT together with clinical T stage (cT) led to highest prognostic value for the prediction of nCRT, while CT-based SOT features performed well in external validation for FFDM.

4 External validation of published radiomics models for patient prognosis in locally advanced rectal cancer

4.1 Motivation

Current evaluation of patient response and long-term outcomes in LARC depends on visual interpretation of imaging data obtained after nCRT or by analysing resected specimens obtained after surgery. Thus, there is a need for non-invasive biomarkers that can aid in treatment personalization based on individual chance for response. Many radiomics-based studies have been published recently that demonstrate encouraging results for predicting patient outcome after treatment. For LARC the most interesting and most explored area in radiomics analysis is predicting patient's response to nCRT, while relatively fewer studies have explored radiomics for predicting long-term outcomes including FFDM, overall survival and progression free survival (PFS) (Bundschuh et al., 2014; Chee et al., 2017; Dinapoli et al., 2018; Meng et al., 2018; Park et al., 2020).

Radiomics models were commonly developed on features extracted from T2-w MRI (Caruso et al., 2018; Cusumano et al., 2018; Dinapoli et al., 2018; Antunes et al., 2020; Petkovska et al., 2020), and mpMRI (De Cecco et al., 2016; Nie et al., 2016; Giannini et al., 2019; Zhou et al., 2019). Few studies have considered radiomic features extracted from CT imaging (Chee et al., 2017; Bibault et al., 2018), PET (Bang et al., 2016; Van Helden et al., 2018), or a combination of CT and MRI features (Li et al., 2020b). However, it is difficult to evaluate the findings of these studies due to the complexity, and heterogeneity of the literature that has been published in the past decade. Furthermore, different studies have proposed different imaging biomarkers, and there is lack of consensus on which features are more relevant or generalizable for predicting treatment outcomes in LARC. A comprehensive review of radiomics studies published for prediction of patient response and long-term outcomes in LARC have appeared (Davey et al., 2021), but none of them have externally validated previously published radiomics signatures.

Therefore, in this chapter, we aimed to externally validate radiomics signatures that were previously developed by other researchers for predicting tumour response to nCRT or FFDM in LARC. The work presented within this chapter has been published in an international journal (Shahzadi et al., 2022b).

4.2 Materials and methods

4.2.1 Patient cohort

The evaluation in this external validation study is based on multicentre, retrospective data of 190 patients as described previously in Chapter 3, Section 2.2.1. The data was collected from four partner sites and divided into training and validation data based on the site (122 and 68 patients, respectively). The details of the patient characteristics for both cohorts are summarized in Chapter 3 Table 3.1.

The considered endpoints for external validation were tumour response to nCRT and FFDM. A detailed description of tumour response, tumour regression grade and FFDM was described in Chapter 3. For the external validation study, patients were stratified into response groups according to the grading scheme indicated in the respective manuscript. Regardless of grading scheme, Dworak (Dworak et al., 1997) criteria of TRG (as explained in Appendix Table A.1) were used as reference standard for response prediction. The patients were stratified into either (i) responders (corresponding to TRG 3 and 4) and non-responders (corresponding to TRG 0-2) or (ii) complete responders (corresponding to TRG 4) and non-responders (corresponding to TRG 0-3) (iii) complete responders as in (ii) and non-responders (corresponding to TRG 0 and 1 thus excluding partial responders).

For the validation of published signatures, non-contrast-enhanced treatment planning CT and T2-w MRI data were used. Where necessary, imaging information was also combined with patient clinical information, as described below.

4.2.2 Literature search

The design for literature search is shown in Figure 4.1. A free search was conducted using Google Scholar and PubMed until October 2021 to identify the relevant radiomics-based LARC studies for the validation of biomarkers. The following free search keywords were used: ‘rectal cancer’ OR ‘Locally advanced rectal cancer’, ‘radiomics’, ‘response prediction’ OR ‘response to neoadjuvant chemoradiotherapy’, ‘distant metastases prediction’ OR ‘prognosis’, ‘deep learning’, ‘machine learning’.

The studies were then reviewed for their eligibility. Studies that met the following criteria were included: (1) patients with LARC, (2) radiomics analysis on pre-treatment T2w MRI or non-contrast-enhanced CT, (3) radiomics features extracted from primary tumour or GTV, (4) normo-fractionated nCRT (dose 45–55 Gy) followed by surgery, (5) clear radiomics workflow and definition of finally used features available, (6) prediction of FFDM as time to event endpoint. The search and inclusion of studies were supervised by two reviewers with expertise in radiomics modelling.

The following data were extracted from the included studies: (1) sample size and distribution of training and validation dataset (if any), (2) nature of study, i.e. single centre or multicentre, (3) clinical characteristics of patient cohort (4) used imaging modality, (5) reference standard for TRG, (6) image pre-processing workflow, (7) feature extraction geometry, i.e. 3D, 2D, or largest slice, (8) applied feature extraction framework, (9) final classification/regression model or statistical test, (10) features included in final model, (11) final model parameters (if any), and (12) reported results. The studies were arranged in chronological order of year of publication.

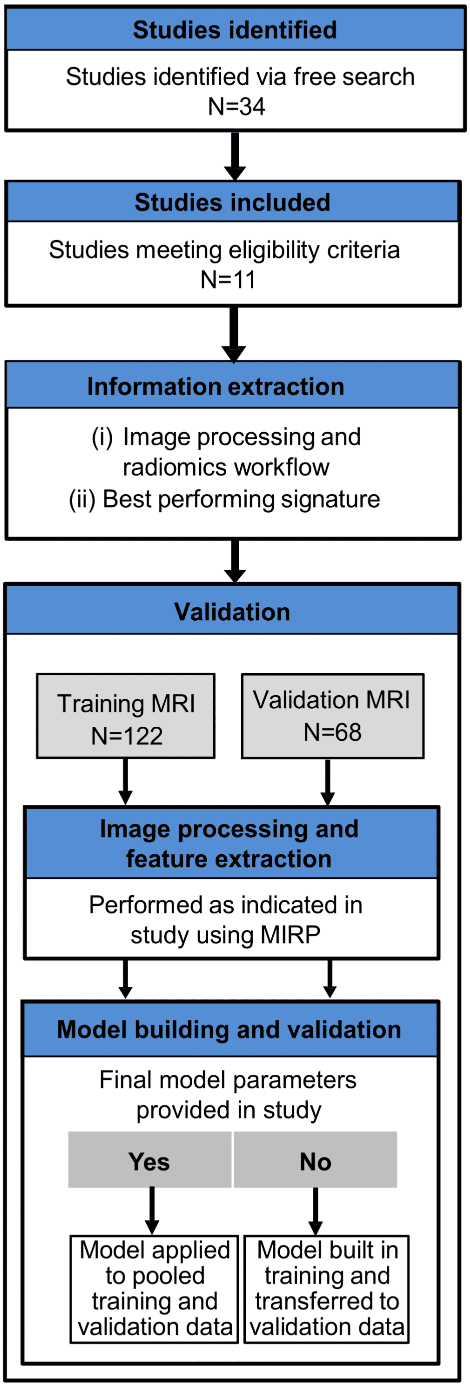

Studies identified
Studies identified via free search
N=34
Studies included
Studies meeting eligibility criteria
N=11
Information extraction
(i) Image processing and radiomics workflow
(ii) Best performing signature
Validation
Training MRI
N=122
Validation MRI
N=68
Image processing and feature extraction
Performed as indicated in study using MIRP
Model building and validation
Final model parameters provided in study
Yes
No
Model applied to pooled training and validation data
Model built in training and transferred to validation data

Figure 4.1: Design of the external validation study. Studies were identified via free search using Google Scholar and PubMed and excluded if the inclusion criteria were not fulfilled. Information regarding image processing, radiomics workflow, and the best performing radiomics signature was extracted as reported. Image processing and feature extraction was reproduced using MIRP (Zwanenburg et al., 2019b). Finally, validation was performed either on the pooled training and validation data if model parameters were reported in the study or the model was re-trained on the training data and validated on the validation data.

4.2.3 Data pre-processing

Feature extraction was carried out in accordance with IBSI guidelines for the studies that qualified for inclusion in the validation analysis, using the MIRP Python toolkit (version 1.1.3). The reported features in each study were mapped to the IBSI manual's closest-matching synonyms. A feature was excluded from validation analysis if (i) it was not defined in the IBSI manual or (ii) MIRP cannot extract it, and the remaining features were considered candidates for validation investigation. Image pre-processing (e.g. image interpolation, image normalization, bias correction) and feature extraction parameters (e.g. feature extraction in 2D, 3D or from the largest tumour area, discretization used for histogram or texture features, LoG or wavelet transformations) were replicated for each study if indicated. If specified, feature extraction parameters from the study were repeated; if not, the MIRP default settings were applied.

4.2.4 Radiomics modelling

For this external validation study, the pooled training (DD and MTU cohort) and validation data (F and FR) were used for biomarker validation if final model coefficients or model training parameters were provided in the respective study, or a statistical test, e.g. t-test or Wilcoxon test, was performed for associating the considered biomarker to the endpoint of interest. On the other hand, if model coefficients or model training parameters were not provided, the given radiomic features were used to re-train a predictive model on the training data, and was subsequently validated on the validation data. Clinical features were combined with imaging biomarkers if mentioned in the study.

4.2.5 Statistical analysis

As shown in Table 3.1, the following baseline clinical parameters were available: gender, age, tumour localization, UICC stage, grading, T stage, N stage, surgery type, chemotherapy type. Categorical variables of the clinical data were compared between the training and validation data by the χ^2 test, whereas continuous variables were compared using the Mann-Whitney-U test. The majority of the studies evaluated the association between final model predictions and the tumour response using the AUC metric, and so did we. The estimated value and the 95% confidence interval of these metrics were computed. For the studies with acceptable performance for tumour response to nCRT, model calibration was also assessed via a calibration plot and the HL test (Hosmer & Lemesbow, 1980). Where required, two-sided statistical tests were performed with a significance level of 0.05. Statistical analysis and model building for validation analysis was performed in R (version 4.0.3).

4.3 Results

In total, 34 studies were identified as relevant based on their titles and abstracts. All identified studies were performed on patients with LARC that were treated with nCRT followed by surgery with the aim of predicting tumour response using radiomics. 23 studies were excluded after full text review due to following reasons: 3 studies used contrast enhanced CT data that was not available in our dataset (Chee et al., 2017; Li et al., 2020a; Zhuang et al., 2021), 4 studies used both pre and\or post treatment data (Aker et al., 2019; Boldrini et al., 2019; Jeon et al., 2019; Li et al., 2021), 5 studies used pre-treatment multiparametric MRI (mpMRI) to develop a final signature with no standalone T2-w MRI signature being reported (Liu et al., 2017; Giannini et al., 2019; Zhou et al., 2019; Bulens et al., 2020; van Griethuysen et al., 2020), 2 studies did not report any final signature (Bibault et al., 2018; Delli Pizzi et al., 2021), 3 studies could not be reproduced as the radiomics workflow or feature definition was not clearly explained (Yi et al., 2019; Li et al., 2020b; Yuan et al., 2020), 1 study was excluded as the considered ROI was not the primary tumour (Shaish et al., 2020), 3 studies were excluded as authors reported failure of radiomics to predict the outcome of interest (Rao et al., 2016; Hamerla et al., 2019; Crimı et al., 2020), 2 studies were excluded as the reported signature was computed from feature maps, which are currently not supported by MIRP (Ferrari et al., 2019; Coppola et al., 2021). Finally, eleven studies were included for external validation analysis. All of them used T2-w MRI for predicting tumour response and were published between 2015 and 2020. One study was prospective, nine were retrospective, and three were multicentric. Two of these multicentre studies considered both clinical features and imaging biomarkers.

The clinical characteristics of the 11 included studies are given in Appendix Table B.1. Our external validation results are summarized in Table 4.1. The considered biomarkers and their corresponding synonyms together with image processing and feature extraction details for included studies are summarized in Appendix Table B.2. Except for one study, none of the included studies could be validated, i.e. they showed p-values above 0.05 and/or a training/validation AUC significantly below the reported value in the study with a 95% confidence interval including the value 0.5. The only study that could be validated is by Petkovska et al. (Petkovska et al., 2020). An acceptable performance was observed on our pooled data (AUC=0.64 [0.51-0.77]). In a study by Chidbaram et al. (Chidambaram et al., 2017), pathological complete responders showed a significant association with tumour volume delineated on T2-w image (Mann-Whitney-U test p=0.013). This was somewhat confirmed in our analysis, where we observed a statistical trend (p=0.061). However, radiomics analyses are not needed to assess the tumour volume. For the study by Antunes et al. (Antunes et al., 2020), the random forest model created on a single feature was not successful on our training data but achieved an acceptable performance on the validation data (AUC: Train, Validation = 0.48, 0.63). Still, on the pooled training and validation data, the selected feature was insignificant (Mann-Whitney-U test p=0.12). Below we show details for the validation of each study and describe the results as summarized in Table 4.1.

Table 4.1: Overview of studies included in validation analysis. For all included studies, patients were treated with nCRT followed by resection. Radiomics analysis was reported on pre-treatment T2-w MRI with features extracted from the primary tumour region. The column Validation approach indicates whether model coefficients or statistical tests were applied on the pooled training and validation data (Pooled) or the model was re-trained on the training data and validated on the validation data (Train/valid). AUC: area under a curve (with 95% confidence interval in brackets), MRI: magnetic resonance imaging, nCRT: neoadjuvant chemoradiotherapy.

Study	**Study type**	**Validation approach**	**Final results from study**	**Results from validation analysis (Unadjusted p-value)**

De Cecco (2015-16)	Prospective, single centre	Pooled	AUC = 0.91, 0.86
			p-value = 0.01, 0.01
				AUC=0.56 (0.44-0.68)
				p-value=0.31

Chidbaram (2017)	Retrospective, single centre	Pooled	p-value = 0.013	p-value=0.061
Caruso (2018)	Retrospective, single centre	Pooled	p-values <0.05 for all features	p-values >0.05 for all features
Casumano (2018)	Retrospective, multicentre	Pooled	AUC=0.79	AUC=0.58 (0.46-0.70)
Dinapoli (2018)	Retrospective, multicentre	Pooled	AUC=0.75	AUC=0.59 (0.47-0.71)
Meng (2018)	Retrospective, single centre	Pooled	p-value=0.02	p-value=0.098
Cui (2019)	Retrospective, single centre	Pooled	AUC=0.73	AUC=0.52 (0.38-0.64)
Antunes (2020)	Retrospective, multicentre	Train/valid	Train\Valid AUC= 0.699\0.712 Skewness-Laws Wave-Ripple (p-value Train=1.6×10^{-4})	Results on Skewness-Laws Wave-Ripple Train\valid AUC= 0.48 (0.36-0.57) \0.63 (0.52-0.76) p-value Train\valid=0.71\0.055 p-value Pooled=0.12
Petkvoska (2020)	Retrospective, single centre	Pooled	AUC=0.75	AUC=0.64 (0.51-0.77)
Petresc (2020)	Retrospective, single centre	Pooled	AUC=0.80	AUC=0.48 (0.38-0.57)

Included studies

The two consecutive single centre studies by De Cecco et al. (De Cecco et al., 2015; De Cecco et al., 2016) extracted first-order intensity features from the tumour ROI delineated on the largest slice for the prediction of pathological complete responders (TRG=4) and non-responders (TRG=0-3). Images were transformed using the spatial scale filter (SSF4) filter (alternate name for LoG filter). In both studies SSF4 kurtosis was reported to be significant (Wilcoxon rank-sum test) in predicting response groups (AUC=0.91, 0.86; p-value= 0.01, 0.01). To replicate this study, we extracted features from the largest tumour slices in our pooled cohort. Images were transformed using the LoG filter and validation was performed for stat_kurt (IBSI: IPH6). We adapted TRG status to match the study (TRG=4 vs TRG<4 following Dworak et al. (Dworak et al., 1997)). The study was not validated successfully, as stat_kurt achieved an AUC of 0.56 and a p-value of 0.31.

The study by Chidambaram et al. (Chidambaram et al., 2017) as no high-dimensional features were extracted from imaging data. However, this study analysed some basic morphological and statistical features for tumour response prediction using ADC maps of T2-w MRI. Pre-treatment

MRI volume was found to be significantly associated with tumour response (complete responders vs incomplete/non-responders following American Joint Committee on Cancer (AJCC)) with p-value of 0.013 from a Mann–Whitney-U test. We extracted morph_vol (IBSI: RNU0) from the 3D GTV on the pooled cohort and analysed its significance via the Mann–Whitney-U test for tumour response (TRG=4 vs TRG<4 following Dworak et al. (Dworak et al., 1997)). The study was not validated successfully, however, we observed a statistical trend for morph_vol feature (p-value=0.061)

In a retrospective study conducted by Caruso et al. (Caruso et al., 2018) on a small cohort of 8 patients, the directional GLCM (no discretization mentioned) features extracted from T2-w MRI were shown to be significantly different between pathological complete and incomplete responders, while partial responders were excluded from the study. The logistic regression model followed by a Wald test for feature importance was used to analyse the predictive performance of features in the model. The study reported a p-value<0.05 for included features. For validation, we excluded partial responders from our pooled training and validation data, thus including only 65 patients (TRG=4 vs TRG=0 following Dworak et al. (Dworak et al., 1997)). Directional features are not supported by MIRP. Thus, we extracted 2D GLCM features using the average method, i.e. features were computed from all matrices and then averaged using a fixed bin number of 64. IBSI synonyms for validated features are mentioned in Appendix Table B.2. Validation was performed by fitting a multivariable logistic regression model followed by the Wald test to obtain p-values for each feature. The study was not validated successfully, as all included features showed an insignificant p-value (> 0.05) on our pooled cohort.

Two multicentre retrospective studies (Cusumano et al. (Cusumano et al., 2018) and Dinapoli et al. (Dinapoli et al., 2018)) proposed statistical and intensity histogram features extracted from LoG transformed images together with clinical T and N stage for tumour response prediction (TRG=1 vs TRG>1, Mandard et al. (Mandard et al., 1994)). The study by Cusumano et al. (Cusumano et al., 2018) showed an AUC of 0.79, while the study by Dinapoli et al. (Dinapoli et al., 2018) showed an AUC of 0.75. The signature presented by Cusumano et al. also included an additional fractal feature that could not be extracted by MIRP. The studies did not report discretization for intensity histogram features. Final model coefficients were reported. To validate the study by Cusumano et al. (Cusumano et al., 2018) pixel intensities inside the GTV were normalized by the 99th percentile. The fractal feature was excluded, and the remaining features were extracted from 2D slices from the discretized intensity histogram (25 bins). The model provided in each study was then applied to our pooled cohort. Details of features and their corresponding IBSI synonyms are presented in Appendix Table B.2. Both studies were not validated successfully as they showed AUC <0.60 on our pooled cohort.

The study by Meng et al. (Meng et al., 2018) analysed statistical and intensity histogram features extracted from the largest tumour slices for tumour response prediction. The image intensities were discretized. However, the study did not report the number of bins used for discretization. Further response groups were created as responders (TRG=1-2) and non-responders (TRG=3-5) following Mandard et al. (Mandard et al., 1994). The study showed image kurtosis to be significantly different between responders and non-responders (Mann-Whitney-U test p-value = 0.02). For validation, we extracted intensity histogram features from the largest tumour slices after discretizing image intensities into 25 uniform bins from the pooled cohort. Finally, the ih_kurt (IBSI: C3I7) feature was tested for tumour response prediction (TRG=3-4 vs TRG=0-2, Dworak et al. (Dworak et al., 1997)) using the Mann-Whitney-U test. The study was not validated successfully, since the ih_kurt feature showed a p-value of 0.098 on our pooled cohort.

The study by Cui et al. (Cui et al., 2019) used mpMRI to develop a radiomic signature. However, a standalone T2-w MRI signature comprising directional GLCM, statistical, and morphological features extracted from the GTV for tumour response prediction (TRG=1 vs TRG>1 following Mandard et al. (Mandard et al., 1994)) was also presented. The coefficients of the final model built on z-score normalized features were reported by the study. The model was shown to achieve an AUC of 0.73. However, feature extraction parameters including discretization, merge method, and feature extraction plane, i.e. 2D/3D for GLCM features were not reported by. Since directional features are not supported by MIRP, we extracted 3D GLCM features with a fixed bin number of 64 using the average method, i.e. features were computed from all matrices and then averaged, thus compensating for directional texture features. We then applied the model coefficients provided in the study to compute the radiomics score presented in the study and assessed the AUC on our pooled cohort. Since we can get only closely related features for this study, we also fitted the logistic regression model on the training data and applied it to the validation dataset. However, this validation was not successful (AUC: train/valid = 0.72/0.32). We excluded 'HaralickCorrelation_angle90_offset7' from the signature, as by definition Haralick features are no different from the GLCM feature 'Correlation_angle135_offset7' except for the difference in angle (Zwanenburg et al., 2020).

A retrospective study conducted by Antunes et al. (Antunes et al., 2020) reported 4 (1 Haralick co-occurrence, 2 Gradient organization, 1 Laws energy response) features extracted from the largest tumour slice for tumour response prediction (TRG=4 vs TRG<4 following Dworak et al. (Dworak et al., 1997)). Pre-processing applied before feature extraction in this study includes (i) image interpolation to 0.781 × 0.781 × 4.0 mm, (ii) N4 bias correction, and (iii) intensity normalization with a reference to the mean intensity of the obturator internus muscle. The signature comprised of above-mentioned 4 features achieved a training and validation AUC of 0.699 and 0.712 respectively. To validate this study, we replicated steps (i) and (ii) of pre-processing, while for step (iii) relative range intensity normalization [0, 90] was performed. Furthermore, gradient organization features assessed in the study are not IBSI compliant. Therefore, none of the organization features could be extracted. Finally, a random forest model was created on training data using one feature, i.e. Skewness-Laws Wave-Ripple w5s5, and subsequently transferred to the validation data. The model achieved training and validation AUC of 0.48 and 0.63, respectively. Feature importance was computed via the Mann-Whitney-U test. On pooled data, feature was not significant (p-value Pooled=0.12).

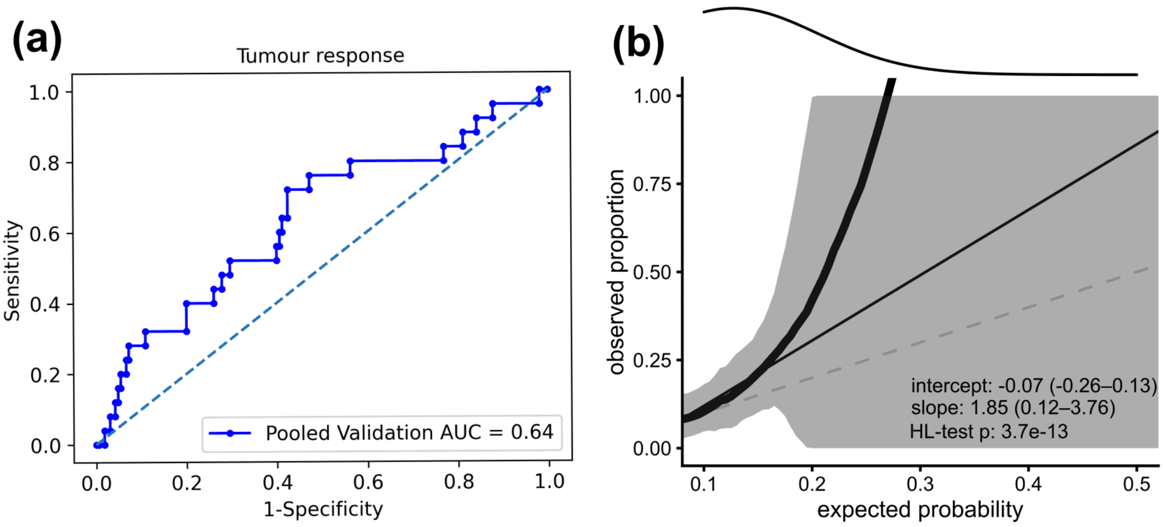


Figure 4.2: (a) Receiver operating characteristics (ROC) curve and (b) calibration plot for tumour response prognosis in our pooled training and validation data based on the study by Petkovska et al. (Petkovska et al., 2020). For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines). The density of expected probabilities is shown above the calibration plot. The calibration line roughly follows the optimal expectation for most of the observations at small probabilities.

Petkovska et al. (Petkovska et al., 2020) reported 6 features (2 morphological, 2 grey level texture, 2 directional Gabor) extracted from the 3D tumour volume using T2-w MRI. The proposed radiomics signature achieved an AUC of 0.75 for tumour response prediction. In the study, images and corresponding tumour mask were interpolated 1 × 1 × 1 mm prior to feature extraction. The study reported discretisation of normalized image intensities using fixed bin size=128, however the normalization step was not clearly explained. Moreover, to unambiguously specify Gabor filters at least two out of three parameters are required (scale (sigma), wavelength (lambda), bandwidth). The study reported only sigma values and initial angles for Gabor filters. Model coefficients were provided in the study for the T2-w signature. In our validation analysis, we interpolated MR images to isotropic 1 mm resolution using cubic interpolation followed by standard normalization of image intensities within the soft tissue region. Grey level intensities were discretized using fixed bin size=128 for texture features. Gabor transformed features were extracted using angle and sigma values as reported in the study, however we used lambda=4 to complete feature extraction. Finally, the radiomics score was computed by applying model coefficients using our pooled cohort and subsequently the AUC was computed for tumour response prediction (TRG=4 vs TRG <4 following Dworak et al. (Dworak et al., 1997)). The study was somewhat validated, as we observed acceptable performance on our pooled data (AUC=0.64 [0.51-0.77]). Figure 4.2 shows the ROC and calibration plot for this validation.

A retrospective single-centre study conducted by Petresc et al. (Petresc et al., 2020) proposed a signature comprising second-order texture features on LoG and wavelet transformed images to predict tumour response (TRG=3 vs TRG=1,2 following Ryan et al. (Ryan et al., 2005)). These features achieved an AUC of 0.80. Image intensities within the GTV were discretized using a fixed-bin width of 5. However, discretization of wavelet features, the merging method for texture features, and the neighbourhood distance for GLCM features were not documented. Model coefficients were provided in the study. In our validation analysis, we applied pre-processing steps as indicated in the study, including image standardization (mean=0, std=100), B-spline interpolation, re-segmentation of segmentation mask. For feature extraction we used fixed bin number=64, fixed bin size=5, merge method=average, and GLCM neighbourhood distance=1. Further, we excluded the 'wavelet_hhl_glcm_MCC' feature, as it is not standardized by the IBSI. For the remaining features, we computed a radiomics score by applying model coefficients on z-score normalized features using our pooled cohort and subsequently computed the AUC for tumour response prediction (TRG=3,4 vs TRG=0-2 following Dworak et al. (Dworak et al., 1997)). The study was not validated successfully and achieved an AUC of 0.48.

4.4 Summary and discussion

In general, the external validation of previously published radiomics signatures developed for tumour response prediction based on our multicentre data was not successful. Remarkably, no significant results were obtained, except for one study by Petkovska et al. (Petkovska et al., 2020) (AUC=0.64), which overall indicates a potential lack of reproducibility for radiomics studies.

The results shown by the included studies are promising, however most of these studies are based on data from a single centre without any external validation, which can be one of the factors leading to low reproducibility of radiomics models. Considering MRI-based multicentre radiomic studies with an independent validation for patients with LARC, Antunes et al. (Antunes et al., 2020) used features extracted from laws kernels and gradient organization responses. In our validation analysis, only skewness-laws features could be validated. The corresponding feature used by Antunes et al. (Antunes et al., 2020) was not significant in training and showed a statistical trend in validation (p=0.055). Dinapoli et al. (Dinapoli et al., 2018) used first-order intensity histogram-based features, while the study by Cusumano et al. (Cusumano et al., 2018) additionally used fractal features in the final signature to build the model. Both studies also combined clinical features (cT and cN) with the radiomics signature. In our validation study, these signatures did not show a good performance (AUC < 0.60).

Single centre retrospective studies have shown promising results for tumour response prediction in LARC. De Cecco et al. (De Cecco et al., 2016) and Caruso et al. (Caruso et al., 2018) showed a significant association (p<0.05) of FO statistical and GLCM features, respectively, with tumour response to nCRT on small cohorts ($\leq$15 subjects). However, in our validation analysis, no significant association has been found for these features (p>0.05). Coppola et al. (Coppola et al., 2021) showed that heterogeneity of local skewness is associated to tumour response (AUC=0.90). Ferrari et al. (Ferrari et al., 2019) showed that complete responders have higher GLCM energy and good responders have high expression of histogram features (AUC=0.87).

These studies could not be validated as the features were extracted from feature maps, which are currently not supported in MIRP. More recent studies showed the association of SOT features with tumour response prediction. The studies by Delli Pizzi et al. (Delli Pizzi et al., 2021) and Petresc et al. (Petresc et al., 2020) showed an AUC of 0.79 and 0.80 in internal validation, respectively. However, validating the results of Petresc et al. (Petresc et al., 2020) on our multicentre data was not successful (AUC=0.48). Fewer studies have investigated the performance of CT imaging for tumour response prediction to nCRT using patient populations treated with standard procedures, i.e. nCRT followed by TME (Rao et al., 2016; Chee et al., 2017; Bibault et al., 2018; Hamerla et al., 2019), or combined CT and MR imaging (Li et al., 2020b; Zhang et al., 2020). Bibault et al. (Bibault et al., 2018) developed a model for the prognosis of tumour response with radiomics features extracted from treatment plan CT data using DNN with an AUC of 0.72. However, the study did not mention any finally selected features for building the deep learning model and hence it could not be validated. Contrast enhanced CT images provide better evaluation of tumours compared to unenhanced images and some studies have also utilized contrast enhanced CT. For example, Chee et al. (Chee et al., 2017) demonstrated that pre-treatment contrast enhanced CT-based FO features were associated with tumour response prediction (responders showed low entropy, high uniformity, and low standard deviation). However, the validation of this study was also not possible as our multicentre data contains treatment planning CT without contrast enhancement.

In this work a literature search was also performed to identify studies that handle FFDM as time-to-event endpoint. However, to the best of our knowledge, none of the studies has used radiomics modelling for FFDM prognosis in LARC patients.

One major issue in radiomics analyses is feature reproducibility and the lack of consensus on which features should be extracted from clinical imaging data. In our validation study, we experienced limited reproducibility of published literature. Only 32% of the eligible literature could be assessed for their validation performance with our data and methods, mostly due to the use of different software implementations and underreporting of methods employed for radiomics analysis of LARC. Important details such as image processing for feature extraction (e.g. discretization for intensity and texture features), final signatures together with their interpretation and final models were not always provided. Thus, there is a strong need of standard radiomics process for signature definition for both reproducibility and progression of radiomics towards clinical application.

Although some studies have used large cohorts for radiomics analyses in LARC, external validation was rarely performed. Only 4 studies (Cusumano et al., 2018; Dinapoli et al., 2018; Antunes et al., 2020; Shaish et al., 2020) have used retrospective multicentre cohorts with a maximum of 3 data centres involved, which may lead to a low generalizability of the presented radiomic signatures. To tackle such problems, in our multicentre study, we have established and externally validated radiomics signatures in accordance with the IBSI guidelines, and we report parameters and algorithms used for their extraction, transformation, stability analysis, and modelling. In addition to the lack of standardization in the radiomics workflow, there is lack of standardized imaging protocols as well. This can obstruct the successful validation of radiomics models, e.g. for imaging from MR scanners of different vendors or different magnetic field strengths, because such differences may lead to the extraction of differently distributed features (Cusumano et al., 2021b). Standardization at hardware level is costly, thus there is a need to develop generalizable models by incorporating data from different scanners and protocols. We addressed this issue by using multicentre data independent of vendor and imaging protocols for training and validation. Furthermore, we observed significant differences between the clinical characteristics of our pooled cohort and the external cohorts included in the validation study (mainly clinical T and N stage). These differences may explain part of the observed reduced performance of the published models in our external validation analysis.

In conclusion, we observed low performance of published radiomics literature in our external validation analysis, which indicates an overall lack of reproducibility and the need for standardization in radiomics procedure and reporting before its prospective clinical application.

5 Radiomics for the detection of tumour residuals after surgery of glioblastoma based on [^{11}C] methionine PET and T1c-w MRI

5.1 Motivation

The current standard of care for newly diagnosed GBM is maximum safe surgical resection followed by CRT. Despite multimodal treatment, patients with GBM still face an overall poor prognosis, with a high recurrence rate and 5-year survival probability of only 5% (Alexander & Cloughesy, 2017). Gross total resection of GBM has been associated with improved local control and survival compared to subtotal or partial resection (Coburger et al., 2017; Wang et al., 2019). However, due to infiltrative growth patterns, total resection cannot always be achieved, and residual tumour cells may persist after resection. These residual tumours are widely held to be responsible for recurrence of the tumour, leading to overall poor prognosis of GBM patients (Nazzaro & Neuwelt, 1990).

The residual tumour status is usually evaluated by T1c-w MRI. The area of gadolinium-diethylenetriaminepentaacetic acid (Gd-DTPA) enhancement is generally assumed to correspond well to the main mass of active tumour tissue. However, the reliability of T1c-w MRI in distinguishing tumour tissue from unspecific treatment effects such as post-surgical blood-brain barrier breakdown is limited. For example, reactive transient blood–brain barrier alterations with consecutive contrast enhancement can mimic tumour progression. This phenomenon, so-called pseudoprogression, is seen in 20%–30% of cases (Kumar et al., 2000).

Amino acid PET, e.g. L-[methyl-^{11}C] methionine (MET) has been shown to be particularly useful for determining the extent of cerebral gliomas more precisely than MRI alone (Galldiks et al., 2010; Galldiks et al., 2012; Harat et al., 2016). The high uptake of radiotracer in residual tumour reflects an increased expression of amino acid transporters (Jager et al., 2001).

The accurate detection of tumour residuals in post-surgical imaging can help to identify patients with a poor prognosis, who may benefit from escalated treatment (Nazzaro & Neuwelt, 1990). Commonly, the imaging-based assessment of the residual tumour status is done visually by experienced radiation oncologists, nuclear medicine experts, and radiologists in a complex evaluation procedure that is at risk for inter-rater variability (Kubben et al., 2010). Thus, automatic methods of detecting residual tumour status may be helpful to support the clinical decision.

Currently, only few studies have evaluated the automatic detection of residual tumours. A study by Chow et al. (Chow et al., 2014) examined a semi-automated computer aided volumetry (CAV) approach to quantify residual disease in GBM and noted no significant difference compared with manual volumetric measurements, which are time-consuming and impractical in a busy clinical practice. Meier et al. (Meier et al., 2017) used a fully automated end-to-end machine learning based algorithm for segmentation of tumour residuals. Similarly, Krivoshapkin et al. (Krivoshapkin et al., 2019) used an automated tool based on a mathematical model to segment residual tumours. However, the analyses in these studies have been performed only on post-surgical contrast enhanced MRI using small cohorts without independent validation.

