Investigating Treatment Response in Glioblastoma Using radiomic Evaluation (INTRIGUE)

Kavi Fatania

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Leeds Institute of Medical Research

August 2024

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The following joint publications have been written as a result of the work in this thesis:

i. Fatania, K., Mohamud, F., Clark, A., Nix, M., Short, S.C., O'Connor, J., Scarsbrook, A.F. & Currie, S. Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma-a systematic review. *Eur Radiol* **32**, 7014–7025. (Oct. 2022)

   Contribution: K Fatania supervised the initial systematic database search, updated it and repeated the literature database searches, reviewed all search results and reviewed the included studies for risk of bias, extracted data from included studies, and drafted the manuscript.

   Contribution from other authors: F Mohamud conducted the initial literature search, independently reviewed search results and the included studies for risk of bias, and independently extracted data from included papers and helped to draft the manuscript. A Clark, M Nix, SC Short, J O'Connor, AF Scarsbrook and S Currie supervised the review and proof-read and edited the manuscript.

ii. Fatania, K., Frood, R., Mistry, H., Short, S.C., O'Connor, J., Scarsbrook, A.F. & Currie, S. Tumour Size and Overall Survival in a Cohort of Patients with Unifocal Glioblastoma: A Uni- and Multivariable Prognostic Modelling and Resampling Study. *Cancers (Basel)* **16**, 1301. (Mar. 2024).

   Contribution: K Fatania collected the data, carried out the data processing and statistical analysis, manuscript drafting and editing.

   Contribution from other authors: R Frood contributed to data collection. H Mistry supervised the statistical analysis. R Frood, H Mistry, SC Short, J O'Connor, AF Scarsbrook and S Currie proof-read and edited the manuscript and supervised the project.

iii. Fatania, K., Frood, R., Mistry, H., Short, S.C., O'Connor, J., Scarsbrook, A.F. & Currie, S. Impact of intensity standardisation and ComBat batch size on clinical-radiomic prognostic models performance in a multi-centre study of patients with Glioblastoma. Submitted to *Eur Radiol* Epub ahead of print. (Nov. 2024)

Contribution: K Fatania collected the data, carried out the data processing and statistical analysis, manuscript drafting and editing.

Contribution from other authors: R Frood contributed to data collection. H Mistry supervised the statistical analysis. R Frood, H Mistry, SC Short, J O'Connor, AF Scarsbrook and S Currie proof-read and edited the manuscript and supervised the project.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Kavi Fatania to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

©The University of Leeds and Kavi Fatania

Acknowledgements

I would like to acknowledge the support and guidance of my PhD supervisors Dr Hitesh Mistry, Prof. James O'Connor, Prof. Susan C. Short and Dr Stuart Currie. Their expertise and guidance has been invaluable over these past three years.

I have to express particular gratitude to Stuart, who has been more than generous with his time, mentorship and advice whenever I have needed it.

Although not a named PhD supervisor, I would also like to thank Prof. Andrew Scarsbrook for all of his encouragement and tutelage.

I would like to thank Dr Russell Frood, Dr Joanna Start, Dr Ming-Te Lee and Dr Ruth Whitehead for their assistance in collecting the data upon which this PhD is based. Dr Frood has also provided assistance with the coding required for the experiments.

Lastly, I could not have completed this work without the unwavering support of my partner Clara or the years of nurturing provided by my mother, Sudha. Thank you.

Abstract

Glioblastoma is the most aggressive primary brain tumour; despite maximal oncological management, median survival is at most 17 months, rising to 19 months if there is methylation of 6-O-Methylguanine-DNA Methyltransferase (MGMT) promoter (15 months otherwise) [1]. Presently, imaging biomarkers (IBs) used for prognostic stratification at the time of pre-operative, and suspected, diagnosis of glioblastoma are limited, principally, to the characteristics of the tumour on T1-weighted post-contrast (T1CE) magnetic resonance imaging (MRI) and the presence or absence of multifocal lesions [2]. There is also a paucity of clinically-applicable prognostic imaging biomarkers (IBs) that characterise the peritumoural tumour habitat of glioblastoma, which is not typically the target of surgical debulking. Radiomics, a quantitative high throughput approach to image analysis, combined with machine learning (ML) analysis, has shown great promise in non-invasively characterising the whole (enhancing and non-enhancing) tumour volume in glioblastoma [3, 4]. However, the translation of radiomics-based models has been hampered by several factors including data heterogeneity due to multi-centre MRI acquisition.

This PhD thesis outlines three studies undertaken to address some of the concerns regarding translation barriers of multi-centre radiomics models in Glioblastoma. (1) A systematic review of the evidence surrounding intensity standardisation techniques (ISTs) of MRI prior to the extraction of radiomic features was conducted. (2) A resampling study was conducted to investigate tumour volume as a prognostic radiomic IB in Glioblastoma, and examine whether non-linear transformation or sample size might contribute to heterogeneous results in other studies. (3) A modelling study was conducted to investigate the impact of ISTs and also of ComBat, a statistical model for realigning multi-centre radiomic features, on the performance of prognostic models. This included assessment of model stability and calibration, which are typically not assessed in proposed Glioblastoma survival models.

The main findings from the studies included: (1) Three techniques, WhiteStripe (WS), Nyul's histogram matching (HM) and Z-Score (ZS) were the most commonly applied ISTs in the studies ($n = 12$) included in the systematic review. There was no consensus on the optimal IST. (2)

In a multi-centre cohort of patients with glioblastoma ($n = 259$), whole tumour volume (WTV) and tumour diameter were found to be prognostic of overall survival (OS) in multivariable Cox proportional hazards models. Log-transformation of WTV and increasing sample size increased the chances of detecting a prognostic relationship during the bootstrap resampling experiment. (3) Increased batch size for ComBat realignment improved discrimination, relative model fit and explained variation of clinical and radiomic prognostic models. However, the calibration accuracy and model stability deteriorated. HM and WS tended to improve discrimination, fit and explained variation.

There was limited evidence from the published literature for an optimal IST. In our multi-centre dataset HM and WS tended to improve some model performance metrics but this was inconsistent and model stability and calibration were not improved. ComBat also improved prognostic model performance but required larger batch sizes, which discarded a large proportion of data in this heterogeneous real-world dataset, and degraded model calibration and stability. Resampling experiments also suggest that variation in sample size and ignoring the possibility of non-linear relationships could be two reasons that prognostic studies show inconsistent prognostic relationship for tumour size and this could also be the case for other radiomic IB discovery studies. Future work will focus on exploring prognostic radiomic IBs in large, multi-centre and heterogeneous imaging data and evaluate any potential IBs across multiple performance metric including stability and calibration.

REFERENCES

1. Karschnia, P. *et al.* Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO resect group. *Neuro-Oncology* **25,** 940–954 (May 2023).

2. Currie, S. *et al.* A Comprehensive Clinical Review of Adult-Type Diffuse Glioma Incorporating the 2021 World Health Organization Classification. *Neurographics* **12,** 43–70 (Apr. 2022).

3. Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C. & Mohan, S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology* **290,** 607–618 (Mar. 2019).

4. Forghani, R. Precision Digital Oncology: Emerging Role of Radiomics-based Biomarkers and Artificial Intelligence for Advanced Imaging and Characterization of Brain Tumors. *Radiology: Imaging Cancer* **2,** e190047 (2020).

Table of Contents

List of Tables

List of Figures

Abbreviations

ATRX - Alpha Thalassaemia/mental Retardation Syndrome X-linked Gene

ADC - Apparent diffusion coefficient

AUC - Area under the receiver-operator curve

AI - Artificial intelligence

BraTs - Brain tumour image segmentation benchmark

CaPTk - Cancer Imaging Phenomics Toolkit

CSC - Cancer stem cell

CNS - Central nervous system

CT - Computed tomography

CNN - Convolutional Neural Network

CV - Core volume

CPH - Cox proportional hazards

CycleGAN - Cycle-consistent adversarial network

CDKN2A/B – Cyclin Dependent Kinase Inhibitor 2A/B

CpG - Cytosine-Guanine dinucleotide pairs

DL - Deep learning

DNA - Deoxyribonucleic acid

DSC - Dice similarity coefficient

DWI - Diffusion weighted imaging

DICOM - Digital Imaging and Communications in Medicine

EGFR - Epidermal growth factor receptor

EPP - Events per predictor parameter

ECM - Extracellular matrix

FS - Feature selection

FeTS - Federated Tumor Segmentation

FBN - Fixed bin number

FLAIR - Fluid Attenuated Inversion Recovery

GPC - Glial progenitor cell

GLCM - Grey level co-occurrence matrix

GLDM - Grey level dependence matrix

GLRLM - Grey level run length matrix

GLSZM - Grey level size zone matrix

HR - Hazard ratio

HM - Histogram matching

HS-GS - Histogram specification-grid search

HSASR - Histogram specification with automated selection of reference frames

IBSI - Image Biomarker Standardisation Initiative

IB - Imaging biomarker

IS - Intensity standardisation

IST - Intensity standardisation technique

ICC - Intra-class correlation coefficient

IDH1 - Isocitrate Dehydrogenase 1

IDHm - mutant Isocitrate Dehydrogenase

IDHwt - wild-type Isocitrate Dehydrogenase

KPS - Karnofsky Performance Status

KS - Kolmogorov-Smirnov

LASSO - Least Absolute Shrinkage and Selection Operator

ML - Machine learning

MRI - Magnetic resonance imaging

MBS - Minimum ComBat batch size

mpMRI - Multiparametric magnetic resonance imaging

MGMT - 6-O-Methylguanine-DNA Methyltransferase

NCRS - National Cancer Registration Service

NGTDM - Neighbouring grey tone difference matrix

NSC - Neural stem cell

NF1 - Neurofibromin 1

NIfTI - Neuroimaging Informatics Technology Initiative

NAWM - Normal-appearing white matter

OS - Overall survival

PTEN - Phosphatase and tensin homolog

PACS - Picture Archive and Communication System

PDGRA - Platelet Derived Growth Factor Receptor A

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analysis

PCA - Principle Component Analysis

QUADAS-2 - Quality Assessment of Diagnostic Accuracy Studies 2

RF - Radiomic feature

RFo - Random forests

RSF - Random survival forests

ROS - Reactive oxygen species

ROI - Region of interest

RECIST - Response Evaluation Criteria in Solid Tumors

RNA - Ribonucleic acid

SPM12 - Statistical Parametric Mapping 12

TERT – Telomerase Reverse Transcriptase

TCGA - The Cancer Genome Atlas

TCIA - The Cancer Imaging Archive

TMZ - Temozolomide

TME - Tumour microenvironment

TNM - Tumour, Node, Metastasis staging

TP53 - Tumour protein P53

T1W - T1-weighted imaging

T1CE - T1-weighted post-contrast MR imaging

T2W - T2-weighted imaging

VASARI - Visually AccesSAble Rembrant Images

VOI - Volume of interest

WS - WhiteStripe

WTV - Whole tumour volume

WV - Whole volume

WHO - World Health Organisation

ZS - Z-score

CHAPTER

ONE

INTRODUCTION

1.1 Overview

Glioblastoma (**Figure 1.1**) is a highly aggressive brain tumour and patients have up to 17 months median survival from diagnosis [1]. First line treatment consists of maximal safe surgical resection, adjuvant radiotherapy (60 Gray in 30 fractions) with concomitant temozolomide (TMZ) chemotherapy followed by further six cycles of adjuvant TMZ ('Stupp protocol') [2]. Resection, or biopsy if debulking is not feasible due to the tumour location or patient fitness, provides the tissue necessary for classifying tumour type, which includes identification of diagnostic and prognostic molecular markers [3, 4].

Multiparametric MRI (mpMRI) is a key assessment tool in the patient's journey; they will usually have imaging to aid diagnosis, plan surgery, assess the extent of resection and extent of residual tumour, plan radiotherapy and monitor for treatment response and/or disease progression, the latter potentially leading to second- or third-line chemotherapeutic agents. MRI also characterises the tumour beyond the surgically-targeted core. Many qualitative and quantitative imaging biomarkers

(IBs) using structural or advanced MRI sequences have been proposed to non-invasively characterise glioblastoma at these stages of the patient journey [5]. The putative IBs could act as surrogate diagnostic and prognostic indicators when surgery is contraindicated, or they may provide information predicting disease response or progression [5]. Despite a myriad of research, robust and clinically-translated IBs in glioblastoma are lacking and this limits individualised patient management and the potential for personalised risk estimation.

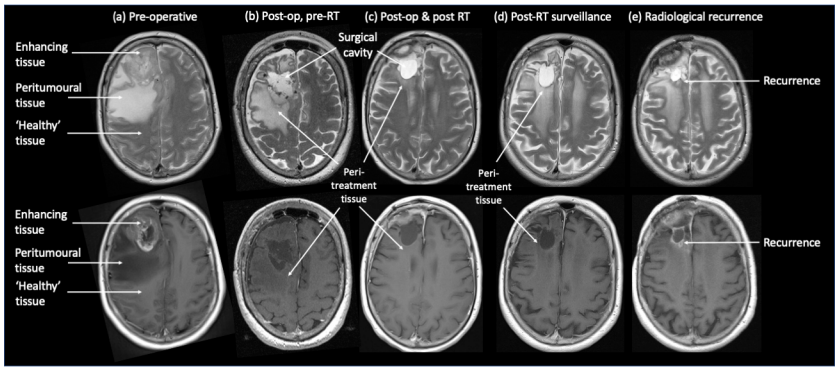

Figure 1.1: Representative images of key milestones in the treatment pathway of a patient with glioblastoma. Top row – T2-weighted MRI, bottom row – T1-weighted MRI post-contrast. RT – radiotherapy. Column (a) pre-operative assessment : establishes the likely diagnosis of a primary high grade neoplasm, determines key features that will impact surgical management such as tumour location and extent, including proximity to eloquent structures and number of lesions (multifocal or unifocal tumour). (b) Post-operative assessment: extent of surgical debulking, including determining the extent of residual tumour, and allows planning of radiotherapy treatment volumes. (c & d) Post-adjuvant chemoradiotherapy: Following adjuvant radiotherapy with concomitant chemotherapy and further adjuvant chemotherapy, regular surveillance (at 3-monthly intervals in first year and 6-monthly thereafter) imaging helps assess baseline appearances and monitor for signs of response, progression or recurrence of tumour. (e) recurrence: Despite maximal therapy, glioblastoma will eventually recur. In this case with new enhancement posterior to the surgical resection margin.

There has been much interest in augmenting medical imaging through artificial intelligence (AI) and quantitative image analysis, which may yield novel IBs. This thesis aims to examine how radiomics, a process of high throughput extraction of large numbers of quantitative imaging features with the purpose of creating mineable datasets [6], has the potential to provide novel prognostic IBs in glioblastoma. The translation from promising results in retrospective studies of radiomic IBs, to clinically-applicable prediction models has so far been lacking, and this PhD aims to examine

some of the key image processing and statistical decisions that are crucial for producing stable and robust prognostic models for future validation studies. During this opening chapter, I will introduce important topics that link this research, namely (1) glioblastoma – diagnosis and classification, (2) radiomics and (3) how statistical modelling, including machine learning (ML) can be used to produce prognostic IBs in glioblastoma. Finally, research questions will be addressed.

1.2 Adult-type diffuse glioma

1.2.1 Primary neoplasm of glial cells

Adult-type diffuse glioma (henceforth referred to as 'diffuse glioma'), of which glioblastoma is the most aggressive type, is a primary brain malignancy which shares certain histological, functional and molecular characteristics in common with glial cells [7]. Glia or glial cells, along with neurons, form one of the main constituent cell types within the brain and are subdivided into astrocytes, oligodendrocytes, NG2 cells and microglia. These cells have a key role in normal brain development and function including homeostatic and defensive functions [8].

Diffuse gliomas were once thought to arise from fully differentiated glial cells, for example an oligodendroglioma originating from differentiated oligodendrocytes [9]. However, the discovery of multipotent, self-renewing neural stem cells (NSCs) and glial progenitor cells (GPCs), which can produce astrocytes and oligodendrocytes within various locations of the adult human and mammalian brain have changed this perspective [9]. Evidence points to the NSCs in the subventricular zone as the cell of origin of glioblastoma in humans - shared mutations were observed between tumours and NSCs remote from the tumour epicentre [10].