Conventional and DL-based radiomics have been widely used for medical image analysis, as non-invasive methods for diagnosis support and biomarker discovery. However, to the best of our knowledge, no studies have evaluated the diagnostic performance of post-surgical T1c-w MRI and methionine (MET)-PET for the detection of tumour residuals.

Therefore, in this study, we developed and independently validated conventional radiomics and 3D-CNN models to detect the residual tumour status in postoperative MET-PET and gadolinium-enhanced T1-w MRI in patients with newly diagnosed GBM.

5.2 Materials and Methods

5.2.1 Patient cohort

Imaging and clinical data of 132 adult patients were collected from the PETra trial, which is a prospective one-arm, single-centre, nonrandomized biomarker study as described elsewhere (Seidlitz et al., 2021) and from an additional retrospective validation cohort. All patients were newly diagnosed with histologically confirmed GBM and were treated at the University Hospital and Faculty of Medicine Carl Gustav Carus. 85 consecutive patients from the PETra trial (ethics id. EK41022013) were allocated to the training data, while 47 consecutive patients from the validation trial (ethics id. EK390072021) were allocated to an independent test data. Patients underwent standard CRT with standard radiotherapy dose of 60 Gy and temozolomide, starting within 7 weeks after surgery. The inclusion criteria for this study were: T1c-w MRI acquired contemporaneously with MET-PET before RCT with sufficient imaging quality and availability of considered endpoints. Patient characteristic of training and test cohort are summarized in Table 5.1.

5.2.2 Experimental design

We developed and independently validated conventional radiomics and deep learning (DL) models for the detection of the residual tumour status in patients with, GBM based on MET-PET and T1c-w MRI data acquired before RCT. Figure 5.1 summarizes the design of this study. For the conventional radiomics analysis, we used the CTV to separately compute imaging features in MET-PET and T1c-w MRI. These features included first-order features (local, statistical, intensity histogram and intensity volume histogram), second-order texture features, and Laplacian of Gaussian (LoG) transformed intensity features. The features were filtered for stability under

Table 5.1: Patient, tumour, and treatment characteristics for the training and test data.

Variable		**Training (85)**		**Test (47)**		
		Median	**Range**	**Median**	**Range**	**p-value**
Age	years	58	23-82	61	24-77	0.049
TTR	months	7.43	0-73.0	9.76	1.15.58.0	0.60
OS	months	16.6	1.54-73.0	13.9	1.94-58.0	0.10
		Number	**%**	**Number**	**%**	
Gender	Male/female	51/34	60.0/40.0	31/16	66.0/34.0	0.63
ECOG	0/1/2/unknown	45/35/5/0	52.9/41.2/5.9/0	21/19/3/4	44.7/40.4/6.4/8.5	0.054
MGMT	Wildtype/methylated/ unknown	56/29/0	65.9/34.1/0	20/26/1	42.6/55.3/2.1	0.019
Resection	GTR/STR/BIO	49/29/7	57.6/34.1/8.2	26/21/0	55.3/44.7/0.0	0.09
IDH	Wildtype/mutated/ unknown	75/6/4	88.2/7.1/4.7	44/2/1	93.6/4.3/2.1	0.60
PET status	0/1 (negative, positive)	28/57	32.9/67.1	17/30	36.2/63.8	0.85
MRI status	0/1 (negative, positive)	49/36	57.6/42.4	23/24	48.9/51.1	0.44
TTR status	0/1 (censored, event)	11/74	12.9/87.1	12/35	25.5/74.5	0.11
OS status	0/1 (censored, event)	13/72	15.3/84.7	17/30	36.2/63.8	0.011

Abbreviations: BIO, biopsy; ECOG, Eastern Co-operative Oncology Group; GTR, gross total resection; IDH, isocitrate dehydrogenase; MGMT, O6-methylguanine DNA methyltransferase; MRI, magnetic resonance imaging; OS, overall survival; PET, positron emission tomography; STR, subtotal resection; TTR, Time-to-recurrence. Age was compared using Mann-Whitney-U test, TTR and OS were compared using log-rank test and Categorical variables were compared using χ^2 test between training and test data.

small image perturbations and clustered. Separate radiomic models for each imaging modality were developed using the data in the training data. Three different machine-learning algorithms of varying complexity were assessed, including logistic regression (GLM_logistic), Xgboost linear model (XGB_lm), and random forest (RF). The finally selected features were applied to the test data and performance was compared between PET- and MRI-based models. In our DL analysis for detection of residual tumour status in PET and MRI, end-to-end feature extraction and modelling were performed using three different 3D-CNN architectures, i.e. 3D-VGGNet, 3D-Resnet, 3D-DenseNet (see Section 2.6.3 for more details on CNN architectures). Model losses were optimized using the BCE loss. 3D-CNN models were trained from scratch on image patches extracted around the CTV centre of mass individually for each imaging modality. To address the limited data problem and to improve the generalizability of 3D-CNN models, training was performed using two approaches: (i) without data augmentation and (ii) with data augmentation (see Section 2.6.3 for more details on data augmentation). We then applied the developed models to the independent test data and compared their performance. The performance of conventional radiomics and 3D-CNN models was evaluated using the AUC, sensitivity and specificity for the detection of residual tumour status in PET and MRI. Image processing and modelling details are explained in the following paragraphs.

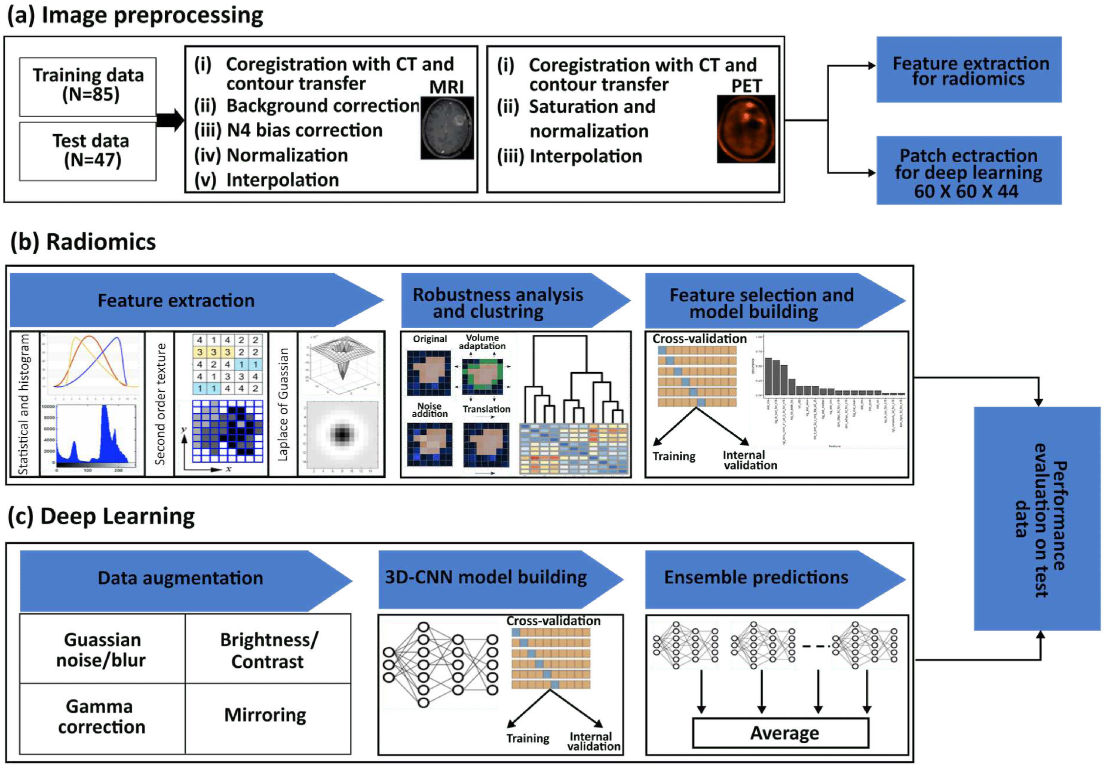


Figure 5.1: Study design. (a) Image preprocessing. (b) Radiomics features were extracted from each imaging modality, analysed for robustness, and clustered. Radiomics signatures for MET-PET and T1c-w MRI were developed in a cross-validation approach and applied to the test data. (c) 3D-CNN models were trained in a cross-validation approach. Subsequently, the performance of ensemble predictions was evaluated on the test data.

5.2.3 Image acquisition, endpoints, and contouring

PET/MRI studies were performed on a 3 Tesla Ingenuity TF PET/MRI scanner (Philips Healthcare, Best, The Netherlands). Image acquisition details for PET and T1c-w MRI of training and test data are summarized in Appendix Table C.1. The considered endpoint was the detection of residual tumour status in PET and MRI individually. To assess the residual tumour status after surgery, PET and MRI data were evaluated qualitatively to form the binary ground truth labels of PET status and MRI status, as follows. Reconstructed MET-PET imaging acquired 20-40 minutes after intravenous injection of the tracer was evaluated by a nuclear medicine expert using the software package ROVER (ABX). Patient's PET status was labelled as positive (1) if focal uptake areas represented the presence of true residual tumour without physiologically enhanced uptake or enhancement in post-surgical alteration. Residual tumour status for MRI was assessed by a radiation oncologist using early post-surgical MRI (24–48 hours after surgery) together with operative reports and the second baseline MRI obtained contemporaneously with PET. The second baseline MRI was acquired a median 23 days after surgery. Therefore, it is prone to unspecific changes. If the second MRI showed no residual tumour, the residual MRI status was set as negative (0). In case of distinct progression between the two MRI scans, MRI status was changed to positive (1). Difficult cases with small residual tumours or laminar enhancement, so that distinction from residual blood in the cavity was difficult, were independently reviewed by an experienced radiologist. A more detailed description of the qualitative analysis performed by clinical experts to evaluate residual tumour status is given in (Seidlitz et al., 2021).

For automatic detection of residual tumour status in PET and MRI with conventional radiomics and DL, PET/MRI images were rigidly co-registered to the planning CT using the treatment planning system RayStation 8B SP2 (RaySearch Laboratories, Stockholm, Sweden) and the clinical target volume that received at least 50 Gy (CTV50) was transferred to MET-PET and T1c-w MRI. After registration, imaging datasets were retrieved from the RayStation for further analysis.

5.2.4 Image pre-processing, and feature extraction

Figure 5.1(a) illustrates the process of image pre-processing used before radiomics and DL modelling. T1c-w MR imaging was bias-corrected using the N4ITK bias correction algorithm (Tustison et al., 2010) after masking the soft tissue region in the image using the Canny Edge detection algorithm (Canny, 1986). After bias correction, intensity values of T1c-w data were z-score normalized. PET imaging was converted SUV. SUV values were truncated to the range [0, 10] to remove potential outlying intensities. Subsequently, the entire volume was normalized to the [0,1] range (see Section 2.3.2 for more details on image processing). Further pre-processing was specific to radiomics or DL analysis. For the DL analysis, we aligned the orientation of all MET-PET and T1c-w MR images and resampled these to isotropic 2.0 × 2.0 × 2.0 mm^3 voxels using trilinear interpolation. A single image volume of size 60 × 60 × 44, centred around the CTV centre of mass, was extracted in the axial plane for both imaging modalities.

For the radiomics analysis, further image pre-processing followed by feature extraction was carried out using the MIRP Python toolkit (version 1.1.3) (Zwanenburg et al., 2019b). MET-PET and T1c-w MR image voxels were resampled to 2.0 × 2.0 × 2.0 mm^3 and 1.0 × 1.0 × 1.0 mm^3, respectively, using trilinear interpolation. LoG filters with kernel widths σ = 2 mm for MET-PET and σ = 1 mm for T1c-w MRI were applied to the base images. The choice of kernel width was based on the original slice thickness of each imaging modality. A total of 270 and 152 intensity-based and texture-based features were extracted from the 3D CTV on the baseline MET-PET and T1c-w MRI, respectively. In addition, 57 first-order intensity-based features were extracted from the CTV on the LoG transformed images for both imaging modalities. This resulted in a total of 327 and 209 features extracted from MET-PET and T1c-w MRI, respectively. Further details on feature classes are summarized in Appendix Table C.2.

Image pre-processing and feature extraction in MIRP were implemented according to the recommendations of the IBSI (Zwanenburg et al., 2020). The definitions used to calculate the features can be found in the IBSI reference manual. Image processing parameters are summarized in Appendix Table C.3.

In order to obtain reproducible results, imaging features have to be stable under small image perturbations, such as those caused by slight variations in acquisition parameters or positioning uncertainties (see Section 2.3.2 for more details). We evaluated feature robustness by applying the following image augmentations based on the training data: adding Gaussian noise (mean 0, standard deviation as present in the image), random volume changes of the CTV (0%, -15%, 15%), and translations (0.0, 0.25, and 0.75 mm) in all three spatial dimensions. All combinations of these perturbations were considered, leading to 81 perturbed images for each original dataset. The intra-class correlation coefficient (ICC) was calculated with a 95% confidence interval. Features with the lower boundary of the 95% confidence interval of the ICC below 0.8 were removed (Zwanenburg et al., 2019a). Feature redundancy was reduced through clustering. The Spearman correlation coefficient (ρ) was used as a similarity metric, with average linkage as a criterion for merging two clusters. The feature with the highest mutual information with the endpoint was selected as the representative for each cluster. The clustering process was done separately for MET-PET and T1c-w MRI-based feature sets.

5.2.5 Conventional radiomics modelling

Figure 5.1(b) illustrates the workflow for the conventional radiomics analysis. We implemented a workflow containing four major processing steps to derive radiomics signatures from the pre-processed feature sets: (i) feature pre-processing, (ii) feature selection, (iii) model building with internal validation, and (iv) testing. This workflow was implemented using the open-source end-to-end statistical learning software package familiar (1.0.0) in R (version 4.0.3). Steps (i)-(iii) were first performed using 5 repetitions of 5-fold stratified cross-validation (CV) nested in the training dataset to identify an optimal signature, i.e. the steps were repeatedly performed on the internal training part and validated on the internal validation part of the CV folds. After identifying the final signature, a final model was developed on the entire training data and validated on the test data. The following procedure was performed for each of the 25 CV runs:

1. Features were transformed using the Yeo-Johnson transformation to align their distribution to a normal distribution. Afterwards, features were z-transformed to mean zero and standard deviation one. Both transformations were performed on the internal training part and applied unchanged to the features of the internal validation part.

2. Four supervised feature-selection algorithms were considered: mRMR, MIM, EN, and univariate regression (UR). To avoid potential overfitting, only the five most relevant features were selected in each CV fold.

3. The selected features were used by three different classifiers: logistic regression (GLM_logistic), Xgboost linear model (XGB_lm) and RF for detection of residual tumour status in PET and MRI. Model hyperparameters were tuned automatically using a variant of the sequential model-based optimisation (SMBO) algorithm based on bootstrap sampling of the train-

ing data (Hothorn & Lausen, 2003). Each classifier was built on the internal training part, which was validated on the internal validation part.

For every feature selection method, average model performance was assessed by the median AUC for the detection of PET and MRI status. After cross-validation, features were ranked according to their occurrence across the 25 CV folds for each of the feature-selection methods. The top 5 most commonly occurring features that appeared in at least 75% (i.e. 3 out of 4) of feature-selection methods were selected. If a subset of these features showed a Spearman correlation $\rho>$ 0.5 with each other on the entire training data, the most relevant feature, i.e. the one showing higher association with endpoint on the training data, was considered.

After feature selection has been performed, the resulting radiomics signature is used to build a prognostic model on the entire training data and (iv) the trained model was applied to the independent validation data.

Feature selection criteria for final signature in conventional radiomics model

Here, we explain an example of feature selection for residual tumour status prediction on MET-PET imaging. The same technique applies to residual tumour status prediction on T1c-w MRI as well. Appendix Table C.4 shows 39 MET-PET features with the highest mutual information (measured by the AUC) with residual tumour status on MET-PET selected after hierarchical clustering. These features were then used to build a diagnostic model. As mentioned above, feature selection and model building with internal validation was first performed within 5 repetitions of 5-fold cross-validation (CV) nested in the training data to identify an optimal signature, with model performance evaluated in terms of median AUC across all CV folds. For each of the above-mentioned feature selection methods, the occurrence of every feature in the 25 modelling steps (5 repetitions of 5-fold CV) was counted, and features were ranked according to their occurrences across the cross-validation folds. Appendix Table C.5 shows features with top 5 ranks across each feature selection method that were further considered. Finally, features that showed repeated occurrences across at least 75% of the feature selection methods were selected. Two features, i.e. log_ih_kurt_fbn_n16 and log_stat_skew occurred in all 4 feature-selection methods, thus meeting the 75% occurrence criteria for candidate features. Both features showed a Spearman correlation (ρ) >0.5 on the entire training data, as shown in Appendix Figure C.1. Finally, log_ih_kurt_fbn_n16 was selected as a one-feature signature due to the stronger association of this feature with the endpoint (p-value=$2.58 \times 10^{(-5)}$) as compared to log_stat_skew (p-value=$8.37 \times 10^{(-5)}$). The finally selected signature and the average AUC (average of AUC across all feature selection methods) in internal training and external test are reported in the results section.

5.2.6 Deep learning radiomics modelling

Three different 3D-CNN architectures, i.e. 3D-VGGNet, 3D-ResNet, and 3D-DenseNet, were trained from scratch (see Section 2.6.3 for details). Architectures were adapted to get the best performance on the internal validation data.

The 3D-VGGNet network consists of 3 convolution blocks with 2 convolution layers in the first two blocks and 3 convolution layers in the third block (filter size = 3 × 3 × 3, activation = ReLU) followed by max-pooling (pool-size = 2 × 2 × 2) and dropout layer (rate = 0.4). The first block comprised 64 filters. The number of filters were doubled in each subsequent block. A batch normalization and flattening operation followed the last convolutional block.

The 3D-ResNet network was based on a vanilla ResNet18 implementation for 3D image data, adapted from (Ju, 2019). The first convolutional layer was modified to use a filter size of 3 × 3 × 3, a stride of 2 and global average pooling after the last residual block followed by the flattening layer.

The 3D-DenseNet121 was adapted from (Dudovitch, 2019). Instead of using 4 dense blocks as in the original DenseNet implementation (Huang et al., 2017), only 3 dense blocks (6, 12, 24 layers per block) were used. Like the 3D-ResNet18 adaptation, we used 3 × 3 × 3 convolutions with a stride of 2 in the first convolution layer and global average pooling after the last residual block followed by a flattening layer.

All the above-mentioned architectures were further adapted for improved performance on 3D data by appending a set of 4 fully connected (FC) layers with 512, 512, 256, and 128 neurons respectively at the end of the network. To reduce potential overfitting, a dropout rate of 0.4 was applied between those FC layers. Lastly, the model output was given by a single dense neuron with tanh activation. We used a batch size of 16 and Adam optimizer to estimate model parameters while training of all three architectures. Training was done for a maximum of 300 epochs, while doing early stopping (patience=100) with an adaptive learning rate using exponential decay (initial learning rate = 1.10-4, decay steps = 1000, decay rate = 0.96) via Keras callbacks. Model losses were optimized using binary cross entropy loss function. Final model output was given by a single dense neuron with sigmoid activation.

For the analysis of each endpoint with the aforementioned 3D-CNN architectures, network training was performed within 5 repetitions of 5-fold cross-validation (CV), stratified by the event status on the training dataset. For each of the CV splits, training volumes were augmented by changing contrast, brightness, Gamma correction, Gaussian noise, and Gaussian blur using the open-source Python package batchgenerators for data augmentation (Isensee et al., 2020). To assess the benefit of data augmentation on model generalization to unseen data, the above pipeline was also implemented without augmenting the training data.

Model training was performed on the training folds of the CV splits, and model losses were evaluated at the end of each epoch on the internal validation split. Since each of the 25 CV runs resulted in a trained model, an ensemble prediction was created by averaging outputs for each patient. Training ensemble prediction was obtained by averaging the predicted output for each patient across the 20 models for which that patient was part of the training fold. Similarly, internal validation ensemble prediction was computed by averaging the predicted output using the remaining five models for which the patient was assigned to the internal test fold. All trained 25 models were then applied to external validation data, and a patient's ensemble prediction was computed by averaging over all 25 model predictions. Below, we describe the details of data augmentation applied to 3D imaging data.

3D data augmentation

In this work, we used random flipping to create mirror reflection of input image volume along only the x and y-axis. Mirroring in the batchgenerators package is evenly distributed, i.e. the probability of mirroring along each axis is 0.5. The rest of augmentations used in this work belong to pixel-level transformations. We used additive Gaussian noise with a variance of the noise uniformly sampled from the range (0, 0.05). We also used Gaussian blur with a standard deviation (σ) selected randomly from the range (1, 1.75). Further, we used a gamma correction to improve luminance of input volumes with gamma values selected randomly from the range (0.5, 2). Finally, we applied the brightness multiplicative transform, where the multiplier is randomly sampled from the range (0.7, 1.5), and the random contrast transform, where contrast values were randomly sampled from the interval (1, 1.75) for augmenting MRI data only. We did not use brightness and contrast transform for PET data, as the effect of these transformations was found to be less effective for improving model performance. The hyperparameters for pixel-level transforms were selected manually by visually inspecting the images so that each transformation creates an image that is representative of real perturbations and by avoiding extreme transformations with very high or low values of transformation parameters. All transformations were applied to image volumes extracted around the CTV. Each augmentation was applied with the probability of 0.15, which limits the number of original images shown to the network. The percentage of original images used during the training was 40% and 10%, combining 4 and 6 different augmentation techniques for MET-PET and MRI respectively. Data augmentation parameters are summarized in Appendix Table C.6.

5.2.7 Statistical analysis

The following baseline clinical parameters were available: gender, age, Eastern Cooperative Oncology Group (ECOG) score, MGMT promoter methylation status, IDH mutation status, and resection type. Categorical clinical features were compared between training and test data by the χ^2 test, whereas continuous features were compared using the Mann-Whitney-U test.

Associations between the final model predictions and the endpoints were evaluated by the area under the curve (AUC). Its estimated value and 95% confidence interval were reported. The importance of individual features in the final signature was assessed by the univariate fitting of a logistic regression and computing Wald-test p-values. For creating a confusion matrix based on the final predictions, an optimal cutoff was selected on the training data using the Youden index and transferred to the test data.

Model calibration was assessed via the HL test (Hosmer & Lemesbow, 1980) for the detection of residual tumour status in PET and MRI, and by creating calibration plots. Correlations between features were assessed by the Spearman correlation coefficient (ρ). All tests were two-sided with a significance level of 0.05.

Radiomics analysis was performed in R version 4.0.3, while DL analysis was performed in Python 3.7.0 and Keras (v2.3.1) with TensorFlow (v2.1.0) on NVIDIA GeForce RTX 2080 Max-Q. Our code is publicly available from https://github.com/oncoray/cnn-petra

5.3 Results

Patient characteristics are summarized in Table 5.1. MGMT status and age was significantly different between training and test data. Patients in the training data had a higher percentage of wildtype and a lower percentage of methylated MGMT status (p-value=0.019), and a slightly lower median age (p-value=0.049) compared to the test data.

For radiomics modelling, 327 and 209 radiomic features were extracted from the CTV MET-PET and T1c-w MRI, respectively. Stability analysis reduced these features to 258 and 134 in MET-PET and T1c-w MRI data. Clustering of correlated features further reduced their number (MET-PET = 39, T1c-w MRI = 36). Based on these reduced feature sets, radiomic signatures were developed and validated for both imaging modalities to detect tumour residual status based on PET and MRI.

Table 5.2 presents the results for the classification of residual tumour status in PET and MRI using conventional radiomics, including the model names and the finally selected features. In internal CV, overall higher performance for detection of residual tumour status in PET was observed for all considered machine-learning approaches (AUC PET=0.93) compared to MRI-status classification (AUC MRI=0.66-0.68). Similarly, on the test data we observed higher performance for detection of residual tumour status in PET with all machine learning models (AUC PET=0.90-0.91), while for detection of residual tumour status in MRI, logistic regression and linear Xgboost model showed a similar performance (AUC=0.78) and random forest showed a relatively low performance (AUC=0.73). Corresponding confusion matrices for a logistic regression model are shown in Appendix Figure C.2(a) with a sensitivity of 0.73 and 0.54 and a specificity of 0.88 and 0.87 on the test data for residual tumour status in PET and MRI, respectively. At a threshold of 0.77, the signature developed for PET status prediction was able to accurately classify 22/30 PET-positive and 15/17 PET-negative patients in the test data. At a threshold of 0.38 the signature developed for residual tumour status in MRI was able to accurately classify 13/24 MRI-positive and 20/23 of MRI-negative patients.

Table 5.2: Median AUC values for CV and for the final signature on the test data for PET-status prediction based on MET-PET and for MRI-status prediction based on T1c-w MRI using conventional radiomics. Values in parentheses represent the 95% confidence interval. Best test performance is marked in bold.

Modality	Model	CV train AUC	CV valid AUC	Features	Final training AUC	Final test AUC
MET-PET	**GLM logistic**	0.95	0.93	log_ih_kurt_fbn_n16	0.92 (0.86-0.97)	**0.91** **(0.81-0.98)**
	RF	0.97	0.93		0.93 (0.87-0.97)	0.90 (0.80-0.97)
	XGB_lm	0.94	0.93		0.92 (0.86-0.97)	0.91 (0.81-0.98)
T1c-w MRI	**GLM logistic**	0.78	0.66	dzm_ldhge_3d_fbn_n32, ih_rmad_fbn_n32	0.76 (0.65-0.87)	**0.78** **(0.64-0.89)**
	RF	0.87	0.68		0.86 (0.78-0.94)	0.73 (0.58-0.87)
	XGB_lm	0.76	0.66		0.77 (0.63-0.87)	0.78 (0.64-0.90)

The selected MET-PET feature was log_ih_kurt_fbn_n16 (IBSI: C317). It represents the kurtosis of the discretized histogram (16 bins) on the LoG transformed images. High values indicate the presence of high intensities within the CTV with pronounced peaks of MET uptake, which was related to the positive PET residual tumour status, in comparison to the PET-negative group with relatively low values of this feature. Box plots of this feature (Yeo-Johnson transformed and z-score normalized) in the two PET residual tumour status groups of the training data are shown in Appendix Figure C.3. The feature showed a significant contribution both in training and test (p<0.01). The definition of the selected features for the PET and MRI signatures is presented in Appendix Table C.7 and the logistic regression model and transformation parameters for the best performing signatures are given in Appendix Table C.8.

For detection of residual tumour status in PET, 3D-CNN architectures trained with data augmentation showed a higher performance in internal CV folds compared to 3D-CNN models trained without data augmentation, as shown in Appendix Table C.9. Therefore, models with data augmentation were evaluated on the test data. Table 5.3 presents the results of the 3D-CNN architectures with data augmentation, including the name of the 3D-CNN architecture and the model performance in internal CV and external test. In internal CV, DenseNet showed a higher AUC for both imaging modalities. As for conventional radiomics, detection of residual tumour status with DL based radiomics in PET was more accurate than MRI (AUC PET=0.96, AUC MRI=0.77). On the test data, the highest performance was achieved by DenseNet for detection of residual tumour status in PET, while VGGNet showed a better performance for MRI (AUC PET=0.95, AUC MRI=0.71). At a threshold of 0.56 the 3D-DenseNet model trained to predict PET residual tumour status was able to accurately classify 29/30 MET-positive and 12/17 MET-negative patients in test data. At a threshold of 0.40 3D-VGGNet trained to predict MRI residual tumour status was able to accurately classify 9/24 MRI-positive and 20/23 of MRI-negative patients. Corresponding con-

Table 5.3: Ensemble AUC values for CV and the final test data for PET-status prediction based on MET-PET and for MRI-status prediction based on T1c-w MRI using deep learning. Values in parentheses represent the 95% confidence interval. Best test performance is marked in bold.

Modality	Model	CV train AUC	CV valid AUC	Final test AUC
MET-PET	**DenseNet**	1.00 (0.99-1.00)	0.96 (0.93-0.99)	**0.95** **(0.89-1.00)**
	ResNet	1.00 (1.00-1.00)	0.92 (0.87-0.98)	0.81 (0.70-0.94)
	VGGNet	1.00 (1.00-1.00)	0.95 (0.90-1.00)	0.93 (0.86-1.00)
T1c-w MRI	DenseNet	1.00 (0.99-1.00)	0.77 (0.68-0.87)	0.63 (0.47-0.80)
	ResNet	1.00 (1.00-1.00)	0.73 (0.63-0.84)	0.61 (0.44-0.78)
	VGGNet	0.99 (0.98-1.00)	0.71 (0.59-0.82)	**0.71** **(0.55-0.86)**

fusion matrices are presented in Appendix Figure C.2(b) showing a sensitivity of 0.97 and 0.38 and a specificity of 0.71 and 0.87 for PET and MRI-based classification, respectively.

Figure 5.2 compares the receiver operating characteristic (ROC) curves of the best performing conventional radiomics and DL-based radiomics model for residual tumour status in (a, b) MRI, and (c, d) PET. The corresponding calibration plots are shown in Appendix Figure C.3.

5.4 Summary and discussion

We investigated radiomics-based machine learning models and 3D-CNNs for detection of residual tumour status based on MET-PET and T1c-w MRI in patients with newly diagnosed GBM. Overall, classification on MET-PET was possible with a higher accuracy than on T1c-w MRI. For PET residual tumour status detection, the best performance was achieved by the 3D DenseNet (AUC=0.95), while logistic regression using radiomics features performed best for MRI residual tumour status detection (AUC=0.78).

For MET-PET, the best performing DenseNet model showed a high sensitivity but lower specificity. We visually assessed false-positive predictions on the test data and observed that MET uptake appeared vague or patchy in falsely classified images, possibly due to infiltrative invasion of tracer in white matter (Figure 5.3(a)). Furthermore, MET-negative volumes used in training have an overall smoother appearance with no physiological uptake or patchy appearance (Figure 5.3), making neural networks blind to such images during training, thus leading to false-positive classifications.

The observed lower performance of MRI-based classification can be attributed to the fact that in the clinical setting, the extent of resection and residual disease was also assessed on early post-operative MRI performed within 24-48 hours after surgery, due to the confounding effects

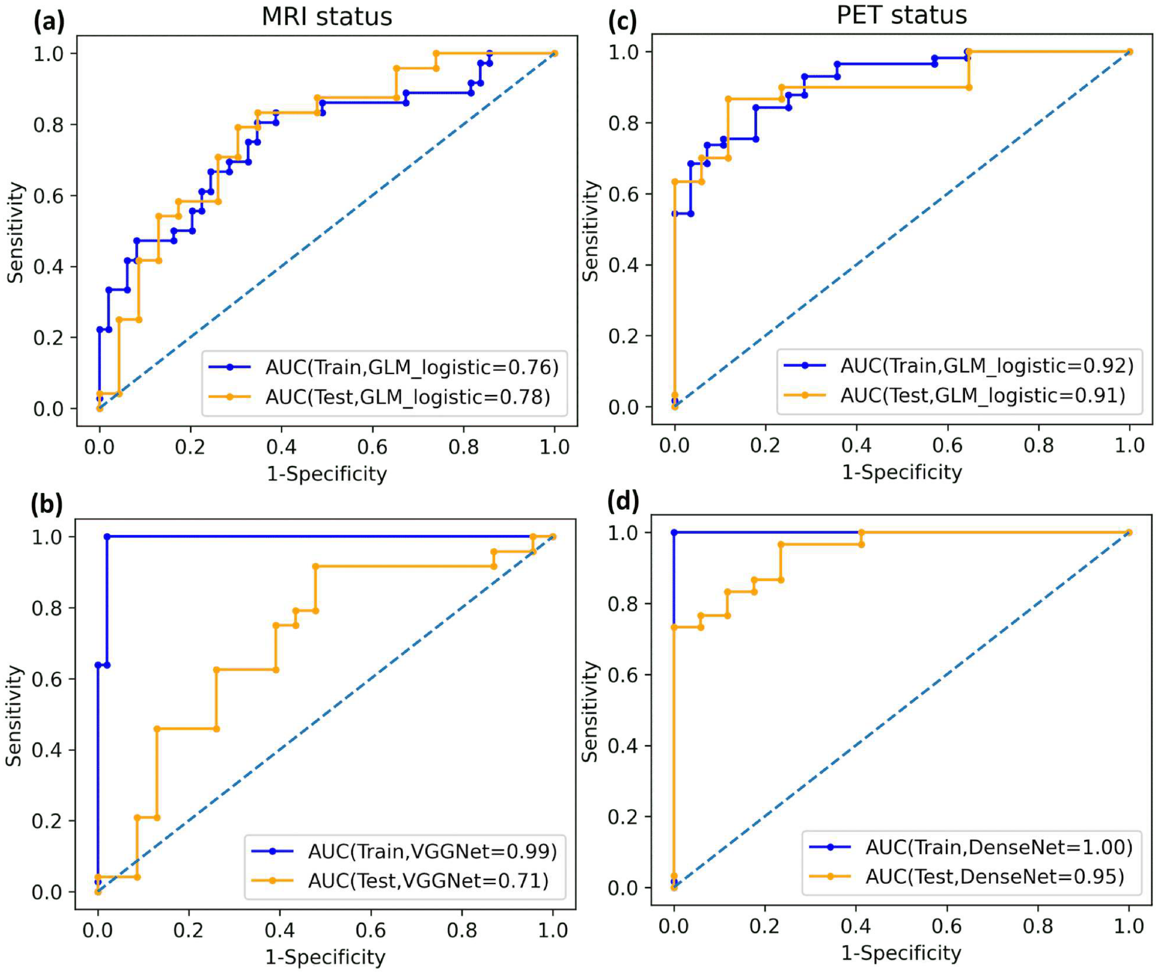


Figure 5.2: Receiver operating characteristics (ROC) curves of the best performing radiomics and deep-learning-based model for (a-b) MRI-status and (c-d) PET-status classification on the training and test data.

of surgically induced contrast enhancement (Wen et al., 2010). Thus, the diagnostic accuracy of using only the second baseline MRI (at least 3 weeks after surgery), which was the only MRI available for the present analysis, is limited as inflammatory/repair-related changes can result in non-tumour-related contrast enhancement that can be misinterpreted as tumour remnants. An example of such case is shown in (Figure 5.3(c-d)). Figure 5.3(c) shows second baseline T1c-w MR image with confounding effects of surgically induced contrast enhancement. This contrast enhancement led to a misclassification of the MR residual tumour status by the 3D-DenseNet model. The true clinical decision of MR-negative status was made on early post-surgical MRI shown in Figure 5.3(d), which does not show contrast enhancement on surgical boundaries. This indicates that the inclusion of early post-operative MRI may help to improve predictions.

Due to the volumetric nature of medical imaging, 3D-CNNs are more promising than 2D alternatives as they incorporate potentially relevant spatial information (Singh et al., 2020). It is, however, generally unclear if 3D-CNN architectures achieve improved results compared to

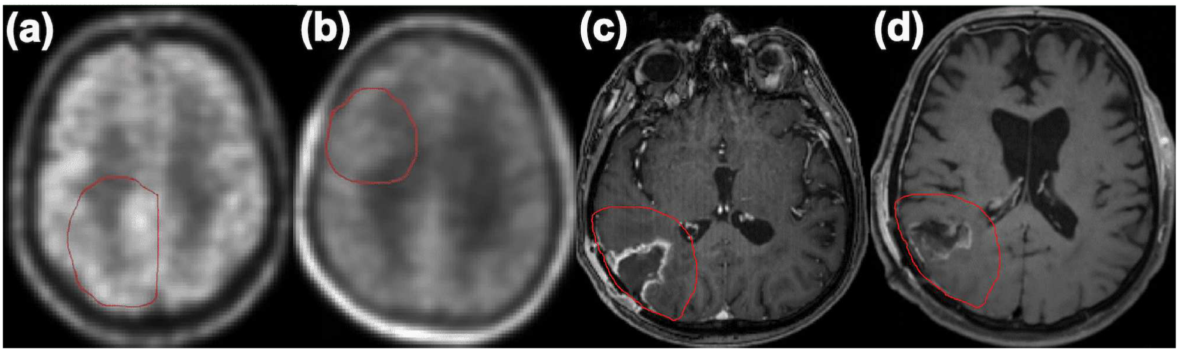


Figure 5.3: (a) Example PET image of a false-positive case. The PET image appears patchy, compared to (b) a true negative PET image with overall homogenous appearance. (c) Example T1c-w MR image of a false-positive case with confounding effects of surgically induced contrast enhancement in the second baseline MR image. The clinical decision of MR-negative status was made on early post-surgical MRI shown in (d), which does not show contrast enhancement on surgical boundaries. Red contours represent the CTV.

radiomics-based machine learning models for limited datasets. In our analysis, the performance benefit observed for PET-status prediction via the 3D-DenseNet may be attributed to the fact that multi-layer feature concatenation improves the representation capacity of CNNs (Zhang et al., 2021). In the radiomics analysis, we observed a limited gain from complex classifiers such as Xgboost_lm and RF compared to simple logistic regression, probably due to elaborate feature selection.

Limitations of this study are the relatively low number of patients in the training and validation data. In addition, there is somewhat a class imbalance due to the smaller number of PET and MRI negative instances. We aimed to mitigate this problem by internal cross-validation (CV) on the training data for both conventional and DL based radiomics analysis. A 5-fold CV approach was used and repeated 5 times, to ensure that each fold contained sufficient number of PET and MRI negative instances for training and validation and that the finally considered ensemble model performance was sufficiently robust. Further, in order to improve model generalization on unseen data and compensate for low number of training instances, we used data augmentation.

In conclusion, we developed and independently tested radiomics approaches and 3D-CNNs for predicting the residual tumour status of patients with GBM after surgery based on MET-PET and T1c-w MRI. Overall, classification on MET-PET was possible with a higher accuracy than on MRI. While the 3D-DenseNet showed the best performance on PET data, feature-based logistic regression performed best on MRI. In the future, the presented models should be prospectively validated to eventually support clinicians in the diagnosis of the residual tumour status and potentially reduce the required resources and inter-rater variability.

6 Radiomics for patient outcome prediction in glioblastoma using [^{11}C] methionine PET and T1c-w MRI

6.1 Motivation

Treatment personalization is a major objective for radiation oncology research, particularly for diseases with poor prognosis. Patients with GBM show poor prognosis despite the use of a standard aggressive treatment regime that includes tumour resection, followed by concurrent chemoradiation with temozolomide and radiation therapy (CRT), and subsequent adjuvant temozolomide (Stupp et al., 2005; Mangla et al., 2010). Radiation therapy (RT) plays an important role in GBM treatment, as it may help to delay further progression of disease and improve patient survival (Nieder et al., 2008). However, a general increase of radiation doses is not possible due to associated risks. Therefore, biomarkers are needed for the selection of patients that may benefit from targeted RT.

Survival statistics for GBM are well-described at the population level, and many factors that impact survival have been identified including age, Karnofsky performance status (KPS), MGMT promotor methylation status, isocitrate dehydrogenase (IDH), neurological deficit, extent of resection, and tumour multifocality and tumour location among others (Lutterbach et al., 2003; Sizoo et al., 2010; Zhou et al., 2018). However, none of these markers has been used in stratification of GBM patients for treatment personalization in clinical settings (Zhou et al., 2018).

Conventionally, the diagnosis and chemoradiotherapy (RCT) treatment planning in GBM includes MRI comprising T1c-w, T2w and fluid-attenuated inversion recovery (FLAIR) images. Concurrent PET/MRI is not a widely available method for diagnosis in GBM. However, studies have shown that post-surgical PET such as [^{11}C] methionine (MET) and [18F]-fluorodeoxyglucose (FDG) PET can have a superior prognostic value compared to MRI as it can predict tumour progression with a higher accuracy. Therefore, PET offers potential for treatment personalization and guided therapy intensification (Wang et al., 2018b; Seidlitz et al., 2021).

For GBM, various studies have evaluated radiomics features extracted from multiparametric MRI including T1-w, T2-w, T1c-w, FLAIR and DWI to evaluate OS and progression free survival (PFS) (Yang et al., 2015; Kickingereder et al., 2016a; Chaddad et al., 2018; Lee et al., 2019). Integrating MRI radiomics features with patient clinical and molecular profils was shown to further improve the prognostic performance (Osman, 2018). Fewer studies have also evaluated CNN based methods for OS prediction using MRI data (Lao et al., 2017; Tang et al., 2020). Only few recent studies have involved radiomics analysis for evaluating the prognostic role of post-radiotherapy f-fluoro-ethyl-l-tyrosine (FET) PET imaging (Lohmann et al., 2020; Carles et al., 2021). However, to the best of our knowledge, a comparative analysis of conventional radiomics and deep learning to evaluate the prognostic role of pre-RCT MET-PET/MRI in patients with GBM has not yet been evaluated.

Therefore, in this study, we developed and independently validated conventional radiomics and 3D-CNN models for prognosis and risk stratification of TTR and OS in pre-RCT MET-PET and T1c-w MRI in patients with newly diagnosed GBM. A part of the work presented within this chapter has been presented at an international conference (Shahzadi et al., 2022a).

6.2 Materials and methods

6.2.1 Patient cohort

In this work, imaging and clinical data of 132 adult patients were collected from the same cohort as outlined in Chapter 5, Section 5.2.1. Patient characteristics are given in Table 5.1.

6.2.2 Experimental design

We developed and independently validated radiomics signatures and deep learning models for the prognosis of TTR and OS in patients with GBM based on MET-PET and T1c-w MRI data acquired before RCT. Figure 6.1 summarizes the design of the radiomics analysis. Radiomics features were extracted from the CTV separately for MET-PET and T1c-w MRI. Details concerning feature classes, stability analysis, and clustering are reported in Chapter 5, Section 5.2.2. Three different survival-based machine learning algorithms of varying complexity were assessed, including Cox regression (Cox), Xgboost linear model (XGB_lm), and random survival forest (RSF) for the prognosis of TTR and OS. The finally selected MET-PET and T1c-w MRI features were applied to the test data for prognosis and risk stratification, and the performance was compared. In our deep learning analysis, end-to-end feature extraction and modeling were performed using three different 3D-CNN architectures, i.e. 3D-VGG, 3D-Resnet, 3D-DenseNet. Model losses were optimized using the CPHM, which is a survival-specific regression model for assessing time-dependent endpoints (see Section 2.6.4 for more details). For each imaging modality, models were individually trained on the training folds of each CV split for prognosis of TTR and OS, both with and without data augmentation. The finally developed models were then applied to the independent test cohort and model performances were compared. Predictions from best performing 3D-CNN models on the training data were integrated with important clinical/molecular features in a multivariable Cox model, which was then validated on the test dataset. The performance of developed models was assessed using the C-index.

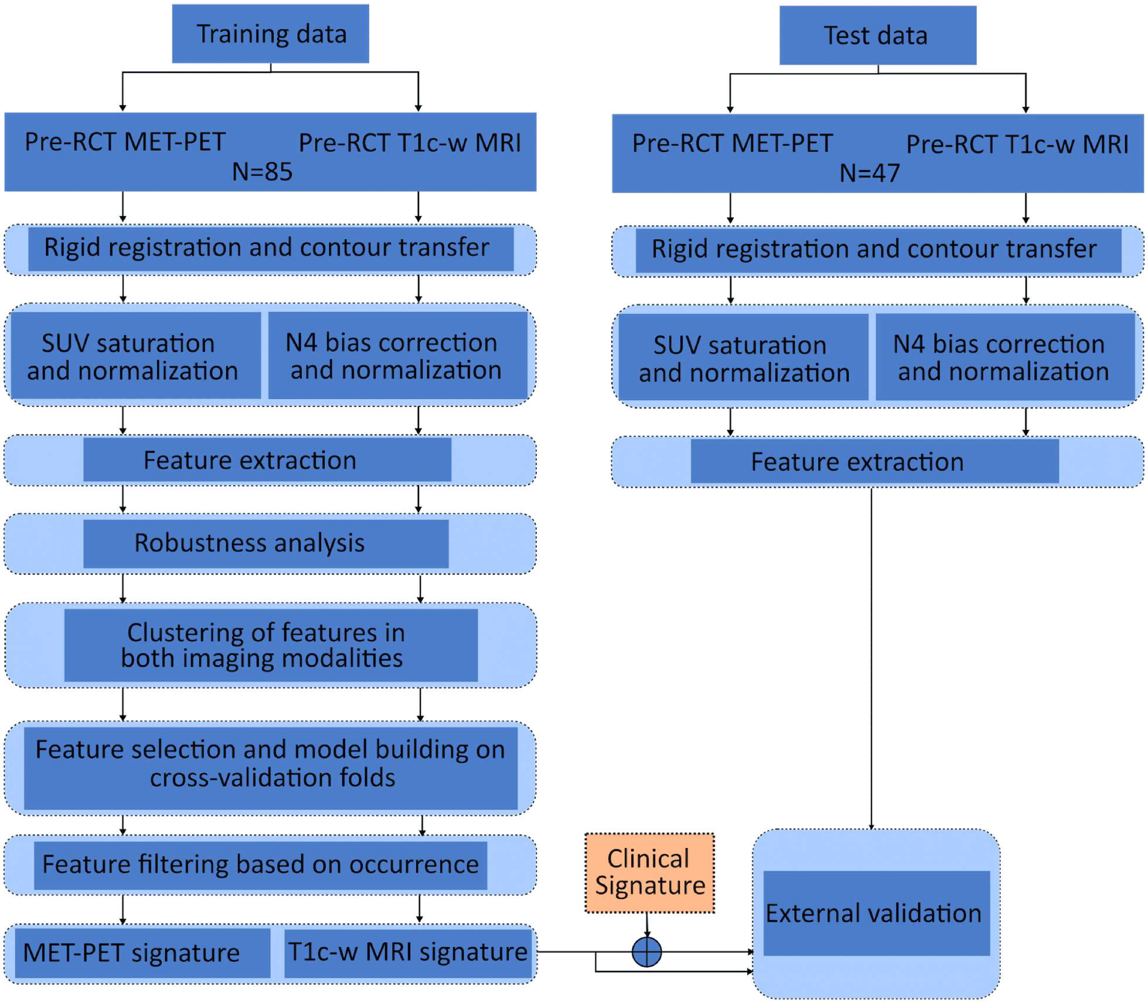

Training data
Pre-RCT MET-PET
Pre-RCT T1c-w MRI
N=85
Rigid registration and contour transfer
SUV saturation and normalization
N4 bias correction and normalization
Feature extraction
Robustness analysis
Clustering of features in both imaging modalities
Feature selection and model building on cross-validation folds
Feature filtering based on occurrence
MET-PET signature
T1c-w MRI signature
Test data
Pre-RCT MET-PET
Pre-RCT T1c-w MRI
N=47
Rigid registration and contour transfer
SUV saturation and normalization
N4 bias correction and normalization
Feature extraction
Clinical Signature
External validation

Figure 6.1: Conventional radiomics workflow for outcome prediction in GBM. Images in both MET-PET and T1cw-MRI were co-registered with the treatment planning CT, and CTV contours were transferred. MET-PET images were saturated and normalized and whereas T1c-w were N4 bias corrected and normalized. Radiomics features were extracted from each imaging modality, analysed for robustness, and clustered. Radiomics signatures for MET-PET and T1c-w MRI were developed in a CV approach and applied to the test data.

6.2.3 Image acquisition, endpoints, and contouring

The details of PET/MRI acquisition and contouring of imaging data for this study are outlined in Chapter 5, Section 5.2.4 and image acquisition parameters are summarized in Appendix Table C.1 for both training and test data. In this study, the considered endpoints were the prognosis of TTR and OS in patients with GBM. The survival endpoints TTR and OS were calculated from the first day of RCT to the day of the event (local recurrence for TTR and death for OS) or censoring. For the patients with the observed event, the event time was accompanied by an event indicator variable of 1, whereas for patients without an event, the last follow-up time was used together with an event indicator variable of 0.

6.2.4 Image pre-processing, and feature extraction

Image pre-processing and feature extraction details are explained previously in Chapter 5, Section 5.2.4. T1c-w MR imaging was bias-corrected and normalized, while PET imaging was converted to SUV values followed by SUV truncation. Further, for DL analysis, image orientation for both modalities was aligned and resampled to isotropic voxel resolution. Images patches centred around CTV were then extracted for training DL models. For conventional radiomics analysis, both imaging modalities were resampled to isotropic voxel resolution and feature extraction from 3D CTV MIRP Python toolkit (version 1.1.3) (Zwanenburg et al., 2019b).

6.2.5 Conventional radiomics modelling

For the prognosis of TTR and OS, we implemented a radiomics modelling workflow that is similar to the one previously described in Chapter 5, Section 5.2.5. As mentioned in this section, radiomics modelling consisted of four major steps, i.e. (i) feature pre-processing, (ii) feature selection, (iii) model building with internal validation, and (iv) testing. However, the following changes were made to adapt the workflow:

1. In model building with internal validation, we used three different survival specific regression models: Cox regression (Cox), Xgboost linear model (XGB_lm) and random survival forest (RSF) for the prognosis of TTR and OS.

2. Model performances were assessed by the median C-index.

After step (i)-(iv) of radiomics modelling were executed on cross validation (5 times 5 fold CV) splits, a final signature was defined based on feature ranking. Features were ranked according to their occurrence across the 25 CV folds for each of the feature-selection methods, as explained previously in Chapter 5. After feature selection is complete, the resulting radiomic signature was then used to build prognostic models on the entire training data and (iv) the trained model was applied to the independent test data.

Feature selection criteria for the final signature in conventional radiomics model

Here we explain an example of feature selection for prognosis of TTR on MET-PET imaging. The same technique applies to prognosis of OS on MET-PET and prognosis of TTR and OS on T1c-w-MRI as well. Appendix Table D.1 shows 39 MET-PET features with the highest mutual information with TTR on MET-PET selected after hierarchical clustering. These features were then used to build a prognostic model. Feature selection and model building with internal validation was first performed within 5 repetitions of 5-fold CV nested in the training dataset to identify an optimal signature. Four supervised feature-selection algorithms were considered: mRMR (Peng et al., 2005), MIM (Gelfand & I A glom, 1959), EN (Zou & Hastie, 2005), and UR (Cox & Oakes, 1984). To avoid potential overfitting, only the five most relevant features were selected in each CV fold. These features were then used to build a prognostic model on the internal training part, and validated on the internal validation part. For each of the above-mentioned feature selection methods, the occurrence of every feature in the 25 modelling steps was counted, and features were ranked according to their occurrences across the CV folds. Appendix Table D.2 shows features with top 5 ranks across each feature selection method that were further considered. Finally, features that showed repeated occurrences across at least 75% of the feature selection methods were selected. Four features, i.e. dzm_sdhge_3d_fbn_n16, log_stat_min, log_ivh_i90, and log_ih_skew_fbn_n16 occurred in all 4 feature selection methods thus meeting the 75% occurrence criteria for candidate features. All 4 features showed a Spearman correlation of >0.5 on the entire training data, as presented in Appendix Figure D.1. Finally, log_stat_min was selected as a signature due to the stronger association with the endpoint (p-value<0.001) as compared to other features, thus forming the MET-PET based one-feature radiomic signature as shown in Appendix Table D.3. The finally selected signature and the average C-index in internal training and external test are reported in the results section.

6.2.6 Deep learning radiomics modelling

Three different 3D-CNN architectures, i.e. 3D-VGGNet, 3D-ResNet, and 3D-DenseNet, were trained from scratch for the prognosis of OS and TTR using MET-PET and T1c-w MRI. The details of model architecture and model training are explained in Chapter 5, Section 5.2.6. The following adaptations were implemented:

1. The Cox proportional hazard loss was applied instead of the binary cross-entropy loss for optimizing the model.
2. The tanh activation function was used instead of the sigmoid for the prediction of the final output layer.
3. Model performance was evaluated using the C-index instead of the AUC.

6.2.7 Statistical analysis

The following baseline clinical parameters were available: gender, age, ECOG, MGMT promoter methylation status, IDH mutation status, and resection type. Categorical clinical features were compared between training and test cohorts by the χ^2 test whereas continuous features were compared using the Mann-Whitney-U test. Associations between the final model predictions and the endpoints were evaluated by the C-index. Its estimated value and 95% confidence interval were reported. The importance of individual features in the final signature was assessed by the univariate fitting of a Cox regression and computing Wald-test p-values. For association with TTR and OS, patients were stratified into an optimally separated low and a high-risk group using an optimal cut-off on the training data that was based on maximally selected rank statistics (Hothorn & Lausen, 2003). The cut-off was transferred to the validation data. TTR and OS of stratified groups were assessed with Kaplan Meier curves, which were compared with the log-rank test. Model calibration was assessed via the GND test (Demler et al., 2015) and by creating calibration plots. Correlations between features were assessed by the Spearman correlation coefficient (ρ). All tests were two-sided with a significance level of 0.05.

The radiomics analysis was performed in R version 4.0.3, while the deep learning analysis was performed in Python 3.7.0 and Keras (v2.3.1) with TensorFlow (v2.1.0) on NVIDIA GeForce RTX 2080 Max-Q. Our code for prognostic modelling of TTR and OS using 3D-CNNs is publicly available from https://github.com/oncoray/cnn-petra

6.3 Results

Patients in the training data had a higher percentage of wildtype and a lower percentage of methylated MGMT status (p-value=0.019), and a slightly lower median age (p-value=0.049) compared to the test data. In univariate Cox analysis, a significant association of TTR and OS was observed for MGMT status (TTR, OS p-value<0.001), age (TTR: p-value=0.034, OS p-value=0.001) and IDH status (p-value=0.018) in the training cohort (Table 6.1). However, due to the large number of missing values, IDH status was not considered in signature development.

For radiomics modelling, 327 and 209 radiomic features were extracted from the CTV MET-PET and T1c-w MRI, respectively. Stability analysis reduced these features to 258 and 134 in MET-PET and T1c-w MR data. Clustering of correlated features further reduced their number (in MET-PET = 39, T1c-w MR = 36), as reported in Section 5.3. Based on these reduced feature sets, radiomic signatures were developed and validated for both imaging modalities to prognosticate TTR and OS.

Table 6.2 presents the results for the prognosis of TTR and OS using radiomics, including the model names and the finally selected features. For the prognosis of TTR, XGB_lm and the Cox model showed slightly better performance than the RSF on MET-PET data in internal CV (C-index: XGB_lm=0.61, Cox=0.60, RSF=0.58). Furthermore, T1c-w MRI showed a lower performance than MET-PET, with comparable results obtained for all considered machine learning models (C-index: RSF=0.53, Cox=0.51, XGB_lm=0.51). This also translated to the test cohort, where the signatures developed on MET-PET showed a better performance than the signatures developed on T1c-w MRI. No significant difference was observed between different machine learning models. For the prognosis of OS, all considered machine learning models showed overall lower performance in internal CV for both MET-PET and T1c-w MRI data (C-index MET-PET: XGB_lm=0.54, Cox=0.52, RSF=0.51, C-index T1c-w MRI: RSF=0.51, Cox=0.49, XGB_lm=0.49).

Table 6.1: Univariable analysis of TTR, and OS using Cox regression, in the training data. ci: confidence interval. Significant p-values of patient's clinical characteristics are marked in bold.

Clinical feature		**TTR**		**OS**	
		Hazard ratio (95% CI)	**p-value**	**Hazard ratio (95% CI)**	**p-value**
Age / years		1.017 (1.001-1.033)	**0.034**	1.028 (1.011-1.045)	**0.001**
Gender (female vs male)		1.085 (0.679-1.733)	0.733	0.919 (0.57-1.481)	0.728
MGMT (Methylated vs Wildtype)		0.251 (0.147-0.429)	**<0.001**	0.251 (0.144-0.437)	**<0.001**
ECOG	(1 vs 0)	1.135 (0.705-1.826)	0.603	1.147 (0.702-1.873)	0.584
	(2 vs 0)	1.624 (0.635-4.149)	0.311	2.243 (0.868-5.796)	0.095
IDH (Mutated vs Wildtype)		0.243 (0.076-0.78)	**0.018**	0.319 (0.099-1.022)	**0.054**
PET status (0 vs 1)		2.578 (1.518-4.379)	<0.001	2.119 (0.472-1.251)	0.005
MRI status (0 vs 1)		2.755 (1.717-4.421)	<0.001	2.129 (1.331-3.406)	0.002
Abbreviations: ECOG, Eastern Co-operative Oncology Group; IDH, isocitrate dehydrogenase; MGMT, O6-methylguanine DNA methyltransferase; MRI, magnetic resonance imaging; OS, overall survival; PET, positron emission tomography; TTR, Time-to-recurrence.					

On the test cohort, the selected signature developed on MET-PET data achieved acceptable performance in terms of C-index with all considered machine learning models (C-index: Cox=0.60; RSF=0.60; XGB_lm=0.60) while the signature developed on T1c-w MRI showed slightly better performance (C-index: Cox=0.63, RSF=0.62, XGB_lm=0.62). However, none of the above-mentioned models achieved significant stratification of patients in low and high-risk groups of TTR and OS on the test cohort (p-value>0.05).

The clinical model containing age and MGMT status showed a decent performance for prognosis of TTR on the test cohort with significant risk group stratification, while the performance for prognosis of OS was relatively low (TTR: C-index=0.59, p-value = 0.004; OS: C-index=0.55, p-value=0.32). Combining this clinical signature with MET-PET and T1c-w MRI-based imaging signatures showed improved performance, with significant stratification of the patients into low and high-risk groups of TTR (Clinical+MET-PET: C-index=0.66, p-value<0.001; Clinical+T1cw-MRI: C-index=0.62, p-value=0.008). Figure 6.2 shows the Kaplan-Meier curves for the clinical model (Figure 6.2(a)), the clinical+MET-PET model (Figure 6.2(b)), and the clinical+T1cw-MRI model (Figure 6.2(c)) for prognosis of TTR. The corresponding calibration plots are shown in Appendix Figure D.2(a-c).

For the prognosis of OS, the joint clinical and imaging signature (both MET-PET and T1cw-MRI) showed improved performance on the test cohort compared to the clinical signature alone but reduced performance compared to the imaging signature alone in terms of C-index (Clinical+MET-PET: C-index=0.59, Clinical+T1cw-MRI: C-index=0.57), however, patient stratification into low and high risk groups of OS still remained insignificant. As no significant benefit of using complex machine learning models was observed, we used the simple Cox regression model for building the signatures. Appendix Table D.4. contains model and transformation parameters for the best performing signatures (highlighted in bold in Table 6.2) developed for the prognosis of TTR and OS.

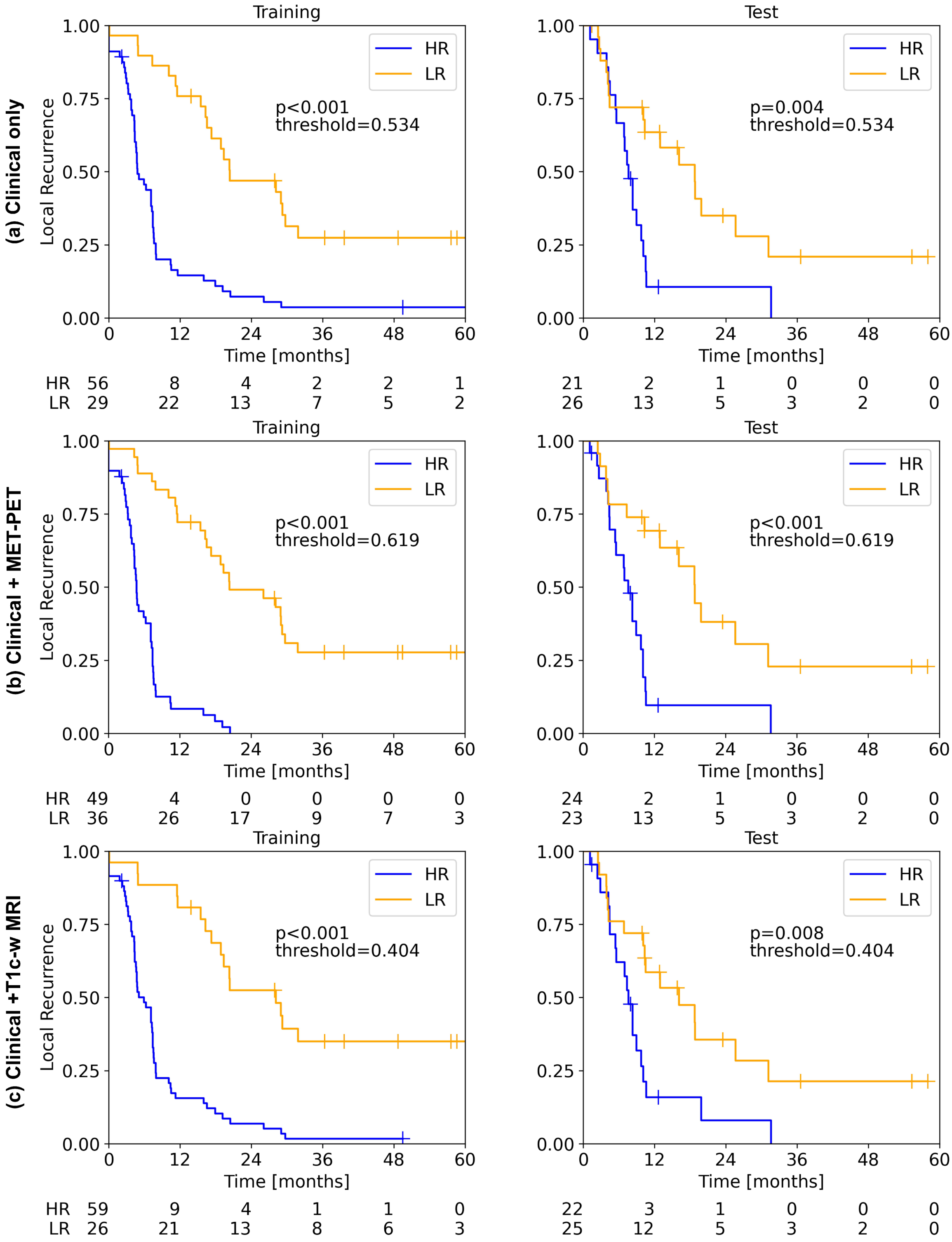


Figure 6.2: Kaplan-Meier plots for the prognosis of TTR on the training and test cohort using the Cox regression model based on (a) the clinical signature, (b) the clinical + MET-PET signature and (c) the clinical + T1c-w MRI signature. Imaging-based signatures were developed using conventional radiomics. All models resulted in significant patient stratification into low and high-risk groups (p<0.01) on the test set.

Table 6.2: C-index for the endpoints TTR and OS based on MET-PET imaging and T1c-w MRI data using radiomics. Values in parenthesis represent the 95% confidence interval. Best performance is marked in bold.

Endpoint	Modality	Model	CV train C-index	CV valid C-Indedx	Features	Final training C-index	Final test C-index	p-value test
TTR	**MET-PET**	Cox	0.66	0.60	log_stat_min	0.64 (0.57-0.71)	0.59 (0.48-0.70)	0.25
		RSF	0.64	0.58		0.64 (0.57-0.51)	0.58 (0.47-0.69)	0.23
		XGB_lm	0.66	0.61		0.64 (0.56-0.71)	0.59 (0.48-0.70)	0.25
	T1c-w MRI	Cox	0.62	0.51	ivh_diff_i25_i75 dzm_zd_var_3d_fbn_n32 loc_peak_glob	0.60 (0.52-0.67)	0.54 (0.42-0.64)	0.58
		RSF	0.64	0.53		0.64 (0.58-0.72)	0.53 (0.42-0.65)	0.89
		XGB_lm	0.62	0.51		0.59 (0.52-0.66)	0.54 (0.44-0.65)	0.23
	Clinical	Cox	-	-	Age MGMT	**0.72** **(0.65-0.79)**	**0.59** **(0.49-0.71)**	**0.004**
	Clinical + MET-PET	Cox	-	-	Age MGMT log_stat_min	**0.74** **(0.68-0.79)**	**0.66** **(0.56-0.76)**	**<0.001**
	Clinical + T1c-w MRI	Cox	-	-	Age MGMT ivh_diff_i25_i75 dzm_zd_var_3d_fbn_n32 loc_peak_glob	**0.74** **(0.67-0.79)**	**0.62** **(0.51-0.73)**	**0.008**
OS	**MET-PET**	Cox	0.63	0.52	stat_max	0.60 (0.53-0.68)	0.60 (0.46-0.74)	0.84
		RSF	0.65	0.51		0.60 (0.52-0.68)	0.60 (0.48-0.70)	0.85
		XGB_lm	0.63	0.54		0.60 (0.53-0.68)	0.60 (0.47-0.73)	0.84
	T1c-w MRI	Cox	0.62	0.49	ivh_diff_i25_i75 dzm_zd_var_3d_fbn_n32	0.60 (0.53-0.67)	0.63 (0.49-0.73)	0.86
		RSF	0.65	0.51		0.61 (0.53-0.70)	0.62 (0.48-0.74)	0.63
		XGB_lm	0.62	0.49		0.59 (0.52-0.66)	0.62 (0.50-0.73)	0.3
	Clinical	Cox	-	-	Age MGMT	0.74 (0.69-0.81)	0.55 (0.45-0.66)	0.32
	Clinical + MET-PET	Cox	-	-	Age MGMT stat_max	0.75 (0.60-0.80)	0.59 (0.48-0.69)	0.21
	Clinical + T1c-w MRI	Cox	-	-	Age MGMT ivh_diff_i25_i75 dzm_zd_var_3d_fbn_n32	0.76 (0.70-0.81)	0.57 (0.46-0.67)	0.25

The selected MET-PET feature for the prognosis of TTR and OS was log_stat_min and stat_max (IBSI:1GSF), respectively. Both these features are intensity-based statistical features that describe how intensities (or SUV values in case of MET-PET imaging) within the ROI are distributed. The highest SUV present within the CTV on baseline MET-PET is captured by the stat_max feature, which is closely related to the minimum SUV on LoG transformed MET-PET images. High values of stat_max and consequently low values of log_stat_min indicate MET uptake in the residual tumour. Image-based interpretation of these features is presented in Figure 6.3. Patients in the high-risk group of TTR showed relatively low values of log_stat_min. which translates to the existence of bright voxels or alternatively high values of stat_max in the CTV (Figure 6.3(a)). In

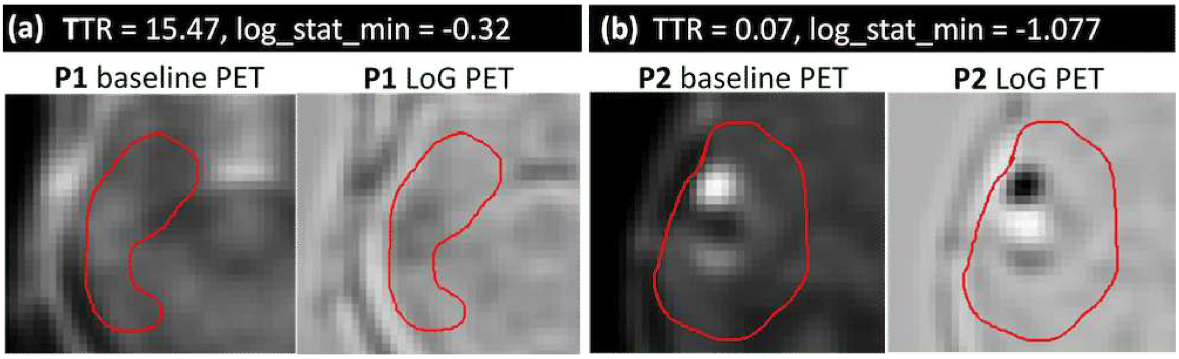

(a) TTR = 15.47, log_stat_min = -0.32
P1 baseline PET
P1 LoG PET
(b) TTR = 0.07, log_stat_min = -1.077
P2 baseline PET
P2 LoG PET

Figure 6.3: Representative images from MET-PET imaging with corresponding Laplacian of Gaussian (LoG) transformed images and the selected signature, i.e. log_stat_min value from two patients (P) in the two risk groups (P1 from low risk group and P2 from high risk group) of training data. The red contours mark the clinical target volume (CTV). P1 (TTR status = 0, TTR = 15.47 months) showed an overall homogenous appearance on the baseline MET-PET with higher log_stat_min value (a). On the contrary, P2 (TTR status = 1, TTR = 0.07 (months)) showed a more heterogeneous CTV with a low log_stat_min , which corresponds to high pixel intensities (stat_max) on the baseline PET (b).

comparison, patients with a longer time to recurrence on average showed lower maximum SUV values (6.3(b)).