1.2.2 Heterogeneity

Each patient with a diffuse glioma will have unique genetic and microenvironment features in their tumour, which will result in physiological and structural differences between a patient population

with a nominally equivalent diagnosis. This is termed 'intertumour' heterogeneity. Within each tumour there will be spatially distinct areas of malignant cells and stroma, termed 'intratumour' heterogeneity. Intratumour heterogeneity can change over time in natural evolution, in response to therapy and with relapse [11, 12]. Therefore, a single glioma may be composed of different cancer cell populations, each with varying sensitivity to a particular treatment and may have the capacity to develop resistance and recurrence [13, 14]. Two different models of tumourigenesis have been proposed to explain how cellular heterogeneity develops.

A clonal evolution model describes a process whereby different genetic and epigenetic modification occurs within clones at random, and through a process of natural selection, fitter clonal populations outgrow their competitors [15]. Selection pressures may include changes induced by treatments and, although 'fitter' clones may proliferate more, the result is a diverse population of cells. Alternatively in a cancer stem cell (CSC) model, a hierarchical pattern of tumourigenesis describes how CSCs asymmetrically divide to maintain their population and to generate differentiated daughter cells, which make up the tumour bulk but do not have the capacity to proliferate [16]. Such CSCs have been isolated from diffuse gliomas, including glioblastoma, and appear relatively resistant to treatments such as radiotherapy and chemotherapy [17, 18]. Additionally, phenotypic alterations that arise from the response of tumour cells to their microenvironment, for example variations in oxygen tension, from changes induced by non-tumour cells or local growth factors is another source of heterogeneity [19].

Heterogeneity is evident in the identification of distinct subtypes of glioblastoma based on gene expression profiles: classical, mesenchymal and proneural subtypes using bulk ribonucleic acid (RNA) sequencing techniques [11, 12, 20]. The classical subtype exhibits high levels of epidermal growth factor receptor (EGFR) expression, higher likelihood of point (vIII) mutation of EGFR and chromosome 7 amplification alongside chromosome 10 loss. The mesenchymal subtype is characterised by low expression or focal deletion of the region containing the neurofibromin 1 (NF1) gene, and higher proportion of necrosis and inflammatory infiltrates. High expression of platelet derived growth factor receptor A (PDGFRA) paired with focal amplification at the gene's locus is associated with

the proneural type. The proneural subtype is more commonly present in tumours with isocitrate dehydrogenase 1 (IDH1) mutation, a gene with important diagnostic and prognostic significance in glioblastoma, see below) [11, 21].

Methylation of deoxyribonucleic acid (DNA) cytosine-guanine dinucleotide pairs (CpG), an epigenetic modification that typically occurs in sections of DNA with higher CpG concentration (CpG islands), can affect the levels of gene expression when this occurs within gene promotor regions [22]. Genome-wide analysis of DNA methylation patterns in gene promotor regions of glioblastoma samples have allowed the identification of genes that are hypo and hypermethylated compared to normal controls, and allowed grouping of glioblastomas based on distinct genome-wide methylation profiles [23]. Distinct clusters of DNA methylation [24] and transcription profiles have been demonstrated across patients [25] and different gene expression and genome wide methylation subtypes of glioblastoma have been shown to co-exist within the same tumour, with cells changing from one type to another over time [11, 26, 27].

Single cell RNA sequencing studies have contributed further evidence of heterogeneity in glioblastoma; using 430 cells isolated from five patients, diverse populations of cells from different expression subtypes were identified within each tumour [28]. Highly heterogeneous tumours or tumours with more mesenchymal expression signatures were associated with reduced survival compared to those with proneural signatures. Features associated with aggressive tumour behaviour such as necrosis and microvascular proliferation have also been correlated with a mesenchymal molecular profile and this may contribute to the relatively poor prognosis of this pattern [29]. Multiple expression subtypes co-exist throughout a single glioma specimen, with varying proportions of each demonstrated depending on whether the sample is taken from the leading edge of the tumour, in a site of microvascular proliferation or from pseudopallisading cells around areas of necrosis [11, 29].

1.2.3 Invasion

Diffuse gliomas are not only heterogeneous, they are also highly invasive and therefore challenging to control with localised treatments such as resection and radiotherapy. Using immunohistochemi-

cal staining of autopsy specimens from patients with glioblastoma, microscopic tumour cells could be detected in remote regions (ie. including contralateral brain and in different lobes not connected to the tumour epicentre by any macroscopic abnormalities) of the brain parenchyma that appeared macro- and microscopically otherwise normal [30]. Despite evidence of diffuse microscopic brain invasion in glioma, most recurrences occur at the surgical margins, even following complete resection of enhancing tumour and Stupp protocol adjuvant treatment [31]. This discrepancy may be explained by a greater density of glioma cells in the region closest to the enhancing component ('peritumoural' zone) [32] and the interaction of tumour cells and growth promoting signalling factors in the microenvironment [33].

The invasiveness of diffuse glioma, and glioblastoma in particular, is thought to be driven by complex interactions between the tumour cells and the tumour microenvironment (TME), with the other non-cancer cellular constituents of tumour, and other pro-migratory changes occurring in the extracellular matrix (ECM) [34]. Tumour-associated macrophages (TAMs) represent one of the largest group of non-neoplastic cells within the TME [35]. TAMs can promote invasion through release of epidermal growth factor [36], stress-inducible protein 1 [37], interleukin-6 [38], and transforming growth factor-β [39]. It has also been suggested that TAMs in the leading edge of glioblastoma have a anti-inflammatory effect, with the degree of anti-inflammatory pathway expression in these TAMs correlating with worse survival [40].

Glioma cells may express or upregulate processes that increase motility and invasiveness as well as remodel the ECM [34]. For instance, tenascin C is an ECM protein that promotes invasion through activation of cellular motility, and has been shown to be produced by glioma cells at the tumour margin [41]. The marginal tumour cells have demonstrated differences in the subunits of integrin molecules at their surface, a protein that manages cell-cell and cell-matrix interactions and activates cellular motility [42]. Glioma cells also demonstrate increased proteolysis capacity, which helps to degrade the ECM and promote invasion. Proteases such as matrix metalloproteinases, cathepsin B and urokinase plasminogen activator help to remodel and breakdown the ECM and facilitate tumour invasion [43].

Knowing the constituency of the tumour core and peritumoural zone in glioblastoma allows speculation of the macroscopic and cellular variation in the tumour which causes variation in MR signal intensity and change in imaging phenotype, which could form the basis of imaging biomarkers (IBs). MR resolution is typically measured in millimetres, rather than the micrometre scale used in light microscopy and therefore it is unlikely that signal is composed of just cellular variation across a tumour, and instead features such as cellular density, necrosis, microhaemorrhage will influence the MR signal returned from each voxel [44]. The described changes in the ECM and the volume of TAMs in the tissue for example, could influence the macroscopic appearances of the tissue, which then impact on MR signal. Studies have, however, correlated certain radiomic features with variations in bulk tumour gene expression profiles between patients [45], but if there is a biological link between image phenotype and tumour gene expression, the change is most likely to be evident on a macroscopic level.

The invasive potential of glioma cells, facilitated by the complex interplay between tumour and TME, suggests that there is potential prognostic information contained within the peritumoural zone, and that biomarkers that characterise the whole tumour volume (WTV) may provide a more personalised prognostic model. Heterogeneity is another characteristic of diffuse glioma, and biomarkers that help to capture and quantify inter and intra-tumoural heterogeneity, as well as changes over time could also play a key role in better risk stratification for patients. Having discussed the biological characteristics of glioblastoma, the next section will outline key issues regarding the classification and diagnosis.

1.2.4 Integrated diagnosis and classification

The classification of diffuse glioma and glioblastoma has changed over the past decade, and successive iterations of the World Health Organisation (WHO) Classification of Tumours of the Central Nervous System (CNS) reflect the shift from solely histological to a combination of phenotypic, genetic and molecular characteristics [4, 46]. Before 2016, the classification of diffuse glioma was based on light microscopy resemblance to one of the differentiated glial cells. For example, astrocytomas

resemble astrocytes, and oligoastrocytomas contain a mixture of cells that resemble astrocytes and oligodendrocytes [47].

Grading, from 2 to 4, with increased aggressiveness and worse prognosis, was based on histological features that signify dedifferentiation and increased proliferation such as increased mitotic activity, cellular atypia and anaplasia, microvascular proliferation, and necrosis. The change to this classification system was driven by a number of factors that included increasing bodies of evidence that showed diagnostic and prognostic significance of genetic and molecular alterations, the variable prognosis within diagnostic groups, and interobserver variability when histological diagnosis alone is used [48, 49].

The 2016 and 2021 updates to the WHO CNS classification of tumours integrated a host of genotypic and molecular alterations; the diagnostic process for several adult-type diffuse astrocytic tumours are outlined in **Figure 1.2**.

Assessment of the IDH mutation status distinguishes glioblastoma from the IDH mutant (IDHm) astrocytomas. IDH1 and IDH2 encode the cytoplasmic and mitochondrial subtypes of IDH, respectively [50] and the normal function of wild type IDH (IDHwt) is the conversion of isocitrate to α-ketoglutarate, a reaction within the Krebs cycle. IDHm converts α-ketoglutarate to 2-hydroxyglutarate (2HG), resulting in inhibition of dioxygenases including histone demethylases and leads to methylation of CpG islands in the genome-wide DNA promoter. This results in the silencing of some and increased expression of other genes that promotes oncogenesis [21].

IDHm has also been associated with the "Glioma CpG Island Methylation Phenotype" (G-CIMP), a distinct subgroup of glioma, with widespread hypermethylation at multiple loci, leading to an undifferentiated cellular state [51] and characterised by a younger age of tumour onset [21]. The most common IDH mutation is a missense mutation of IDH1 at arginine 132 to histidine (R132H) and the mutated protein can be detected by H09 antibody immunohistochemical staining [50]. IDH status also has a prognostic role, with significantly higher survival in those with IDHm diffuse gliomas compated to their IDHwt counterparts [50].

IDHwt plays an important role in promoting the aggressive clinical phenotype of glioblastoma

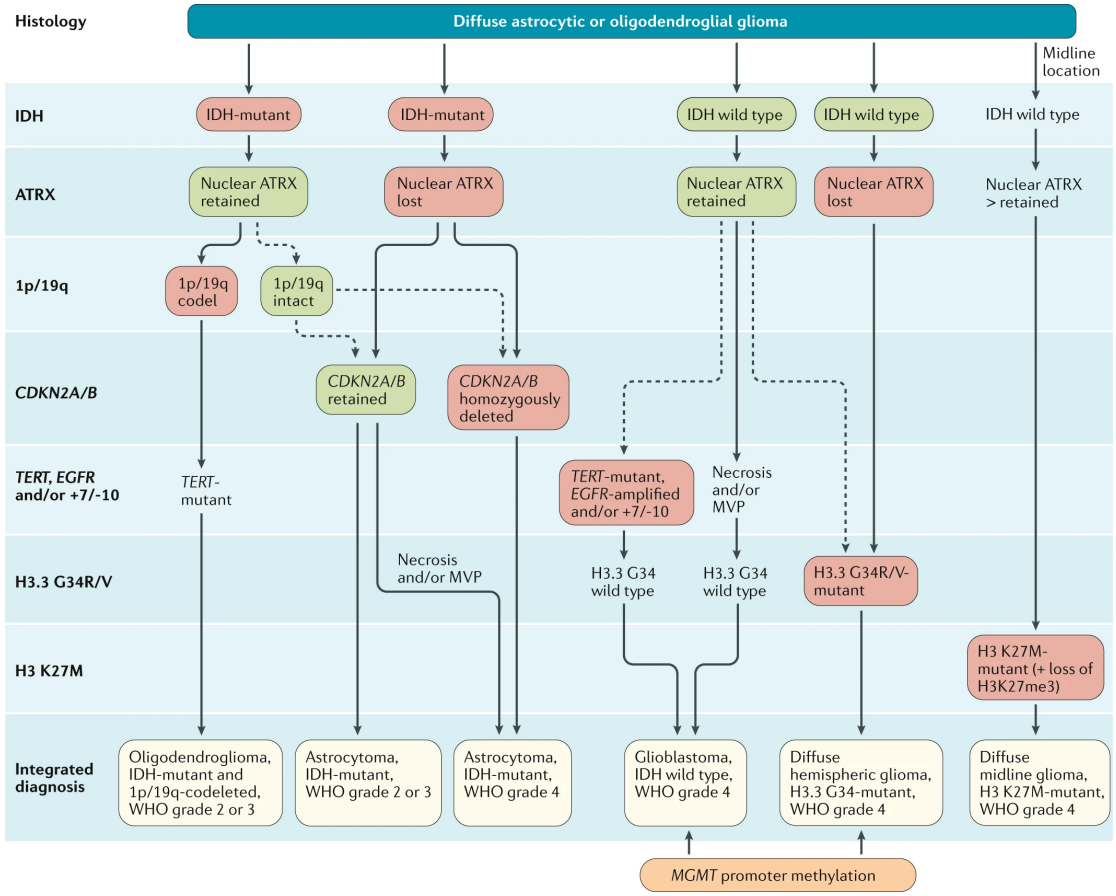

Figure 1.2: Outline of the diagnostic pathway for integrated diagnosis of the major adult-type diffuse astrocytic gliomas. IDH – Isocitrate dehydrogenase, ATRX – alpha thalassaemia/mental retardation syndrome X-linked gene, CDKN2A/B – cyclin dependent kinase inhibitor 2A/B, TERT – telomerase reverse transcriptase, EGFR – epidermal growth factor receptor. H3.3 G34R/V – glycine to arginine or valine at codon 34 of the H3F3A histone gene, H3 K27M – lysine to methionine substitution at codon 27 in histone genes H3F3A or HIST1H3B/C, H3K27me3 – trimethylation at H3K27. Figure 1.2 is reproduced under Creative Commons CC BY license [3].

[52]. IDH1 was overexpressed in most glioblastomas, and had higher activity than other enzymes responsible for the production of NADPH, a molecule used in fatty acid synthesis and scavenging of reactive oxygen species (ROS), in analysis of glioblastoma samples by The Cancer Genome Atlas (TCGA) Consortium [53]. IDH1 expression was also higher in glioblastoma specimens compared to tumours diagnosed as lower grade IDHwt astrocytomas according to the 2016 WHO classification [53]. Accepting that some of these other IDHwt might now be reclassified as IDHwt glioblastoma if they harboured the necessary molecular alterations, the evidence still supports the concept that IDH1 expression may be linked to biological aggressiveness of these tumours. Immunosuppressed murine models that received glioma-initiating cells with reduced IDH1 expression had reduced tumour growth and increased survival, supporting the role of IDH1 overexpression in progression of glioblastoma [53]. IDH1 expression can also increase after ionizing radiation, and its silencing increases radiosensitivity of in vitro and xenograft glioblastoma models [54].

Proliferation under hypoxic conditions is also an important feature of glioblastoma, which may be facilitated through IDH1 upregulation; in vitro studies have shown reduced cell proliferation under hypoxia if IDH1 was silenced [55]. Branched-chain amino acid transaminase 1 is also overexpressed in IDH1wt gliomas [56], which can lead to increased excretion of glutamate by tumour cells, which promotes progression and invasion of tumour cells and may also play a role in seizures experienced by patients with glioblastoma [57]. IDH2 and IDH3 (an IDH subtype also found in mitochondria), are thought to help tumour cells to continue proliferation in the face of respiratory chain disruption (important in ATP synthesis) or oxidative stress induced by chemo- or radiotherapy [52]. IDH2 can help cells survive and proliferate under hypoxic conditions, and IDH3 increases nucleotide synthesis and DNA methylation in vitro and in murine models [58, 59]. All three IDH subtypes, therefore, can play a crucial role in oncogenesis, as well as promoting progression, protection against hypoxia and treatments such as radio- or chemotherapy.