Table 6.3 presents the results for the prognosis of TTR and OS using 3D-CNN architectures trained with data augmentation, including the name of the architectures and their performance in internal CV and external test. In internal CV, the highest performance was achieved by the 3D-DenseNet and the 3D-VGG model for prognosis of TTR using MET-PET imaging (C-index: 3D-DesneNet=0.68, 3D-ResNet=0.63, 3D-VGG=0.69). Like in our radiomics analysis, T1c-w MRI showed lower performance than MET-PET for the prognosis of TTR in internal CV, with 3D-DenseNet and 3D-ResNet models showing better performance than 3D-VGG (C-index: 3D-DenseNet=0.63, 3D-ResNet=0.60, 3D-VGG=0.53). This also translated to the test cohort, where the 3D-DenseNet model developed on MET-PET imaging showed a higher performance than other 3D-CNN models developed on either MET-PET or T1c-w MRI (C-index MET-PET: 3D-DenseNet=0.66, 3D-ResNet=0.61, 3D-VGG=0.55; C-index T1c-w MRI: 3D-VGG=0.56, 3D-ResNet=0.55, 3D-DenseNet=0.50) with significant stratification for TTR (p-value 0.027). For the prognosis of OS, in internal CV the 3D-VGG model built on MET-PET showed the best performance (C-index: 3D-VGG=0.70, 3D-DenseNet=0.61, 3D-ResNet=0.51), while for T1c-w MRI, the 3D-DenseNet model performed best (C-index: 3D-DenseNet=0.62, 3D-ResNet=0.58, 3D-VGG=0.49). However, in the external test cohort, only the 3D-DenseNet model built on MET-PET showed decent performance with significant patient stratification (3D-DenseNet: C-index=0.64, p-value=0.033). Overall, for the prognosis of TTR and OS, 3D-CNN models trained with data augmentation showed better performance in internal validation compared to 3D-CNN models trained without data augmentation as shown in Appendix Table D.5.

Finally, we integrated the clinical signature (age and MGMT status) with the best performing 3D-DenseNet ensemble prediction based on MET-PET imaging to build a Cox regression model.

Table 6.3: Ensemble C-index values for CV on the training and test data for TTR and OS prediction based on MET-PET and T1c-w MRI data using DL.

Endpoint	Modality	Model	C-index train	C-index valid	C-index test	p-value test
TTR	**MET-PET**	DenseNet	**0.84 (0.79-0.88)**	**0.68 (0.60-0.75)**	**0.66 (0.51-0.81)**	**0.027**
		ResNet	0.90 (0.85-0.93)	0.63 (0.56-0.71)	0.61 (0.43-0.79)	0.168
		VGGNet	0.84 (0.79-0.89)	0.69 (0.62-0.76)	0.55 (0.44-0.67)	0.763
	T1cw-MRI	DenseNet	0.86 (0.82-0.90)	0.63 (0.56-0.71)	0.50 (0.43-0.58)	0.406
		ResNet	0.82 (0.78-0.85)	0.60 (0.51-0.70)	0.55 (0.46-0.64)	0.096
		VGGNet	0.66 (0.60-0.73)	0.53 (0.46-0.60)	0.56 (0.45-0.68)	0.857
	Clinical + DenseNet MET-PET	**Cox**	**0.85 (0.81-0.88)**	**0.74 (0.67-0.79)**	**0.68 (0.53-0.83)**	**0.017**
OS	**MET-PET**	DenseNet	**0.82 (0.77-0.87)**	**0.61 (0.53-0.69)**	**0.64 (0.43-0.86)**	**0.033**
		ResNet	0.87 (0.84-0.91)	0.55 (0.47-0.62)	0.61 (0.44-0.77)	0.227
		VGGNet	0.88 (0.82-0.93)	0.70 (0.64-0.76)	0.53 (0.42-0.65)	0.426
	T1cw-MRI	DenseNet	0.84 (0.80-0.89)	0.62 (0.55-0.69)	0.60 (0.43-0.77)	0.067
		ResNet	0.87 (0.82-0.92)	0.58 (0.50-0.65)	0.59 (0.49-0.70)	0.191
		VGGNet	0.59 (0.51-0.66)	0.49 (0.42-0.57)	0.65 (0.55-0.76)	-
	Clinical + DenseNet MET-PET	**Cox**	**0.82 (0.77-0.87)**	**0.69 (0.63-0.75)**	**0.65 (0.51-0.78)**	**0.039**

This joint clinical and imaging model further improved the prognosis of TTR and OS in terms of C-index in the external test cohort, with significant stratification of the patients into low and high-risk groups (TTR: Clinical + 3D DenseNet: C-index=0.68, p-value=0.017, OS: Clinical + 3D DenseNet: C-index=0.65, p-value=0.039). Figure 6.4 shows the Kaplan-Meier curves for the best performing Clinical + 3D DenseNet model for prognosis of TTR (Figure 6.4(a)) and Clinical+3D DenseNet model for prognosis of OS (Figure 6.4(b)) using MET-PET imaging. The corresponding calibration plots are shown in Appendix Figure D.3(a, b).

6.4 Summary and discussion

We investigated radiomics-based machine learning models and 3D-CNNs for the prognosis of TTR and OS based on MET-PET and T1c-w MRI for both endpoints in patients with newly diagnosed GBM. Overall, MET-PET allowed for better prognosis than T1c-w MRI. The best per-

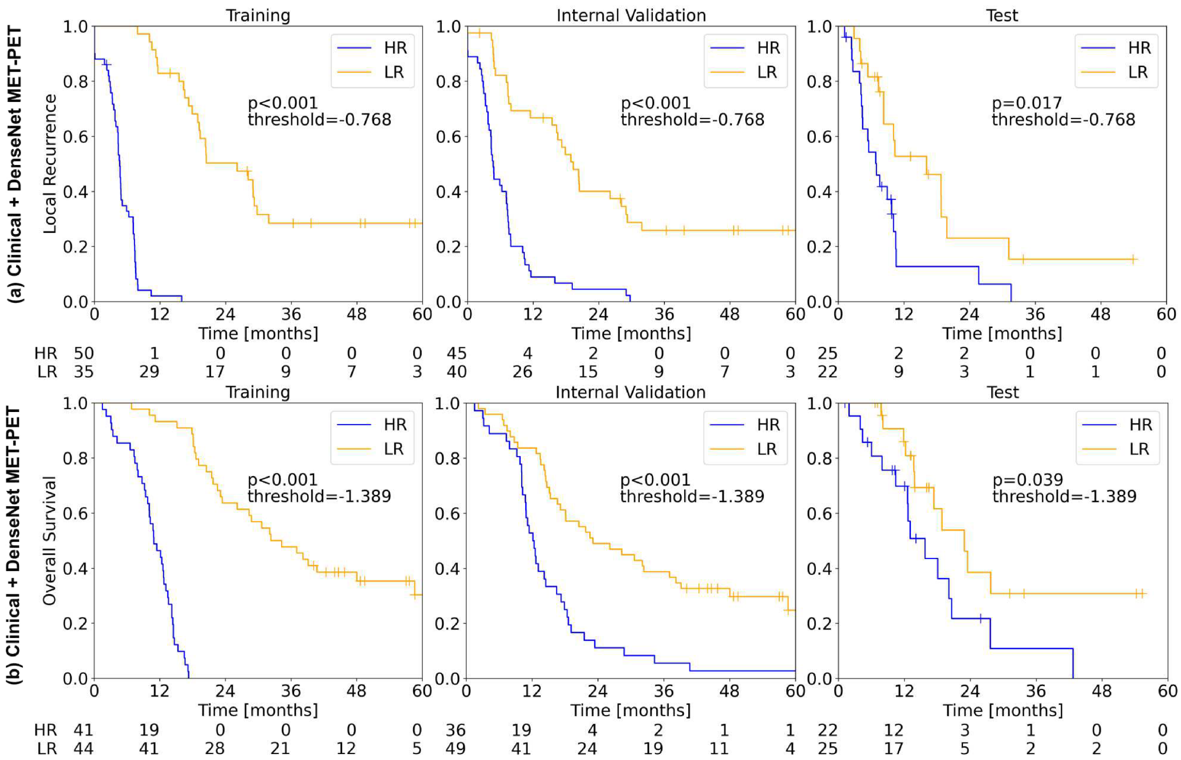


Figure 6.4: Kaplan-Meier estimates for risk-group stratification for (a) TTR and, (b) OS in training, internal validation and external test data based on the respective joint clinical + ensemble predictions (3D-DenseNet model) on MET-PET data.

formance, for both endpoints, in the external test data was achieved by combining a clinical signature (age and MGMT) with a 3D-DenseNet ensemble model based on MET-PET imaging, with significant stratification of the patient in a low and high-risk group (TTR C-index: Clinical + 3D DenseNet: C-index=0.68, p-value=0.017, OS C-index: Clinical + 3D DenseNet=0.65, p-value=0.039).

To the best of our knowledge, this is the first radiomics and deep learning-based evaluation of the prognostic role of pre-RCT [^{11}C] MET-PET and T1c-w MRI in adult patients with newly diagnosed GBM. Similar studies focused on prognosis of patient outcome using radiomics based on MRI and [18F]-FDG PET. Li et al. (Li et al., 2017) showed that a prognostic model built on multiparametric MRI radiomics features achieved a C-index of 0.70 for OS prediction on an independent validation cohort (32 patients). Lao et al. (Lao et al., 2017) showed that a model combining clinical risk factors (age, Karnofsky Performance Score) with deep features extracted from pre-treatment multiparametric MRI achieved a C-index of 0.74 on an independent validation cohort (37 patients) for OS prediction. Kickingereder et al. (Kickingereder et al., 2016b) showed that radiomics features extracted from pre-treatment multiparametric MRI achieved a C-index of 0.65 for OS prognosis. However no external validation was performed. Carles et al. (Carles et al., 2021) observed that second-order texture features extracted from [18F]-FDG PET acquired before re-irradiation can predict OS (p=0.038). Overall, the performance of our best performing conventional radiomics signature for the prognosis of OS based on features extracted from MET-PET and T1c-w MRI was somewhat lower than other validated results (MET-PET C-index=0.60, T1c-w MRI C-index=0.63). However, a full comparison with previous studies is not possible as we used post-surgical imaging instead of pre-treatment imaging for prognostic modelling.

The finding that MET-PET allowed for better prognosis than T1c-w MRI mirrors the finding for the residual tumour detection analysis presented in Chapter 5. Residual tumour burden is a prognostic imaging biomarker in GBM (Matsuo et al., 2012), however, postoperative MRI is prone to confounding effects that can lead to misinterpretation of residual tumours and alternatively reduced prognostic performance (Grosu et al., 2005; Matsuo et al., 2012). On the other hand, MET-PET is capable of providing better differentiation of nonspecific postoperative changes in GBM and therefore provides improved prognostic and diagnostic performance (Palanichamy & Chakravarti, 2017).

In order to build generalizable DL models, a large amount of data is required (LeCun et al., 2015). It is particularly challenging to build deep learning models in medical image analysis, where high-quality data is expensive and dependent on human resources for collection and labelling. To deal with the problem of limited training data, synthetic training examples are created using data augmentation techniques that can help large-capacity learners to benefit from more representative training data. Data augmentation can increase robustness of a deep learning model by increasing its ability to correctly predict unseen examples that are noisy or slightly perturbed. Therefore, it is necessary to perform data augmentation to reduce model overfitting. Further, due to the volumetric nature of medical imaging data 3D-CNN models are preferred over 2D-CNN models as explained in Section 2.6.3.

In this work, we were able to show that CNNs, despite being highly parametrized models, were able to achieve a somewhat better performance than conventional radiomics for prognostic modelling. The improved performance of deep learning models can be attributed to the use of 3D-CNN models together with extensive data augmentation, as explained in the methodology of this study. The generalizability of our 3D-CNN models was validated using an independent test cohort.

Limitations of this study are the relatively low number of patients in the training and test cohorts, which leads to model overfitting and wide confidence intervals. To overcome the problem of potential model overfitting, we used extensive feature selection approach in over conventional radiomics analysis and to diversify the training we used data augmentation approach in deep learning analysis.

For future studies, we plan to use both T1c-w and MET-PET as a two-channel input in a 3D-CNN to get a joint estimation of the prognostic performance from single 3D-CNN architectures to improve model performance. The inclusion of early post-surgical MRI into the analysis may also help to get improved prognostic performance, as it is less prone to nonspecific surgical changes in GBM.

7 Conclusion and further perspective

Personalized treatment is an evolving field in translational oncology that aims to determine the optimal treatment for each individual patient to improve treatment outcome. To do so, diagnostic, prognostic and predictive biomarkers are being developed based on each patient's clinical, imaging and/or molecular information. Molecular characterization of tumour utilizes biopsies or invasive surgeries to extract and analyse small portions of tumour tissue. However, radiomics attempts to develop biomarkers using imaging data which is noninvasive process and provides comprehensive view of entire tumour region. Further, radiomics analyses allows for tumour characterization at several time points because imaging is often repeated during treatment in routine clinical practice.

This thesis aims to develop diagnostic and prognostic radiomics biomarkers for predicting treatment response and long-term survival outcomes in patients with LARC and generalized linear model (GLM). The first part of this thesis is focused on a detailed analysis addressing challenges in field of radiomics for LARC, while the second part addresses the goal of treatment personalization in GBM by performing comparative analysis of deep learning and conventional radiomics for biomarker development.

One of the major challenges in radiomics analyses is that normally numerous quantitative imaging features of different complexity are extracted from multi-modality imaging data. In conventional radiomics analyses, all feature classes are pooled together to identify the prognostic signature. However, the association of individual feature classes to the endpoint of interest is not often investigated. In this thesis, we aimed to address this issue for predicting tumour response to nCRT and FFDM in LARC patients using diagnostic T2-w MRI and treatment planning CT data in a multicentre cohort. For tumour response prediction after nCRT, a novel signature based on LoG intensity features based on diagnostic T2-w MRI and treatment planning CT combined with cT stage was developed and validated, while for FFDM prediction, a signature based on a SOT feature based on treatment planning CT was developed and validated.

In the past decade, there has been an exponential growth in the radiomics literature. Several radiomics studies for LARC have shown the potential of radiomics-based prognostic modelling, however, most of these studies lack independent validation, which is an important step towards their clinical application. Therefore, in this thesis, an extensive literature search was performed to validate radiomics signatures developed by other researchers to predict tumour response to nCRT or FFDM in LARC using our multicentre cohort. Remarkably, only one out of 11 studies could be validated, indicating a lack of reproducibility of published radiomics models. We observed that studies use different software and methods for feature extraction and often do not report all required modelling details, which makes it difficult to reproduce or compare results.

Thus, for successful application of radiomics in clinical management of patients, it is necessary to standardize the radiomics workflow.

Among primary brain tumours, GBM is the most frequently occurring malignant brain tumour with a poor prognosis. The median OS time is 12-15 months, and there is a high recurrence rate after initial treatment. Most patients experience recurrence as it is difficult to completely excise the tumour during surgery. The residual tumour burden after surgery is an established imaging biomarker in GBM, which is commonly assessed on T1c-w MRI. PET imaging with radiotracers, such as [^{11}C] MET, provides greater insight into image-specific pathophysiological changes that extend beyond conventional T1c-w MRI. PET can be used for delineating the extent of the residual tumour, for radiotherapy planning, patient follow-up monitoring and prognosis. However, the accurate detection of residual tumour on MET-PET and T1c-w MRI is a complex evaluation procedure that involves expertise from radiologists, radiation oncologists and nuclear medicine experts, and it is at the risk of inter-rater variabilities. In this work, we leveraged the potential of radiomics for detection of residual tumours on MET-PET and T1c-w MRI acquired after surgery. Models were validated on an independent test cohort, and we could show that residual tumour status was easier to detect on MET-PET imaging than on T1c-w MRI. The best results were observed for a CNN-based 3D-DenseNet model. This model has the potential of increasing the clinician's confidence in residual tumour detection and reducing inter-rater variability.

GBM is associated with poor prognosis, and one of the major aims is to provide tailored clinical management that fits to the needs of individual patient and thereby improve patient survival. To achieve this aim, the development of noninvasive biomarkers for patient stratification into survival risk groups is decisive. In this work, we assessed whether conventional radiomics and deep learning-based imaging model developed on MET-PET and T1c-w MRI allow for the prognosis and stratification of patients with newly diagnosed GBM. The considered endpoints were TTR and OS. We compared machine-learning-based radiomics and 3D-CNN models of different levels of complexity to evaluate the prognostic performance of both imaging modalities. As are result, important clinical features combined with ensemble predictions from the CNN-based 3D-DenseNet model developed using MET-PET provided improved prognostic performance compared to machine learning models developed using a conventional radiomics approach. After further prospective validation, the proposed models may be considered for the treatment personalization in GBM.

The field of conventional radiomics and DL-base radiomics for treatment personalization in LARC and GBM still offers many interesting challenges and open research questions for the future. Some of them are shortly discussed in the following paragraphs.

Standardization and validation of radiomics models for GBM

One of the current challenges in the field of radiomics is a lack of reproducibility and standardization across heterogeneous acquisition protocols, multiple institutions, patient populations and radiomics workflows. Most of the radiomics-based studies develop their own models using different software tools. This makes it difficult to reproduce the results. A recent study by Fornacon-Wood et al. (Fornacon-Wood et al., 2020) showed that the use of different software platforms may result in different values of radiomics features on the same imaging data. One of the efforts towards the standardization of the radiomics workflow is the IBSI. The IBSI is an international collaboration that is focused on providing recommendations concerning feature calculation and standardized feature definition. In this thesis we focused on validation of previously published radiomics models for LARC, however the validation of GBM radiomics models is still an open question. As the current field of radiomics for treatment personalization GBM is moving towards more complex models, future research should focus on the validation of previously published models with standardized methods. This will not only help to identify the promising biomarkers, but also help to evaluate the potential of radiomics.

CNN for prognostic modeling in LARC

In this thesis, we developed and validated DL-based radiomics models for diagnosis and prognosis in GBM. Our analysis showed overall improved performance of DL over conventional radiomics modelling. Similarly, the prognostic performance of radiomics risk models in LARC can be improved by using a DL approach. Specifically, CNN-based models that are specialized to learn spatial features from imaging data have shown a higher performance than human raters for diagnostic tasks, e.g. for interpreting chest radiography and mammography (Rajpurkar et al., 2018; Ardila et al., 2019; Wu et al., 2019). However, CNNs are also capable of predicting patient prognosis by learning subtle difference in tumour properties related to outcome and risk. Existing studies for predicting patient response to nCRT and prognosis of long-term survival outcomes in LARC focused mainly on using deep features extracted using 2D-CNNs, which were then used by an external learner. Fu et al. (Fu et al., 2020) used the 2D-VGGNet model to extract deep features from a CNN. These features were further post-processed and 105 deep features were used to train a least absolute shrinkage and selection operator (LASSO) model. A recent study by Liu et al. (Liu et al., 2021) used a pre-trained 2D-ResNet model for prognosis of FFDM in LARC, however, model losses were optimized using the binary cross-entropy loss on FFDM status without accounting for time to FFDM and finally model predictions were mapped to time-to-event data using a Cox regression model. Thus, there is a need of end-to-end CNN models such as 3D-CNNs and transformers (He et al., 2022), to prognosticate patient outcome in LARC.

Delta radiomics for prognostic modelling in LARC and GBM

Radiomics modelling is normally based on quantitative imaging features extracted from single or multiple imaging modalities acquired at a single time point. However, the radiomics features acquired at a single time point may be insufficient to describe all characteristics of prognostic outcome. Incorporating information derived from several time points during treatment may contain additional prognostic information. Based on this idea, delta radiomics analyses feature variation between different time points before, during and after treatment. In principle, this approach allows for adapting the ongoing treatment. Chiloiro et al. (Chiloiro et al., 2020) used delta radiomics to predict two year distant metastases in LARC using the ratio of pre-nCRT and post-nCRT MRI based features. Other prognostic studies have shown significant associations between delta radiomics features and patient OS in non-small cell lung cancer (Khorrami et al., 2020; Shi et al., 2020) and locally advanced pancreatic cancer (Cusumano et al., 2021a). However, large multicentric prognostic studies with external validation utilizing delta radiomics are sparse in LARC and GBM. Usage of delta radiomics in LARC and GBM may yield useful biomarkers for evaluating different treatment strategies, which could be investigated in the future.

Deep learning for GBM

There are several open research questions concerning decision support and prognostication of outcome in newly diagnosed GBM that can be explored for future research with the help of deep learning. Firstly, a prognostic model based on single-modality medical imaging only partially reflects the available tumour information. Similar to clinicians, who perform diagnoses and give prognostic suggestions, predictive models should be based on multimodal imaging data to extract more diverse aspects of phenotypical tumour information and integrate them in model development. This is relatively simple to implement in conventional radiomics, where features from multiple modalities are extracted and combined for model building (Tewarie et al., 2021). CNNs have a specialized architecture that allows for processing multichannel data, e.g. the use of 3 (red, green, and blue) channels for RGB images. CNNs can be extended to process any number of channels at the cost of increased complexity and computational resources. This feature of CNN architectures can be utilized to process multimodal medical imaging data, where each imaging modality is fed to a designated channel. Thus, final prediction will be made on features extracted from all modalities. Nie et al. (Nie et al., 2019) used a multichannel 3D-CNN to extract deep features to prognosticate long and short-term OS (i.e. less than or more than 650 days after surgery) in high grade glioma patients and achieved accuracy of 90.66%. However, to the best of our knowledge, no study has been performed for the prognosis of TTR and OS in GBM patients using PET/MRI data in multichannel CNN models. Therefore, this question can be investigated in the future.

One of the most time-consuming tasks in radiotherapy (RT) treatment planning is defining the target volume, which is currently done manually by a human expert. Especially in GBM, the unclear margins of the residual tumour on post-operative images makes it difficult to delineate even for highly experienced experts (Visser et al., 2019). Over the last decade, auto-segmentation of brain tumours with machine learning and deep learning methods has been a widely explored area in medical imaging (Ghaffari et al., 2019). Promising results have been published for glioma segmentation using CNNs (Kamnitsas et al., 2017; Wang et al., 2018a), however, automatic residual tumour definition with CNNs on postoperative PET/MRI has not been widely explored. Zeng et al. (Zeng et al., 2016) used a generative model based on expectation maximization algorithm (Gooya et al., 2012) for residual tumour segmentation in high grade gliomas on post-operative MRI. The study did not report results on test data, however showed acceptable performance of algorithm on training data (Dice coefficient on training data=0.7). Miere et al. (Meier et al., 2017) reported a machine learning based auto-segmentation method for GBM residual tumour segmentation on MRI. The study showed a good agreement of volumetric estimate of residual tumour between automatic segmentation and human raters delineation (coefficient of concordance=0.693). In this thesis, we developed a method based on 3D-CNN models to detect the presence of residual tumours in MET-PET and T1c-w MRI. This work can be further extended to the auto-segmentation of residual tumour with CNNs that will help to facilitate the clinical workflow of GBM patient management.

Finally, it can be interesting to utilize MET-PET to assess the colocalization of MET uptake in pre-RCT PET images with recurrence sites. Seidlitz et al. (Seidlitz et al., 2021) showed that tumour recurrence occurred in the region of MET accumulation on pre-RCT images for 86.0% of cases. Identification of regions of recurrence would help to deliver higher radiation doses to the targeted volume and sparing the surrounding tissues. CNNs could be used to predict the regions, where a recurrence is most likely to occur.

In summary, we focused on the development and external validation of radiomics-based prognostic biomarkers for treatment personalization in LARC and GBM using multimodal data. We utilized 3D CNNs and compared their performance with conventional radiomics approaches using machine learning methods. Moreover, we focused on the requirement for standardization and external validation of radiomics biomarkers for their clinical application by conducting an external validation study of published radiomics signatures for tumour response prediction in LARC. The results may initiate future steps towards personalization of LARC and GBM treatment, e.g. by their application in interventional clinical trials after prospective validation.

8 Summary

Background

The standard treatment for locally advanced rectal cancer (LARC) is neoadjuvant chemoradiotherapy (nCRT) followed by total mesorectal excision (TME) and postoperative adjuvant chemotherapy, while the standard treatment for glioblastoma multiforme (GBM) is surgical resection followed by chemoradiotherapy (CRT). Despite intense multimodal treatment, local and distant progression remain leading problems in current patient management. The personalization of treatment is a central aim in cancer therapy to improve the outcome of patient populations with heterogeneous treatment response. LARC patients with a high chance of achieving pathologically complete response (pCR) after nCRT may benefit from the adaptation of low-morbidity surgeries or watch-and-wait strategies. Further, the assessment of patient prognosis in GBM before the start of treatment and the detection of residual tumours after surgery may help to identify patients that would benefit from escalated treatment strategies. Defining such sub-populations of patients requires the identification of biomarkers. Several studies have been analysing clinical, molecular, and imaging data to identify potential biomarkers for patient prognosis in LARC and GBM. However, few of these markers are currently considered for treatment personalization in clinical routine. Thus, the development of reliable biomarkers and the validation of existing studies may help to identify subgroups of patients for treatment adaptation.

Objectives

The main objective of this thesis is to identify and independently validate multimodal imaging biomarkers for outcome prediction after treatment in patients with LARC and newly diagnosed GBM using conventional feature-based radiomics and deep learning (DL) based radiomics approaches. Multimodal radiomics signatures are developed and validated for tumour response prediction after nCRT and freedom from distant metastases (FFDM) in LARC and for predicting time-to-recurrence (TTR) and overall survival (OS) in GBM. In addition, we perform an external validation study to validate previously published radiomics signatures for the prediction of tumour response to nCRT in LARC on our multi-centre cohort.

Material and methods

Imaging and clinical data of 190 LARC patients of the DKTK-ROG treated with nCRT followed by surgery was evaluated for developing radiomics signatures and for our external validation study. For treatment outcome prediction, a conventional feature-based radiomics approach using machine learning techniques was employed to develop multimodal signatures based on CT, T2w-MRI and clinical parameters. For each imaging modality, different feature classes were analysed,

i.e. morphological and first order (MFO), second-order texture (SOT) and Laplacian of Gaussian transformed (LoG) features. For the external validation study, the radiomics pipelines from included studies were replicated and validated on our multi-centre data. For the analysis of patients with GBM, imaging and clinical data of 132 adult patients were collected from the PETra trial and from an additional retrospective validation cohort. Conventional radiomics and 3D-CNN-based approaches were used to detect the residual tumour status in postoperative [^{11}C] MET-PET and in gadolinium-enhanced T1-w MRI. For the prognosis of TTR and OS in GBM patients, additional clinical parameters were included in the final models.

Results

For LARC, we developed and validated a radiomics signature based on LoG features extracted from pre-treatment T2-w MRI and treatment-planning CT combined with cT stage for the prediction of tumour response to nCRT, while SOT features were extracted from CT for the prediction of FFDM. Our external validation study of previously published radiomics signatures developed for tumour response prediction after nCRT in LARC patients based on our multi-centre data showed limited success. Of 11 studies that qualified for final validation, only one study achieved acceptable performance, which indicates a potential lack of reproducibility for radiomics studies. For patients with GBM, MET-PET allowed for a better classification of the residual tumour status and prognosis of TTR and OS than T1c-w MRI. For MET-PET-based residual tumour status detection, the best performance was achieved by 3D-CNNs, while for MRI, the best performance was given by logistic regression using a conventional feature-based radiomics approach. Finally, for the prognosis of TTR and OS in GBM, the best performance with a significant stratification of patients in groups at low and high risk was observed when combining clinical parameters with a 3D-CNN ensemble model based on MET-PET imaging.

Conclusion

In this thesis, novel radiomics signatures were identified by combining multimodal imaging and clinical information to predict tumour response to nCRT and FFDM in LARC as well as TTR and OS in GBM patients. Furthermore, this thesis provides valuable insight into unaddressed issues in the radiomics workflow. Firstly, the interpretability of features in radiomics is poorly understood, as a large number of features of different complexity are commonly extracted. In this thesis, we addressed this issue by evaluating the performance of different feature classes. Secondly, the standardization of the radiomics workflow is generally overlooked in various radiomics studies. In this thesis, we emphasize on the need of reproducibility and standardization in the radiomics process by conducting an external validation study to validate previously published radiomics signatures for LARC. This thesis also provides valuable insight into conventional radiomics and DL-based radiomics analyses. Our GBM analysis for the detection of residual tumour status and prognosis of TTR and OS using MET-PET and T1c-w MRI data revealed that DL-based radiomics

may provide improved diagnostic and prognostic performance. The analyses in this thesis can be further extended by combining genomics and molecular signatures with radiomics signatures. This might help to improve the prognostic performance for LARC and GBM. Overall, the radiomics signatures identified in this thesis have to be validated in prospective studies before their potential application in interventional clinical trials to improve personalized treatments.

9 Zusammenfassung

Hintergrund

Die Standardbehandlung für lokal fortgeschrittene Rektumkarzinome (LARC) ist eine neoadjuvante Radiochemotherapie (nCRT), gefolgt von einer totalen mesorektalen Exzision (TME) und einer postoperativen adjuvanten Chemotherapie. Die Standardbehandlung für Glioblastoma multiforme (GBM) besteht aus einer chirurgischen Resektion, gefolgt von einer CRT. Trotz intensiver multimodaler Behandlung limitieren lokale Rezidive und Fernmetastasen das Behandlungsergebnis. Die Personalisierung der Behandlung ist ein zentrales Ziel in der Krebstherapie, um das Outcome von Patientenpopulationen mit heterogenem Therapieansprechen zu verbessern. LARC-Patienten mit einer hohen Wahrscheinlichkeit einer pathologisch vollständigen Remission (pCR) können von angepassten Operationsformen oder Watch-and-Wait-Strategien profitieren. Darüber hinaus kann die Beurteilung der Patientenprognose bei GBM vor Beginn der Behandlung und die Erkennung von Resttumoren nach der Operation dazu beitragen, Patienten zu identifizieren, die von eskalierten Behandlungsstrategien profitieren würden. Die Definition solcher Patientengruppen erfordert den Einsatz von Biomarkern. Mehrere Studien haben klinische, molekulare und bildgebende Daten analysiert, um potenzielle Biomarker für die Patientenprognose bei LARC und GBM zu identifizieren. Allerdings werden derzeit nur wenige dieser Marker für die Personalisierung der Behandlung in der klinischen Routine in Betracht gezogen. Daher kann die Entwicklung zuverlässiger Biomarker und die Validierung bestehender Studien dazu beitragen, Patientengruppen für eine Behandlungsanpassung zu identifizieren.

Fragestellung

Das Hauptziel dieser Arbeit ist die Identifizierung und unabhängige Validierung multimodaler bildgebender Biomarker für die Outcomevorhersage von Patienten mit LARC und GBM unter Verwendung konventioneller merkmalsbasierter Radiomics-Verfahren und Deep Learning (DL)-basierter Radiomics-Ansätze. Multimodale Radiomics-Signaturen werden für die Vorhersage des Tumoransprechens nach nCRT bei LARC und für die Vorhersage der Zeit bis zum Rezidiv (TTR) und des Gesamtüberlebens (OS) bei GBM entwickelt und validiert. Darüber hinaus wird eine externe Validierungsstudie durchgeführt, um veröffentlichte Radiomics-Signaturen für die Vorhersage des Tumoransprechens auf nCRT in LARC basierend auf unserer multizentrischen Kohorte zu validieren.

Material und methoden

Bildgebende und klinische Daten von 190 LARC-Patienten des DKTK-ROG, die mit nCRT und TME behandelt wurden, wurden für die Entwicklung von Radiomics-Signaturen und für die externe Validierungsstudie ausgewertet. Zur Vorhersage des Behandlungsergebnisses wurde ein konventioneller merkmalsbasierter Radiomics-Ansatz unter Verwendung von maschinellen Lerntechniken eingesetzt, um multimodale Signaturen auf der Grundlage von CT, T2w-MRT und klinischen Parametern zu entwickeln. Für jede Bildgebungsmodalität wurden verschiedene Merkmalsklassen analysiert, d. h. morphologische Merkmale und Merkmale erster Ordnung (MFO), Texturmerkmale zweiter Ordnung (SOT) und Laplace-transformierte (LoG) Merkmale. Für die externe Validierungsstudie wurden die Radiomics-Pipelines aus eingeschlossenen Studien repliziert und anhand der vorliegenden multizentrischen Daten validiert. Für die Analyse von Patienten mit GBM wurden bildgebende und klinische Daten von 132 erwachsenen Patienten aus der PETra-Studie und aus einer zusätzlichen retrospektiven Validierungskohorte verwendet. Konventionelle Radiomics-Verfahren und 3D-CNN-basierte Ansätze wurden genutzt, um den Resttumorstatus in der postoperativen [^{11}C] MET-PET und in der T1-w-MRT zu erkennen. Für die Prognose der TTR und des OS wurden zusätzlich klinische Parameter in die endgültigen Modelle aufgenommen.

Ergebnisse

Für Patienten mit LARC wurde eine Radiomics-Signatur entwickelt und validiert, die LoG-Merkmale der T2-w-MRT sowie des Bestrahlungsplanungs-CT in Kombination mit dem cT-Stadium für die Vorhersage des Tumoransprechens auf nCRT enthält, während SOT-Merkmale des CT zur Vorhersage der Metastasenfreiheit (FFDM) verwendet wurden. Die externe Validierung bereits veröffentlichter Radiomics-Signaturen für die Vorhersage des Ansprechens von LARC zeigte auf Basis der vorliegenden Patientenkohorte nur begrenzten Erfolg. Von 11 Studien, die sich für die endgültige Validierung qualifizierten, erreichte nur eine Studie eine akzeptable Güte, was auf einen potenziellen Mangel an Reproduzierbarkeit für Radiomics-Studien hindeutet. Bei Patienten mit GBM ermöglichte die MET-PET eine bessere Klassifizierung des Resttumorstatus und eine bessere Prognose von TTR und OS als die T1c-w MRT. Bei der MET-PET-basierten Erkennung des Resttumorstatus wurde die beste Vorhersage durch 3D-CNNs erzielt, während bei der MRT die beste Güte durch logistische Regression unter Verwendung eines konventionellen merkmalsbasierten Radiomics-Ansatzes erzielt wurde. Schließlich wurde für die Prognose von TTR und OS bei GBM die beste Vorhersage mit einer signifikanten Stratifizierung von Patienten in Gruppen mit niedrigem und hohem Risiko beobachtet, wenn klinische Parameter mit einem 3D-CNN-Ensemble-Modell basierend auf MET-PET-Bildgebung kombiniert wurden.

Schlussfolgerung

In dieser Arbeit wurden neuartige Radiomics-Signaturen durch die Kombination multimodaler Bildgebung und klinischer Informationen identifiziert, um das Ansprechen des Tumors auf nCRT und die FFDM bei LARC-Patienten sowie TTR und OS bei GBM-Patienten vorherzusagen. Darüber hinaus bietet diese Arbeit wertvolle Einblicke in einige Probleme des Radiomics-Workflows.

Erstens ist die Interpretierbarkeit von Radiomics-Signaturen häufig schwierig, da üblicherweise eine große Anzahl von Merkmalen unterschiedlicher Komplexität extrahiert wird. In dieser Arbeit wurde dieses Problem adressiert, indem der prognostische Beitrag verschiedener Merkmalsklassen einzeln untersucht wurde. Zweitens findet die Standardisierung des Radiomics-Workflows in zahlreichen Radiomics-Studien wenig Beachtung. Anhand einer externen Validierungsstudie von zuvor veröffentlichten Radiomics-Signaturen für LARC wird die Notwendigkeit der Reproduzierbarkeit und Standardisierung des Radiomics-Prozesses in dieser Arbeit herausgestellt. Die Arbeit bietet weiterhin wertvolle Einblicke in konventionelle Radiomics-Verfahren und DL-basierte Radiomics-Analysen. Die GBM-Analysen zur Erkennung des Resttumorstatus und zur Prognose von TTR und OS ergaben, dass DL-basierte Radiomics-Ansätze eine verbesserte diagnostische und prognostische Güte aufweisen können. Die in dieser Arbeit entwickelten Radiomics-Signaturen können in der Zukunft mit zusätzlichen Omics-Ebenen, wie beispielsweise Gensignaturen, kombiniert werden um die prognostische Güte weiter zu verbessern. Insgesamt müssen die in dieser Arbeit identifizierten Radiomics-Signaturen in prospektiven Studien validiert werden, bevor sie in interventionellen klinischen Studien zur Verbesserung der personalisierten Behandlung eingesetzt werden können.

Bibliography

Aerts, H. J. (2016). The potential of radiomic-based phenotyping in precision medicine: A review. *JAMA oncology*, *2*. 1636–1642. https://doi.org/10.1001/jamaoncol.2016.2631

Aker, M., Ganeshan, B., Afaq, A., Wan, S., Groves, A. M., & Arulampalam, T. (2019). Magnetic resonance texture analysis in identifying complete pathological response to neoadjuvant treatment in locally advanced rectal cancer. *Diseases of the Colon & Rectum*, *62*. 163–170. https://doi.org/10.1097/DCR.0000000000001224

Alexander, B. M., & Cloughesy, T. F. (2017). Adult glioblastoma. *Journal of Clinical Oncology*, *35*. 2402–2409. https://doi.org/10.1200/JCO.2017.73.0119

Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, *19*. 1264–1274. https://doi.org/10.1109/21.44046

Amin, M. B., Edge, S. B., Greene, F. L., Byrd, D. R., Brookland, R. K., Washington, M. K., Gershenwald, J. E., Compton, C. C., Hess, K. R., Sullivan, D. C., et al. (2017). *Ajcc cancer staging manual* (Vol. 1024). Springer. https://doi.org/10.1007/978-1-4757-3656-4

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., & Aggarwal, B. B. (2008). Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research*, *25*. 2097–2116. https://doi.org/10.1007/s11095-008-9661-9

Antunes, J. T., Ofshteyn, A., Bera, K., Wang, E. Y., Brady, J. T., Willis, J. E., Friedman, K. A., Marderstein, E. L., Kalady, M. F., Stein, S. L., et al. (2020). Radiomic features of primary rectal cancers on baseline t2-weighted mri are associated with pathologic complete response to neoadjuvant chemoradiation: A multisite study. *Journal of Magnetic Resonance Imaging*, *52*. 1531–1541. https://doi.org/10.1002/jmri.27140

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, *25*. 954–961. https://doi.org/10.1038/s41591-019-0536-x

Bailey, O. T. (1985). Genesis of the percival bailey-cushing classification of gliomas. *Pediatric Neurosurgery*, *12*. 261–265. https://doi.org/10.1159/000120262

Balaban, R. S., & Peters, D. C. (2019). Basic principles of cardiovascular magnetic resonance. In *Cardiovascular magnetic resonance* (pp. 1–14). Elsevier. https://doi.org/10.1016/B978-0-323-41561-3.00001-X

Balagurunathan, Y., Gu, Y., Wang, H., Kumar, V., Grove, O., Hawkins, S., Kim, J., Goldgof, D. B., Hall, L. O., Gatenby, R. A., et al. (2014a). Reproducibility and prognosis of quantitative features extracted from ct images. *Translational oncology*, *7*. 72–87. https://doi.org/10.1593/tlo.13844

Balagurunathan, Y., Kumar, V., Gu, Y., Kim, J., Wang, H., Liu, Y., Goldgof, D. B., Hall, L. O., Korn, R., Zhao, B., et al. (2014b). Test–retest reproducibility analysis of lung ct image features. *Journal of digital imaging*, *27*. 805–823. https://doi.org/10.1007/s10278-014-9716-x

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*. 803–821. https://doi.org/10.2307/2532201

Bang, J.-I., Ha, S., Kang, S.-B., Lee, K.-W., Lee, H.-S., Kim, J.-S., Oh, H.-K., Lee, H.-Y., & Kim, S. E. (2016). Prediction of neoadjuvant radiation chemotherapy response and survival using pretreatment [18 f] fdg pet/ct scans in locally advanced rectal cancer. *European journal of nuclear medicine and molecular imaging*, *43*. 422–431. https://doi.org/10.1007/s00259-015-3180-9

Barnholtz-Sloan, J. S., Ostrom, Q. T., & Cote, D. (2018). Epidemiology of brain tumors. *Neurologic clinics*, *36*. 395–419. https://doi.org/10.1016/j.ncl.2018.04.001

Becquerel, H. (1896). Sur les radiations invisible emises par les corps phosphorescents. *Comptes rendus*, *122*. 501–503. http://gallica.bnf.fr/ark:/12148/bpt6k30780.image.f503

Bengio, Y., et al. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, *2*. 1–127. https://doi.org/10.1561/2200000006

Bibault, J.-E., Giraud, P., Housset, M., Durdux, C., Taieb, J., Berger, A., Coriat, R., Chaussade, S., Dousset, B., Nordlinger, B., et al. (2018). Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Scientific reports*, *8*. 12611. https://doi.org/10.1038/s41598-018-35359-7

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer. https://link.springer.com/book/9780387310732

Bloch, F. (1946). Nuclear induction. *Physical review*, *70*. 460. https://doi.org/10.1016/0031-8914(51)90068-7

Bloch, O., Han, S. J., Cha, S., Sun, M. Z., Aghi, M. K., McDermott, M. W., Berger, M. S., & Parsa, A. T. (2012). Impact of extent of resection for recurrent glioblastoma on overall survival. *Journal of neurosurgery*, *117*. 1032–1038. https://doi.org/10.3171/2012.9.JNS12504

Boige, V., Mendiboure, J., Pignon, J.-P., Loriot, M.-A., Castaing, M., Barrois, M., Malka, D., Trégouët, D.-A., Bouché, O., Le Corre, D., et al. (2010). Pharmacogenetic assessment of toxicity and outcome in patients with metastatic colorectal cancer treated with lv5fu2, folfox, and folfiri: Ffcd 2000-05. *Journal of Clinical Oncology*, *28*. 2556–2564. https://doi.org/10.1200/JCO.2009.25.2106

Boldrini, L., Cusumano, D., Chiloiro, G., Casà, C., Masciocchi, C., Lenkowicz, J., Cellini, F., Dinapoli, N., Azario, L., Teodoli, S., et al. (2019). Delta radiomics for rectal cancer response prediction with hybrid 0.35 t magnetic resonance-guided radiotherapy (mrgrt): A hypothesis-generating study for an innovative personalized medicine approach. *La radiologia medica*, *124*. 145–153. https://doi.org/10.1007/s11547-018-0951-y

Bondy, M. L., Scheurer, M. E., Malmer, B., Barnholtz-Sloan, J. S., Davis, F. G., Il'Yasova, D., Kruchko, C., McCarthy, B. J., Rajaraman, P., Schwartzbaum, J. A., et al. (2008). Brain tumor epidemiology: Consensus from the brain tumor epidemiology consortium. *Cancer*, *113*. 1953–1968. https://doi.org/10.1002/cncr.23741

Bouguer, P. (1729). *Essai d'optique sur la gradation de la lumière*. Claude Jombert. https://doi.org/10.1259/jrs.1922.0026

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*. 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (2001). Random forests. *Machine learning*, *45*. 5–32. https://doi.org/10.1023/A:1010933404324

Bulens, P., Couwenberg, A., Intven, M., Debucquoy, A., Vandecaveye, V., Van Cutsem, E., D'Hoore, A., Wolthuis, A., Mukherjee, P., Gevaert, O., et al. (2020). Predicting the tumor response to chemoradiotherapy for rectal cancer: Model development and external validation using mri radiomics. *Radiotherapy and Oncology*, *142*. 246–252. https://doi.org/10.1016/j.radonc.2019.07.033

Bundschuh, R. A., Dinges, J., Neumann, L., Seyfried, M., Zsótér, N., Papp, L., Rosenberg, R., Becker, K., Astner, S. T., Henninger, M., et al. (2014). Textural parameters of tumor heterogeneity in 18f-fdg pet/ct for therapy response assessment and prognosis in patients with locally advanced rectal cancer. *Journal of Nuclear Medicine*, *55*. 891–897. https://doi.org/10.2967/jnumed.113.127340

Candido, S., Lupo, G., Pennisi, M., Basile, M. S., Anfuso, C. D., Petralia, M. C., Gattuso, G., Vivarelli, S., Spandidos, D. A., Libra, M., et al. (2019). The analysis of mirna expression profiling datasets reveals inverse microrna patterns in glioblastoma and alzheimer's disease. *Oncology Reports*, *42*. 911–922. https://doi.org/10.3892/or.2019.7215

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*. 679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Carles, M., Popp, I., Starke, M. M., Mix, M., Urbach, H., Schimek-Jasch, T., Eckert, F., Niyazi, M., Baltas, D., & Grosu, A. L. (2021). Fet-pet radiomics in recurrent glioblastoma: Prognostic value for outcome after re-irradiation? *Radiation Oncology*, *16*. 1–10. https://doi.org/10.1186/s13014-020-01744-8

Caruso, D., Zerunian, M., Ciolina, M., de Santis, D., Rengo, M., Soomro, M. H., Giunta, G., Conforto, S., Schmid, M., Neri, E., et al. (2018). Haralick's texture features for the prediction of response to therapy in colorectal cancer: A preliminary study. *La radiologia medica*, *123*. 161–167. https://doi.org/10.1007/s11547-017-0833-8

Chaddad, A., Sabri, S., Niazi, T., & Abdulkarim, B. (2018). Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. *Medical & biological engineering & computing*, *56*. 2287–2300. https://doi.org/10.1007/s11517-018-1858-4

Chalkidou, A., O'Doherty, M. J., & Marsden, P. K. (2015). False discovery rates in pet and ct studies with texture features: A systematic review. *PloS one*, *10*. e0124165. https://doi.org/10.1371/journal.pone.0124165

Chau, I., Brown, G., Cunningham, D., Tait, D., Wotherspoon, A., Norman, A. R., Tebbutt, N., Hill, M., Ross, P. J., Massey, A., et al. (2006). Neoadjuvant capecitabine and oxaliplatin followed by synchronous chemoradiation and total mesorectal excision in magnetic resonance imaging–defined poor-risk rectal cancer. *Journal of Clinical Oncology*, *24*. 668–674. https://doi.org/10.1200/JCO.2005.04.4875

Chee, C. G., Kim, Y. H., Lee, K. H., Lee, Y. J., Park, J. H., Lee, H. S., Ahn, S., & Kim, B. (2017). Ct texture analysis in patients with locally advanced rectal cancer treated with neoadjuvant chemoradiotherapy: A potential imaging biomarker for treatment response and prognosis. *PloS one*, *12*. e0182883. https://doi.org/10.1371/journal.pone.0182883

Cheeseman, P. C., Stutz, J. C., et al. (1996). Bayesian classification (autoclass): Theory and results. *Advances in knowledge discovery and data mining*, *180*. 153–180. https://www.eecis.udel.edu/~shatkay/Course/papers/CheesemanStutzAutoclass.pdf

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, Y., Luo, Y., Hu, Y., Zhang, Z., Wang, X., Yu, Q., Liu, G., Cui, E., Yu, T., & Jiang, X. (2021). Multiparametric mri-based radiomics approaches on predicting response to neoadjuvant chemoradiotherapy (ncrt) in patients with rectal cancer. *Abdominal Radiology*, *46*. 5072–5085. https://doi.org/10.1007/s00261-021-03219-0

Chhikara, B. S., & Parang, K. (2023). Global cancer statistics 2022: The trends projection analysis. *Chemical Biology Letters*, *10*. 451–451. https://pubs.thesciencein.org/journal/index.php/cbl/article/view/451

Chidambaram, V., Brierley, J. D., Cummings, B., Bhayana, R., Menezes, R. J., Kennedy, E. D., Kirsch, R., & Jhaveri, K. S. (2017). Investigation of volumetric apparent diffusion coefficient histogram analysis for assessing complete response and clinical outcomes following pre-operative chemoradiation treatment for rectal carcinoma. *Abdominal Radiology*, *42*. 1310–1318. https://doi.org/10.1007/s00261-016-1010-6

Chiloiro, G., Rodriguez-Carnero, P., Lenkowicz, J., Casà, C., Masciocchi, C., Boldrini, L., Cusumano, D., Dinapoli, N., Meldolesi, E., Carano, D., et al. (2020). Delta radiomics can predict distant metastasis in locally advanced rectal cancer: The challenge to personalize the cure. *Frontiers in Oncology*, *10*. 595012. https://doi.org/10.3389/fonc.2020.595012

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, *14*. e1006076. https://doi.org/10.1371/journal.pcbi.1006076

Cho, H.-h., Lee, S.-h., Kim, J., & Park, H. (2018). Classification of the glioma grading using radiomics analysis. *PeerJ*, *6*. e5982. https://doi.org/10.7717/peerj.5982

Chow, D., Qi, J., Guo, X., Miloushev, V., Iwamoto, F., Bruce, J., Lassman, A., Schwartz, L., Lignelli, A., Zhao, B., et al. (2014). Semiautomated volumetric measurement on postcontrast mr imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *American Journal of Neuroradiology*, *35*. 498–503. https://doi.org/10.3174/ajnr.A3724

Coburger, J., Wirtz, C. R., & Konig, R. (2017). Impact of extent of resection and recurrent surgery on clinical outcome and overall survival in a consecutive series of 170 patients for glioblastoma in intraoperative high field magnetic resonance imaging. *J Neurosurg Sci*, *61*. 233–244. https://doi.org/10.23736/S0390-5616.16.03284-7

Coppola, F., Mottola, M., Lo Monaco, S., Cattabriga, A., Cocozza, M. A., Yuan, J. C., De Benedittis, C., Cuicchi, D., Guido, A., Rojas Llimpe, F. L., et al. (2021). The heterogeneity of skewness in t2w-based radiomics predicts the response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Diagnostics*, *11*. 795. https://doi.org/10.3390/diagnostics11050795

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*. 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Cox, D., & Oakes, D. (1984). Analysis of survival data chapman and hall. *New York*. https://doi.org/10.1002/bimj.4710290119

Crimı, F., Capelli, G., Spolverato, G., Bao, Q. R., Florio, A., Milite Rossi, S., Cecchin, D., Albertoni, L., Campi, C., Pucciarelli, S., et al. (2020). Mri t2-weighted sequences-based texture analysis (ta) as a predictor of response to neoadjuvant chemo-radiotherapy (ncrt) in patients with locally advanced rectal cancer (larc). *La radiologia medica*, *125*. 1216–1224. https://doi.org/10.1007/s11547-020-01215-w

Cui, Y., Yang, X., Shi, Z., Yang, Z., Du, X., Zhao, Z., & Cheng, X. (2019). Radiomics analysis of multiparametric mri for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *European radiology*, *29*. 1211–1220. https://doi.org/10.1007/s00330-018-5683-9

Cui, Y., Tha, K. K., Terasaka, S., Yamaguchi, S., Wang, J., Kudo, K., Xing, L., Shirato, H., & Li, R. (2016). Prognostic imaging biomarkers in glioblastoma: Development and independent validation on the basis of multiregion and quantitative analysis of mr images. *Radiology*, *278*. 546–553. https://doi.org/10.1148/radiol.2015150358

Cusumano, D., Boldrini, L., Yadav, P., Casà, C., Lee, S. L., Romano, A., Piras, A., Chiloiro, G., Placidi, L., Catucci, F., et al. (2021a). Delta radiomics analysis for local control prediction in pancreatic cancer patients treated using magnetic resonance guided radiotherapy. *Diagnostics*, *11*. 72. https://doi.org/10.3390/diagnostics11010072

Cusumano, D., Dinapoli, N., Boldrini, L., Chiloiro, G., Gatta, R., Masciocchi, C., Lenkowicz, J., Casà, C., Damiani, A., Azario, L., et al. (2018). Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. *La radiologia medica*, *123*. 286–295. https://doi.org/10.1007/s11547-017-0838-3

Cusumano, D., Meijer, G., Lenkowicz, J., Chiloiro, G., Boldrini, L., Masciocchi, C., Dinapoli, N., Gatta, R., Casà, C., Damiani, A., et al. (2021b). A field strength independent mr radiomics model to predict pathological complete response in locally advanced rectal cancer. *La radiologia medica*, *126*. 421–429. https://doi.org/10.1007/s11547-020-01266-z

Dahlbom, S., & Cherry, S. (2006). Pet: Physics, instrumentation and scanners. https://doi.org/10.1007/978-0-387-22529-6_1

Davey, M. S., Davey, M. G., Ryan, E. J., Hogan, A. M., Kerin, M. J., & Joyce, M. (2021). The use of radiomic analysis of magnetic resonance imaging in predicting distant metastases of rectal carcinoma following surgical resection: A systematic review and meta-analysis. *Colorectal Disease*, *23*. 3065–3072. https://doi.org/10.1111/codi.15919

De Cecco, C. N., Ciolina, M., Caruso, D., Rengo, M., Ganeshan, B., Meinel, F. G., Musio, D., De Felice, F., Tombolini, V., & Laghi, A. (2016). Performance of diffusion-weighted imaging, perfusion imaging, and texture analysis in predicting tumoral response to neoadjuvant chemoradiotherapy in rectal cancer patients studied with 3t mr: Initial experience. *Abdominal Radiology*, *41*. 1728–1735. https://doi.org/10.1007/s00261-016-0733-8

De Cecco, C. N., Ganeshan, B., Ciolina, M., Rengo, M., Meinel, F. G., Musio, D., De Felice, F., Raffetto, N., Tombolini, V., & Laghi, A. (2015). Texture analysis as imaging biomarker of tumoral response to neoadjuvant chemoradiotherapy in rectal cancer patients studied with 3-t magnetic resonance. *Investigative radiology*, *50*. 239–245. https://doi.org/10.1097/RLI.0000000000000116

Delli Pizzi, A., Chiarelli, A. M., Chiacchiaretta, P., d'Annibale, M., Croce, P., Rosa, C., Mastrodicasa, D., Trebeschi, S., Lambregts, D. M. J., Caposiena, D., et al. (2021). Mri-based clinical-radiomics model predicts tumor response before treatment in locally advanced rectal cancer. *Scientific Reports*, *11*. 1–11. https://doi.org/10.1038/s41598-021-84816-3

Demler, O. V., Paynter, N. P., & Cook, N. R. (2015). Tests of calibration and goodness-of-fit in the survival setting. *Statistics in medicine*, *34*. 1659–1680. https://doi.org/10.1002/sim.6428

Depeursinge, A., Andrearczyk, V., Whybra, P., van Griethuysen, J., Müller, H., Schaer, R., Vallières, M., & Zwanenburg, A. (2020). Standardised convolutional filtering for radiomics. *arXiv preprint arXiv:2006.05470*. https://doi.org/10.48550/arXiv.2006.05470

Dinapoli, N., Barbaro, B., Gatta, R., Chiloiro, G., Casà, C., Masciocchi, C., Damiani, A., Boldrini, L., Gambacorta, M. A., Dezio, M., et al. (2018). Magnetic resonance, vendor-independent, intensity histogram analysis predicting pathologic complete response after radiochemotherapy of rectal cancer. *International Journal of Radiation Oncology* Biology* Physics*, *102*. 765–774. https://doi.org/10.1016/j.ijrobp.2018.04.065

Dossa, F., Chesney, T. R., Acuna, S. A., & Baxter, N. N. (2017). A watch-and-wait approach for locally advanced rectal cancer after a clinical complete response following neoadjuvant chemoradiation: A systematic review and meta-analysis. *The lancet Gastroenterology & hepatology*, *2*. 501–513. https://doi.org/10.1016/S2468-1253(17)30074-2

Dudovitch, G. (2019). A 3d implementation of densenet & densenetfcn [Accessed: 2021-09-30]. URL: https://github.com/GalDude33/DenseNetFCN-3D

Duffy, M. J., & Crown, J. (2008). A personalized approach to cancer treatment: How biomarkers can help. *Clinical chemistry*, *54*. 1770–1779. https://doi.org/10.1373/clinchem.2008.110056

Duldulao, M. P., Lee, W., Streja, L., Chu, P., Li, W., Chen, Z., Kim, J., & Garcia-Aguilar, J. (2013). Distribution of residual cancer cells in the bowel wall after neoadjuvant chemoradiation in patients with rectal cancer. *Diseases of the colon and rectum*, *56*. 142. https://doi.org/10.1097/DCR.0b013e31827541e2

Dworak, O., Keilholz, L., & Hoffmann, A. (1997). Pathological features of rectal cancer after preoperative radiochemotherapy. *International journal of colorectal disease*, *12*. 19–23. https://doi.org/10.1007/s003840050072

Edge, S. B., & Compton, C. C. (2010). The american joint committee on cancer: The 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of surgical oncology*, *17*. 1471–1474. https://doi.org/10.1245/s10434-010-0985-4

Efron, B., & Hastie, T. (2013). Computer age statistical inference. https://doi.org/10.1017/CBO9781316576533

Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in medicine*, *14*. 73–82. https://doi.org/10.1002/sim.4780140108

Ferguson, S., & Lesniak, M. S. (2005). Percival bailey and the classification of brain tumors. *Neurosurgical focus*, *18*. 1–6. https://doi.org/10.3171/foc.2005.18.4.8

Ferrari, R., Mancini-Terracciano, C., Voena, C., Rengo, M., Zerunian, M., Ciardiello, A., Grasso, S., Mare, V., Paramatti, R., Russomando, A., et al. (2019). Mr-based artificial intelligence model to assess response to therapy in locally advanced rectal cancer. *European journal of radiology*, *118*. 1–9. https://doi.org/10.1016/j.ejrad.2019.06.013

Fisher, B., Wolmark, N., Rockette, H., Redmond, C., Deutsch, M., Wickerham, D. L., Fisher, E. R., Caplan, R., Jones, J., Lerner, H., et al. (1988). Postoperative adjuvant chemotherapy or radiation therapy for rectal cancer: Results from nsabp protocol r-011. *JNCI: Journal of the National Cancer Institute*, *80*. 21–29. https://doi.org/10.1093/jnci/80.1.21

Fisher, J. L., Schwartzbaum, J. A., Wrensch, M., & Wiemels, J. L. (2007). Epidemiology of brain tumors. *Neurologic clinics*, *25*. 867–890. https://doi.org/10.1016/B978-0-7216-8148-1.50004-8

Fornacon-Wood, I., Mistry, H., Ackermann, C. J., Blackhall, F., McPartlin, A., Faivre-Finn, C., Price, G. J., & O'Connor, J. P. (2020). Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European radiology*, *30*. 6241–6250. https://doi.org/10.1007/s00330-020-06957-9

Franke, A. J., Skelton IV, W. P., George, T. J., & Iqbal, A. (2021). A comprehensive review of randomized clinical trials shaping the landscape of rectal cancer therapy. *Clinical Colorectal Cancer*, *20*. 1–19. https://doi.org/10.1016/j.clcc.2020.07.009

Fu, J., Zhong, X., Li, N., Van Dams, R., Lewis, J., Sung, K., Raldow, A. C., Jin, J., & Qi, X. S. (2020). Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer. *Physics in Medicine & Biology*, *65*. 075001. https://doi.org/10.1088/1361-6560/ab7970

Galldiks, N., Langen, K.-J., Holy, R., Pinkawa, M., Stoffels, G., Nolte, K. W., Kaiser, H. J., Filss, C. P., Fink, G. R., Coenen, H. H., et al. (2012). Assessment of treatment response in patients with glioblastoma using o-(2-18f-fluoroethyl)-l-tyrosine pet in comparison to mri. *Journal of Nuclear Medicine*, *53*. 1048–1057. https://doi.org/10.2967/jnumed.111.098590

Galldiks, N., Ullrich, R., Schroeter, M., Fink, G. R., & Kracht, L. W. (2010). Volumetry of [11 c]-methionine pet uptake and mri contrast enhancement in patients with recurrent glioblastoma multiforme. *European journal of nuclear medicine and molecular imaging*, *37*. 84–92. https://doi.org/10.1007/s00259-009-1219-5

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer graphics and image processing*, *4*. 172–179. https://doi.org/10.1016/S0146-664X(75)80008-6

Gelfand, I. M., & I A glom, A. (1959). *Calculation of the amount of information about a random function contained in another such function*. American Mathematical Society Providence. https://doi.org/10.1090/trans2/012

Gensheimer, M. F., & Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, *7*. e6257. https://doi.org/10.7717/peerj.6257

Ghaffari, M., Sowmya, A., & Oliver, R. (2019). Automated brain tumor segmentation using multimodal brain scans: A survey based on models submitted to the brats 2012–2018 challenges. *IEEE reviews in biomedical engineering*, *13*. 156–168. https://doi.org/10.1109/RBME.2019.2946868

Giannini, V., Mazzetti, S., Bertotto, I., Chiarenza, C., Cauda, S., Delmastro, E., Bracco, C., Di Dia, A., Leone, F., Medico, E., et al. (2019). Predicting locally advanced rectal cancer response to neoadjuvant therapy with 18 f-fdg pet and mri radiomics features. *European* *journal of nuclear medicine and molecular imaging*, *46*. 878–888. https://doi.org/10.1007/s00259-018-4250-6

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*. 563–577. https://doi.org/10.1148/radiol.2015151169

Giunchiglia, E., Nemchenko, A., & van der Schaar, M. (2018). Rnn-surv: A deep recurrent model for survival analysis. *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. 23–32. https://doi.org/10.1007/978-3-030-01424-7_3

Gooya, A., Pohl, K. M., Bilello, M., Cirillo, L., Biros, G., Melhem, E. R., & Davatzikos, C. (2012). Glistr: Glioma image segmentation and registration. *IEEE transactions on medical imaging*, *31*. 1941–1954. https://doi.org/10.1109/TMI.2012.2210558

Grosu, A.-L., Weber, W. A., Riedel, E., Jeremic, B., Nieder, C., Franz, M., Gumprecht, H., Jaeger, R., Schwaiger, M., & Molls, M. (2005). L-(methyl-11c) methionine positron emission tomography for target delineation in resected high-grade gliomas before radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, *63*. 64–74. https://doi.org/10.1016/j.ijrobp.2005.01.045

Group, M. S., et al. (2007). Extramural depth of tumor invasion at thin-section mr in patients with rectal cancer: Results of the mercury study. *Radiology*, *243*. 132–139. https://doi.org/10.1148/radiol.2431051825

Häfner, M. F., & Debus, J. (2016). Radiotherapy for colorectal cancer: Current standards and future perspectives. *Visceral medicine*, *32*. 172–177. https://doi.org/10.1159/000446486

Hajnal, J. V., Saeed, N., Soar, E. J., Oatridge, A., Young, I. R., & Bydder, G. M. (1995). A registration and interpolation procedure for subvoxel matching of serially acquired mr images. *Journal of computer assisted tomography*, *19*. 289–296. https://doi.org/10.1097/00004728-199503000-00022

Hamerla, G., Meyer, H.-J., Hambsch, P., Wolf, U., Kuhnt, T., Hoffmann, K.-T., & Surov, A. (2019). Radiomics model based on non-contrast ct shows no predictive power for complete pathological response in locally advanced rectal cancer. *Cancers*, *11*. 1680. https://doi.org/10.3390/cancers11111680

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*. 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Harat, M., Małkowski, B., & Makarewicz, R. (2016). Pre-irradiation tumour volumes defined by mri and dual time-point fet-pet for the prediction of glioblastoma multiforme recurrence: A prospective study. *Radiotherapy and Oncology*, *120*. 241–247. https://doi.org/10.1016/j.radonc.2016.06.004

Harrell, F. E., et al. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. 608). Springer. https://doi.org/10.1007/978-1-4757-3462-1

Hartigan, J. A., Wong, M. A., et al. (1979). A k-means clustering algorithm. *Applied statistics*, *28*. 100–108. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. springer series in statistics. *New York, NY, USA*. https://link.springer.com/book/10.1007/978-0-387-84858-7

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://doi.org/10.1109/CVPR.2016.90

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., & Shen, D. (2022). Transformers in medical image analysis: A review. *Intelligent Medicine*. https://doi.org/10.1016/j.imed.2022.07.002

Herman, G. T. (2009). *Fundamentals of computerized tomography: Image reconstruction from projections*. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-1-84628-723-7

Horvat, N., Carlos Tavares Rocha, C., Clemente Oliveira, B., Petkovska, I., & Gollub, M. J. (2019). Mri of rectal cancer: Tumor staging, imaging techniques, and management. *Radiographics*, *39*. 367–387. https://doi.org/10.1148/rg.2019180114

Horvat, N., Veeraraghavan, H., Khan, M., Blazic, I., Zheng, J., Capanu, M., Sala, E., Garcia-Aguilar, J., Gollub, M. J., & Petkovska, I. (2018). Mr imaging of rectal cancer: Radiomics analysis to assess treatment response after neoadjuvant therapy. *Radiology*, *287*. 833–843. https://doi.org/10.1148/radiol.2018172300

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, *9*. 1043–1069. https://doi.org/10.1080/03610928008827941

Hothorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, *43*. 121–137. https://doi.org/10.1016/S0167-9473(02)00225-6

Hounsfield, G. N. (1980). Nobel award address. computed medical imaging. *Medical physics*, *7*. 283–290. https://doi.org/10.1118/1.594709

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Hutchings, M., Loft, A., Hansen, M., Pedersen, L. M., Buhl, T., Jurlander, J., Buus, S., Keiding, S., D'Amore, F., Boesen, A.-M., et al. (2006). Fdg-pet after two cycles of chemotherapy predicts treatment failure and progression-free survival in hodgkin lymphoma. *Blood*, *107*. 52–59. https://doi.org/10.1182/blood-2005-06-2252

Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Roß, T., Wirkert, S., et al. (2020). Batchgenerators—a python framework for data augmentation. *Zenodo*. https://github.com/MIC-DKFZ/batchgenerators

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. https://doi.org/10.1002/9781118445112.stat08188

Jager, P. L., Vaalburg, W., Pruim, J., de Vries, E. G., Langen, K.-J., & Piers, D. A. (2001). Radiolabeled amino acids: Basic aspects and clinical applications in oncology. *Journal of nuclear medicine*, *42*. 432–445. https://jnm.snmjournals.org/content/42/3/432.short

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc. https://dl.acm.org/doi/abs/10.5555/42779

Jeon, S. H., Song, C., Chie, E. K., Kim, B., Kim, Y. H., Chang, W., Lee, Y. J., Chung, J.-H., Chung, J. B., Lee, K.-W., et al. (2019). Delta-radiomics signature predicts treatment outcomes after preoperative chemoradiotherapy and surgery in rectal cancer. *Radiation oncology*, *14*. 1–10. https://doi.org/10.1186/s13014-019-1246-8

Ju, J. (2019). Keras-resnet3d [Accessed: 2021-09-30]. URL: https://github.com/JihongJu/keras-resnet3d#keras-resnet3d

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, *36*. 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kapiteijn, E., & van de Velde, C. J. (2000). European trials with total mesorectal excision. *Seminars in Surgical Oncology*, *19*. 350–357. https://doi.org/10.1002/ssu.5

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, *18*. 1–12. https://doi.org/10.1186/s12874-018-0482-1

Khorrami, M., Prasanna, P., Gupta, A., Patil, P., Velu, P. D., Thawani, R., Corredor, G., Alilou, M., Bera, K., Fu, P., et al. (2020). Changes in ct radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non–small cell lung cancertil-related radiomics predicts os and immunotherapy response. *Cancer immunology research*, *8*. 108–119. https://doi.org/10.1158/2326-6066.CIR-19-0476

Kickingereder, P., Bonekamp, D., Nowosielski, M., Kratz, A., Sill, M., Burth, S., Wick, A., Eidel, O., Schlemmer, H.-P., Radbruch, A., et al. (2016a). Radiogenomics of glioblastoma: Machine learning–based classification of molecular characteristics by using multiparametric and multiregional mr imaging features. *Radiology*, *281*. 907–918. https://doi.org/10.1148/radiol.2016161382

Kickingereder, P., Burth, S., Wick, A., Götz, M., Eidel, O., Schlemmer, H.-P., Maier-Hein, K. H., Wick, W., Bendszus, M., Radbruch, A., et al. (2016b). 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. https://doi.org/10.1148/radiol.2016160845

Kickingereder, P., Götz, M., Muschelli, J., Wick, A., Neuberger, U., Shinohara, R. T., Sill, M., Nowosielski, M., Schlemmer, H.-P., Radbruch, A., et al. (2016c). Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment responseradiomic profiling of bev efficacy in glioblastoma. *Clinical Cancer Research*, *22*. 5765–5771. https://doi.org/10.1158/1078-0432.CCR-16-0702

Kong, D.-S., Kim, J., Lee, I.-H., Kim, S. T., Seol, H. J., Lee, J.-I., Park, W.-Y., Ryu, G., Wang, Z., Ma'ayan, A., et al. (2016). Integrative radiogenomic analysis for multicentric radiophenotype in glioblastoma. *Oncotarget*, *7*. 11526. https://doi.org/10.18632/oncotarget.7115

Korfiatis, P., Kline, T. L., Coufalova, L., Lachance, D. H., Parney, I. F., Carter, R. E., Buckner, J. C., & Erickson, B. J. (2016). Mri texture features as biomarkers to predict mgmt methylation status in glioblastomas. *Medical physics*, *43*. 2835–2844. https://doi.org/10.1118/1.4948668

Krivoshapkin, A. L., Sergeev, G. S., Gaytan, A. S., Kalneus, L. E., Kurbatov, V. P., Abdullaev, O. A., Salim, N., Bulanov, D. V., & Simonovich, A. E. (2019). Automated volumetric analysis of postoperative magnetic resonance imaging predicts survival in patients with glioblastoma. *World Neurosurgery*, *126*. e1510–e1517. https://doi.org/10.1016/j.wneu.2019.03.142

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger ([Eds]), *Advances in neural information processing systems*. Curran Associates, Inc. https://doi.org/10.1145/3065386

Kubben, P. L., Postma, A. A., Kessels, A. G., van Overbeeke, J. J., & van Santbrink, H. (2010). Intraobserver and interobserver agreement in volumetric assessment of glioblastoma multiforme resection. *Neurosurgery*, *67*. 1329–1334. https://doi.org/10.1227/NEU.0b013e3181efbb08