As well as making IDH mutation status more central to the diagnosis of glioblastoma, the 2021 WHO classification also introduced a new 'molecular' diagnosis of glioblastoma [4]. In IDHwt adult diffuse astrocytic tumours, three molecular alterations can now confer grade 4 status regardless of whether

there is microvascular proliferation or necrosis on histological examination: (i) telomerase reverse transcriptase (TERT) promoter mutation, (ii) amplification of EGFR or (iii) whole chromosome 7 gain with loss of chromosome 10 (+7/- 10) [4, 60].

Telomere shortening occurs after each mitosis, and after continuous cell division, this eventually promotes cells to cease proliferation or to undergo apoptosis. Gliomas overcome this by reactivating the telomere lengthening enzyme, TERT. TERT promoter mutations are more common in IDHwt tumours and associated with aggressive behaviour [60]. EGFR is a cell surface receptor tyrosine kinase, which promotes cell proliferation and its gene or expressions is frequently amplified in glioblastoma [61, 62]. The vIII variant (EGFRvIII), capable of ligand-independent activation, is also frequently associated with amplification of EGFR and is associated with more aggressive phenotype [63, 64]. The genes for EGFR and phosphatase and tensin homolog (PTEN), a tumour suppressor gene are located on chromosomes 7 and 10 respectively [65].

1.2.5 Demographics characteristics and prognosis

Glioblastoma is the most common type of primary brain malignancy in adults. English National Cancer Registration Service (NCRS) and Hospital Episode Statistics data from 2007-2011 reported an average annual incidence of glioblastoma of 4.64 per 100,000 people [66] and US incidence is estimated at 3.2 per 100,000 people annually [67]. More recent NCRS data has demonstrated an increase in the incidence of glioblastoma in England [68, 69] - incidence of glioblastoma, across all ages, increased from 2.39 per 100,000 in 1995 to 5.02 per 100,000 in 2015, for example [68]. Improved imaging and diagnostic pathways, improved registration and reporting of cases and changes to classification schemes have been put forwards as possible explanations for the rise [68, 69].

At diagnosis, patients with glioblastoma have a median age of 64 years and there is a male and Caucasian predominance [70, 71]. The majority ($> 90\%$) of glioblastoma occur in supratentorial locations, within the frontal lobe most commonly [70, 72]. Known predisposing factors are limited to previous exposure to ionising radiation; radiation-induced glioblastoma occurs mostly within 20 years following the exposure to therapeutic doses [73, 74]. A tiny fraction of patients with

diffuse glioma has a predisposing hereditary condition such as neurofibromatosis 1 or 2, tuberous sclerosis, Li Fraumeni or Turcot syndrome, genetic conditions which increase the individual's risk of developing malignancies [71].

Oncological management according to the 'Stupp protocol' has been shown to extend median overall survival (OS) by 2 months, and increase 5-year survival by a factor of 5 compared to adjuvant radiotherapy only [2, 75]. TMZ is an DNA-alkylating agent that interferes with cell division and whose efficacy is altered by methylation of the 6-O-methylguanine-DNA methyltransferase (MGMT) gene promoter [76, 77]. Hypermethylation of the MGMT promoter results in reduced expression of the gene products, which are involved in the repair of DNA damage caused by TMZ [76] and it is therefore unsurprising that MGMT promoter methylation predicts response to TMZ chemotherapy. A retrospective study of trial data demonstrated increased median progression free survival (PFS) in patients with IDHwt glioblastoma and MGMT promoter methylation who received chemotherapy (27 months) compared to patients with unmethylated promoters (PFS 9 months) [78]. Patients that received radiotherapy alone also showed no difference between groups stratified by MGMT promoter methylation [78].

Despite maximal treatment, median OS for patients remains poor; historically quoted at around 12-15 months median survival [2, 75], and more recently 17 months (19 months if MGMT promoter is methylated, 15 months otherwise) in trial patients [1]. Median OS is typically lower depending on the type of study. It ranges between 6 - 13 months in observational, less selective cohorts that necessarily include patients that would not be fit for intervention [66, 79, 80], and it is suggested that the poor outcomes could be due to the intra- and intertumour heterogeneity and tissue invasion outlined above.

Recent observational cohorts of patients diagnosed with glioblastoma according to 2021 criteria confirm this picture, and also demonstrate a wide range in OS for patients, for example between 0 and 80 months [79]. Hence there has been much interest in developing potential prognostic models for patients with glioblastoma, which incorporate clinical, genetic, oncological and imaging characteristics of the patient and tumour to try to risk stratify patients and attempt to tailor risk

prediction to the individual [81, 82]. Such information may help patients and clinicians make more informed decisions from the outset, with entry into pre-surgical/neoadjuvant therapy trials, more aggressive resection or even palliation. With this in mind, IBs in patients with glioblastoma will be explored in the next sections.

1.3 Imaging biomarkers in patients with glioblastoma

1.3.1 Imaging biomarkers

A biomarker is "a characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions" [83, 84]. Biomarkers must be measurable, but they can be numerical or categorical and can be derived from measurements from imaging and they can be used at any point along a patient's treatment pathway [85]. Biomarkers used in neuro-oncology can be categorised based on the information they provide. For example, susceptibility biomarkers indicate increased risk to developing disease and include presence of germline mutation in genes such as the tumour suppressor gene p53, which increases the chance of developing malignancy. Other types of biomarker include diagnostic (detects or confirms presence of a disease), monitoring (used serially to assess the status of a disease), prognostic (indicates the likelihood of an event), predictive (indicates if a patient might be higher risk of developing unwanted effects of an agent), response (indicates that a particular biological response has occurred) or safety biomarker (indicates presence or degree of toxicity following an agent) [85]. An example of a diagnostic biomarker is the presence of IDH mutation and 1p/19 co-deletion in a histologically confirmed glioma, indicating that it is a oligodendroglioma [4], and a prognostic biomarker could be the presence of MGMT promotor methylation.

O'Connor et al., however, stipulate a more constrained definition of IBs that need to be distinguished from modality, technique, and image signal, which describe the methods or processes related to image acquisition. Their definition of IBs consist of specific measurements derived from the image acquisition process [84]. For instance, MRI is an imaging modality that is based upon the imaging

signal derived from the physical process of free induction decay of protons within the body. Diffusion weighted imaging (DWI) is an MRI technique, which indicates the diffusivity of water molecules in the body.

A putative IB might be derived from the mean apparent diffusion coefficient (ADC) value in a defined portion of an image (for example a tumour) and the mean ADC value above or below a prespecified threshold may have a prognostic or diagnostic implication for a patient. 'Putative' in this context is taken to mean IBs that have only been used in research settings and that have not been validated or translated to clinical practice. IBs henceforth will be used to refer to both putative and validated IBs.

Biomarkers can be quantitative, measurable on an interval or ratio scale and expressed as a quantity value such as tumour diameter, which may be used in a staging system such as Tumour, Node and Metastasis (TNM) staging, or the measurement of maximum standardised uptake value (SUV_{max}) on positron-emission tomography (PET) for example [84]. Quantitative IBs can be expressed on a continuous scale, with each unit change in the IB value being meaningful in terms of predicted outcome or therapeutic response, or they can be expressed as a categorical IB. Cagney et al. suggest that tumour histology is an example of a qualitative diagnostic biomarker, and such as tumour staging in non-small cell lung cancer (NSCLC) TNM staging [86]. TNM staging itself is an example of a categorical, prognostic biomarker by Cagney et al.'s definition given it can be "used to identify likelihood of a clinical event, disease recurrence, or progression" [85].

Qualitative IBs cannot be expressed as a quantity value and examples include the presence of mediastinal invasion in a NSCLC, which would upstage a tumour to T4 according to current TNM staging [86] or the American College of Radiology Breast Imaging Reporting and Data System, a structured mammography reporting system with a ordinal five point score that positively correlates to the risk of breast malignancy [87].

1.3.2 IBs used to characterise glioblastoma

As discussed in preceding sections, patients with diffuse gliomas, and particularly glioblastoma, have a poor prognosis in part due to the highly invasive nature of the tumours and high levels of heterogeneity. Routine clinical assessment is performed with multiparametric MRI (mpMRI, combining several different MRI techniques into a single scan episode), and radiological examination will typically include an assessment of tumour location, measurement of size and assessment of selected qualitative features [88]. **Figure 1.3** shows the typical radiological appearances of glioblastoma, with an area of enhancement in the right frontal lobe, which is thought to represent the tumour core, with central areas of non-enhancing necrotic tumour, and is surrounded by high T2W signal that is labelled 'peritumoural tissue'.

A commonly used diagnostic, qualitative IB is the presence or absence of enhancement on T1-weighted post-contrast MRI (T1CE), which is a marker of breakdown of the blood brain barrier, and is a predictor of higher grade tumours [8, 89, 90] and therefore also a prognostic IB. The specificity of this IB is reduced by the association between lower grade diffuse gliomas, especially those with oligodendroglial histology, and intra-tumoural enhancement [91], but it nevertheless represents one of the most important IBs in patients with diffuse glioma and glioblastoma [89].

The widely used qualitative descriptors of imaging in patients with glioblastoma, such as presence of enhancement, tumour location, side of tumour, presence of satellite or multifocal lesions have been investigated for their association to patient prognosis in glioblastoma through retrospective cohort studies [92–94]. Some potential pitfalls with relying solely on qualitative analysis might be inter-observer variation in application of descriptors and also variation in the meaning ascribed to descriptors such as 'extensive perilesional oedema' or 'predominantly necrotic tumour' for example.

In an attempt to standardise qualitative assessment and also explore the association between descriptor of glioblastoma with outcome, the Visually AccesSAble Rembrant Images (VASARI) feature set was developed by neuroradiologists. The VASARI feature set consists of 24 imaging characteristics derived from standard anatomical MRI, including T1-weighted imaging pre- (T1W) and T1CE,

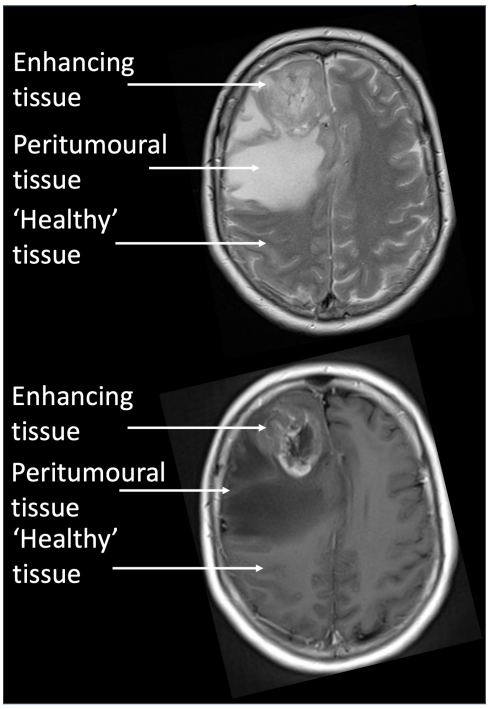

Figure 1.3: Representative images from T2-weighted (top) and T1-weighted post-contrast (bottom) MRI sequences in a patient with glioblastoma. There is a right frontal irregularly rim-enhancing tumour with central non-enhancing necrosis, together taken to represent tumour core. This core is surrounded by high T2 signal peritumoural tissue, and beyond that, radiologically normal tissue is seen in the right parietal lobe and left cerebral hemisphere.

T2-weighted imaging (T2W), which may include Fluid Attenuated Inversion Recovery (FLAIR) sequences, and also DWI [95]. The aims were to provide a controlled, easy to use lexicon that would allow easier comparison between research study findings, be robust to inter-observer variation in descriptions, and provide prognostic relevance for patients with glioblastoma [88, 95]. The VASARI features are a combination of qualitative features, for example tumour location, side of tumour epicentre or involvement of eloquent brain, but also categorical, ordinal IBs such as proportion of enhancing tumour (**Figure 1.4**).

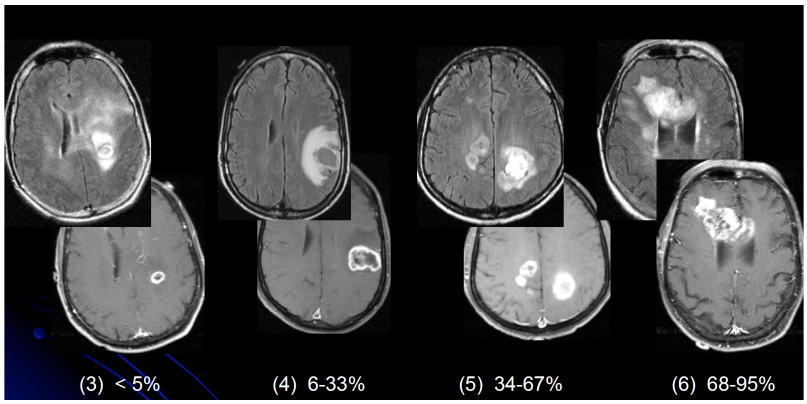

Figure 1.4: From left to right, representative MRIs that typify four ordinal categories of the proportion of tumour that enhances, one of the VASARI features. The proportion of each patient's tumour that enhances is determined by comparing the volume of enhancing tumour, which is assessed on the T1-weighted post-gadolinium (T1CE) image (bottom image of each image pair) to the overall volume of tumour using the Fluid Attenuated Inversion Recovery (FLAIR) sequence (top image of each image pair).

The inter-observer concordance for VASARI features such as proportion of enhancing tumour has been shown to be moderate to high, although other features such as calvarial remodelling or cortical involvement had low concordance [88]. The proportion of enhancing tumour has been associated with poorer OS in glioblastoma, after adjusting for Karnofsky Performance Status (KPS) - Hazard Ratio (HR) 7.75 95% Confidence Interval (CI) 1.14 - 52.87 [88]. The presence of synchronous tumour sites, unconnected by any high T2 signal - 'multifocal' glioblastoma [8] - is another prognostic IB associated with lower median OS (median OS 8 months typically for multifocal glioblastoma [96]). Midline, or bilateral unifocal lesions, and tumours or peritumoural tissue extending into deep brain

structures such as the brain stem are also associated with poor relative outcomes [94, 97]. Hence, many IBs that qualitatively capture the radiological phenotype of glioblastoma on anatomical MRI have been shown to be prognostic IBs.

This overview on IBs in glioblastoma has focused on 'conventional' MRI sequences, T1CE or FLAIR for example, which provide anatomical macrostructural information about tissues and are distinguished from 'advanced' sequences, such as DWI or perfusion-weighted imaging, which characterise cellular or microscopic tissue properties [90]. DWI, for example, semiquantitatively measures the diffusivity of extracellular water molecules and its voxel intensities are measured in units such as mm^2/sec, whereas 'conventional' MRI intensity is an arbitrary scale measured relative to other tissues and structures in the image rather than to some estimated biophysical property [90].

VASARI does include one DWI feature, although this is a qualitative assessment of whether there is, or is not, reduced diffusivity within the tumour relative normal grey matter and not a quantitative measurement of DWI signal [88]. This project will focus solely on 'conventional', anatomic MRI, not due to its perceived superiority over advanced MRI or positron emission tomography [90, 98, 99], but due to the ubiquity of conventional MRI in clinical practice [100] and as a substrate for radiomic analysis. Extracting quantitative IBs from sequences that are typically qualitatively assessed has been one of the key motivators for radiomic IB development [101].

1.4 Development of novel IBs with radiomics

1.4.1 Radiomics

Radiomics is a process of high throughput extraction of large numbers of quantitative imaging features, creating mineable datasets from standard of care medical images that can be used for IB discovery and ultimately improve clinical decision making [6]. Quantitative radiomic imaging features may be combined with clinical and laboratory information to provide clinical decision support [101]. Radiomic features (RFs) may provide good substrates for novel IBs in oncology,

as it is thought that the features may quantify the heterogeneity of the imaging phenotype in glioblastoma ('multiforme' was dropped from the nomenclature in the WHO 2021 classification but attested to the variety of macroscopic appearances) and this may correlate with the considerable intra- and intertumoural heterogeneity demonstrated on a microscopic, genetic and epigenetic level [101, 102].

Glioblastoma can be divided into physiologically distinct regions or 'habitats' that are possible to visualise on MRI (**Figure 1.3**) - (1) the necrotic core (central region of high T2 and low T1W, surrounded by enhancing margin), (2) enhancing rim (high T1CE) and, (3) non-enhancing peritumoural tissue (high T2W beyond the enhancing rim). The shape, signal intensity, proportions of each habitat to one another, the heterogeneity of the signal across the habitat can all be quantitatively described through the mineable radiomic data extracted from these habitats on conventional MRI sequences [103]. By potentially capturing the heterogeneity of the radiological phenotype, it has been suggested that links can be derived between tumour biology and macroscopic appearances, which have traditionally thought to be more feasible only with advanced MRI sequences or PET [90, 101, 102].

The process of extracting RFs from MRI and developing novel IBs can be broken down into discrete steps: 1) image acquisition, 2) image processing and segmentation of the region or volume of interest (ROI, VOI respectively), 3) intensity standardisation, 4) RF extraction including grey-level discretization, 5) post-extraction feature realignment (typically using ComBat realignment), 6) feature selection and modelling and model assessment (**Figure 1.5**).

1.4.2 The radiomics workflow

1.4.2.1 Acquisition

Most referral systems for tertiary neurosciences centres in England operate using a 'hub-and-spoke' model. Initial diagnostic imaging for patients with (suspected) glioblastoma is acquired across multiple, geographically different sites before referral to a central treatment centre in the region. This

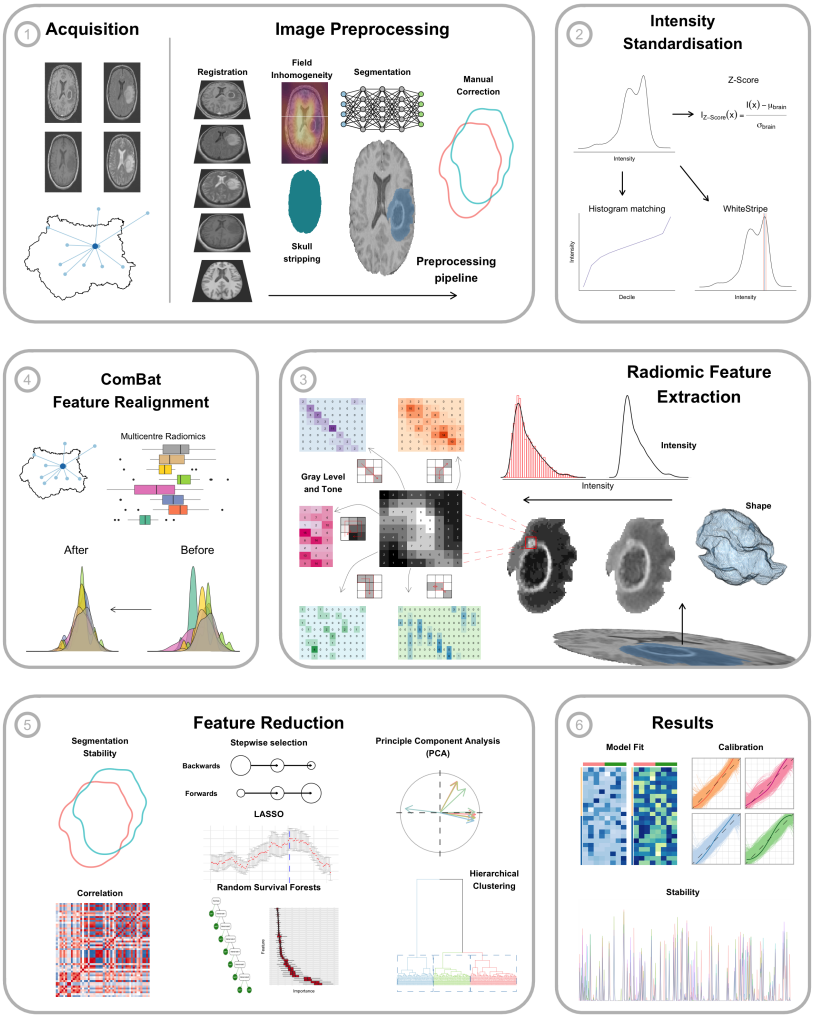

Figure 1.5: Panels 1-6 outline the main steps of the workflow: 1) The acquisition of MRIs usually occurs across multiple geographic sites and pre-processing includes registration, skull stripping and field inhomogeneity reduction; 2) Intensity standardisation of MRI signal intensities using one of three popular techniques (WhiteStripe, Nyul's histogram matching or Z-score); 3) Radiomic feature extraction, including calculation of shape, intensity and higher level features; 4) post-extraction realignment of multi-centre radiomics using ComBat; 5) Application of multiple feature reduction techniques to reduce the dimensionality of the data; 6) Calculation of results and data analysis. GLCM = grey level co-occurrence matrix, GLDM = grey level dependence matrix, GLRLM = grey level run length matrix, GLSZM = grey level size zone matrix, LASSO = Least Absolute Shrinkage and Selection Operator, NGTDM = neighbouring grey tone difference matrix, T1 = T1-weighted, T1CE = T1-weighted, contrast-enhanced, T2 = T2-weighted.

lends itself to heterogeneous acquisition protocols, at least before care is transferred to the neuroscience centre. Because large imaging datasets are often analysed retrospectively, harmonisation of image protocols in such referral systems can be challenging [104]. MRI is the mainstay of glioblastoma characterisation (**Figure 1.1**), and, for the purpose of most radiomics studies, consists of at least four anatomical sequences - T1W, T2W, FLAIR and T1CE. Recommended acquisition protocols for these sequences in the context of brain tumour imaging have been published, highlighting the importance of homogeneous data for subsequent analyses [105, 106].

Protocols can vary significantly between sites, not least by slice thickness, field strength, voxel size, echo and repetition times and all these parameters have an impact on RF values [104, 107–109]. Some radiomics analyses choose to limit their approach by only including patients imaged using the same protocol or site [110] but this can lead to data loss, particularly with diverse acquisition across a hub-and-spoke imaging model. By limiting radiomic analyses to imaging acquired only at one centre, a more homogenised acquisition protocol can be obtained, but this can result in an oversimplification of diagnostic imaging acquisition. Whilst subsequent data analysis is made easier, the generalisability and applicability of the models derived from homogeneous datasets is more challenging [104].

IBs need to be widely available across geographical locations to be translated to clinical use [84], and therefore it may pose a challenge to radiomic IBs if they are heavily dependent on homogenised imaging protocols. Even with harmonised imaging protocols across sites, it has been shown that scanning the same patient at the same site but at different times (within minutes or days) using the same protocols can also introduce detectable differences into the images (and RFs) that will need to be removed or adjusted for through image processing or statistical techniques [111, 112].

1.4.2.2 Pre-processing and segmentation

MRI preparation for brain tumour radiomics extraction requires file conversion, and often requires additional preprocessing steps to prepare the images so that they can be segmented using automated or semi-automated approaches. MRI is acquired as Digital Imaging and Communications

in Medicine (DICOM) images, which are first converted to Neuroimaging Informatics Technology Initiative (NIfTI). Images are often rigidly co-registered to one another, and often to a standardised brain atlas. Image registration also spatially resamples voxels, and an isotropic voxel size of $1mm^3$ is commonly used [113], although there is no consensus position on whether upsampling or downsampling of the native image resolution is preferable [114, 115]. Magnetic field inhomogeneity within the image is reduced [116], and the skull voxels removed to improve subsequent processing by segmentation algorithms.

Accurate segmentation of the tumour VOI can be performed manually or with the assistance of deep-learning (DL) segmentation models with or without manual correction. Even with the use of expert segmentors, manual mask delineation will vary, and the task is time intensive [117]. Hence, the use of automated or semi-automated methods using accurate and fast DL networks has increased in popularity, making the annotation of large datasets more feasible [118]. Variability of segmentations between two (or more) observers or DL models can be a benefit to radiomics as it allows assessment of RF reproducibility; i.e. the consistency of the IB to changes in experimental settings [84]. Multiple segmentation has also been incorporated as a criterion in the radiomics quality score (RQS), a 36-point checklist for assessing the quality of a radiomics study [119, 120].

1.4.2.3 Intensity standardisation

As conventional MRI signal intensity is relative and measured in arbitrary units [90], changes in scanner parameters, vendors, models can all impact the image intensity and therefore RF values [111]. Prior to RF extraction, the intensity of MRI can be standardised, so that there is a similar range and distribution of values across patients. There are many approaches available; three commonly used intensity standardisation techniques (ISTs) in glioblastoma are WhiteStripe (WS), Z-score (ZS) and Nyul's histogram matching (HM) [111, 121].

Nyul's HM is a two-step process: (i) an average intensity histogram is developed from a sub-set of images, and (ii) the intensity values in all images are linearly mapped to the average histogram, per decile of the intensity range [122, 123]. To produce the standard histogram scale, the intensity values

of the training images are averaged at set percentiles (the default percentiles are 1, 10, 20,...90, 99) of the intensity range. The intensities below the 1st and above the 99th centiles are discarded to minimise the impact of outliers. To transform each image, a histogram of its signal intensity is calculated, the intensity range divided into deciles and the voxels that fall into each intensity decile are transformed separately. For each decile, the corresponding voxel intensities are mapped linearly to the standardised 'training' histogram and the intensity altered (**Figure 1.5**, panel 2). ZS and WS standardisation adjust the signal intensity of images by subtracting the mean and dividing by the standard deviation of either the whole image or of normal appearing white matter, respectively [111, 124]. All three of the ISTs could be applied to single images at the point of testing, including HM, but HM requires a 'training' step as outlined before a standardised histogram can be created and therefore, ZS and WS can work on a per image basis in both training and testing steps.

ISTs have been applied to brain and other body imaging with promising results for the techniques described. HM and WS improved the segmentation of multiple sclerosis lesions [123, 125], WS increased the consistency of intensities across brain MRIs in public datasets of healthy volunteers [124], and all three ISTs described, increased the performance of DL networks that synthesise T2W or FLAIR from T1W images [126]. Despite the promise of each technique, it is not clear whether there is an optimal approach in brain MRI standardisation [127].

1.4.2.4 Radiomic feature extraction

Four categories of RF can be extracted from the tumour VOI:

1. Size and shape features such as volume, surface area, sphericity
2. First-order or intensity features - derived from the histogram of the VOI intensities
3. Second-order or texture features - describe the relations of discretized voxel intensities to their neighbours
4. Filtered features - different image filters can be applied to the image, for example edge-enhancing filters, and spatial patterns can be extracted.

A comprehensive description of RFs and their calculation can be found in the Image Biomarker Standardisation Initiative (IBSI) documentation [114]. IBSI's aim is to improve the reproducibility of IB research by providing reference feature definitions, image processing workflow and verification methods for radiomics extraction software [114].

Size and shape features are calculated from a mesh created from the VOI, and features such as the mesh volume, surface area and surface area to volume ratio are derived. Intensity or first-order features are calculated from the histogram of all the intensity values in the VOI and include summary statistics of the histogram distribution such as mean, median, interquartile range, standard deviation and range. The shape of the histogram is described by the kurtosis and skewness and the homogeneity of intensity values is described by the uniformity [128].

Second-order features are calculated after 'intensity discretization', a process of dividing the intensity range of the VOI into discrete, non-overlapping bins and assigning each voxel to a bin. Discretization can either be 'relative', using a fixed number of bins, or 'absolute' whereby the bins have a fixed width. Discretization serves to reduce the impact of noise on feature calculation (and reduces computation time) and may reduce the multi-centre effects on MRI signal intensity variation [111], but it also leads to loss of image detail. For images with arbitrary intensity units such as conventional MRI, IBSI recommends a relative discretization (fixed bin number) approach [114].

Texture features describe the heterogeneity in the grey-levels throughout the VOI, and this is achieved through the calculation of various matrices, such as the grey level co-occurrence matrix (GLCM). GLCMs describe the probability that voxels of specific intensities are found in neighbouring voxels for each of the unique 13 angles of travel that are possible in 3-dimensions. The matrices are used to calculate the texture features [114]. Finally, filtered features are created by first passing the image through a image convolutional filter, such as one that enhances edges and RFs are then extracted in the same manner as non-filtered images. Examples of filters that can be applied include fractal analyses (patterns added to the image and elements of a certain intensity value are evaluated), Minkowski functionals (assesses patterns of voxels above a certain intensity threshold) and wavelets (apply a filter of radial or linear waves to the image and then extract features) [101].

These features, which can be prespecified and calculated using known formulae are distinguished from a related but distinct set of features that are extracted from the images using a convolutional neural network (CNN), a type of DL model, in a process called 'deep radiomics' [44]. The main point of difference between these approaches is how and when the extracted features are defined and calculated, with 'traditional' radiomics approaches using predefined equations and deep radiomics using the outputs from a DL model and how the images are prepared. For traditional radiomics approaches, the image preprocessing and segmentation steps outlined above are typical, whereas the CNN used for deep radiomics-based approaches typically will not require tumour segmentation or intensity standardisation [44].

1.4.2.5 Feature realignment - ComBat

Once extracted, RFs can be harmonised using ComBat, a statistical model that was first applied in the field of genomics, and that helps to realign imaging features extracted under different experimental settings or 'batches' [110, 129]. The batch effects can include, but are not limited to, hospital site, scanner manufacturer, vendor, MR acquisition protocol, or tumour segmentation method [130]. Underlying the method is a linear model, which assumes the following relationship:

$$y_{ij} = \alpha + X_{ij}\beta + \gamma_i + \delta_i \varepsilon_{ij}$$

Where y is the measured RF, i is the experimental setting, for example scanner, vendor or geographic site, j is the measurement used to derive the image feature, for example VOI for RF calculation, α is the average value of the RF of interest, γ_i is an additive batch effect, δ_i is a multiplicative batch effect and ε_{ij} is an error term [130–132]. This form of the ComBat (also referred to as neuroComBat) equation also includes biological co-variates, where X is the matrix for the biological co-variates, and β is the regression coefficient for each co-variate. The original implementation of ComBat did not account for these biological effects, whereas the neuroComBat approach used in this project (henceforth referred to as ComBat) will use biological covariates. It is important to take into

account any clinical variables that might confound the values of the RFs between batches.

Imaging sites that vary by geographic location could have different compositions of patient and tumour type, for example the grades of tumour may vary significantly between a tertiary centre and district hospital. Tumour grade is a confounder as it could impact radiological phenotype, and hence RF values, so it needs to be accounted for if batch effects are to be accurately estimated [130].

ComBat has been shown to reduce the effects of multi-centre acquisition in radiomic studies with promising results [104, 127, 129], but it does have limitations. The distribution of each feature is assumed to be similar across sites, adjusting for biological co-variates, and the only difference should be shift or spread of the data, which can be judged by visual inspection of histograms. For this to be accurately assessed, a sufficiently large sample size is required. If covariates are to be included, the sample size should be increased; 20-30 patients per site, per covariate is suggested [130].

In the context of multi-centre glioblastoma prognostic radiomic studies, in which many clinical predictors might impact on OS, and there may be a multitude of experimental settings, it would be difficult to amass enough data in one tertiary referral centre to accurately estimate the batch effects for every scenario. Some have suggested as few as five patienwts per experimental setting might be enough to accurately estimate batch effects [104]. However, additional public datasets or longer data collection periods would be beneficial to create adequately powered studies that can produce accurate radiomic feature distributions. Modified versions of ComBat have been suggested, which allow users to shift the radiomic feature values to a reference location, rather than averaging the results of all centres and this may be beneficial if operating in a hub-and-spoke referral system [133]. Also, data acquired from a new site, not included in the initial ComBat model, might require recalculation of the ComBat model coefficients and batch effects if the feature distribution is felt to be different to those used in the original model building process [130]. If the new centre happened to use very similar protocols for image acquisitions, and the patient and feature distributions were similar to the original cohort, then the original model may still perform well but there may be

unadjusted batch effects that have not been adequately removed.