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. n taylor & francis group. https://doi.org/10.1080/00031305.2020.1790217

Kumar, A. J., Leeds, N. E., Fuller, G. N., Van Tassel, P., Maor, M. H., Sawaya, R. E., & Levin, V. A. (2000). Malignant gliomas: Mr imaging spectrum of radiation therapy-and chemotherapy-induced necrosis of the brain after treatment. *Radiology*, *217*. 377–384. https://doi.org/10.1148/radiology.217.2.r00nv36377

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J., Dekker, A., Fenstermacher, D., et al. (2012). Radiomics: The process and the challenges. *Magnetic resonance imaging*, *30*. 1234–1248. https://doi.org/10.1016/j.mri.2012.06.010

Laigle-Donadey, F., & Delattre, J.-Y. (2006). Glioma in the elderly. *Current opinion in oncology*, *18*. 644–647. https://doi.org/10.1097/01.cco.0000245324.19411.19

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., & Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, *7*. 10353. https://doi.org/10.1038/s41598-017-10649-8

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*. 436–444. https://doi.org/10.1038/nature14539

Lee, M. H., Kim, J., Kim, S.-T., Shin, H.-M., You, H.-J., Choi, J. W., Seol, H. J., Nam, D.-H., Lee, J.-I., & Kong, D.-S. (2019). Prediction of idh1 mutation status in glioblastoma using machine learning technique based on quantitative radiomic data. *World neurosurgery*, *125*. e688–e696. https://doi.org/10.1016/j.wneu.2019.01.157

Lee, Y.-T. (1995). Local and regional recurrence of carcinoma of the colon and rectum: I. tumour-host factors and adjuvant therapy. *Surgical oncology*, *4*. 283–293. https://doi.org/10.1016/S0960-7404(10)80040-4

Leijenaar, R. T., Carvalho, S., Velazquez, E. R., Van Elmpt, W. J., Parmar, C., Hoekstra, O. S., Hoekstra, C. J., Boellaard, R., Dekker, A. L., Gillies, R. J., et al. (2013). Stability of fdg-pet radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta oncologica*, *52*. 1391–1397. https://doi.org/10.3109/0284186X.2013.812798

Li, G., Li, L., Li, Y., Qian, Z., Wu, F., He, Y., Jiang, H., Li, R., Wang, D., Zhai, Y., et al. (2022). An mri radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. *Brain*, *145*. 1151–1161. https://doi.org/10.1093/brain/awab340

Li, M., Li, J., Zhao, A., & Gu, J. (2007). Colorectal cancer or colon and rectal cancer? *Oncology*, *73*. 52–57. https://doi.org/10.1159/000120628

Li, M., Zhu, Y.-Z., Zhang, Y.-C., Yue, Y.-F., Yu, H.-P., & Song, B. (2020a). Radiomics of rectal cancer for predicting distant metastasis and overall survival. *World journal of gastroenterology*, *26*. 5008. https://doi.org/10.3748/wjg.v26.i33.5008

Li, Q., Bai, H., Chen, Y., Sun, Q., Liu, L., Zhou, S., Wang, G., Liang, C., & Li, Z.-C. (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. https://doi.org/10.1038/s41598-017-14753-7

Li, Z.-Y., Wang, X.-D., Li, M., Liu, X.-J., Ye, Z., Song, B., Yuan, F., Yuan, Y., Xia, C.-C., Zhang, X., et al. (2020b). Multi-modal radiomics model to predict treatment response to neoadjuvant chemotherapy for locally advanced rectal cancer. *World journal of gastroenterology*, *26*. 2388. https://doi.org/10.3748/wjg.v26.i19.2388

Li, Z., Ma, X., Shen, F., Lu, H., Xia, Y., & Lu, J. (2021). Evaluating treatment response to neoadjuvant chemoradiotherapy in rectal cancer using various mri-based radiomics models. *BMC Medical Imaging*, *21*. 1–10. https://doi.org/10.1186/s12880-021-00560-0

Liu, X., Zhang, D., Liu, Z., Li, Z., Xie, P., Sun, K., Wei, W., Dai, W., Tang, Z., Ding, Y., et al. (2021). Deep learning radiomics-based prediction of distant metastasis in patients with locally advanced rectal cancer after neoadjuvant chemoradiotherapy: A multicentre study. *EBioMedicine*, *69*. 103442. https://doi.org/10.1016/j.ebiom.2021.103442

Liu, Z., Zhang, X.-Y., Shi, Y.-J., Wang, L., Zhu, H.-T., Tang, Z., Wang, S., Li, X.-T., Tian, J., & Sun, Y.-S. (2017). Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Clinical Cancer Research*, *23*. 7253–7262. https://doi.org/10.1158/1078-0432.CCR-17-1038

Lohmann, P., Elahmadawy, M. A., Gutsche, R., Werner, J.-M., Bauer, E. K., Ceccon, G., Kocher, M., Lerche, C. W., Rapp, M., Fink, G. R., et al. (2020). Fet pet radiomics for differentiating pseudoprogression from early tumor progression in glioma patients post-chemoradiation. *Cancers*, *12*. 3835. https://doi.org/10.3390/cancers12123835

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., & Kleihues, P. (2007). The 2007 who classification of tumours of the central nervous system. *Acta neuropathologica*, *114*. 97–109. https://doi.org/10.1007/s00401-007-0243-4

Lutterbach, J., Sauerbrei, W., & Guttenberger, R. (2003). Multivariate analyse prognostischer faktoren bei patienten mit glioblastom. *Strahlentherapie und Onkologie*, *179*. 8–15. https://doi.org/10.1007/s00066-003-1004-5

Magee, J. F. (1964). *Decision trees for decision making*. Harvard Business Review Brighton, MA, USA. https://hbr.org/1964/07/decision-trees-for-decision-making

Mandard, A.-M., Dalibard, F., Mandard, J.-C., Marnay, J., Henry-Amar, M., Petiot, J.-F., Roussel, A., Jacob, J.-H., Segol, P., Samama, G., et al. (1994). Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. clinicopathologic correlations. *Cancer*, *73*. 2680–2686. https://doi.org/10.1002/1097-0142(19940601)73:11<2680::AID-CNCR2820731105>3.0.CO;2-C

Mangla, R., Singh, G., Ziegelitz, D., Milano, M. T., Korones, D. N., Zhong, J., & Ekholm, S. E. (2010). Changes in relative cerebral blood volume 1 month after radiation-temozolomide therapy can help predict overall survival in patients with glioblastoma. *Radiology*, *256*. 575–584. https://doi.org/10.1148/radiol.10091440

Matsuo, M., Miwa, K., Tanaka, O., Shinoda, J., Nishibori, H., Tsuge, Y., Yano, H., Iwama, T., Hayashi, S., Hoshi, H., et al. (2012). Impact of [11c] methionine positron emission tomography for target definition of glioblastoma multiforme in radiation therapy planning. *International Journal of Radiation Oncology* Biology* Physics*, *82*. 83–89. https://doi.org/10.1016/j.ijrobp.2010.09.020

McCarthy, K., Pearson, K., Fulton, R., & Hewitt, J. (2012). Pre-operative chemoradiation for non-metastatic locally advanced rectal cancer. *Cochrane database of systematic reviews*. https://doi.org/10.1002/14651858.CD008368.pub2

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967*. http://projecteuclid.org/euclid.bsmsp/1200512992

Meier, A., Nekolla, K., Hewitt, L. C., Earle, S., Yoshikawa, T., Oshima, T., Miyagi, Y., Huss, R., Schmidt, G., & Grabsch, H. I. (2020). Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. *The Journal of Pathology: Clinical Research*, *6*. 273–282. https://doi.org/10.1002/cjp2.170

Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., Slotboom, J., Wiest, R., & Reyes, M. (2017). Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma. *Journal of neurosurgery*, *127*. 798–806. https://doi.org/10.3171/2016.9.JNS16146

Meneghetti, A. R., Zwanenburg, A., Leger, S., Leger, K., Troost, E. G., Linge, A., Lohaus, F., Schreiber, A., Kalinauskaite, G., Tinhofer, I., et al. (2021). Definition and validation of a radiomics signature for loco-regional tumour control in patients with locally advanced head and neck squamous cell carcinoma. *Clinical and Translational Radiation Oncology*, *26*. 62–70. https://doi.org/10.1016/j.ctro.2020.11.011

Meng, Y., Zhang, Y., Dong, D., Li, C., Liang, X., Zhang, C., Wan, L., Zhao, X., Xu, K., Zhou, C., et al. (2018). Novel radiomic signature as a prognostic biomarker for locally advanced rectal cancer. *Journal of Magnetic Resonance Imaging*, *48*. 605–614. https://doi.org/10.1002/jmri.25968

Nazzaro, J. M., & Neuwelt, E. A. (1990). The role of surgery in the management of supratentorial intermediate and high-grade astrocytomas in adults. *Journal of neurosurgery*, *73*. 331–344. https://doi.org/10.3171/jns.1990.73.3.0331

Newman, N. B., Sidhu, M. K., Baby, R., Moss, R. A., Nissenblatt, M. J., Chen, T., Lu, S.-E., & Jabbour, S. K. (2016). Long-term bone marrow suppression during postoperative chemotherapy in rectal cancer patients after preoperative chemoradiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, *94*. 1052–1060. https://doi.org/10.1016/j.ijrobp.2015.12.374

Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., Liu, L., Wang, Q., Wu, J., & Shen, D. (2019). Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific reports*, *9*. 1103. https://doi.org/10.1038/s41598-018-37387-9

Nie, K., Shi, L., Chen, Q., Hu, X., Jabbour, S. K., Yue, N., Niu, T., & Sun, X. (2016). Rectal cancer: Assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric mriprediction of pathological response for larc using mri. *Clinical cancer research*, *22*. 5256–5264. https://doi.org/10.1158/1078-0432.CCR-15-2997

Nieder, C., Astner, S. T., Mehta, M. P., Grosu, A. L., & Molls, M. (2008). Improvement, clinical course, and quality of life after palliative radiotherapy for recurrent glioblastoma. *American journal of clinical oncology*, *31*. 300–305. https://doi.org/10.1097/COC.0b013e31815e3fdc

Ohgaki, H. (2009). Epidemiology of brain tumors. *Cancer Epidemiology: Modifiable Factors*. 323–342. https://doi.org/10.1007/978-1-60327-492-0_14

Osman, A. F. (2018). Automated brain tumor segmentation on magnetic resonance images and patient's overall survival prediction using support vector machines. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*. 435–449. https://doi.org/10.1007/978-3-319-75238-9_37

Ostrom, Q. T., Price, M., Neff, C., Cioffi, G., Waite, K. A., Kruchko, C., & Barnholtz-Sloan, J. S. (2022). Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2015–2019. *Neuro-oncology*, *24*. v1–v95. https://doi.org/10.1093/neuonc/noac202

Palanichamy, K., & Chakravarti, A. (2017). Diagnostic and prognostic significance of methionine uptake and methionine positron emission tomography imaging in gliomas. *Frontiers in oncology*, *7*. 257. https://doi.org/10.3389/fonc.2017.00257

Park, H., Kim, K. A., Jung, J.-H., Rhie, J., & Choi, S. Y. (2020). Mri features and texture analysis for the early prediction of therapeutic response to neoadjuvant chemoradiotherapy and tumor recurrence of locally advanced rectal cancer. *European radiology*, *30*. 4201–4211. https://doi.org/10.1007/s00330-020-06835-4

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. (2015). Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, *5*. 13087. https://doi.org/10.1038/srep13087

Partridge, S. C., Gibbs, J. E., Lu, Y., Esserman, L. J., Tripathy, D., Wolverton, D. S., Rugo, H. S., Hwang, E. S., Ewing, C. A., & Hylton, N. M. (2005). Mri measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival. *AJR Am J Roentgenol*, *184*. 1774–1781. https://doi.org/10.2214/ajr.184.6.01841774

Paschke, S., Jafarov, S., Staib, L., Kreuser, E.-D., Maulbecker-Armstrong, C., Roitman, M., Holm, T., Harris, C. C., Link, K.-H., & Kornmann, M. (2018). Are colon and rectal cancer two different tumor entities? a proposal to abandon the term colorectal cancer. *International journal of molecular sciences*, *19*. 2577. https://doi.org/10.3390/ijms19092577

Peng, H., Long, F., & 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. https://doi.org/10.1109/TPAMI.2005.159

Petkovska, I., Tixier, F., Ortiz, E. J., Golia Pernicka, J. S., Paroder, V., Bates, D. D., Horvat, N., Fuqua, J., Schilsky, J., Gollub, M. J., et al. (2020). Clinical utility of radiomics at baseline rectal mri to predict complete response of rectal cancer after chemoradiation therapy. *Abdominal Radiology*, *45*. 3608–3617. https://doi.org/10.1007/s00261-020-02502-w

Petresc, B., Lebovici, A., Caraiani, C., Feier, D. S., Graur, F., & Buruian, M. M. (2020). Pre-treatment t2-wi based radiomics features for prediction of locally advanced rectal cancer non-response to neoadjuvant chemoradiotherapy: A preliminary study. *Cancers*, *12*. 1894. https://doi.org/10.3390/cancers12071894

Qi, J., & Leahy, R. M. (2006). Iterative reconstruction techniques in emission computed tomography. *Physics in Medicine & Biology*, *51*. R541. https://doi.org/10.1088/0031-9155/51/15/R01

Radon, J. (1986). On the determination of functions from their integral values along certain manifolds. *IEEE transactions on medical imaging*, *5*. 170–176. https://doi.org/10.1109/TMI.1986.4307775

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., et al. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, *15*. e1002686. https://doi.org/10.1371/journal.pmed.1002686

Rao, S.-X., Lambregts, D. M., Schnerr, R. S., Beckers, R. C., Maas, M., Albarello, F., Riedl, R. G., Dejong, C. H., Martens, M. H., Heijnen, L. A., et al. (2016). Ct texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? *United European gastroenterology journal*, *4*. 257–263. https://doi.org/10.1177/2050640615601603

Rees, J. H., Smirniotopoulos, J. G., Jones, R. V., & Wong, K. (1996). Glioblastoma multiforme: Radiologic-pathologic correlation. *Radiographics*, *16*. 1413–1438. https://doi.org/10.1148/radiographics.16.6.8946545

Renehan, A. G., Malcomson, L., Emsley, R., Gollins, S., Maw, A., Myint, A. S., Rooney, P. S., Susnerwala, S., Blower, A., Saunders, M. P., et al. (2016). Watch-and-wait approach versus surgical resection after chemoradiotherapy for patients with rectal cancer (the oncore project): A propensity-score matched cohort analysis. *The Lancet Oncology*, *17*. 174–183. https://doi.org/10.1016/S1470-2045(15)00467-2

Rich, J. N., Hans, C., Jones, B., Iversen, E. S., McLendon, R. E., Rasheed, B. A., Dobra, A., Dressman, H. K., Bigner, D. D., Nevins, J. R., et al. (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer research*, *65*. 4051–4058. https://doi.org/10.1158/0008-5472.CAN-04-3936

Rimkus, C., Friederichs, J., Boulesteix, A.-.-l., Theisen, J., Mages, J., Becker, K., Nekarda, H., Rosenberg, R., Janssen, K.-.-P., & Siewert, J. R. (2008). Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. *Clinical gastroenterology and hepatology*, *6*. 53–61. https://doi.org/10.1016/j.cgh.2007.10.022

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American statistician*. 59–66. https://doi.org/10.2307/2685263

Rozsa, A., Gunther, M., & Boult, T. E. (2016). Towards robust deep neural networks with bang. *arXiv preprint arXiv:1612.00138*. https://doi.org/10.1109/WACV.2018.00093

Ryan, R., Gibbons, D., Hyland, J., Treanor, D., White, A., Mulcahy, H., O'Donoghue, D., Moriarty, M., Fennelly, D., & Sheahan, K. (2005). Pathological response following long-course neoadjuvant chemoradiotherapy for locally advanced rectal cancer. *Histopathology*, *47*. 141–146. https://doi.org/10.1111/j.1365-2559.2005.02176.x

Sanai, N., & Berger, M. S. (2009). Operative techniques for gliomas and the value of extent of resection. *Neurotherapeutics*, *6*. 478–486. https://doi.org/10.1016/j.nurt.2009.04.005

Sanai, N., Eschbacher, J., Hattendorf, G., Coons, S. W., Preul, M. C., Smith, K. A., Nakaji, P., & Spetzler, R. F. (2011). Intraoperative confocal microscopy for brain tumors: A feasibility analysis in humans. *Operative Neurosurgery*, *68*. ons282–ons290. https://doi.org/10.1227/NEU.0b013e318212464e

Sauer, R., Becker, H., Hohenberger, W., Rödel, C., Wittekind, C., Fietkau, R., Martus, P., Tschmelitsch, J., Hager, E., Hess, C. F., et al. (2004). Preoperative versus postoperative chemoradiotherapy for rectal cancer. *New England Journal of Medicine*, *351*. 1731–1740. https://doi.org/10.1056/NEJMoa040694

Scharstein, D., & Pal, C. (2007). Learning conditional random fields for stereo. *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. https://doi.org/10.1007/s11263-010-0385-z

Seeram, E. (2015). *Computed tomography-e-book: Physical principles, clinical applications, and quality control*. Elsevier Health Sciences. https://www.elsevier.com/books/computed-tomography/seeram/978-0-323-79063-5

Seidlitz, A., Beuthien-Baumann, B., Löck, S., Jentsch, C., Platzek, I., Zöphel, K., Linge, A., Kotzerke, J., Petr, J., van den Hoff, J., et al. (2021). Final results of the prospective biomarker trial petra:[11c]-met-accumulation in postoperative pet/mri predicts outcome after radiochemotherapy in glioblastomabiomarker trial: Met-pet predicts outcome after rctx in glioblastoma. *Clinical Cancer Research*, *27*. 1351–1360. https://doi.org/10.1158/1078-0432.CCR-20-1775

Shahzadi, I., Lattermann, A., Linge, A., Zwanenburg, A., Baldus, C., Peeken, J. C., Combs, S. E., Baumann, M., Krause, M., Troost, E. G., et al. (2021). Do we need complex image features to personalize treatment of patients with locally advanced rectal cancer? *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. 775–785. https://doi.org/10.1007/978-3-030-87234-2_73

Shahzadi, I., Seidlitz, A., Zwanenburg, A., Beuthien-Baumann, B., Platzek, I., Kotzerke, J., Baumann, M., Krause, M., & Löck, S. (2022a). 3d convolutional neural networks for outcome prediction in glioblastoma using methionine pet and t1w mri. *Medical Imaging with Deep Learning*. https://openreview.net/forum?id=BLXlChVgVb5

Shahzadi, I., Zwanenburg, A., Lattermann, A., Linge, A., Baldus, C., Peeken, J. C., Combs, S. E., Diefenhardt, M., Rödel, C., Kirste, S., et al. (2022b). Analysis of mri and ct-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. *Scientific Reports*, *12*. 10192. https://doi.org/10.1038/s41598-022-13967-8

Shaish, H., Aukerman, A., Vanguri, R., Spinelli, A., Armenta, P., Jambawalikar, S., Makkar, J., Bentley-Hibbert, S., Del Portillo, A., Kiran, R., et al. (2020). Radiomics of mri for pretreatment prediction of pathologic complete response, tumor regression grade, and neoadjuvant rectal score in patients with locally advanced rectal cancer undergoing neoadjuvant chemoradiation: An international multicenter study. *European radiology*, *30*. 6263–6273. https://doi.org/10.1007/s00330-020-06968-6

Shi, L., Rong, Y., Daly, M., Dyer, B., Benedict, S., Qiu, J., & Yamamoto, T. (2020). Cone-beam computed tomography-based delta-radiomics for early response assessment in radiotherapy for locally advanced lung cancer. *Physics in Medicine & Biology*, *65*. 015009. https://doi.org/10.1088/1361-6560/ab3247

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., & Michalski, M. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. 1–11. https://doi.org/10.1007/978-3-030-00536-8_1

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, *86*. 420. https://doi.org/10.1037/0033-2909.86.2.420

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3d deep learning on medical images: A review. *Sensors*, *20*. 5097. https://doi.org/10.3390/s20185097

Sizoo, E. M., Braam, L., Postma, T. J., Pasman, H. R. W., Heimans, J. J., Klein, M., Reijneveld, J. C., & Taphoorn, M. J. (2010). Symptoms and problems in the end-of-life phase of high-grade glioma patients. *Neuro-oncology*, *12*. 1162–1166. https://doi.org/10.1093/neuonc/nop045

Sobin, L. (2009). International union against cancer (uicc) tnm classification of malignant tumours. *Oesophagus including Oesophagogastric Junction*. 66–72. https://doi.org/10.1002/(sici)1097-0142(19971101)80:9<1803::aid-cncr16>3.0.co;2-9

Soerjomataram, I., & Bray, F. (2021). Planning for tomorrow: Global cancer incidence and the role of prevention 2020–2070. *Nature reviews Clinical oncology*, *18*. 663–672. https://doi.org/10.1038/s41571-021-00514-z

Song, J., Yin, Y., Wang, H., Chang, Z., Liu, Z., & Cui, L. (2020). A review of original articles published in the emerging field of radiomics. *European journal of radiology*, *127*. 108991. https://doi.org/10.1016/j.ejrad.2020.108991

Spearman, C. (1910). Correlation calculated from faulty data. *British journal of psychology*, *3*. 271. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Staal, F. C., Van Der Reijd, D. J., Taghavi, M., Lambregts, D. M., Beets-Tan, R. G., & Maas, M. (2021). Radiomics for the prediction of treatment outcome and survival in patients with colorectal cancer: A systematic review. *Clinical Colorectal Cancer*, *20*. 52–71. https://doi.org/10.1016/j.clcc.2020.11.001

Starke, S., Leger, S., Zwanenburg, A., Leger, K., Lohaus, F., Linge, A., Schreiber, A., Kalinauskaite, G., Tinhofer, I., Guberina, N., et al. (2020). 2d and 3d convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Scientific reports*, *10*. 1–13. https://doi.org/10.1038/s41598-020-70542-9

Stupp, R., Hegi, M. E., Mason, W. P., Van Den Bent, M. J., Taphoorn, M. J., Janzer, R. C., Ludwin, S. K., Allgeier, A., Fisher, B., Belanger, K., et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-ncic trial. *The lancet oncology*, *10*. 459–466. https://doi.org/10.1016/S1470-2045(09)70025-7

Stupp, R., Mason, W. P., Van Den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., Belanger, K., Brandes, A. A., Marosi, C., Bogdahn, U., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England journal of medicine*, *352*. 987–996. https://doi.org/10.1097/01.COT.0000289242.47980.f9

Sun, C., & Wee, W. G. (1983). Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, *23*. 341–352. https://doi.org/10.1016/0734-189X(83)90032-4

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tang, Z., Xu, Y., Jin, L., Aibaidula, A., Lu, J., Jiao, Z., Wu, J., Zhang, H., & Shen, D. (2020). Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. *IEEE transactions on medical imaging*, *39*. 2100–2109. https://doi.org/10.1109/TMI.2020.2964310

Tewarie, I. A., Senders, J. T., Kremer, S., Devi, S., Gormley, W. B., Arnaout, O., Smith, T. R., & Broekman, M. L. (2021). Survival prediction of glioblastoma patients—are we there yet? a systematic review of prognostic modeling for glioblastoma and its clinical potential. *Neurosurgical review*, *44*. 2047–2057. https://doi.org/10.1007/s10143-020-01430-z

Thibault, G., Angulo, J., & Meyer, F. (2013). Advanced statistical matrices for texture characterization: Application to cell classification. *IEEE Transactions on Biomedical Engineering*, *61*. 630–637. https://doi.org/10.1109/TBME.2013.2284600

Thies, S., & Langer, R. (2013). Tumor regression grading of gastrointestinal carcinomas after neoadjuvant treatment. *Frontiers in oncology*, *3*. 262. https://doi.org/10.3389/fonc.2013.00262

Titterington, D. M., Afm, S., Smith, A. F., Makov, U., et al. (1985). *Statistical analysis of finite mixture distributions* (Vol. 198). John Wiley & Sons Incorporated. https://doi.org/doi.org/10.2307/2981482

Tixier, F., Hatt, M., Le Rest, C. C., Le Pogam, A., Corcos, L., & Visvikis, D. (2012). Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18f-fdg pet. *Journal of Nuclear Medicine*, *53*. 693–700. https://doi.org/10.2967/jnumed.111.099127

Traverso, A., Wee, L., Dekker, A., & Gillies, R. (2018). Repeatability and reproducibility of radiomic features: A systematic review. *International Journal of Radiation Oncology* Biology* Physics*, *102*. 1143–1158. https://doi.org/10.1016/j.ijrobp.2018.05.053

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE transactions on medical imaging*, *29*. 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Van Helden, E., Vacher, Y., Van Wieringen, W., Van Velden, F., Verheul, H., Hoekstra, O., Boellaard, R., & Menke-van der Houven van Oordt, C. (2018). Radiomics analysis of pretreatment [18 f] fdg pet/ct for patients with metastatic colorectal cancer undergoing palliative systemic treatment. *European Journal of Nuclear Medicine and Molecular Imaging*, *45*. 2307–2317. https://doi.org/10.1007/s00259-018-4100-6

van Griethuysen, J. J., Lambregts, D. M., Trebeschi, S., Lahaye, M. J., Bakers, F. C., Vliegen, R. F., Beets, G. L., Aerts, H. J., & Beets-Tan, R. G. (2020). Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging mri in rectal cancer. *Abdominal Radiology*, *45*. 632–643. https://doi.org/10.1007/s00261-019-02321-8

van Velden, F. H., Kramer, G. M., Frings, V., Nissen, I. A., Mulder, E. R., de Langen, A. J., Hoekstra, O. S., Smit, E. F., & Boellaard, R. (2016). Repeatability of radiomic features in non-small-cell lung cancer [18 f] fdg-pet/ct studies: Impact of reconstruction and delineation. *Molecular imaging and biology*, *18*. 788–795. https://doi.org/10.1007/s11307-016-0940-2

Visser, M., Müller, D., van Duijn, R., Smits, M., Verburg, N., Hendriks, E., Nabuurs, R., Bot, J., Eijgelaar, R., Witte, M., et al. (2019). Inter-rater agreement in glioma segmentations on longitudinal mri. *NeuroImage: Clinical*, *22*. 101727. https://doi.org/10.1016/j.nicl.2019.101727

Wang, G., Li, W., Ourselin, S., & Vercauteren, T. (2018a). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*. 178–190. https://doi.org/10.1007/978-3-319-75238-9_16

Wang, L., Liang, B., Li, Y. I., Liu, X., Huang, J., & Li, Y. M. (2019). What is the advance of extent of resection in glioblastoma surgical treatment—a systematic review. *Chinese Neurosurgical Journal*, *5*. 1–6. https://doi.org/10.1186/s41016-018-0150-7

Wang, Y., Rapalino, O., Heidari, P., Loeffler, J., Shih, H. A., Oh, K., & Mahmood, U. (2018b). C11 methionine pet (met-pet) imaging of glioblastoma for detecting postoperative residual disease and response to chemoradiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, *102*. 1024–1028. https://doi.org/10.1016/j.ijrobp.2018.06.011

Wei, E. K., Giovannucci, E., Wu, K., Rosner, B., Fuchs, C. S., Willett, W. C., & Colditz, G. A. (2004). Comparison of risk factors for colon and rectal cancer. *International journal of cancer*, *108*. 433–442. https://doi.org/10.1002/ijc.11540

Weidman, S. (2019). *Deep learning from scratch: Building with python from first principles*. O'Reilly Media. https://www.oreilly.com/library/view/deep-learning-from/9781492041405/

Wen, P. Y., & Kesari, S. (2008). Malignant gliomas in adults. *New England Journal of Medicine*, *359*. 492–507. https://doi.org/10.1056/NEJMra0708126

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., DeGroot, J., Wick, W., Gilbert, M. R., Lassman, A. B., et al. (2010). Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *Journal of clinical oncology*, *28*. 1963–1972. https://doi.org/10.1200/JCO.2009.26.3541

Wibmer, A., Hricak, H., Gondo, T., Matsumoto, K., Veeraraghavan, H., Fehr, D., Zheng, J., Goldman, D., Moskowitz, C., Fine, S. W., 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. https://doi.org/10.1007/s00330-015-3701-8

Woods, R. P., Mazziotta, J. C., Cherry, S. R., et al. (1993). Mri-pet registration with automated algorithm. *Journal of computer assisted tomography*, *17*. 536–546. https://doi.org/10.1097/00004728-199307000-00004

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., et al. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, *39*. 1184–1194. https://doi.org/10.1109/TMI.2019.2945514

Xi, Y., & Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational oncology*, *14*. 101174. https://doi.org/10.1016/j.tranon.2021.101174

Yamanaka, R., Arao, T., Yajima, N., Tsuchiya, N., Homma, J., Tanaka, R., Sano, M., Oide, A., Sekijima, M., & Nishio, K. (2006). Identification of expressed genes characterizing long-term survival in malignant glioma patients. *Oncogene*, *25*. 5994–6002. https://doi.org/10.1038/sj.onc.1209585

Yang, D., Rao, G., Martinez, J., Veeraraghavan, A., & Rao, A. (2015). Evaluation of tumor-derived mri-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Medical physics*, *42*. 6725–6735. https://doi.org/10.1118/1.4934373

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*. 954–959. https://doi.org/10.1093/biomet/87.4.954

Yi, X., Pei, Q., Zhang, Y., Zhu, H., Wang, Z., Chen, C., Li, Q., Long, X., Tan, F., Zhou, Z., et al. (2019). Mri-based radiomics predicts tumor response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Frontiers in oncology*, *9*. 552. https://doi.org/10.3389/fonc.2019.00552

Yuan, Z., Frazer, M., Zhang, G. G., Latifi, K., Moros, E. G., Feygelman, V., Felder, S., Sanchez, J., Dessureault, S., Imanirad, I., et al. (2020). Ct-based radiomic features to predict pathological response in rectal cancer: A retrospective cohort study. *Journal of Medical Imaging and Radiation Oncology*, *64*. 444–449. https://doi.org/10.1111/1754-9485.13044

Zeng, K., Bakas, S., Sotiras, A., Akbari, H., Rozycki, M., Rathore, S., Pati, S., & Davatzikos, C. (2016). Segmentation of gliomas in pre-operative and post-operative multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2*. 184–194. https://doi.org/10.1007/978-3-319-55524-9_18

Zhang, C., Benz, P., Argaw, D. M., Lee, S., Kim, J., Rameau, F., Bazin, J.-C., & Kweon, I. S. (2021). Resnet or densenet? introducing dense shortcuts to resnet. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3550–3559. https://doi.org/10.1109/WACV48630.2021.00359

Zhang, Y., He, K., Guo, Y., Liu, X., Yang, Q., Zhang, C., Xie, Y., Mu, S., Guo, Y., Fu, Y., et al. (2020). A novel multimodal radiomics model for preoperative prediction of lymphovascular invasion in rectal cancer. *Frontiers in oncology*, *10*. 457. https://doi.org/10.3389/fonc.2020.00457

Zhou, G., Li, M. H., Tudor, G., Lu, H. T., Kadirvel, R., & Kallmes, D. (2018). Remote ischemic conditioning in cerebral diseases and neurointerventional procedures: Recent research progress. *Frontiers in neurology*, *9*. 339. https://doi.org/10.3389/fneur.2018.00339

Zhou, X., Yi, Y., Liu, Z., Cao, W., Lai, B., Sun, K., Li, L., Zhou, Z., Feng, Y., & Tian, J. (2019). Radiomics-based pretherapeutic prediction of non-response to neoadjuvant therapy in locally advanced rectal cancer. *Annals of surgical oncology*, *26*. 1676–1684. https://doi.org/10.1245/s10434-019-07300-3

Zhuang, Z., Liu, Z., Li, J., Wang, X., Xie, P., Xiong, F., Hu, J., Meng, X., Huang, M., Deng, Y., et al. (2021). Radiomic signature of the fowarc trial predicts pathological response to neoadjuvant treatment in rectal cancer. *Journal of Translational Medicine*, *19*. 256. https://doi.org/10.1186/s12967-021-02919-x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*. 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x

Zwanenburg, A., Leger, S., Agolli, L., Pilz, K., Troost, E. G., Richter, C., & Löck, S. (2019a). Assessing robustness of radiomic features by image perturbation. *Scientific reports*, *9*. 614. https://doi.org/10.1038/s41598-018-36938-4

Zwanenburg, A., Leger, S., & Starke, S. (2019b). Medical image radiomics processor [Accessed: 2021-07-30]. URL: https://github.com/oncoray/mirp

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*. 328–338. https://doi.org/10.1148/radiol.2020191145

Appendix

A MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer

Table A.1: Tumour regression grade (TRG) system following Dowark et al (Dworak et al., 1997).

Grade	Description
Complete regression (TRG 4)	No tumour cells
Near complete regression (TRG 3)	Very few tumour cells
Moderate regression (TRG 2)	Dominantly fibrotic changes with few tumour cells or groups
Minimal regression (TRG 1)	Dominant tumour mass with obvious fibrosis
No regression (TRG 0)	No regression

Table A.2: Image acquisition parameters of diagnostic T2-w MRI and treatment planning CT data.

MRI			CT		
Imaging parameters	**Training (122)**	**Validation (68)**	**Imaging parameters**	**Training (122)**	**Validation (68)**
Voxel spacing / mm			**Voxel spacing / mm**	119/2/0/1	57/3/8
1.0/0.9/0.8/0.7/0.6/0.5/0.4/0.3	1/5/33/50/27/4/1/1	4/1/15/17/17/9/1/4	1/0.9/0.8/0.7		
Slice thickness / mm			**Slice thickness / mm**		
7/6/5/4/3	0/4/53/35/30	1/5/8/3/51	3/5	70/52	56/12
Flip angle / °			**Reconstruction kernel**		
180/160/150/141-147/132-137/120-127/90	69/1/22/2/4/6/18	4/15/36/1/2/2/8	B/B20f/B30f/B31s/B40s/B41s/59.10.AB50/unknown	0/1/0/23/86/0/3/9	56/0/1/0/0/11/0/0
Scanning sequence			**Exposure time / ms**		
SE/RM	96/26	68/0	500/800/1000/1200/1500/unknown	1/3/86/0/23/9	1/0/0/1/11/55
Field strength / T			**Tube voltage / kV**	4/109/0/9	56/11/1/0
1.5/3	103/19	51/17	120/130/140/unknown		
Manufacturer			**Manufacturer**		
GE/PHILIPS/SIEMENS/TOSHIBA	3/13/104/2	8/0/60/0	SIEMENS/ PHILIPS/ MDS/Nordion/BrainLAB	113/0/7/2	12/56/0/0

Table A.3: Image pre-processing parameters for both CT and MRI data. All calculations were performed in 3D volume. Detailed configuration settings used in MIRP for MRI and CT are available at the GitHub repo. https://github.com/oncoray/radiomics-rectal_cancer

Parameters	**MRI**	**CT**
Pre-interpolation filter	N4 bias correction	None
Intensity normalization	95th percentile	None
Interpolated isotropic voxel spacing (mm)	1	1
Image interpolation method	linear	linear
ROI interpolation method	linear	linear
Re-segmentation range	None	[-150,180]
Merge method for texture matrices	volume merge	volume merge
Discretisation method: fixed bin number (bins)	32	32

Table A.4: LoG transformed intensity features selected on the training data after clustering (left) extracted from MRI and (right) extracted from CT images. Feature definitions can be found in the IBSI reference manual (Zwanenburg et al., 2020).

MR LoG features	**CT LoG features**
log_loc_peak_loc	log_stat_range
log_loc_peak_glob	log_stat_median
log_stat_rms	log_stat_rmad
log_stat_var	log_ivh_i90
log_stat_skew	log_ih_kurt_fbn_n32
log_ih_rmad_fbn_n32	log_stat_p10
log_stat_min	log_stat_p90
log_stat_p90	log_stat_max
log_stat_max	log_stat_energy
log_stat_cov	log_ivh_v25
log_ih_max_grad_fbn_n32	log_ivh_v75
log_ivh_v25	log_ivh_diff_v25_v75
log_ivh_v50	log_ih_iqr_fbn_n32
log_ivh_i25	log_ih_qcod_fbn_n32
	log_ih_max_grad_fbn_n32

Table A.5: Example of average model performance computation in internal training and validation. CV: cross-validation, AUC: area under a curve, LoG: Laplacian of Gaussian.

Modality	**Feature level**	**CV training**	**CV validation**	**Signature**	**Final training**	**External validation**
MRI	**LoG**	0.70	0.57	MR_log_ih_max_grad_fbn_n32 MR_log_stat_min	0.67 (0.57-0.75)	0.66 (0.51-0.82)
CT	**LoG**	0.73	0.64	CT_log_ih_max_grad_fbn_n32	0.70 (0.60-0.79)	0.61 (0.44-0.76)

Table A.6: Univariable analysis of tumour response (logistic regression) and freedom from distant metastases (FFDM, Cox regression) in the training data. ci: confidence interval. Significant p-values are marked in bold.

Clinical feature		**Tumour response**		**FFDM**	
		Odds ratio (95% ci)	**p-value**	**Hazard ratio (95% ci)**	**p-value**
Age / years		1.00 (0.97-1.03)	0.92	1.00 (0.96-1.04)	0.98
Gender (female vs. male)		1.54 (0.69-3.38)	0.29	1.30 (0.52-3.24)	0.57
UICC stage (3 vs 2)		2.85 (0.46-54.78)	0.34	0.45 (0.10-1.95)	0.29
Grade	(1 vs 0)	1.00 (0.04-14.01)	1.00	*	
	(2 vs 0)	1.92 (0.44-13.38)	0.43	1.58 (0.20-12.34)	0.66
	(3 vs 0)	2.00 (0.42-14.64)	0.42	2.45 (0.31-19.56)	0.40
Localization	(1 vs 0)	1.11 (0.50-2.46)	0.79	0.82 (0.32-2.12)	0.68
	(2 vs 0)	4.84 (0.4-58.06)	0.21	1.89 (0.24-14.63)	0.54
cT	(3 vs 2)	0.24 (0.03-1.31)	0.11	*	
	(4 vs 2)	0.06 (0.004-0.51)	**0.02**	*	
cN (1,2 vs 0)		*		0.47 (0.11-2.04)	0.31
Dose / Gy		0.87 (0.74-1.02)	0.09	0.92 (0.76-1.12)	0.42
Chemotherapy	(2 vs 1)	0.91 (0.22-3.82)	0.90	1.061 (0.24-4.65)	0.94
	(3 vs 1)	0.35 (0.04-3.14)	0.35	*	
	(4 vs 1)	1.28 (0.28-5.78)	0.75	0.86 (0.11-6.52)	0.89
Chemotherapy: 1 = 5 fluorouracil (FU), 2=5FU+oxaliplatine, 3= capecitabine (CAP) , 4=CAP+other Localization (cm): 0 = 3-6, 1= >6-12, 2= >12-16 *: The model did not converge.					

Table A.7: Summary of 3 selected SOT signatures from CT, MRI, and CT+MRI for the FFDM prediction. GLSZM: grey level size zone matrix, NGLDM: neighbouring grey level dependence matrix, GLCM: grey level co-occurrence matrix. 3d_fbn_n32: Features computed from discretized image intensities with fixed bin number 32 from 3D volume. d1: Chebyshev distance=1 around a central voxel for determining neighbourhood in NGLDM and GLCM based features, a0.0: alpha level 0.0 for NGLDM based features.

Signature	**Features**	**Identifier**	**Texture feature type**	**Definition**
CT_SOT	szm_zsnu_3d_fbn_n32	4JP3	GLSZM	This feature assesses the distribution of zone counts over the different zone sizes. Zone size non-uniformity is low when zone counts are equally distributed along zone sizes
MRI_SOT	ngl_dc_var_d1_a0.0_3d_fbn_n32	DNX2	NGLDM	This feature estimates the variance in dependence counts over the different possible dependence counts
	szm_sze_3d_fbn_n32	5QRC	GLSZM	This feature emphasises small zones.
	cm_clust_prom_d1_3d_v_mrg_fbn_n32	AE86	GLCM	This feature describes cluster prominence
CT_SOT + MRI_SOT	CT_ szm_zsnu_3d_fbn_n32	As above		
	MR_ ngl_dc_var_d1_a0.0_3d_fbn_n32			
	MR_ szm_sze_3d_fbn_n32			
	MR_ cm_clust_prom_d1_3d_v_mrg_fbn_n32			

Table A.8: Final models for the prognosis of tumour response and FFDM. Training was performed on the entire training data using multivariable logistic regression for tumour response and Cox regression for freedom from distant metastases. In addition, transformation parameters from the Yeo-Johnson transformation and z-normalization, and optimal cutoff values for Youden index and Kaplan-Meier estimates are given. ci = confidence interval. The R models for prospective use are available on GitHub: https://github.com/oncoray/radiomics-rectal_cancer

Tumour response						
Model	**Feature**	**Coefficient**	**p-value**	**Yeo-Johnson (λ)**	**z-score normalization (mean, sigma)**	**Cutoff**
Clinical+ MRI_LoG+ CT_LoG	MRI_log_stat_min	0.4282	0.027	3.3	(-0.35, 0.054)	0.248
	CT_log_ih_max_grad_fbn_n32	- 0.3088	0.004	0.0	(7.75, 1.12)	
	cT (3 vs 2)	-0.4666	0.017	-	-	
	cT (4 vs 2)	-1.2655				
	Intercept	-0.4337		-	-	
Clinical+ MRI_LoG	MRI_log_stat_min	0.3770	0.027	3.3	(-0.35, 0.054)	0.258
	MR_log_ih_max_grad_fbn_n32	-0.3221	0.008	0.1	(10.33, 2.311)	
	cT (3 vs 2)	-0.5255	0.017	-	-	
	cT (3 vs 4)	-1.3132				
	Intercept	-0.3742		-	-	
Clincial+ CT_LoG	CT_log_ih_max_grad_fbn_n32	-0.4422	0.004	0.0	(7.75, 1.12)	0.321
	cT (3 vs 2)	-0.5802	0.017	-	-	
	cT (3 vs 4)	-1.4047				
	Intercept	-0.3032		-	-	
Freedom from distant metastases						
CT_SOT+ MRI_SOT	MR_ ngl_dc_var_d1_a0.0_3d_fbn_n32	-0.4945	0.071	0.6	(7.85,1.81)	2.249
	MR_ szm_sze_3d_fbn_n32	-0.5044	0.192	10.0	(5.05,2.12)	
	MR_cm_clust_prom_d1_3d_v_mrg_fbn_n32	0.3013	0.176	-0.2	(4.27,0.10)	
	CT_szm_zsnu_3d_fbn_n32	-0.4584	0.046	0.3	(24.78,6.82)	
MRI_SOT	MR_ ngl_dc_var_d1_a0.0_3d_fbn_n32	-0.6337	0.071	0.6	(7.85,1.81)	2.251
	MR_ szm_sze_3d_fbn_n32	-0.4769	0.192	10.0	(5.05,2.12)	
	MR_cm_clust_prom_d1_3d_v_mrg_fbn_n32	0.2189	0.176	-0.2	(4.27,0.10)	
CT_SOT	CT_szm_zsnu_3d_fbn_n32	-0.4790	0.046	0.3	(24.78,6.82)	1.663

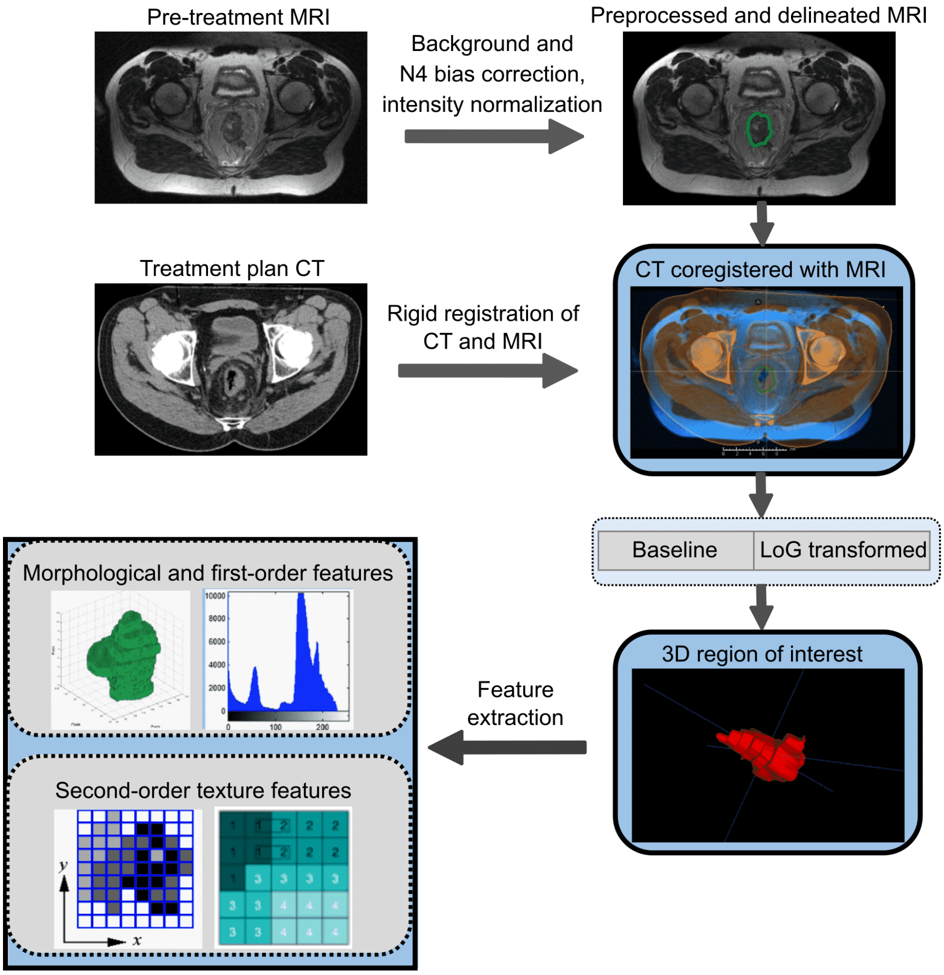
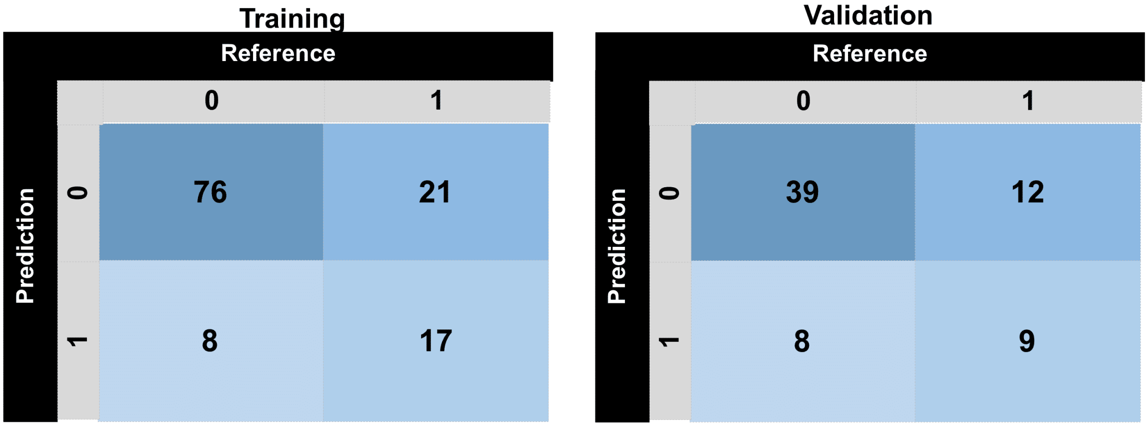


Figure A.1: Image pre-processing and feature extraction pipeline. Magnetic resonance (MR) images were pre-processed, and the gross tumour volume (GTV) was delineated centrally by one experienced radiation oncologist and one radiologist. GTV contours were then transferred to treatment planning computed tomography (CT) after rigid registration. All features were extracted from the GTV on the original and the Laplacian of Gaussian (LoG) transformed CT and MR images using a 3D approach.



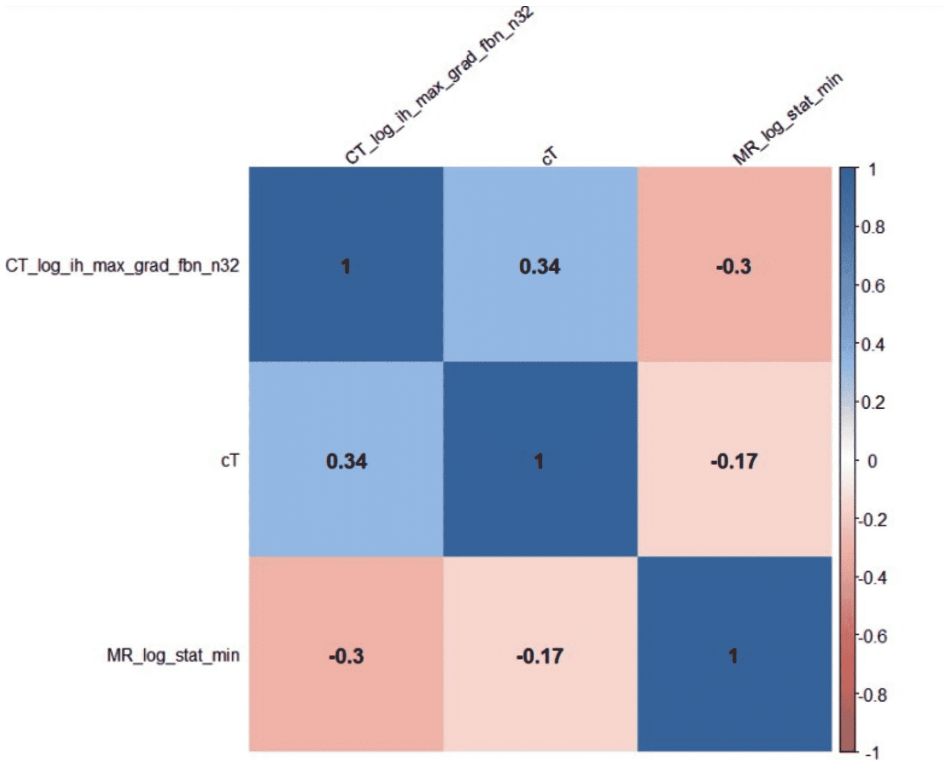
Training

Prediction \ Reference	0	1
0	76	21
1	8	17

Validation

Prediction \ Reference	0	1
0	39	12
1	8	9

Figure A.2: Confusion matrix for the prediction of tumour response to nCRT in LARC patients for the training and validation dataset at an optimal threshold of 0.42 combining clinical T stage and LoG features from MRI and CT.



	CT_log_ih_max_grad_fbn_n32	cT	MR_log_stat_min
CT_log_ih_max_grad_fbn_n32	1	0.34	-0.3
cT	0.34	1	-0.17
MR_log_stat_min	-0.3	-0.17	1

Figure A.3: Correlation plot of finally selected features in the best performing clinical-radiomic signature for prediction of tumour response to nCRT. Selected features were independent predictors, as shown by their low correlations, ρ<0.5.

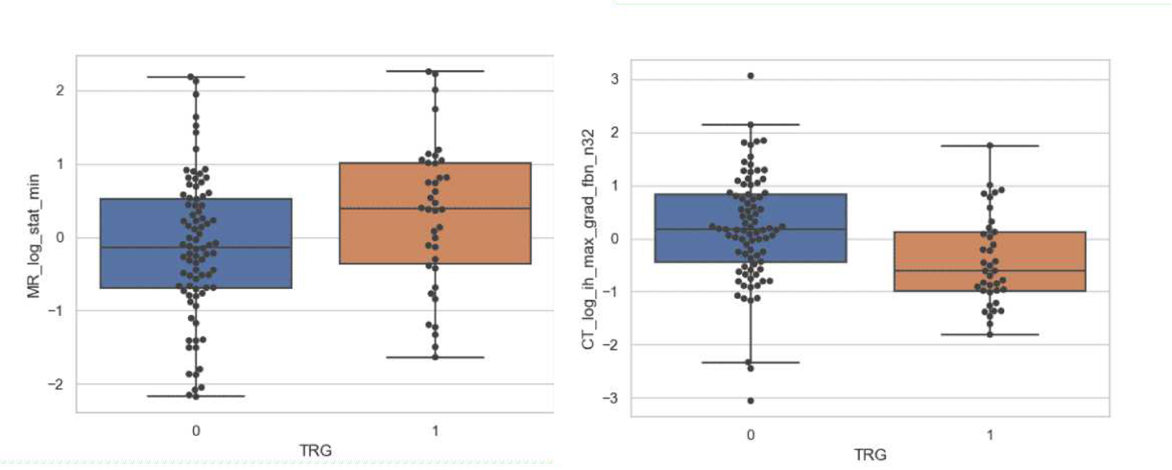


Figure A.4: Box plot of Yeo-Johnson transformed, and z-score normalized features selected in best performing joint CT and MRI model in training data. MRI_log_stat_min showed relatively higher values, while CT_log_ih_max_grad_fbn_n32 showed relatively lower values in responders as compared to non-responders.

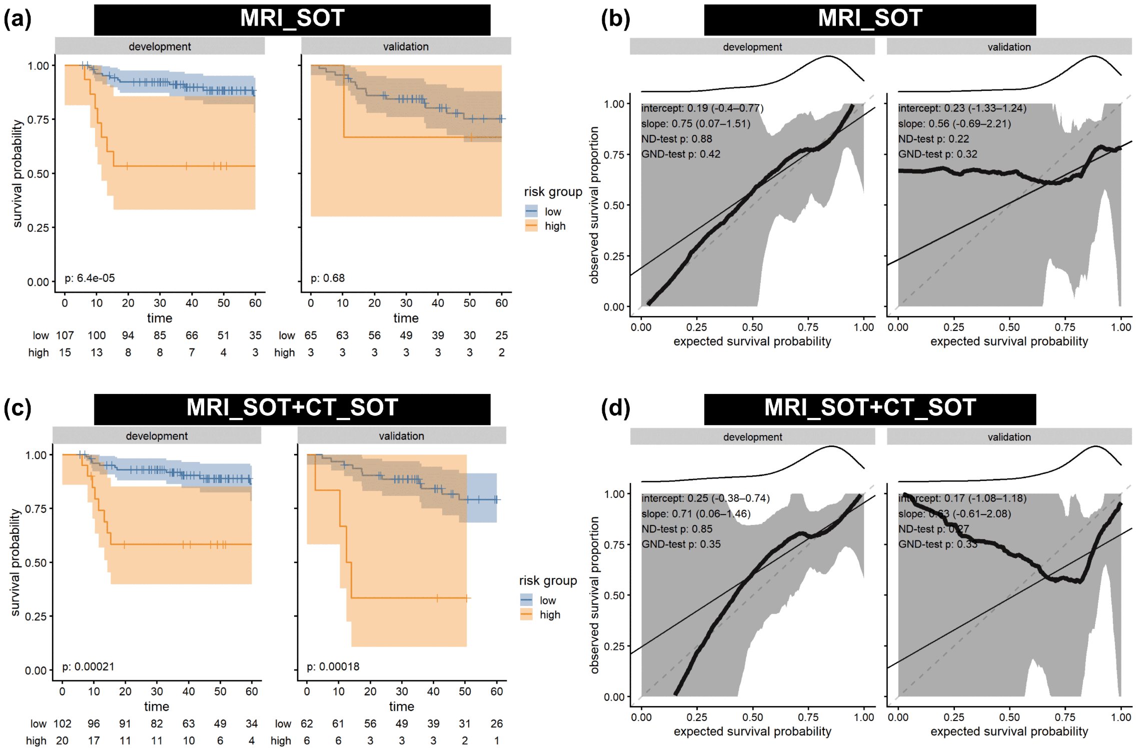


Figure A.5: Kaplan-Meier curves and corresponding calibration plots for the best performing SOT signatures for the prediction of FFDM using (a,b) MRI_SOT only model, (c-d) joint MRI_SOT+CT_SOT model as presented in appendix Table A.5.

B External validation of published radiomics models for patient prognosis in locally advanced rectal cancer

Table B.1: Clinical characteristics of included studies, NP: information not provided in manuscript.

Study	Patient number	Field strength	Sequence	Dose (Gy)	In-plane voxel dimension/ slice thickness (mm)	Male/female average or median age	Responders/ non-responders	cT/cN
De Cecco (2015)	15	3T	T2-w FSE	50.4-54	NP/4.0	9/6 Average age = 63.3	6/9	locally advanced tumour stages II (cT3-4, N0, M0) and III (cT1-4, N+, M0)
De Cecco (2016)	12	3T	T2-w FSE	50.4-54	NP/4.0	4/8 NP	6/6	locally advanced tumour stages II (cT3-4, N0, M0) and III (cT1-4, N+, M0)
Chidbaram (2017)	78	1.5T, 3T	T2-w FSE	45-50	0.7/4.0	NP	8/51 For 8 patients, response was not available	cT 2/3/4/Any= 20/44/6/8 cN 0/1/2/Any = 18/28/24/8
Caruso (2018)	8	3T	T2-w FSE	NP	NP/4.0	6/2 Median age = 65.5	NP	cT 2/3/4 = 16/120/62 cN 0/1/2 = 13/67/118
Casumano (2018)	Total = 198 Train = 173 Valid = 25	NP	T2-w FSE	50-55	NP	NP Average age train = 63	Train 47/126	Train cT 2/3/4= 15/100/58 Train cN 0/1/2= 10/60/103
DiNPpoli (2018)	Total = 226 Train = 162 Valid_1 = 39 Valid_2 = 25	1.5T	T2-w FSE	45-55	0.76/NP	NP Median age train = 65	Train 46/116	Train cT 2/3/4= 15/95/52 Train cN 0/1/2= 9/58/95
Meng (2018)	59	3T	T2-w SE	50	NP/3.0	39/20 Average age = 54	30/29	cT 3/4 = 37 cN 0/+ = 22
Cui (2019)	Total = 186 Train = 131 Valid = 55	3T	NP	50	NP	Train 83/48 Average age train = 53	Train 22/109	Train cT 3/4 = 94/37 Train cN 0/1/2 = 13/71/47
Antunes (2020)	Total = 152 Train = 60 Valid = 44	1.5T,3T	T2-w TSE	45-50.4	0.313-1.172/ 3.0-6.0	Train 50/10 Average age train = 63	Train 13/47	Train cT 1-2/3-4/Any= 10/46/4 Train cN 0/+/Any = 13/43/4
Petkvoska (2020)	102	1.5T,3T	T2-w FSE	50.4	NP/2.0-4.0	60/42 Median age = 61	19/83	cT 2/3/4 = 9/85/8 vascular invasion Yes/No/Any = 20/81/1
Petresc (2020)	Total = 67 Train = 44 Valid = 23	1.5T	T2-w TSE	NP	NP/3.0	Train 33/11 Average age train = 57.4	Train 27/17	cT 2/3/4 = 6/32/6 cN 1/2 = 11/33

Table B.2: Details of validated studies. ROI: Region of interest for feature extraction, NA: not applicable, NS: Not supported.

Study	Grading scheme	ROI	Image processing in study	Image processing applied for validation	Feature	IBSI synonyms	Remarks
De Cecco (2015)	AJCC	Largest	SSF filtering	LoG filter at sigma =4mm	(i) Kurtosis	(i) stat_kurt	The feature extraction pipeline can be completely implemented.
De Cecco (2016)	AJCC	Largest	SSF filtering	LoG filter at sigma =4mm	(i) Kurtosis (ii) Ve from pMRI	(i) stat_kurt (ii) NS	We validated only the first feature.
Chidbaram (2017)	AJCC	3D	NA	NA	(i) Tumour volume	(i) morph_volume	Replicated completely.
Caruso (2018)	Grading system was not mentioned in study.	2D	NA	NA	GLCM at $0^\circ, 45^\circ, 90^\circ, 135^\circ$ (i) Energy (ii) Contrast (iii) Correlation (iv) Inverse difference momentum or homogeneity (v) Entropy	(i) cm_energy (ii) cm_contrast (iii) cm_corr (iv) cm_inv_diff (v) cm_joint_entr	Features extracted using average merge method for all texture metrics, thus including information from all directions.
Meng (2018)	Mandrad	Largest	Voxel intensities were discretized	Voxel intensities were discretized By fixed bin number =25	(i) kurtosis	(i) ih_kurtosis	The study did not mention number of bins used for discretization. We discretized image intensities to 25 bins.
Cusumano (2018)	Mandard	2D	(i) Intensity normalization by 99th percentile within GTV (ii) LoG filter	(i) Intensity normalization by 99th percentile within GTV (ii) LoG filter	(i) cT (ii) cN (iii) Entropy (LoG =0.34) (iv) Skewness (LoG =0.48) (v) Max Fractal Dimension (FD) (40-100)	(iii) ih_entropy (LoG =0.34) (iv) stat_skew (LoG =0.48) (v) NS	Feature (v) cannot be extracted, clinical T and N stage were also included in validation analysis, model parameter used from study for feature (i)–(iv).
Dinapoli (2018)	Mandard	2D	(i) LoG filter	(i) LoG filter	(i) cT (ii) cN (iii) Entropy (LoG =0.344) (iv) skewness (LoG =0.485)	(iii) ih_entropy (LoG =0.344) (iv) stat_skew (LoG =0.485)	Clinical T and N stage were also included in validation analysis, model parameter used from study.
Cui (2018)	Mandard	NA	(i) Features normalization	(i) Features normalization	(i) kurtosis (ii) ClusterProminence _AllDirection_offset7_SD (iii) InverseDifference Moment_angle0_offset7 (iv) GLCMEnergy_angle45 _offset7 (v) HaralickCorrelation_ angle90_offset7 (vi) Correlation_angle135 _offset7 (vii) ClusterShade_angle 135_offset7 (viii) SphericalDisproportion	(i) stat_kurt (ii) cm_clust_prom (iii) cm_inv_diff_mom (iv) cm_energy (v) NA (vi) cm_corr (vii) cm_clust_shade (viii) morph_sph_dispr	We validated T2-w signature by extracting non-directional features using 3D ROI using fixed bin number discretization= 32 bins and merge method = average. We excluded 'HaralickCorrelation' from the signature, as by definition Haralick features are no different from non-directional GLCM'Correlation' features. Model parameters used form study for validation.
Antunes (2020)	Dworak	Largest	(i) Interpolation= 0.781×0.781× 4.0 mm (ii) N4 bias correction (iii) Intensity normalization reference to the mean intensity of the obturator internus muscle	(i) Interpolation= 0.781×0.781× 4.0mm (ii) N4 bias correction (iii) Intensity normalization within the range of 0.0-0.90	(i) Skewness-Laws Wave-Ripple ws = 5 (ii)Kurtsosis-Haralick SumEntropy ws = 9 (iii) Skewness-CoLIAGe Correlation ws = 5 (iv)Kurtosis-CoLIAGe InformationMetric1 ws = 3	(i) stat_skew (on energy map of the W5R5 Laws kernel) (ii) NS (iii) NS (iv) NS	Our data does not contain delineation for obturator internus muscle. Therefore, in order to replicate image processing step (iii), relative range intensity normalization was performed within masked. Organization feature are not IBSI compliant, therefore none of organization features could be validated.
Petkvoska (2020)	Histopathology	3D	(i) Interpolation =1×1×1 mm (ii) Normalized voxel intensities were discretized (normalization was not explained) (ii) Gabor filter	(i) Interpolation= 1×1×1 mm (ii) Standard normalization (iii) Voxel intensities were discretized (ii) Gabor filter	(i) shape surface area (ii) shape compactness (iii) GLCM difference variance (iv) GLSZM size zone low-gray level emphasis (v) std of Gabor (sigma=2, theta=30) (vi) kurtosis of Gabor (sigma=2sqrt2, theta=30)	(i) morph_area (ii) morph_comp_1 (iii) cm_diff_var (iv) szm_lgze (v) $\sqrt{\text{stat.var}}$ (Gabor, σ=2, λ=4, θ=30) (vi) stat.kurt (Gabor ,σ= 2sqrt2, λ=4, θ=30)	Normalization process was not clearly mentioned in study therefore we used standard normalization of MRI intensities before feature extraction. The study did not report lambda and/or bandwidth for Gabor features. Thus, to complete feature extraction we used lambda=4 model parameters used form study for validation.
Petresc (2020)	Ryan	3D	(i) Image normalization (mean=0, std=100) (ii) B-spline interpolation (x=y=z=2mm) (iii) Resegmentation of segmentation mask (iii) z-score normalization of extracted features before feature selection	(i) Image normalization (mean=0, std=100) (ii) B-spline interpolation (x=y=z=2mm) (iii) Resegmentation of segmentation mask (iii) z-score normalization of extracted features before feature selection.	(i) log_sigam_5.0_mm_3D _glszm_SmallArea-Emphasis (ii) wavelet_lhl_glcm_ correlation (iii) wavelet_lhl_ firstorder_10Perecntile (vi) wavelet_hhl_glcm_Imc1 (v) wavelet_hhl_ firstorder_kurtosis (vi) wavelet_hhl_glszm_ SmallAreaHighGray-LevelEmphasis (vii) wavelet_hhl_glcm_MCC	(i) szm_sze (LoG, σ=5.0) (ii) cm_corr (wavelet filter=lhl) (iii) stat_p10 (wavelet filter=lhl) (vi) cm_info_corr1 (wavelet filter=hhl) (v) stat_kurt (wavelet filter=hhl) (vi) szm_szhge (wavelet filter=hhl) (vii) NS	Feature (vii) is not IBSI standardized therefore model was validated using features (i)-(vi). Model parameters used form study for validation.

C Radiomics for the detection of tumour residuals after surgery of glioblastoma based on [^{11}C] methionine PET and T1c-w MRI

Table C.1: Image acquisition parameters of T1c-w MRI and MET-PET data for both training (N=85) and test (N=47) cohort.

Imaging parameters MRI		**Imaging parameters PET**	
Voxel spacing / mm	1.0	**Voxel spacing /mm**	2.0
Slice thickness / mm	1.0	**Slice thickness / mm**	2.0
Flip angle / °	8	**Field of view**	[903,180]
Scanning sequence	Gradient recalled	**Attenuation Correction**	MARC
Field strength / T	3	**Reconstruction method**	LOR-RAMLA
Manufacturer	Philips Medical Systems	**Manufacturer**	Philips Medical Systems

Table C.2: Feature classes extracted from MET-PET and T1c-w MRI. LoG transformations used for intensity-based features. fbs: fixed bin size, fbn: fixed bin number, MET: [^{11}C] methionine, PET: positron emission tomography, IBSI: image biomarker standardization initiative, LoG: Laplacian of Gaussian.

Features	**Number of features**	**Modality**	**IBSI Identifier**
(i) Local intensity features	2	PET / MRI	9ST6
(ii) Intensity-based statistical features	18	PET/MRI	UHIW
(iii) Intensity-volume histogram features	14	PET / MRI	P88C
(iv) Intensity histogram features fbn = 16	23	PET / MRI	ZVCW
(iv) Intensity histogram features fbs = 0.25	23	PET	ZVCW
(v) Texture features fbs = 0.25, fbn = 16		PET (fbs,fbn) / MRI (fbn)	
Grey level co-occurrence based features	25		LFYI
Grey level run length based features	16		TP0I
Grey level size zone based features	16		9SAK
Grey level distance zone based features	16		VMDZ
Neighbourhood grey tone difference based features	5		IPET
Neighbourhood grey level based features	17		REK0
(vi) Log transformed features (i)-(iv)	57	PET / MRI	
Total	PET=327 MRI=209		

Table C.3: Image preprocessing parameters for both PET and MRI data, as used in MIRP.

Parameters	**PET/MRI**
Interpolated isotropic voxel spacing (mm)	2/1
Image interpolation method	linear
ROI interpolation method	linear
Merge method for texture matrices	volume merge (IBSI: IAZD and KOBO)
Discretisation method: fixed bin number (bins)	16/32
Discretisation method: fixed bin size (bin width)	0.25/ –
Laplacian of Gaussian sigma	2mm/1mm

Table C.4: Features selected on the training data after clustering (left) extracted from MET-PET (right). These features showed highest mutual information (left) with residual tumor status, measured in terms of AUC. Feature definitions can be found in the IBSI reference manual [5]. MET: [^{11}C] methionine, PET: Positron emission tomography.