1.4.2.6 Feature selection, modelling and machine learning

Hundreds, potentially thousands of RFs could be extracted from one image and hence feature selection strategies are needed to reduce the chance of overfitting and feature redundancy [134, 135]. Radiomics does not necessarily require the use of ML, but ML feature reduction methods are commonplace [136]. Unlike clinical prediction models, it is difficult to specify *a priori* a set of RFs to include in a model based on the results of previous studies, as the field is afflicted by a lack of reproducibility, and there are many experimental differences between studies that influence the measurement of features [111, 135].

Most radiomic studies are examples of exploratory prognostic factor research [137], which aim to narrow down many features into a select few, which are combined to create a putative radiomic IB that needs further validation in larger, external datasets. When conducting such exploratory studies, many more features are typically considered for inclusion into the final model than is recommended, greatly increasing the chance of overfitting [138]. Typically when building prognostic models using clinical predictors, statistical literature recommends that the number of features that ought to be considered for inclusion into the final model ought to be specified prior to analysis and be based on the number of events (the effective sample size) in the dataset and the results of previous models published in the field [138, 139]. In ML analyses and in DL models, this is often not possible or even desireable as neural networks, for instance will generate many thousands of features within an image and use their relationship with the outcome to learn patterns of data that help to predict an outcome. In this context, what is most important is how the model behaves in training, testing and unseen external validation datasets, rather than how many features were used to build the model [44].

This thesis will focus on time-to-event analysis using Cox proportional hazards models, and the discussion of feature selection strategies will be limited to those used in the project but many others are available. Strategies can be divided into two major categories - supervised or unsupervised [136].

Supervised strategies include regression models with forwards or backward stepwise elimination or with least absolute shrinkage and selection operator (LASSO) penalty and random forests (RFo), which all choose predictors based on their association with the outcome of interest and the data labels, which need to be supplied to the feature selection model.

Unsupervised models, conversely, are not supplied with labelled data and determine the structure of the dataset, for example by collapsing multiple predictors into a single predictor, clustering similar features or removing features highly correlated with other features. Examples of unsupervised learning include principle component analysis (PCA) or hierarchical clustering [136].

Stepwise selection methods consider the addition (forwards selection), or the removal (backwards) of one feature at time from either a 'null' model, which contains no predictors (forward) or the 'full' model, containing all candidates (backwards) [140]. Features are included/removed based on a prespecified rule, which may be based on the significance result of the model ($p < 0.05$), or minimisation of the Akaike Information Criterion (AIC). AIC is a relative measure of how well the model fits the data and it also penalises the model for addition of extra predictors [140].

Backwards elimination is generally preferred over forwards, as it allows estimation of the regression coefficients after adjusting for all the other variables, and stepwise selection in general is popular because it is easy to implement, produces relatively consistent results for similar datasets and generally shrinks models to a small number of features [140]. Disadvantages are that it relies on multiple hypothesis testing, can lead to overfitting, inflated regression coefficients and model instability [137, 140].

Cox regression with a LASSO penalty can shrink predictor coefficients to zero and thus eliminate redundant features [137]. The optimal value for the penalty (referred to as lambda or $L1$) needs to be estimated by data resampling (typically with cross-validation). LASSO also has the benefit of shrinking all regression coefficients in the model, thereby reducing the optimism of the model (the difference between prediction accuracy in training versus test data) [137], but there is still the potential for overfitting and instability in small datasets [140].

RFo are a type of decision tree, which are similar in structure to clinical guidelines; starting at

the top of the tree, the prediction is determined using a set of simpler rules based on the available predictors, resulting in binary choices, and leading to an end point or leaf [136]. The advantage of RFo over simple decision trees is that multiple trees are built by only considering a random sample of the available data to make each tree. Novel predictions are based on the consensus result from all trees, and therefore less likely to be overfit. RFos also allow variables to be ranked in terms of their 'importance' by considering the difference in accuracy of trees built with and without that particular predictor [141], and therefore can be used for both prediction and feature selection.

Unsupervised feature reduction can be performed using PCA and hierarchical clustering. PCA is a dimensionality reduction technique that explains the maximum variance in the dataset using linear combinations ('principle components') of the original predictors. The most important principle components, those that explain most variance in the data, can then be examined to determine the contribution ('loadings') of the original predictors, and thereby allow reduction to the most important features for determining data variance without using the outcome or data labels [142].

Hierarchical clustering examines the distance between the feature values for individual patients; patients with similar RF values are grouped closer than those with differing ones, and by iteratively reordering the patients, groups with very similar feature values can be observed [136]. The features that vary most between the groups or clusters can then be determined to select the most important features for explaining variation across the population.

Once the optimal number of features have been chosen, Cox proportional hazards models can be trained with either the RFs, clinical features or both and then compared to see if there is any benefit of adding the novel IBs [134, 135]. Performance of such models can be assessed by the ability to split patients into prognostic groups (discrimination), which is measured using Harrell's concordance index (C-index) or Royston and Sauerbrei's D-statistic [143]. Calibration of the model compares the predicted risk from the model against the observed risk, and can be assessed with calibration plots and the slope of the curve [137]. These were the most commonly assessed metrics of glioblastoma prognostic model performance in a recent literature review of prognostic models for glioblastoma survival prediction [81], however the relative goodness of fit, measured by the AIC,

and the relative explained variation, measured with the model R^2 value also give complementary information and allow comparison of multiple competing models [144].

Model stability, the variability of the model performance due to differences in the patient population, is infrequently assessed but is also important [145]. This can be assessed with resampling of the training data, for example using bootstrapping or repeated cross-validation, and including as many of the model building steps as possible within the random subsample. Each resampled dataset is used to build a training model, and the performance tested against held back 'test' data. Thus, many test predictions are created and the variation in the accuracy across all of them allows researchers to assess the how predictions vary according to small changes in the input data [145].

To summarise, glioblastoma is a condition with poor outcome and published prognostic models try to stratify the risk of survival based on clinical predictors and also IBs derived from conventional MRI such as tumour focality, location, presence and extent of enhancement. Novel IBs might be able to quantify the intra-tumoural heterogeneity of the glioblastoma phenotype using a process of high throughput quantitative analysis, radiomics. The workflow of typical radiomics discovery or exploratory studies involves multiple steps and results in many, potentially thousands of putative IBs that then need to be reduced into a final model that can then be tested and potentially validated in other datasets, once a certain (smaller) number of features have been selected. RFs are then assessed for any added benefit when combined with clinical features used in traditional prognostic models. Many studies have suggested that there is evidence in favour of using RFs for prognostication and the next section will discuss reasons why this has not translated to clinical practice thus far.

1.4.3 Evidence of prognostic role for radiomic IBs

Many studies have reported prognostic value of RFs in survival prediction for patients with glioblastoma. Sun et al. used a Cox regression model with LASSO penalty in institutional and public data from The Cancer Imaging Archive (TCIA), to derive a 13-feature radiomic signature (train $n = 132$) using anatomical MRI sequences (T1W, T1CE, FLAIR and T2W) [146]. A dichotomised radiomics risk score demonstrated HR 3.7 (95% CI 2.1 - 6.5) in the test set (test $n = 66$). A radiogenomics

risk score was also developed (train $n = 95$) by correlating RNA-sequencing data with the 13 RFs. The combined (radiomics-genomics) risk score was dichotomised to label patients as either low- or high-risk and the combined score showed prognostic stratification in external test data (test $n = 78$) - HR 2.0 (95% CI 1.2 - 3.4) and C-index 0.60, although no comparison with a clinical model or assessment of calibration or model stability was performed [146]. 12% of cases were also IDHm tumours.

Kickingereder et al. ($n = 181$) demonstrated improved prognostic performance of Cox models for OS prediction in patients with glioblastoma (all IDHwt) using an 8-feature radiomic signature when combined with clinical predictors (age, KPS, extent of resection, adjuvant treatment, MGMT methylation) [147]. The integrated Brier score of the combined model in the test data ($n = 61$) reduced to 0.103 (0.133 for clinical only model); a lower score indicates better performance.

Choi et al. ($n = 120$, 14% IDHm) showed that including RFs into their prognostic model improved the integrated area under the receiver operator curve (AUC), which allows sensitivity and specificity calculation from censored survival data, from 0.65 (95% CI 0.64 - 0.69) for a clinical (age, gender, type of surgery, tumour location and post-operative treatment) and genetic (IDH mutation and MGMT methylation) model to 0.75 (95% CI 0.70 - 0.76) for a combined radiomic, genetic and clinical model (test $n = 35$) [148]. Chen et al. found similar results in 127 patients with glioblastoma (TCGA and TCIA data), with increased AUC of the combined clinical (age, KPS, radiotherapy status) and radiomic model 0.851 in the test set ($n = 42$) compared to the clinical only model (AUC 0.75) [149].

Using 119 patients from TCIA, Hajianfar et al. used RFs from filtered pre-operative T1CE and FLAIR sequences to produce a number of ML models to predict OS and reported a mean C-index in the test data (test $n = 36$) of 0.77 [150]. They did not compare this to a clinical only model or assess any other metrics of model performance. Verduin et al. demonstrated that a combined clinical, radiomics and VASARI feature model developed in 95 patients had a calibration slope of 0.79 in the external test data ($n = 38$), although no curve was presented for visual analysis and the slope of the clinical model was also not provided [151].

Rathore et al. used RFs to derive a three-category phenotypic classification of glioblastoma (discovery $n = 208$), which improved prognostic separation over using IDHm status alone in the test set (test $n = 53$, C-index 0.752 versus 0.559) [152]. In a separate comparison, the radiomic subtypes improved OS prediction when combined with patient age and tumour location (combined C-index 0.741 versus 0.671 or 0.608 for age and location alone, respectively). Hence, there are indicators that radiomics could have an additive prognostic relationship with OS, which may improve the accuracy of risk stratification in patients with glioblastoma. Attention now turns to the difficulties faced when translating radiomics to clinical practice.

1.5 Difficulties translating radiomic IBs

1.5.1 Robust biomarker development

1.5.1.1 Robustness

In order to translate into the clinic, IBs need to undergo technical, biological and clinical validation [84]. Technical validation means that an "IB measurement can be performed in any geographical location, whenever needed, and given comparable data" and requires assessment of the IBs repeatability, reproducibility, and availability [84]. Repeatability and reproducibility are part of a spectrum of steps used to determine IB precision and refer to consistency of measurements made in the same subject, using the same methods (equipment, processing, software) in a short space of time (repeatability) and consistency of measurement in different subjects, using different methods such as using different sites, scanners and software (reproducibility).

Availability could refer to feasibility and safety of a technique, or in the context of radiomics, could refer to the availability of a particular MR sequence used in a predictive signature or the ability to apply the radiomics signature across heterogeneous imaging acquisition, including differences in image quality (for instance, images degraded by artefact). Other steps before translation include biological and clinical validation, which assesses whether there is any relationship to an underlying

biological mechanism and to a clinical outcome, respectively and it is also important to assess potential clinical utility [84].

Cagney et al. suggest that biological validation is not strictly necessary, that amassing evidence of biomarker association with a therapeutic outcome or other endpoint, in the context of randomised clinical trials can be sufficient show that it is ready for clinical use [85]. Predictive power and evidence of accurate and generalisable model performance in multiple datasets could be sufficient to show that the biomarker should be used in the clinic, and this might be an attractive proposition for DL-based models, in particular as the underlying imaging features that form the basis of the model's output are more opaque than in radiomic analysis [153]. For DL networks, saliency maps, which highlight key areas of the input image that were used to inform the model prediction, may help with providing a biological surrogate, as these parts of a tumour, for instance, can be correlated with histopathological analysis or macroscopic changes in the tumour that might help explain the pattern of the network's predictions[153]. Others have suggested that IBs might be correlated with certain biological features such as gene expression profiles in a tumour and that the biological pathways that are up or downregulated can be correlated with IBs to obtain a surrogate marker of an IBs relationship to increased aggressiveness or shorter patient survival for instance [154].

Information from each of these steps can be collected and the risk and benefit profile of a biomarker can be determined for any given clinical scenario. In general, when an IB has amassed enough evidence, it can be said to have crossed a 'translational gap', either for use in hypothesis testing in medical research, or for making a treatment decision on a patient [84].

1.5.1.2 Barriers to repeatability and reproducibility

Despite the promising findings in exploratory studies of novel radiomic IBs in glioblastoma prognostic modelling, there has been a lack of clinical translation; many radiomic IBs lack repeatability and reproducibility [155, 156]. Radiomic IBs outside of neuro-oncology are, however, showing promise and clinical translation with commercially available tools now available. Software for determining the risk of malignancy associated with pulmonary nodules is undergoing trial evalaution in the

UK and is also being used across the US [157]. Outside of oncology, perivascular fat attenuation is the basis for an AI platform that has demonstrated improved accuracy of cardiovascular event prediction in patients without obstructive coronary artery disease in over 40,000 patients [158, 159].

RF values extracted from MRI are highly dependent upon acquisition and processing parameters including the manufacturer, vendor, geographic site, and even vary for the same person despite controlling all the aforementioned factors [111, 112, 127, 160]. Hence, multi-centre acquisitions will necessarily have a bearing on reproducibility of radiomics, unless imaging protocols can be homogenised [110].

The Quantitative Imaging Biomarker Alliance [161] and the European Imaging Biomarker Alliance [162] are trying to make IBs more robust by standardising image acquisition. If, however, retrospective or heterogeneous imaging data is to be used for RF analysis, reproducibility across sites and protocols, especially for intensity and texture features, is going to be hampered by arbitrary MRI signal intensity units, which lack a direct mapping to a physical tissue property [129]. Heterogeneity in data acquisition might help model building, as models will need to have greater generalisability to other datasets in the training phase, rather than just being produced from homogenous and identical images.

ISTs are usually applied to conventional MRI before RF extraction, so that the scale and distribution of intensities is more homogeneous across all the images (per sequence) acquired across different settings [111]. Although IBSI, which aims to improve RF reproducibility, discusses and provides reference instructions for most image pre-preprocessing steps [114], ISTs and discussion of the optimal approach for MRI is currently beyond the scope of the initiative. Accordingly, there is a lack of consensus on the optimal IST strategy and there are a number of options [111, 127, 163].

1.5.1.3 Inconsistent prognostic modelling

Inconsistent findings between radiomic studies could also be due to choices made during statistical modelling [137]. Dichotomisation of continuous predictors (for example age, tumour volume, RFs or risk scores) is neither necessary nor beneficial, as it reduces statistical power and is usually based

on an arbitrary value (such as the median) or one found through 'data-mining', which is likely to be optimistic and thus harder to replicate [137]. These should be kept as continuous predictors, and if necessary modelled as linear or non-linear variables, either through log-transformation or using spline functions [164].

RFs are often selected based on a statistically significant relationship to outcome ($p < 0.05$) using supervised feature selection techniques such as univariable screening [146], but this leads to a potentially exaggerated estimation of their effects, as they are likely included in the model because they are at a randomly high value [137].

Univariable screening is also problematic because predictors may change their prognostic effect after adjustment for other variables, and therefore univariable selection may discard important predictors, as well as leading to multiple hypothesis testing and exaggeration of their coefficients. The hypothesis test is also greatly dependent on the available sample size, and important predictors might be discarded (type 2 error) or included (type 1 error) due to inadequate statistical power.

Univariable screening also greatly increases the minimum sample size requirements. Several methods exist for calculation of the minimal sample size required based on the number of predictor variables [139] and these assume that the number of predictors in the calculation are all those that are *considered* for entry into the model and not just the ones that end up in the final model. Hence, by testing multiple features against the outcome, supervised strategies require a much higher magnitude of patients to avoid overfitting [139]. An unsupervised feature reduction strategy is therefore generally recommended in these settings, particularly when sample size is limited [137].