MET-PET features	AUC value	MET-PET features	AUC value
log_stat_min	0.90	log_ih_max_grad_fbn_n16	0.87
stat_rms	0.69	ih_max_grad_g_fbn_n16	0.71
stat_var	0.77	cm_info_corr2_d1_3d_v_mrg_fbn_n16	0.65
stat_skew	0.77	cm_info_corr1_d1_3d_v_mrg_fbn_n16	0.77
log_ih_kurt_fbn_n16	0.92	rlm_rl_entr_3d_v_mrg_fbn_n16	0.65
stat_min	0.55	szm_hgze_3d_fbn_n16	0.79
stat_p10	0.58	dzm_zdnu_norm_3d_fbn_n16	0.81
ivh_i90	0.81	dzm_sdhge_3d_fbn_n16	0.89
stat_qcod	0.53	dzm_ldlge_3d_fbn_n16	0.71
log_stat_energy	0.79	ngt_complexity_3d_fbn_n16	0.62
ivh_i75	0.87	ngl_dcnu_norm_d1_a0.0_3d_fbn_n16	0.75
ivh_v75	0.80	ngl_dc_var_d1_a0.0_3d_fbn_n16	0.85
ivh_v90	0.68	log_loc_peak_loc	0.67
rlm_srhge_3d_v_mrg_fbs_w0.25	0.76	log_stat_max	0.77
rlm_rl_entr_3d_v_mrg_fbs_w0.25	0.72	log_stat_mean	0.58
dzm_z_perc_3d_fbs_w0.25	0.73	log_stat_var	0.79
szm_glnu_3d_fbs_w0.25	0.62	log_stat_skew	0.89
dzm_ldlge_3d_fbs_w0.25	0.69	log_stat_median	0-61
ngl_ldhge_d1_a0.0_3d_fbs_w0.25	0.68	log_ih_cov_fbn_n16	0.89
ngl_dc_entr_d1_a0.0_3d_fbs_w0.25	0.56	log_ih_max_grad_fbn_n16	0.87

Table C.5: Median AUC for PET-status prediction based on MET-PET data using cross-validation of the training data with logistic regression. Top 5 features ranked according to their occurrence are shown here. Features with a repeated occurrence across at least 75% (3 out of 4) of the feature selection methods are marked in bold. AUC: area under the curve, CV: cross-validation, EN: elastic-net, MRMR: minimum redundancy maximum relevance, MIM: mutual information maximization, UR: Univariate regression.

Modality	Feature selection	CV training AUC	CV validation AUC	Features	Rank	Selected features
PET	MRMR	0.94	0.90	stat_rms **log_ih_kurt_fbn_n16** ngl_dcnu_norm_d1_a0_0_3d_fbn_n16 log_loc_peak_loc **log_stat_skew**	1 2 3 4 5	**log_ih_kurt_fbn_n16** **log_stat_skew** **Remarks:** Both features occurred in at least 3 out of 4 (75%) feature selection methods. These features showed a correlation >0.5 (Appendix Figure C.1). Finally, **log_ih_kurt_fbn_n16** was selected as a signature due to stronger association with the endpoint compared to **log_stat_skew**. **log_ih_kurt_fbn_n16** was used to build final models using logistic regression (GLM_logistic), Xgboost linear model (XGB_lm) and random forest (RF) learners.
	MIM	0.95	0.93	stat_rms **log_ih_kurt_fbn_n16** ngl_dcnu_norm_d1_a0_0_3d_fbn_n16 log_loc_peak_loc **log_stat_skew**	1 2 3 4 5	
	UR	0.95	0.93	**log_ih_kurt_fbn_n16** log_stat_min log_ih_cov_fbn_n16 dzm_sdhge_3d_fbn_n16 **log_stat_skew**	1 2 3 4 5	
	EN	0.96	0.94	dzm_sdhge_3d_fbn_n16 ivh_v75 **log_ih_kurt_fbn_n16** dzm_zdnu_norm_3d_fbn_n16 **log_stat_skew**	1 1 2 3 4	
Average AUC		0.95	0.93			Average AUC is reported in Table 5.4 for radiomics performance

Table C.6: Data augmentation parameters used for deep learning analysis. Augmentations were carried out using the batchgenerators package, which is an open-source python package for data augmentations.

Augmentation	Modality	Parameters
Mirror	PET/MRI	Axes = (0, 1)
Gamma transform	PET/MRI	gamma range = (0.5, 2)
Gaussian noise	PET/MRI	Noise variance = (0, 0.05)
Gaussian blur	PET/MRI	Blur sigma = (0.5, 1.5)
brightness multiplicative transform	MRI	Range = (0.7, 1.5)
Contrast	MRI	Range = (1, 1.75)

Table C.7: Summary of the selected T1c-w MRI-based radiomics signature for MRI-status detection. GLDZM: grey level size zone matrix, IH: intensity histogram.

Modality	Features	Identifier	Feature type	Definition
MRI	dzm_ldhge_3d_fbn_n32	KLTH	GLDZM	This feature emphasizes runs in the lower right quadrant of the GLDZM, where large zone distances and high grey levels are located. In essence, core regions with high intensity.
	Ih_rmad_fbn_n32	WRZB	IH	The mean absolute deviation is a measure of dispersion from the mean of discretized intensities.

Table C.8: Final models for the PET-status and MRI-status prediction using radiomics. Training was performed on the entire training data using multivariable logistic regression. In addition, transformation parameters from the Yeo-Johnson transformation and z-normalization, and optimal cutoff values from Youden's index are given.

Modality	Feature	Coefficient	p-value	Yeo-Johnson (λ)	z-score normalization (mean, sigma)	Cutoff
PET	log_ih_kurt_fbn_n16	2.65	<0.001	-0.3	(1.14, 0.52)	0.77
	intercept	1.67	-	-	-	
MRI	dzm_ldhge_3d_fbn_n32	1.15	0.001	-0.3	(3.14, 0.04)	0.38
	ih_rmad_fbn_n32	-0.64	0.05	-1.4	(0.52, 0.04)	
	intercept	-0.39	-	-	-	

Table C.9: Ensemble AUC values for training and internal validation CV folds for residual tumour status on MET-PET and T1c-w MRI data using deep learning with and without data augmentation. Models trained with data augmentation showed higher performance in internal validation, compared to models trained without data augmentation. AUC: area under the curve, CV: cross validation, MET: [^{11}C] methionine, PET: positron emission tomography, MRI: magnetic resonance imaging.

Modality	Model	CV train AUC		CV valid AUC	
		Without augmentation	With augmentation	Without augmentation	With augmentation
MET-PET	DenseNet	1.00	1.00	0.88	0.96
	ResNet	1.00	1.00	0.86	0.92
	VGGNet	1.00	1.00	0.96	0.95
T1c-w MRI	DenseNet	0.87	1.00	0.61	0.77
	ResNet	0.91	1.00	0.61	0.73
	VGGNet	0.82	0.99	0.66	0.71

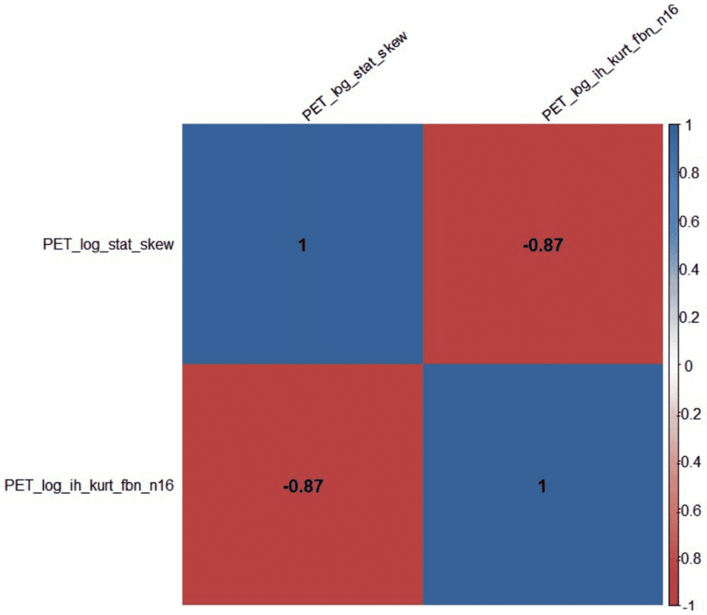


Figure C.1: Correlation plot of features with a repeated occurrence across at least 75% (3 out of 4) of the feature selection methods for prediction of PET status. All features were highly correlated (ρ>0.5).

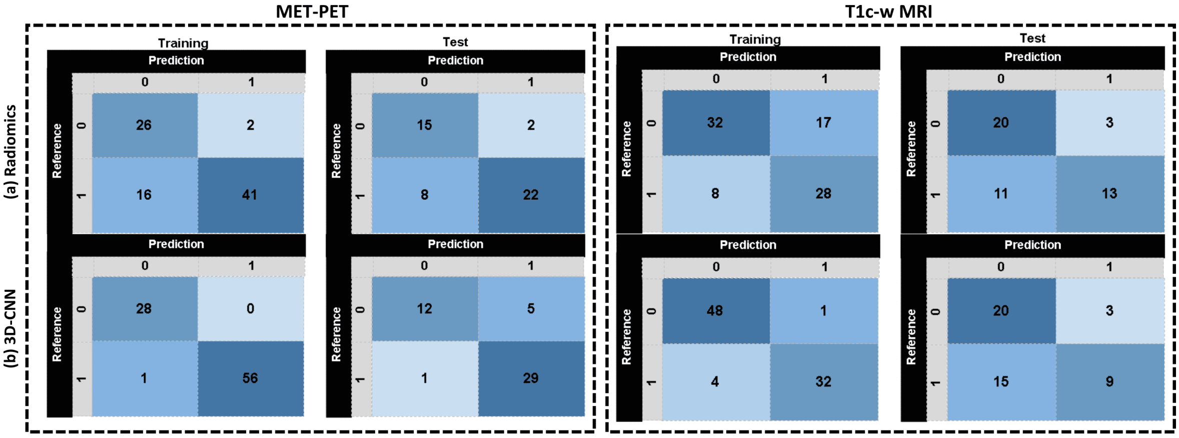


Figure C.2: Confusion matrices for training and test data based on MET-PET and T1c-w MRI data (a) using the final radiomics-based logistic regression model and (b) using the final 3D-CNN (DenseNet for MET-PET and VGGNet for MRI) model.

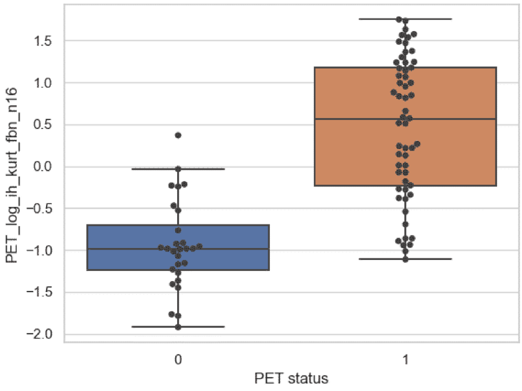


Figure C.3: Box plot of Yeo-Johnson transformed, and z-score normalized features selected in best performing PET signature for prediction of PET status on training data. PET_log_ih_kurt_fbn_16 showed relatively higher values in PET positive patients as compared to PET negative patients.

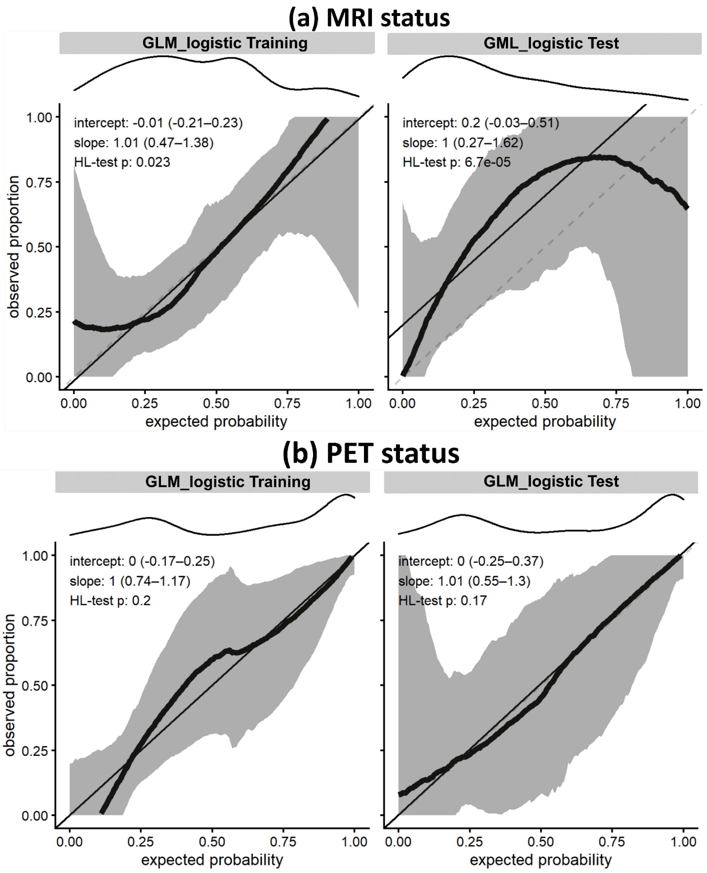


Figure C.4: Calibration plots on training and test data for the residual tumour status detection in (a) T1c-w MRI, and (b) MET-PET using radiomics. For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines) and optimal expectation (dashed lines). Density of expected probabilities is shown above the calibration plot. The residual tumour status detection on MET-PET shows good fit of prediction data.

D Radiomics for patient outcome prediction in glioblastoma using [^{11}C] methionine PET and T1c-w MRI

Table D.1: Features selected on the training data after clustering (left) extracted from MET-PET (right). These features showed the highest mutual information (left) with residual TTR, measured in terms of C-index. Feature definitions can be found in the IBSI reference manual. MET: [^{11}C] methionine, PET: Positron emission tomography.

MET-PET features	C-index value	MET-PET features	C-index value
log_stat_min	0.64	ngl_glnu_d1_a0.0_3d_fbn_n16	0.63
stat_rms	0.61	ih_max_grad_g_fbn_n16	0.53
stat_var	0.60	cm_info_corr2_d1_3d_v_mrg_fbn_n16	0.54
stat_skew	0.60	cm_info_corr1_d1_3d_v_mrg_fbn_n16	0.59
rlm_rl_var_3d_v_mrg_fbn_n16	0.62	rlm_rl_entr_3d_v_mrg_fbn_n16	0.56
stat_min	0.52	szm_hgze_3d_fbn_n16	0.59
stat_p10	0.58	dzm_zdnu_norm_3d_fbn_n16	0.60
rlm_lrlge_3d_v_mrg_fbn_n16	0.59	dzm_sdhge_3d_fbn_n16	0.65
stat_qcod	0.49	dzm_ldlge_3d_fbn_n16	0.53
log_stat_energy	0.61	ngt_complexity_3d_fbn_n16	0.49
ivh_diff_v25_v75	0.61	ngl_dcnu_norm_d1_a0.0_3d_fbn_n16	0.51
ivh_v75	0.58	ngl_dc_var_d1_a0.0_3d_fbn_n16	0.60
ivh_v90	0.49	log_loc_peak_loc	0.55
rlm_srhge_3d_v_mrg_fbs_w0.25	0.62	log_stat_max	0.57
rlm_rlnu_3d_v_mrg_fbs_w0.25	0.60	log_stat_mean	0.58
szm_z_perc_3d_fbs_w0.25	0.59	log_stat_rms	0.61
szm_glnu_3d_fbs_w0.25	0.57	log_ih_skew_fbn_n16	0.63
dzm_ldlge_3d_fbs_w0.25	0.56	log_stat_median	0.50
ngl_ldhge_d1_a0.0_3d_fbs_w0.25	0.60	log_ivh_i90	0.63
ngl_dc_entr_d1_a0.0_3d_fbs_w0.25	0.56		

Table D.2: Median C-index for prognosis of TTR based on MET-PET data using cross-validation of the training data with cox regression (Cox). Top 5 features ranked according to their occurrence are shown here. Features with a repeated occurrence across at least 75% (3 out of 4) of the feature selection methods are marked in bold. C-index: concordance index, CV: cross-validation, EN: elastic net, MRMR: minimum redundancy maximum relevance, MIM: mutual information maximization, TTR: time to recurrence, UR: Univariate regression.

Modality	Feature selection	CV training C-index	CV validation C-index	Features	Rank
MET-PET	**MRMR**	0.66	0.59	**dzm_sdhge_3d_fbn_n16** log_stat_mean **log_stat_min** ngl_glnu_d1_a0_0_3d_fbn_n16 **log_ivh_i90**	1 2 3 4 5
	MIM	0.66	0.58	**log_stat_min** **dzm_sdhge_3d_fbn_n16** **log_ivh_i90** ngl_glnu_d1_a0_0_3d_fbn_n16 **log_ih_skew_fbn_n16**	1 2 3 4 5
	UR	0.67	0.62	**log_stat_min** **dzm_sdhge_3d_fbn_n16** log_ivh_i90 rlm_rl_var_3d_v_mrg_fbn_n16 **log_ih_skew_fbn_n16**	1 2 3 4 5
	EN	0.66	0.60	**dzm_sdhge_3d_fbn_n16** **log_ivh_i90** **log_stat_min** **log_ih_skew_fbn_n16** stat_rms	1 1 2 3 4
-	**Average C-index**	0.66	0.57	**-**	

Table D.3: Wald-test p-values of 4 selected features. Among correlated features, the features with lowest p-value was selected indicating its stronger association with end-point.

Feature	Number of FS method in which feature occur	p-value
log_stat_min	4/4	7.48 e-05
log_ih_skew_fbn_n16	3/4	0.0003
dzm_sdhge_3d_fbn_n16	4/4	0.0005
log_ivh_i90	4/4	0.0007

Table D.4: Final model coefficients for the prognosis of TTR and OS using the clinical only, the Clinical+MET-PET and the Clinical+MRI radiomics models. Training was performed on the entire training data using multivariable logistic regression. In addition, transformation parameters from the Yeo-Johnson transformation and z-normalization, and optimal cutoff values Kaplan Meier plots.

Endpoint	Signature	Features	Coefficient	p-value	Yeo-Johnson (λ)	z-score normalization (mean, sigma)	Cutoff
TTR	Clinical	MGMT	0.23	<0.001	-	-	0.53
		Age	1.32	0.015	1.9	1195.451, 489.816	
	Clinical + PET	MGMT	0.26	<0.001	-	-	0.62
		Age	1.31	0.019	1.9	1195.451, 489.816	
		log_stat_min	0.70	0.004	10	-0.043, 0.018	
	Clinical + MRI	MGMT	0.22	<0.001	-	-	0.40
		Age	1.32	0.037	1.9	1195.451, 489.816	
		ivh_diff_i25_i75	0.76	0.031	-0.4	2.109, 0.039	
		dzm_zd_var_3d_fbn_n32	1.44	0.008	0	3.680,0.512	
		loc_peak_glob	1.12	0.354	0.2	1.163,0.272	
OS	Clinical	MGMT	0.20	<0.001	-	-	0.58
		Age	1.61	<0.001	1.9	1195.451, 489.816	
	Clinical + PET	MGMT	0.21	<0.001			0.68
		Age	1.60	<0.001	1.9	1195.451, 489.816	
		stat_max	1.31	0.038	-4.4	0.140, 0.029	
	Clinical + MRI	MGMT	0.19	<0.001	-	-	0.46
		Age	1.72	<0.001	1.9	1195.451, 489.816	
		ivh_diff_i25_i75	0.79	0.046	-0.4	2.109, 0.039	
		dzm_zd_var_3d_fbn_n32	1.46	0.004	0	3.680, 0.512	

Table D.5: Ensemble C-index values for training and internal validation CV splits for prognosis of TTR and OS based on MET-PET imaging and T1c-w MRI data using deep learning with and without data augmentation. Overall, models trained with data augmentation showed higher performance in internal validation, compared to models trained without data augmentation. C-index: concordance index, CV: cross validation, MET: 11C methionine, PET: positron emission tomography, MRI: magnetic resonance imaging.

Modality	Model	CV train C-index		CV validation C-index	
		Without augmentation	With augmentation	Without augmentation	With augmentation
MET-PET	DenseNet	0.75	0.84	0.68	0.68
	ResNet	0.84	0.90	0.62	0.63
	VGGNet	0.85	0.84	0.66	0.69
T1c-w MRI	DenseNet	0.87	0.86	0.59	0.63
	ResNet	0.82	0.82	0.57	0.60
	VGGNet	0.55	0.66	0.60	0.53
MET-PET	DenseNet	0.63	0.82	0.59	0.61
	ResNet	0.89	0.87	0.58	0.55
	VGGNet	0.87	0.88	0.68	0.70
T1c-w MRI	DenseNet	0.88	0.83	0.62	0.64
	ResNet	0.87	0.87	0.59	0.58
	VGGNet	0.72	0.59	0.53	0.49

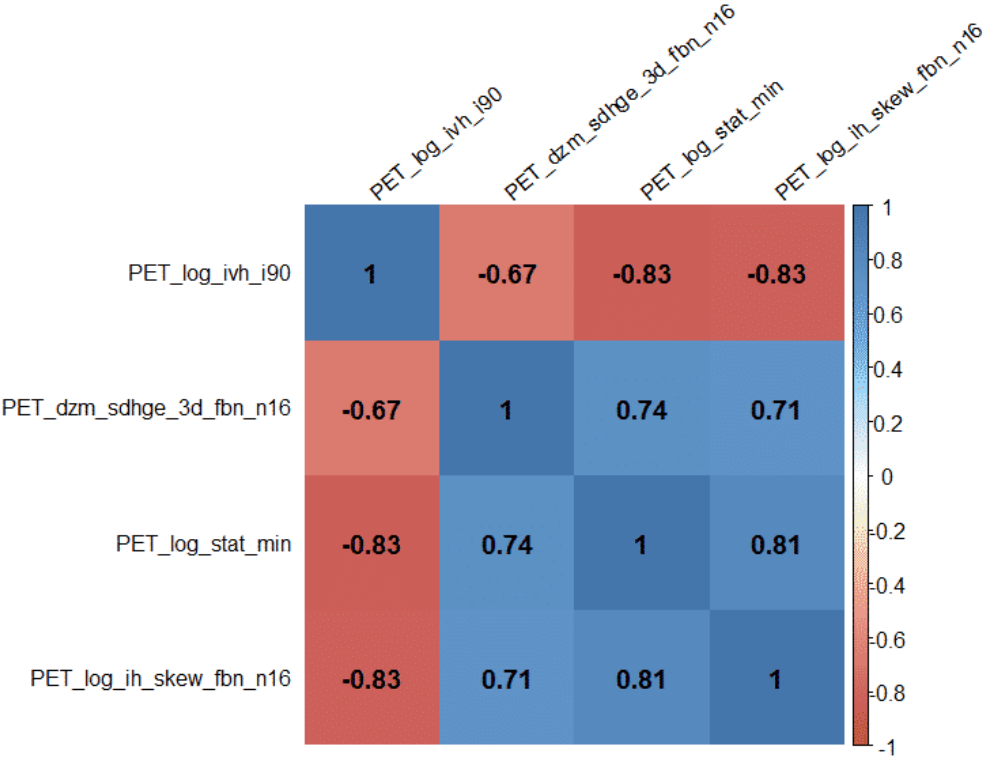


Figure D.1: Correlation plot of features with a repeated occurrence across at least 75% (3 out of 4) of the feature selection methods for prediction of PET status. All features were highly correlated (ρ>0.5).

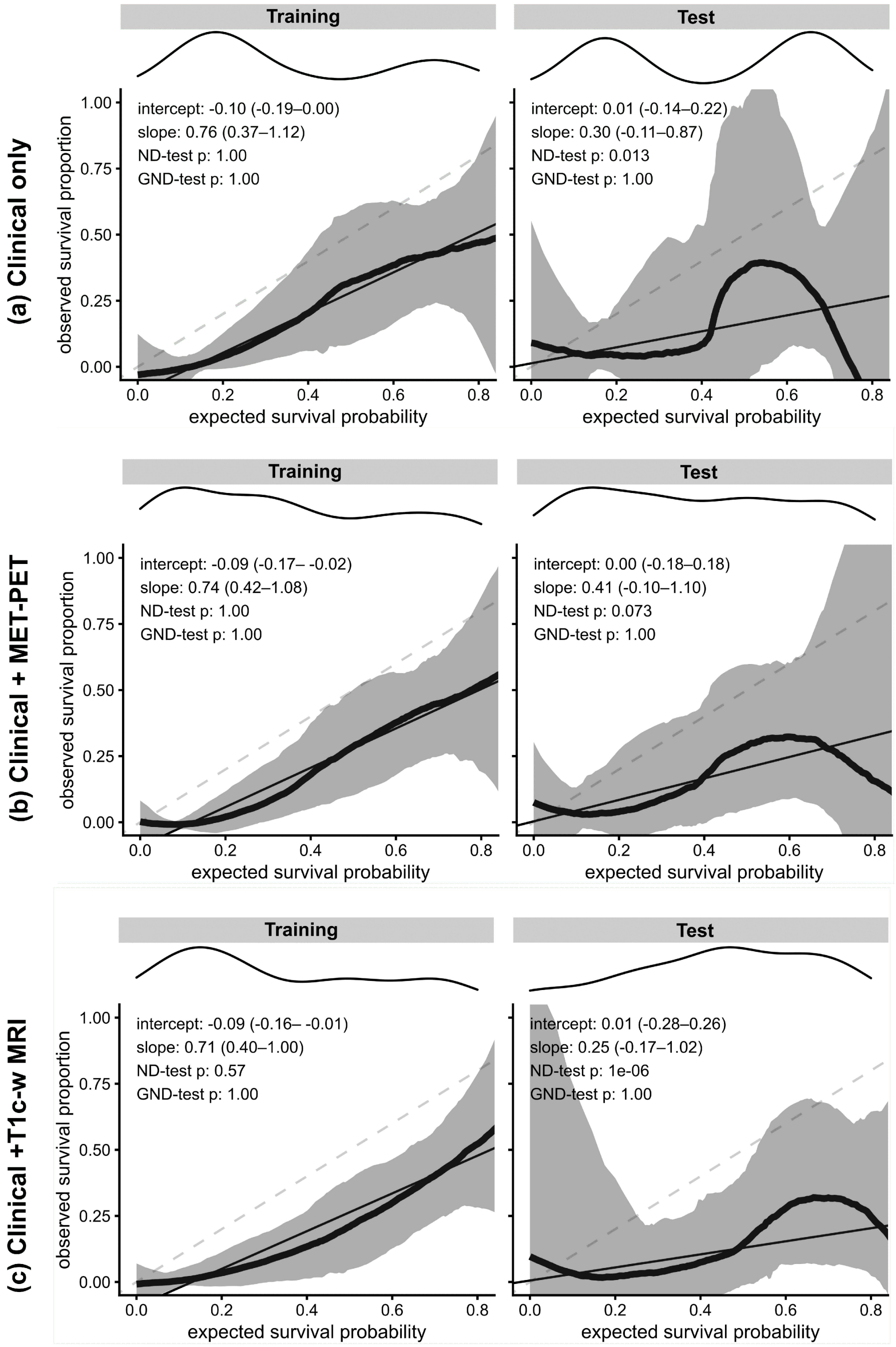


Figure D.2: Calibration plots on training and test data for the prognosis of TTR using Cox regression model based (a) clinical only signature, (b) clinical + MET-PET signature and (c) clinical + T1c-w MRI signature. For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines) and optimal expectation (dashed lines). Density of expected probabilities is shown above the calibration plot.

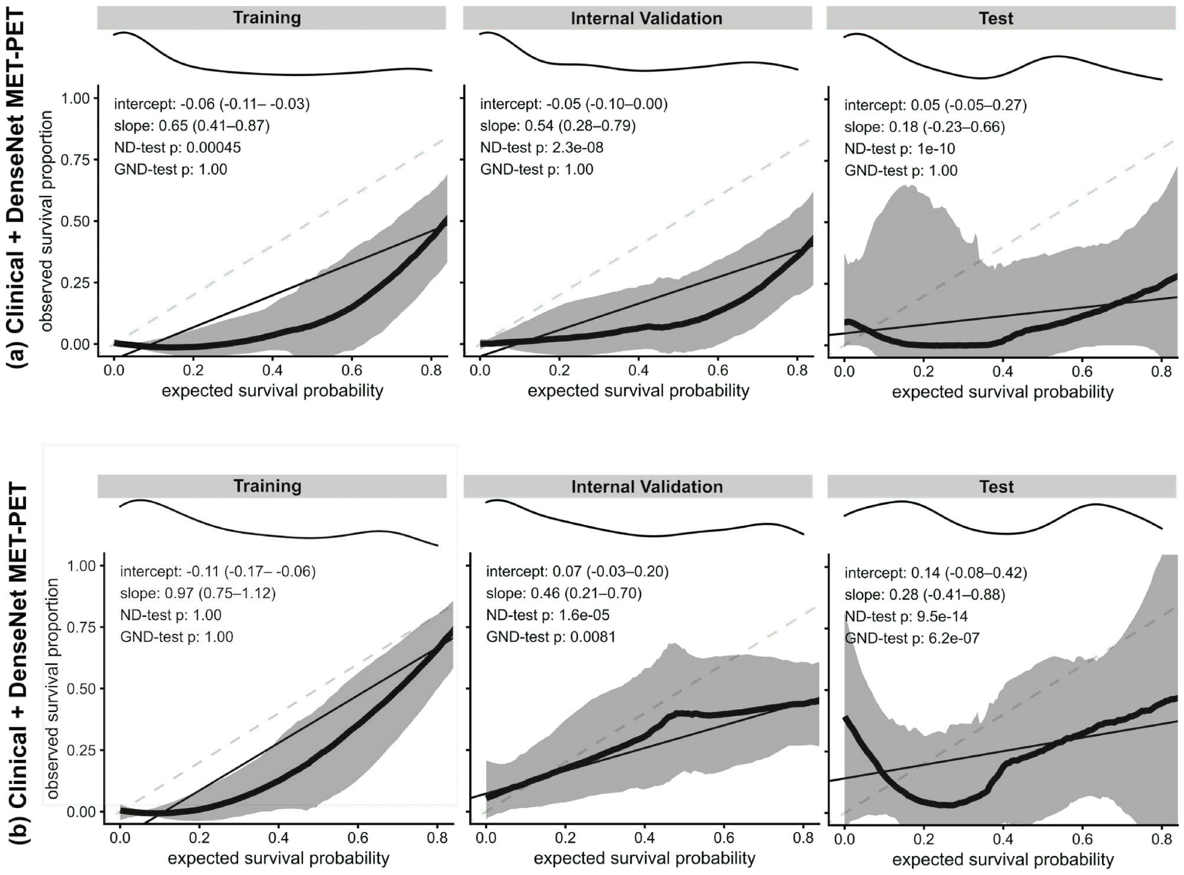


Figure D.3: Calibration plots on training and test data for (a) TTR and, (b) OS in training, internal validation and external test data based on the respective joint clinical + ensemble predictions (3D-DenseNet model) on MET-PET data. For calibration, data (thick lines) and 95% confidence intervals (shaded regions) are shown together with linear regression lines (solid lines) and optimal expectation (dashed lines). Density of expected probabilities is shown above the calibration plot.

Acknowledgement

My special thanks to my supervisor Prof. Dr. Steffen Löck for providing me the opportunity to pursue my doctoral degree in his group, for professional support and valuable scientific exchange, and cooperation, as well as for reviewing my entire theoretical work including this thesis. I am extremely appreciative of his patient guidance, calm nature, and motivation throughout my doctoral candidacy. Next, I would like to extend my sincere thanks to Dr. Alex Zwanenburg-Bezemer for his continuous guidance and support, prompt response to many of my emails and meeting requests, and his motivating and helpful comments on my manuscripts and the detailed review of the very first draft of this thesis. I am also extremely thankful to Prof. Dr. Steffen Löck and Prof. Dr. Daniela Thorwarth for their willingness to evaluate my doctoral thesis. Next, I would like to thank Dr. Annika Lattermann, Dr. Annekatrin Seidlitz, and Dr. Bettina Beuthien-Baumann for providing their expertise to prepare data for the thesis, members of my thesis advisory committee Prof. Dr. Mechthild Krause, PD Dr. Annett Linge, Prof. Dr. Esther G. C. Troost for evaluating my yearly performance and their clinical input to improve my research work and Prof. Dr. Michael Baumann, Prof. Dr. Mechthild Krause and all the other members of the German Cancer Consortium - Radiation Oncology Group and members of German Cancer Research Center for providing data and review processes of my publications and abstracts. I am also grateful to the members of my working group, Dr. Alex Zwanenburg-Bezemer, Dr. Shivaprasad Gangadhar Patil, Asier Rabasco, Dr. Almut Dutz, and Sebastian Starke for the pleasant working atmosphere. Further, my special thanks go to my friends including Aqif Hamid, Michael Amponsah-Offeh, Hannah Lukas, and Carmen Hausmann for their support during my doctorate journey. Last, but not the least, I would like to thank my family for their support and prayers throughout my academic career.

Erklärungen

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: Prof. Dr. Steffen Löck, Dr. Alex Zwanenburg-Bezemer.

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:

   a) Shahzadi I, Zwanenburg A, Lattermann A, Linge A, Baldus C, Peeken JC, Combs SE, Diefenhardt M, Rödel C, Kirste S, et al. (2022). Analysis of MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. Scientific Reports, 12. 10192. https://doi.org/10.1038/s41598-022-13967-8

   b) Shahzadi I, Seidlitz A, Zwanenburg A, Beuthien-Baumann B, Platzek I, Kotzerke J, Baumann M, Krause M, Löck S. (2022a). 3D convolutional neural networks for outcome prediction in glioblastoma using methionine PET and T1w MRI. Medical Imaging with Deep Learning. https://openreview.net/forum?id=BLXlChVgVb5

   c) Shahzadi I, Lattermann A, Linge A, Zwanenburg A, Baldus C, Peeken JC, Combs SE, Baumann M, Krause M, Troost EG, et al. (2021). Do we need complex image features to personalize treatment of patients with locally advanced rectal cancer? Medical

Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. 775–785. https://doi.org/10.1007/978-3-030-87234-273

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.
8. Ich habe die Zitierrichtlinien für Dissertationen an der Medizinischen Fakultät der Technischen Universität Dresden zur Kenntnis genommen und befolgt.
9. Ich bin mit den "Richtlinien zur Sicherung guter wissenschaftlicher Praxis, zur Vermeidung wissenschaftlichen Fehlverhaltens und für den Umgang mit Verstößen" der Technischen Universität Dresden einverstanden.

Dresden, 12. Oktober 2023

Erklärung über die Einhaltung gesetzlicher Bestimmungen

Hiermit bestätige ich die Einhaltung der folgenden aktuellen gesetzlichen Vorgaben im Rahmen meiner Dissertation

☒ das zustimmende Votum der Ethikkommission bei Klinischen Studien, epidemiologischen Untersuchungen mit Personenbezug oder Sachverhalten, die das Medizinproduktegesetz betreffen
Aktenzeichen: EK-385082020, EK-41022013, EK-390072021

☐ die Einhaltung der Bestimmungen des Tierschutzgesetzes: Not applicable

☐ die Einhaltung des Gentechnikgesetzes: Not applicable

☒ die Einhaltung von Datenschutzbestimmungen der Medizinischen Fakultät und des Universitätsklinikums Carl Gustav Carus.

Dresden, 12. Oktober 2023