Stability, first mentioned during the discussion of feature selection, is also relevant to discussions on reproducibility because assessment of model stability should include as many of the model building steps as practically possible [145]. Feature scaling, selection, removal of correlated features and application of ComBat harmonisation for example ought to be considered as part of the model building process. In practice however, this is rarely applied and as a result lead to the creation of optimistic findings that are less easily reproduced [165]. Stability is akin to the term 'generalizability', which is more commonly found in ML literature and refers to the performance of a model on

'unseen' data [44]. A more generalizable model would have higher stability.

Calibration of models is often not reported in glioblastoma prognostic studies, and if it is, a calibration curve not presented [81, 146, 166]. Examining a calibration slope value (ideal value 1), or using a hypothesis test such as the Hosmer-Lemeshow test is not advised as neither inform about potential miscalibration or its extent for certain risk groups. Studying a calibration curve, on the other hand, and the stability of calibration across resampling is generally preferred and gives an easily interpretable result that does not require hypothesis testing [145].

Stability and calibration are also possible to assess in DL model performance. Classification accuracy (i.e. from a confusion matrix), for example, can be used to compare predicted and observed outcome (calibration) using a calibration curve, in a similar manner to evaluating calibration accuracy for survival or time-to-event models. Variability of AUC or any other model performance metric when input data is modified using bootstrapping or data resampling can be used to assess DL model stability.

A recent multicentre study by Dai et al. used approximately 35,000 brain MRIs and their radiology reports to train a DL model that could accurately classify the MRIs into 15 conditions or normal labels and they used bootstrap resampling (with 10,000 repetitions), to assess how changes in the input dataset affected the results of the model and to compute 95% CIs of model performance [167]. The network classified external test data with AUC 0.91 (95% CI $0.88 - 0.93$) and appeared well calibrated for multiple classification labels when visually assessing the calibration plots.

1.6 Project Aims

glioblastoma is a heterogeneous tumour, which has a poor but variable prognosis and accurate prognostic stratification may be helpful for individualised management. MRI is routinely used to characterise glioblastoma throughout a patient's journey and novel IB development using conventional MRI is an intense area of research interest. Extraction of and mining of RFs from conventional MRI has shown promising results for over a decade but has failed to move beyond exploratory, ret-

rospective studies. Recent, DL models have shown the ability to increase the accuracy of outcome prediction in retrospective and prospective data for patients with glioblastoma, albeit not with continuous outcome prediction [168]. Multi-centre, heterogeneous imaging may be one of the key barriers to this lack of reproducibility; MRI signal intensity standardisation and ComBat harmonisation could play a key role in tackling some of the inconsistencies introduced by variation in scan acquisition. Choices made in prognostic modelling stages may also hamper the ability to replicate findings and assess models adequately.

1. To systematically review the literature regarding the use of ISTs in the processing of diffuse glioma and glioblastoma MRI prior to the extraction of RFs, and determine the optimal strategy for this context.

2. To examine prognostic effect of tumour size, given it is one of the simplest quantitative imaging feature, in a large institutional cohort of patients with glioblastoma. A secondary aim will be to examine the effect of varying sample size and non-linear transformation will be assessed on the ability to reproduce these results.

3. To build on steps 1 and 2 and combine radiomic and clinical prognostic models to assess the impact of ISTs and ComBat realignment on prognostic model performance. Model performance will be assessed using various metrics including calibration, and model stability, explained variation and model fit.

Aims 1, 2, and 3 are addressed in Chapters 2, 3 and 4 respectively. Chapter 5 discusses the outcomes generated through Chapters 2-4 and examines the direction of future work.

References

1. Karschnia, P. *et al.* Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO resect group. *Neuro-Oncology* **25,** 940–954 (May 2023).

2. Stupp, R. *et al.* Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine* **352,** 987–996 (Mar. 2005).

3. Weller, M. *et al.* EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nature Reviews Clinical Oncology* **18,** 170–186 (2021).

4. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23,** 1231–1251 (Aug. 2021).

5. Smits, M. Update on neuroimaging in brain tumours. English. *Current Opinion in Neurology* **34,** 497–504 (Aug. 2021).

6. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magnetic Resonance Imaging* **30,** 1234–1248 (Nov. 2012).

7. Laug, D., Glasgow, S. M. & Deneen, B. A glial blueprint for gliomagenesis. *Nature Reviews Neuroscience* **19,** 393–403 (2018).

8. Currie, S. *et al.* A Comprehensive Clinical Review of Adult-Type Diffuse Glioma Incorporating the 2021 World Health Organization Classification. *Neurographics* **12,** 43–70 (Apr. 2022).

9. Sanai, N., Alvarez-Buylla, A. & Berger, M. S. Neural Stem Cells and the Origin of Gliomas. *New England Journal of Medicine* **353,** 811–822 (2005).

10. Lee, J. H. *et al.* Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560,** 243–247 (2018).

11. Varn, F. S. *et al.* Glioma progression is shaped by genetic evolution and microenvironment interactions. *Cell* **185,** 2184–2199.e16 (2022).

12. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178,** 835–849.e21 (2019).

13. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32,** 42–56.e6 (2017).

14. Apostoli, A. J. & Ailles, L. Clonal evolution and tumor-initiating cells: New dimensions in cancer patient treatment. *Critical Reviews in Clinical Laboratory Sciences* **53,** 40–51 (2016).

15. Arnold, C. R., Mangesius, J., Skvortsova, I. I. & Ganswindt, U. The Role of Cancer Stem Cells in Radiation Resistance. *Frontiers in Oncology* **10,** 1–12 (2020).

16. Inda, M. d. M., Bonavia, R. & Seoane, J. Glioblastoma multiforme:A look inside its heterogeneous nature. *Cancers* **6,** 226–239 (2014).

17. Galli, R. *et al.* Isolation and Characterization of Tumorigenic, Stem-like Neural Precursors from Human Glioblastoma. *Cancer Research* **64,** 7011–7021 (Oct. 2004).

18. Cheng, L., Bao, S. & Rich, J. N. Potential therapeutic implications of cancer stem cells in glioblastoma. *Biochemical Pharmacology* **80,** 654–665 (Sept. 2010).

19. Johnson, K. C. *et al.* Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. *Nature Genetics* **53,** 1456–1468 (Oct. 2021).

20. Verhaak, R. G. *et al.* Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).

21. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell. Comment in: Cancer Cell. 2010 May 18;17(5):419-20; PMID: 20478523 [https://www.ncbi.nlm.nih.gov/pubmed/20478523]* **17,** 510–522 (2010).

22. Zhang, P., Xia, Q., Liu, L., Li, S. & Dong, L. Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. *Frontiers in Molecular Biosciences* **7,** 1–13 (2020).

23. Lai, R. K. *et al.* Genome-wide methylation analyses in glioblastoma multiforme. *PLoS ONE* **9** (2014).

24. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155,** 462 (2013).

25. Lee, J. K. *et al.* Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nature Genetics* **49,** 594–599 (2017).

26. Aubry, M. *et al.* From the core to beyond the margin: a genomic picture of glioblastoma intratumor heterogeneity. *Oncotarget* **6,** 12094–12109 (May 2015).

27. Wood, M. D., Reis, G. F., Reuss, D. E. & Phillips, J. J. Protein analysis of glioblastoma primary and posttreatment pairs suggests a mesenchymal shift at recurrence. *Journal of Neuropathology and Experimental Neurology* **75,** 925–935 (2016).

28. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (June 2014).

29. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* **360,** 660–663 (May 2018).

30. Sahm, F. *et al.* Addressing diffuse glioma as a systemic brain disease with single-cell analysis. *Archives of Neurology* **69,** 523–526 (2012).

31. Petrecca, K., Guiot, M. C., Panet-Raymond, V. & Souhami, L. Failure pattern following complete resection plus radiotherapy and temozolomide is at the resection margin in patients with glioblastoma. *Journal of Neuro-Oncology* **111,** 19–23 (2013).

32. Yamahara, T. *et al.* Morphological and flow cytometric analysis of cell infiltration in glioblastoma: A comparison of autopsy brain and neuroimaging. *Brain Tumor Pathology* **27,** 81–87 (2010).

33. Șovrea, A. S. *et al.* Multiple Faces of the Glioblastoma Microenvironment. *International Journal of Molecular Sciences* **23,** 595 (Jan. 2022).

34. Hatoum, A., Mohammed, R. & Zakieh, O. The unique invasiveness of glioblastoma and possible drug targets on extracellular matrix. *Cancer Management and Research* **11,** 1843–1855 (2019).

35. Hambardzumyan, D., Gutmann, D. H. & Kettenmann, H. The role of microglia and macrophages in glioma maintenance and progression. *Nature Neuroscience* **19,** 20–27 (Jan. 2016).

36. Coniglio, S. J. *et al.* Microglial stimulation of glioblastoma invasion involves epidermal growth factor receptor (EGFR) and colony stimulating factor 1 receptor (CSF-1R) signaling. *Molecular medicine (Cambridge, Mass.)* **18,** 519–527 (2012).

37. Da Fonseca, A. C. C. *et al.* Increased expression of stress inducible protein 1 in glioma-associated microglia/macrophages. *Journal of Neuroimmunology* **274,** 71–77 (2014).

38. Saederup, N. *et al.* Selective chemokine receptor usage by central nervous system myeloid cells in CCR2-red fluorescent protein knock-in mice. *PLoS ONE* **5** (2010).

39. Wick, W., Platten, M. & Weller, M. Glioma cell invasion: Regulation of metalloproteinase activity by TGF-β. *Journal of Neuro-Oncology* **53,** 177–185 (2001).

40. Noorani, I. *et al.* Clinical impact of anti-inflammatory microglia and macrophage phenotypes at glioblastoma margins. *Brain Communications* **5,** 1–15 (2023).

41. Xia, S. *et al.* Tumor microenvironment tenascin-C promotes glioblastoma invasion and negatively regulates tumor proliferation. *Neuro-Oncology* **18,** 507–517 (2016).

42. Mahesparan, R. *et al.* Expression of extracellular matrix components in a highly infiltrative in vivo glioma model. *Acta Neuropathologica* **105,** 49–57 (2003).

43. Mentlein, R., Hattermann, K. & Held-Feindt, J. Lost in disruption: Role of proteases in glioma invasion and progression. *Biochimica et Biophysica Acta - Reviews on Cancer* **1825,** 178–185 (2012).

44. Wagner, M. W. *et al.* Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology* **63,** 1957–1967 (2021).

45. Beig, N., Bera, K. & Tiwari, P. Introduction to radiomics and radiogenomics in neuro-oncology: implications and challenges. *Neuro-Oncology Advances* **2,** iv3–iv14 (2020).

46. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica* **131,** 803–820 (2016).

47. Kleihues, P. *et al.* The WHO classification of tumors of the nervous system. *Journal of Neuropathology and Experimental Neurology* **61,** 215–225 (2002).

48. Van Den Bent, M. J. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: A clinician's perspective. *Acta Neuropathologica* **120,** 297–304 (2010).

49. Molinaro, A. M., Taylor, J. W., Wiencke, J. K. & Wrensch, M. R. Genetic and molecular epidemiology of adult diffuse glioma. *Nature Reviews Neurology* **15,** 405–417 (July 2019).

50. Cohen, A. L., Holmen, S. L. & Colman, H. IDH1 and IDH2 Mutations in Gliomas. *Current Neurology and Neuroscience Reports* **13,** 345 (May 2013).

51. Onizuka, H., Masui, K. & Komori, T. Diffuse gliomas to date and beyond 2016 WHO Classification of Tumours of the Central Nervous System. *International Journal of Clinical Oncology* **25,** 997–1003 (June 2020).

52. Alzial, G. *et al.* Wild-type isocitrate dehydrogenase under the spotlight in glioblastoma. *Oncogene* **41,** 613–621 (2022).

53. Calvert, A. E. *et al.* Cancer-Associated IDH1 Promotes Growth and Resistance to Targeted Therapies in the Absence of Mutation. *Cell Reports* **19,** 1858–1873 (May 2017).

54. Wahl, D. R. *et al.* Glioblastoma Therapy Can Be Augmented by Targeting IDH1-Mediated NADPH Biosynthesis. *Cancer Research* **77,** 960–970 (Feb. 2017).

55. Metallo, C. M. *et al.* Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature* **481,** 380–384 (Jan. 2012).

56. Tönjes, M. *et al.* BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH11. Tönjes, M. et al. BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH1. Nat. Med. 19, 901–908 (2013).1. Tönj. *Nature Medicine* **19,** 901–908 (2013).

57. Maus, A. & Peters, G. J. Glutamate and α-ketoglutarate: key players in glioma metabolism. *Amino Acids* **49,** 21–32 (2017).

58. Wise, D. R. *et al.* Hypoxia promotes isocitrate dehydrogenasedependent carboxylation of α-ketoglutarate to citrate to support cell growth and viability. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 19611–19616 (2011).

59. May, J. L. *et al.* IDH3α regulates one-carbon metabolism in glioblastoma. *Science Advances* **5,** 1–15 (2019).

60. Brat, D. J. *et al.* cIMPACT-NOW update 3: recommended diagnostic criteria for "Diffuse astrocytic glioma, IDH-wildtype, with molecular features of glioblastoma, WHO grade IV". *Acta Neuropathologica* **136,** 805–810 (Nov. 2018).

61. Libermann, T. A. *et al.* Amplification, enhanced expression and possible rearrangement of EGF receptor gene in primary human brain tumours of glial origin. *Nature* **313,** 144–147 (1985).

62. Rude Voldborg, B., Damstrup, L., Spang-Thomsen, M. & Skovgaard Poulsen, H. Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. *Annals of Oncology* **8,** 1197–1206 (Dec. 1997).

63. Talasila, K. M. *et al.* EGFR wild-type amplification and activation promote invasion and development of glioblastoma independent of angiogenesis. *Acta Neuropathologica* **125,** 683–698 (2013).

64. Stichel, D. *et al. Distribution of EGFR amplification, combined chromosome 7 gain and chromosome 10 loss, and TERT promoter mutation in brain tumors and their potential for the reclassification of IDHwt astrocytoma to glioblastoma* **5,** 793–803 (2018).

65. Fares, J., Wan, Y., Mair, R. & Price, S. J. Molecular diversity in isocitrate dehydrogenase-wild-type glioblastoma. *Brain Communications* **6,** 1–17 (Mar. 2024).

66. Brodbelt, A. *et al.* Glioblastoma in England: 2007-2011. *European Journal of Cancer* **51,** 533–542 (2015).

67. Research Surveillance Program National Cancer Institute. *SEER*Explorer: An interactive website for SEER cancer statistics* 2024.

68. Philips, A., Henshaw, D. L., Lamburn, G. & O'Carroll, M. J. Brain tumours: Rise in glioblastoma multiforme incidence in England 1995-2015 Suggests an Adverse Environmental or Lifestyle Factor. *Journal of Environmental and Public Health* **2018** (2018).

69. Wanis, H. A., Møller, H., Ashkan, K. & Davies, E. A. The incidence of major subtypes of primary brain tumors in adults in England 1995-2017. *Neuro-Oncology* **23,** 1371–1382 (2021).

70. Thakkar, J. P. *et al.* Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiology Biomarkers and Prevention* **23,** 1985–1996 (2014).

71. Davis, M. E. Glioblastoma: Overview of disease and treatment. *Clinical Journal of Oncology Nursing* **20,** 1–8 (2016).

72. Ostrom, Q. T. *et al.* CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2011-2015. *Neuro-Oncology* **20,** iv1–iv86 (2018).

73. Sadetzki, S. *et al.* Long-Term Follow-up for Brain Tumor Development after Childhood Exposure to Ionizing Radiation for Tinea Capitis. *Radiation Research* **163,** 424–432 (Apr. 2005).

74. Johnson, D. R. *et al.* Case-based review: Newly diagnosed glioblastoma. *Neuro-Oncology Practice* **2,** 106–121 (2015).

75. Stupp, R. *et al.* 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 (2009).

76. Gerson, S. L. MGMT: its role in cancer aetiology and cancer therapeutics. *Nature Reviews Cancer* **4,** 296–307 (Apr. 2004).

77. Wick, W. *et al.* MGMT testing—the challenges for biomarker-based glioma treatment. *Nature Reviews Neurology* **10,** 372–385 (July 2014).

78. Wick, W. *et al.* Prognostic or predictive value of MGMT promoter methylation in gliomas depends on IDH1 mutation. *Neurology* **81,** 1515–1522 (Oct. 2013).

79. Currie, S. *et al.* Imaging Spectrum of the Developing Glioblastoma: A Cross-Sectional Observation Study. *Current Oncology* **30,** 6682–6698 (July 2023).

80. Senders, J. T. *et al.* An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Clinical Neurosurgery* **86,** E184–E192 (2020).

81. Tewarie, I. A. *et al.* 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 (2021).

82. Mowforth, O. D. *et al.* Personalised therapeutic approaches to glioblastoma: A systematic review. *Frontiers in Medicine* **10,** 1–14 (2023).

83. Food and Drug Administration & National Institutes of Health. *BEST (Biomarkers, EndpointS, and other Tools) Resource*

84. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology* **14,** 169–186 (Mar. 2017).

85. Cagney, D. N. *et al.* The FDA NIH Biomarkers, EndpointS, and other Tools (BEST) resource in neuro-oncology. *Neuro-Oncology* **20,** 1162–1172 (2018).

86. Feng, S. H. & Yang, S. T. The new 8th tnm staging system of lung cancer and its potential imaging interpretation pitfalls and limitations with ct image demonstrations. *Diagnostic and Interventional Radiology* **25,** 270–279 (2019).

87. D'Orsi, C., Sickles, E., Mendelson, E., Morris, E. & Al., E. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.* (Reston, VA, American College of Radiology, 2013).

88. Gutman, D. A. *et al.* MR imaging predictors of molecular profile and survival: Multi-institutional study of the TCGA glioblastoma data set. *Radiology* **267,** 560–569 (2013).

89. Wen, P. Y. *et al.* Glioblastoma in adults: A Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro-Oncology* **22,** 1073–1113 (2020).

90. Smits, M. MRI biomarkers in neuro-oncology. *Nature Reviews Neurology* **17,** 486–500 (2021).

91. Smits, M. Imaging of oligodendroglioma. *The British Journal of Radiology* **89,** 20150857 (Apr. 2016).

92. Pope, W. B. *et al.* MR imaging correlates of survival in patients with high-grade gliomas. *American Journal of Neuroradiology* **26,** 2466–2474 (2005).

93. Hammoud, M. A. Prognostic significance of preoperative MRI scans in glioblastoma multiforme. *Journal of Neuro-Oncology* **27,** 65–73 (1996).

94. Wangaryattawanich, P. *et al.* Multicenter imaging outcomes study of The Cancer Genome Atlas glioblastoma patient cohort: Imaging predictors of overall and progression-free survival. *Neuro-Oncology* **17,** 1525–1537 (2015).

95. Wiki for the VASARI feature set. *VASARI Research Project* 2012.

96. Li, Y. *et al.* A systematic review of multifocal and multicentric glioblastoma. *Journal of Clinical Neuroscience* **83,** 71–76 (2021).

97. Lasocki, A., Anjari, M., rs Kokurcan, S. & Thust Arian; ORCID: http://orcid.org/0000-0001-8176-3015, S. C. A. O. .-. L. Conventional MRI features of adult diffuse glioma molecular subtypes: a systematic review. English. *Neuroradiology* **63,** 353–362 (2021).

98. Galldiks, N., Lohmann, P., Fink, G. R. & Langen, K.-J. Amino Acid PET in Neurooncology. *Journal of Nuclear Medicine* **64,** 693–700 (May 2023).

99. Soni, N. *et al.* Amino Acid Tracer PET MRI in Glioma Management: What a Neuroradiologist Needs to Know. *American Journal of Neuroradiology* **44,** 236–246 (Mar. 2023).

100. Thust, S. C. *et al.* Glioma imaging in Europe: A survey of 220 centres and recommendations for best clinical practice. *European Radiology* **28,** 3306–3317 (Aug. 2018).

101. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. English. *Radiology* **278,** 563–577 (Feb. 2016).

102. O'Connor, J. P. *et al.* Imaging intratumor heterogeneity: Role in therapy response, resistance, and clinical outcome. *Clinical Cancer Research* **21,** 249–257 (2015).

103. Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C. & Mohan, S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology* **290,** 607–618 (Mar. 2019).

104. Carré, A. *et al.* AutoComBat: a generic method for harmonizing MRI-based radiomic features. *Scientific Reports* **12,** 1–17 (2022).

105. Ellingson, B. M. *et al.* Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro-Oncology* **17,** 1188–1198 (Aug. 2015).

106. British Society of Neuroradiologists Standards Subcommittee. *Core imaging protocol for brain tumours* tech. rep. (British Society of Neuroradiologists, 2018).

107. Mayerhoefer, M. E., Szomolanyi, P., Jirak, D., Materka, A. & Trattnig, S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study. *Medical Physics* **36,** 1236–1243 (2009).

108. Molina, D. *et al.* Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization. *PLoS ONE* **12,** 1–14 (2017).

109. Huisman, M. & Akinci D'Antonoli, T. What a Radiologist Needs to Know About Radiomics, Standardization, and Reproducibility. *Radiology* **310,** e232459 (Feb. 2024).

110. Da-Ano, R., Visvikis, D. & Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine & Biology* **65,** 24TR02 (Dec. 2020).

111. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10,** 12340 (July 2020).

112. Hoebel, K. V. *et al.* Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence* **3,** e190199 (2021).

113. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine and Biology* **60,** 5471–5496 (2015).

114. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative (2016).

115. Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging* **11** (2020).

116. Tustison, N. J. *et al.* N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging* **29,** 1310–1320 (June 2010).

117. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

118. Pati, S. *et al.* Federated Learning Enables Big Data for Rare Cancer Boundary Detection (2022).

119. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14,** 749–762 (Oct. 2017).

120. Sanduleanu, S. *et al.* Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiotherapy and Oncology* **127,** 349–360 (2018).

121. Fatania, K. *et al.* Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review. *European Radiology* **32,** 7014–7025 (Oct. 2022).

122. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* **19,** 143–150 (2000).

123. Shah, M. *et al.* Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis* **15,** 267–282 (2011).

124. Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9–19 (2014).

125. Sweeney, E., Shinohara, R., Shea, C., Reich, D. & Crainiceanu, C. Automatic Lesion Incidence Estimation and Detection in Multiple Sclerosis Using Multisequence Longitudinal MRI. *American Journal of Neuroradiology* **34,** 68–73 (Jan. 2013).

126. Reinhold, J. C., Dewey, B. E., Carass, A. & Prince, J. L. Evaluating the impact of intensity normalization on MR image synthesis, 126 (2019).

127. Li, Y., Ammari, S., Balleyguier, C., Lassau, N. & Chouzenoux, E. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features. *Cancers* **13,** 1–22 (2021).

128. Van Griethuysen, J. J. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **77,** e104–e107 (Nov. 2017).

129. Orlhac, F. *et al.* How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European Radiology* **31,** 2272–2280 (Apr. 2021).

130. Orlhac, F. *et al.* A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine* **63,** 172–179 (Feb. 2022).

131. Fortin, J.-P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167,** 104–120 (Feb. 2018).

132. Fortin, J. P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161,** 149–170 (2017).

133. Da-ano, R. *et al.* Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports* **10,** 1–12 (2020).

134. Dankers, F. J. W. M., Traverso, A., Wee, L. & van Kuijk, S. M. J. in *Fundamentals of Clinical Data Science* (eds Kubben, P., Dumontier, M. & Dekker, A.) 101–120 (Springer International Publishing, Cham, 2019).

135. Parmar, C., Barry, J. D., Hosny, A., Quackenbush, J. & Aerts, H. J. Data analysis strategies in medical imaging. *Clinical Cancer Research* **24,** 3492–3499 (2018).

136. Traverso, A., Dankers, F. J. W. M., Osong, B., Wee, L. & van Kuijk, S. M. J. in (eds Kubben, P., Dumontier, M. & Dekker, A.) 121–133 (Springer International Publishing, Cham, 2019).

137. Riley, R. D., van der Windt, D., Croft, P. & Moons, K. G. *Prognosis Research in Healthcare: Concepts, Methods, and Impact* (eds Riley, R. D., van der Windt, D. A., Croft, P. & Moons, K. G.) (Oxford University Press, Oxford, UK, 2019).

138. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368,** m441 (Mar. 2020).

139. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* **38,** 1276–1296 (Mar. 2019).

140. Steyerberg, E. W. *Clinical prediction models : a practical approach to development, validation, and updating* (Springer, New York ; 2009).

141. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Annals of Applied Statistics* **2,** 841–860 (2008).

142. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Computational Statistics* **2,** 433–459 (July 2010).

143. Austin, P. C., Pencinca, M. J. & Steyerberg, E. W. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Statistical Methods in Medical Research* **26,** 1053–1077 (2017).

144. Harrell, F. *Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements* 2023.

145. Riley, R. D. & Collins, G. S. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal* **65,** 1–22 (2023).

146. Sun, Q. *et al.* Biologic Pathways Underlying Prognostic Radiomics Phenotypes from Paired MRI and RNA Sequencing in Glioblastoma. *Radiology,* 203281 (2021).

147. Kickingereder, P. *et al.* Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-Oncology* **20,** 848–857 (2018).

148. Choi, Y. *et al.* Radiomics may increase the prognostic value for survival in glioblastoma patients when combined with conventional clinical and genetic prognostic models. *European radiology* **31,** 2084–2093 (2021).

149. Chen, X. *et al.* Development and Validation of a MRI-Based Radiomics Prognostic Classifier in Patients with Primary Glioblastoma Multiforme. *Academic Radiology* **26,** 1292–1300 (2019).

150. Hajianfar, G. *et al.* Time-to-event overall survival prediction in glioblastoma multiforme patients using magnetic resonance imaging radiomics. *Radiologia Medica* **128,** 1521–1534 (2023).

151. Verduin, M. *et al.* Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma. *Cancers* **13,** 1–20 (2021).

152. Rathore, S. *et al.* Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Scientific Reports* **8,** 1–12 (2018).

153. Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine* **140,** 105111 (Jan. 2022).

154. Tomaszewski, M. R. & Gillies, R. J. The biological meaning of radiomic features. *Radiology* **298,** 505–516 (2021).

155. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics* **102,** 1143–1158 (2018).

156. Pinto dos Santos, D., Dietzel, M. & Baessler, B. A decade of radiomics research: are images really data or just patterns in the noise? *European Radiology* **31,** 2–5 (2021).

157. O'Dowd, E. *et al.* Determining the impact of an artificial intelligence tool on the management of pulmonary nodules detected incidentally on CT (DOLCE) study protocol: a prospective, non-interventional multicentre UK study. *BMJ Open* **14,** 1–6 (2024).

158. Chan, K. *et al.* Inflammatory risk and cardiovascular events in patients without obstructive coronary artery disease: the ORFAN multicentre, longitudinal cohort study. *The Lancet* **403,** 2606–2618 (2024).

159. Oikonomou, E. K. *et al.* Non-invasive detection of coronary inflammation using computed tomography and prediction of residual cardiovascular risk (the CRISP CT study): a post-hoc analysis of prospective outcome data. *The Lancet* **392,** 929–939 (2018).

160. Baeßler, B., Weiss, K. & Santos, D. P. D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Investigative Radiology* **54,** 221–228 (2019).

161. Shukla-Dave, A. *et al.* Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *Journal of Magnetic Resonance Imaging* **49,** e101–e121 (June 2019).

162. DeSouza, N. M. *et al.* Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: current status and recommendations from the EIBALL* subcommittee of the European Society of Radiology (ESR). *Insights into Imaging* **10,** 87 (Dec. 2019).

163. Salome, P. *et al.* MR Intensity Normalization Methods Impact Sequence Specific Radiomics Prognostic Model Performance in Primary and Recurrent High-Grade Glioma. *Cancers* **15** (2023).

164. Harrell, F. E. *Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis* Second edi (Springer, Cham, 2015).

165. Demircioğlu, A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights into Imaging* **12,** 172 (Dec. 2021).

166. Compter, I. *et al.* Deciphering the glioblastoma phenotype by computed tomography radiomics. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **160,** 132–139 (2021).

167. Dai, L. *et al.* Boosting Deep Learning for Interpretable Brain MRI Lesion Detection through the Integration of Radiology Report Information. *Radiology: Artificial Intelligence* **6** (2024).

168. Chelliah, A. *et al.* Glioblastoma and radiotherapy: A multicenter AI study for Survival Predictions from MRI (GRASP study). *Neuro-Oncology* **26,** 1138–1151 (June 2024).

CHAPTER

TWO

INTENSITY STANDARDISATION OF MRI PRIOR TO RADIOMIC FEATURE EXTRACTION FOR ARTIFICIAL INTELLIGENCE RESEARCH IN GLIOMA – A SYSTEMATIC REVIEW

2.1 Abstract

2.1.1 Background

Radiomics is a promising avenue in non-invasive characterisation of diffuse glioma. Clinical translation is hampered by lack of reproducibility across centres and difficulty in standardising image intensity in MRI datasets. The study aim was to perform a systematic review of different methods of MRI intensity standardisation prior to radiomic feature extraction.

2.1.2 Methods

MEDLINE, EMBASE, and SCOPUS, were searched for articles meeting the following eligibility criteria: MRI radiomic studies where one method of intensity standardisation was compared with another or no standardisation, and original research concerning patients diagnosed with diffuse gliomas. Using PRISMA criteria, data were extracted from short-listed studies including number of patients, MRI sequences, validation status, radiomics software, method of segmentation and intensity standardisation. QUADAS-2 was used for quality appraisal.

2.1.3 Results

After duplicate removal, 743 results were returned from database and reference searches and from these, 12 papers were eligible. Due to a lack of common pre-processing and different analyses, a narrative synthesis was sought. 3 different intensity standardisation techniques have been studied: histogram matching (5/12), limiting or rescaling signal intensity (8/12), and deep learning (1/12) - only two papers compared different methods. Histogram matching produced the highest area under the receiver-operator curves but these studies lacked direct comparison to other methods.

2.1.4 Conclusion

Multiple methods of intensity standardisation have been described in the literature without clear consensus. Further research that directly compares different methods of intensity standardisation on glioma MRI datasets is required.

2.2 Introduction

Radiomic evaluation of glioblastoma and other adult-type diffuse gliomas has thus far failed to translate to clinical practice in part due to non-biological, scanner-dependent variation in image signal intensity [1–4]. Signal intensity for structural MR, such as T1W or T2W MRI, does not map

easily to a physical tissue property, in contrast to CT, and shows variation between timepoints, vendors, magnetic field strengths and acquisition settings [5–8]. RFs are highly sensitive to the values of the signal intensities in the image, and non-biological alteration must be standardised; the range and distribution of voxel intensity must be similar across patients, prior to radiomic analysis to ensure that the results are reproducible [1]. Despite this, there is a lack of consensus as to the optimal method when characterising diffuse glioma. Although not a specific diagnosis, diffuse glioma is a useful grouping, as they often share the same radiomics pipeline and are a commonly studied group of related tumours [3, 6].

The aim, therefore, was to perform a systematic review of the literature examining the efficacy of different MRI ISTs prior to the extraction of RFs in the setting of adult-type diffuse glioma.

2.3 Materials and Methods

2.3.1 Search strategy and selection criteria

This systematic review was undertaken according to the 'Preferred Reporting Items for Systematic Reviews and Meta-Analysis' (PRISMA) statement [9]. A search of MEDLINE, EMBASE and SCOPUS databases was performed on 5th October 2021 using the following concepts, linked by the "AND" operator, including synonymous terms that were linked with the "OR" operator: (1) MRI, (2) radiomics, artificial intelligence or deep learning, (3) intensity standardisation, and (4) glioma. Full search strategy and PRISMA checklist are available in section 2.7.

No limit was placed on the date, language, location, or type of study. Inclusion criteria were: original research article, adult ($\geq$ 16 years old) patients with diagnosis of adult-type diffuse glioma, application of IST to imaging prior to extraction of RFs, comparison of the effect of one IST to either no standardisation or another IST, and RFs extracted from images. Exclusion criteria were: non-human studies, not regarding adult-type diffuse gliomas, non-original research, non MR radiomics, no mention of IST, or no assessment of the effect of an IST (compared to another or to

no standardisation).

Records were managed using citation management software, and automatic duplicate removal was used to screen the results. Two reviewers independently and manually reviewed the titles and abstracts and subsequently the full texts to determine if they satisfied the inclusion and exclusion criteria. Any disagreement was resolved by consensus. References in the included articles were manually reviewed.

2.3.2 Data-extraction

The primary outcome for the study was assessment of the efficacy of ISTs. The measurement of efficacy was not restricted to any statistic or method due to lack of an agreed standard. If studies developed a predictive or diagnostic model, any reported model performance statistics were extracted. For any studies that did not develop a predictive model, the reported efficacy of standardisation was extracted. All effect measures were agreed by consensus between three reviewers.

Meta-analysis was precluded by heterogeneity of the included studies and therefore a narrative synthesis was presented. In the narrative synthesis, the studies were grouped based on chosen IST. No preparation or processing of data within the included studies was undertaken – results of studies were included as they appeared in published manuscripts.

Since a meta-analysis was not conducted, additional methods such as sensitivity analyses or subgroup analyses were not performed. Similarly, the impact of missing results was not relevant as we included any outcome measure presented in the studies. No additional methods were used to assess certainty or confidence in the outcome, other than qualitative assessment of each study during the narrative synthesis. The systematic review was not prospectively registered.

2.3.3 Quality assessment

Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) was used to assess the risk of bias [10]. QUADAS-2 was used because the objective was to evaluate performance of any given IST,

when compared to either no standardisation or to another IST. QUADAS-2 assesses four domains: (1) patient selection – description of how patients were recruited such as inclusion and exclusion criteria, (2) index test – how the index test was conducted and interpreted, (3) reference standard – how the reference test was conducted and interpreted, and (4) flow and timing – patients that did not have the index or reference test or were excluded from final analysis. Each domain was assessed for risk of bias and the first three domains were also assessed for applicability and categorised as either low risk, high risk, or unclear. The index test was taken to be the IST under investigation, and the reference test was either no standardisation or an alternative IST. Two reviewers independently reviewed each study and any disagreement resolved by consensus.

2.4 Results

2.4.1 Search results

After duplicate removal, 741 results were returned from database searches (**Figure 2.1**). Following title and abstract screening, full text screening was undertaken for 60 articles and 12 articles met the inclusion criteria. Two studies by Florez et al.[11, 12] were both included as separate studies as one used only RFs from FLAIR sequences [12], whereas the other extracted features from mpMRI [11] and this may have an impact upon the results of any IST.

2.4.2 Quality assessment

Risk of bias was assessed for each of the four domains and applicability assessed for the first three domains in the QUADAS-2 framework [10]. Apart from the patient selection domain and applicability concern for the index test, all other domains were low risk for all studies (**Table 2.1**). 10 studies were deemed to have unclear risk due to lack of information on how patients were selected. It was unclear whether institutional patients were selected consecutively or randomly or, if publicly available datasets were used, it was unclear whether any inclusion or exclusion criteria

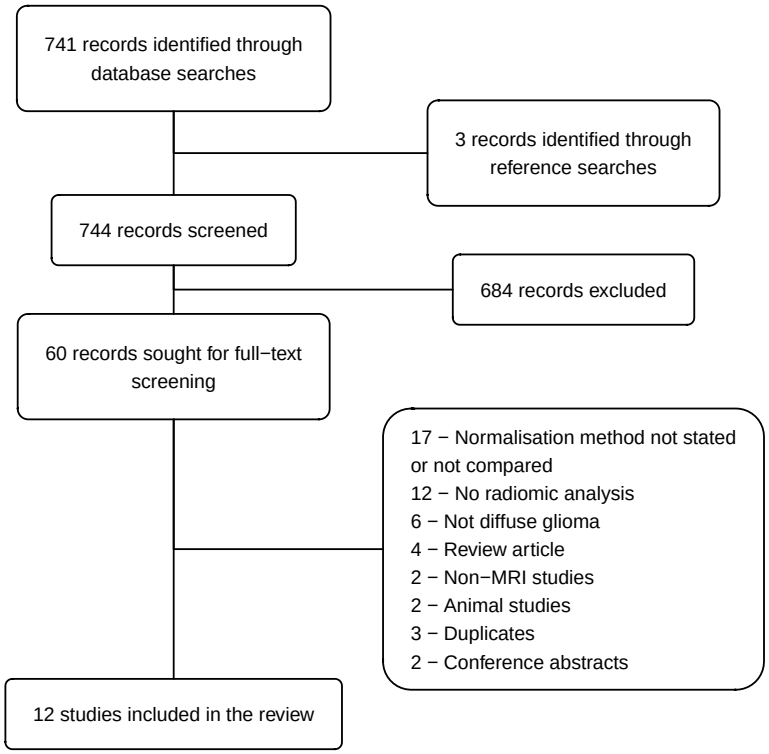

Figure 2.1: Flow chart illustrates the study selection for the systematic review of intensity standardisation techniques of MRI in diffuse glioma radiomic studies.

Table 2.1: Summary of the risk of bias and applicability concerns for the 12 studies included in the systematic review.

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Chen et al. 2019	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Zhao et al. 2020	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Reuze et al. 2018	UNCLEAR	LOW	LOW	LOW	LOW	HIGH	LOW
Um et al. 2019	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Upadhaya et al. 2016	UNCLEAR	LOW	LOW	LOW	LOW	HIGH	LOW
Florez et al. 2018	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Florez et al. 2018	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Hu et al. 2021	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Hoebel et al. 2021	LOW	LOW	LOW	LOW	LOW	LOW	LOW
Vils et al. 2021	LOW	LOW	LOW	LOW	LOW	LOW	LOW
Carré et al. 2020	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW
Orlhac et al. 2020	UNCLEAR	LOW	LOW	LOW	LOW	LOW	LOW

were used to select patients.

For applicability concerns of the index test, two studies [13, 14] were deemed high risk because it was not possible to isolate the effects of standardisation from other pre-processing. Two studies [15, 16] were low risk in all domains. Two studies by Florez et al.[11, 12] also included patients with meningioma, but were not thought to be at risk of bias or an applicability concern as the results for the glioblastoma patients were presented separately.

2.4.3 Characteristics of included studies

Significant heterogeneity in the pre-processing steps and in analysis methodology (**Table 2.2**), precluded a meta-analysis and a narrative synthesis is presented.

Table 2.2: Summary of studies included in the review

First Author	Study aims	Sample size (train/test)[a]	MRI sequences	Standardisation method	Image processing	Segmentation	Radiomics software	Results	Conclusions
Chen et al. 2019	Assess the impact of HSASR standardisation on glioma grading accuracy using radiomics	521 (416:105)	T1CE	HSASR method	Skull stripping and spatial resampling	Manual	Pyradiomics	Glioma grading AUC with HSASR = 0.9934 (0.8512 without). AUCs with HSASR generally increased 15%.	Multicentre data processed by HSASR standardisation improves grading and has value for clinical prediction.
Zhao et al. 2020	Assess the impact of HS-GS standardisation on glioma grading accuracy using radiomics	693 (554:139)	T1CE	HS-GS method	Skull stripping and spatial resampling	Manual	Pyradiomics	Glioma grading AUC with HG-GS = 0.956 (26.96% higher than without)	HS-GS improves accuracy of radiomics diagnostic models for glioma grading.
Reuze et al. 2018	To assess intensity rescaling on robustness of multicentre radiomic analysis	190 (n/a)	T1CE	Intensity rescaling	Spatial resampling and grey level discretisation	Manual	LIFEx freeware	11/31 texture features were robust after the standardisation.	Standardisation was not sufficient for correcting the differences between images.
Um et al. 2019	Assess the impact of pre-processing methods on MRI radiomic feature robustness across multi-institutional datasets	161 (111: 47)	FLAIR, T1W, and T1CE	Histogram standardisation	Co-registration	Semi-automatic	Computational Environment for Radiotherapy Research	Histogram matching had the greatest impact on robustness as shown by a significant decrease of Matthews correlation coefficient.	Histogram standardisation had the biggest contribution on reducing feature dependence on scanner variability.

Upadhaya et al. 2016	Impact of pre-processing steps on the prognostic accuracy of a binary classification model (survival above and below median of 12 months)	58 (58:58)[b]	T1W, T2W, T1CE, and FLAIR	Dynamic intensity limitation	Bias field correction, skull stripping, co-registration, spatial resampling, and grey level discretisation	Automatic	Not identified	Prognostic model sensitivity and specificity respectively increased with pre-processing to 93% (79% and 86% without, respectively)	Acquisition methods from different MR scanners can influence the accuracy of prognostic models and pre-processing steps can help reduce this.
Florez et al. 2018	To assess the accuracy of radiomic features in differentiating GTV from oedema and differentiating vasogenic oedema from tumour cell infiltration	17 (17;n/a)	T1W, T1CE, T2W, FLAIR and ADC	1%-99% normalisation	Segmentation	Semi-automatic	MatLab version 2016a	T1CE with 1%-99% normalisation had the highest accuracy for tumour classification with an AUC 0.97.	Only a small subset of the many radiomic features extracted, showed ability to classify tumour tissue.
Florez et al. 2018	To assess the ability of FLAIR radiomic features to distinguish oedema and infiltrative tumour	20 (20;n/a)	FLAIR	1%-99% normalisation	Segmentation	Semi-automatic	MatLab version 2016a	AUC with standardisation = 0.87 (0.84 without)	Small number of texture features can discriminate oedema from tumour.

Hu et al. 2021	Impact of MIL standardisation on tumour segmentation accuracy and on the accuracy of glioma grading and IDH1 status classification using RFs	800 (533:267)	T1W, T1CE and FLAIR for all of the datasets (and T2W for the BraTs dataset, n=285)	CycleGAN	MIL standardisation	Automatic	Not identified	MIL standardisation improved the AUC of pathological grading and IDH1 status prediction by 32% and 25% ($p<0.001$) respectively. Grading AUC with standardisation = 0.89 (0.69 without) and IDH1 mutation AUC with standardisation = 0.91 (0.70 without).	MIL standardisation results in higher quality data for radiomic analysis.
Hoebel et al. 2021	To assess the impact of z-score and histogram matching, and grey level discretization on the repeatability and reproducibility of features extracted from a scan-rescan GBM cohort	48 (n/a)	T1CE and FLAIR	Z-score and histogram matching	Segmentation, registration, bias field correction and whole brain extraction	Manual	Pyradiomics	Z-score and histogram matching improved the repeatability of FLAIR radiomics ($p=0.003$ for z-score and $p=0.002$ for histogram matching compared to baseline). T1CE radiomics did not show a significant result.	Standardisation methods improved repeatability for FLAIR images but less for T1CE images, and may be due to differences in timing of contrast injection.

Vils et al. 2021	To assess the ability of radiomic features to predict clinical outcome and molecular characteristics such as MGMT status	118 (69:49)	T1CE	Linear intensity interpolation	Segmentation and manual extraction of brain tissue	Manual	Z-Rad	MGMT status prediction using radiomic with standardisation showed diagnostic accuracy in an independent cohort, AUC = 0.670 (95% CI 0.5341-0.8056).	The proposed model may be a non-invasive approach to predict patient response to chemotherapy.
Carré et al. 2020	Impact of three intensity standardisation methods and grey level discretization on glioma grading	263 (195:48)	T1CE and FLAIR	Nyul, WhiteStripe a Z score normalisation method	Bias field correction, spatially resampled, skull-stripping, co-registration and segmentation	Manual	Pyradiomics	Glioma grading using T1CE radiomic features was improved using histogram matching (AUC 0.82), WhiteStripe (0.79) and z-score (0.82) compared to no standardisation (0.67). Relative discretization made intensity standardisation of second-order features unnecessary.	Models based on second-order features do not need intensity standardisation, if relative discretization is adopted

Orlhac et al. 2020	To assess the impact of intensity standadrisation and post-extraction realignment (ComBat) on the statistical distribution of radiomics from diffuse gliomas	18	T1CE and FLAIR	Hybrid WhiteStripe (and ComBat)	Co-registration, bias field correction, spatial resampling	Manual	LIFEx freeware	WhiteStripe reduced the number of features that were significantly different between acquisitions (88% reduced to 69% for normal white matter, and 98% to 60% for tumour radiomic features).	Intensity standardisation results in similar intensity values in images, but significant scanner dependent changes remain.

ADC = Apparent Diffusion Coefficient; AUC = Area under the receiver-operator curve; BraTs = Brain Tumor Segmentation; FLAIR = Fluid Attenuated Inversion Recovery; GAN = generative adversarial network; GBM = Glioblastoma; GTV = Gross Tumour Volume; HSASR = Histogram-specification with automatic selection of reference; HS-GS – Histogram-specification grid search; IDH1 = Isocitrate dehydrogenase 1; MGMT = 6-O-Methylguanine-DNA Methyltransferase; MIL = Modality incompleteness modality incompleteness, uneven intensity distribution and inconsistent layer spacing; RF = radiomic feature; T1W = T1-weighted MRI; T1CE = T1-weighted MRI gadolinium contrast-enhanced; T2W = T2-weighted MRI [a] Train/test numbers are only stated for any predictive model developed in the study; 'n/a' stated if no model was developed [b] Model developed using leave one out cross-validation, according to stated references in the study

All studies were retrospective, although two studies [15, 16] utilised prospectively acquired trial data. Eight included multicentre data, and for one [13], it was unclear whether data comprised single or multicentre data. Six studies used a publicly available multicentre dataset from TCIA [17], or competition data from the brain tumour image segmentation benchmark (BraTs) [18], with five of these also using institutional data. Only one study[13] used solely public data.

Aims of the studies could be divided into two groups:

1. To assess the impact of intensity standardisation on the reproducibility and repeatability of RFs, and/or

2. To assess the impact of intensity standardisation on a predictive radiomics model.

Nine studies assessed the impact of intensity standardisation on a predictive model. Five studies assessed the impact of standardisation on feature robustness (two studies included both aims). Three groups, Hoebel et al. [16], Carré et al. [3], and Orlhac et al. [4] used a 'scan-rescan' method, scanning the same patient after a short interval at different field strengths[3, 4], to test reproducibility, or using identical acquisitions [16], to test RF repeatability. Reproducibility is the consistency of a feature across different experimental settings, whereas repeatability refers to consistency under identical parameters [19]. Two other studies, Um et al. [20] and Reuze et al. [14], assessed differences in the feature distribution between paired scanners or the ability of a classifier to distinguish patients scanned internally *versus* externally [20].

The three main approaches to intensity standardisation can be categorised as histogram matching, deep-learning, or limiting or rescaling the signal intensities. Most of the included studies evaluated one method, however Carré et al. [3] and Hoebel et al. [16] used two or more.

2.4.4 Histogram matching

Histogram matching linearly transforms the signal intensities of an image to produce a match between the histogram of the reference and transformed image [21, 22]. The reference histogram is

calculated by averaging the intensities of training images, at pre-specified intensity landmarks [22]. The image to be transformed is divided into deciles of signal intensity and each decile is linearly mapped to the new intensity using the reference histogram.

Um et al. [20] assessed RF robustness after the following pre-processing steps: 8-bit rescaling, bias field correction, histogram matching, and isotropic resampling. A Random Forest classifier was used to predict whether images were from internal or external datasets and classification accuracy was measured using the Matthews correlation coefficient. A value of 1 means perfect prediction and 0 no better than chance, and therefore no scanner-dependency and > 0.2 was taken to mean that images could still retain scanner-dependence. Multiple classes of features were extracted. For edge features, different image filters (Sobel, Laplacian of Gaussian, Gabor, wavelet) were applied and first order features extracted. Haralick features were calculated from the GLCMs. For baseline images, the Matthews correlation coefficients were 0.36, 0.22 and 0.39 (measured from the provided bar chart) for Haralick and the Sobel and Laplacian of Gaussian features, respectively. Histogram matching significantly decreased these to 0.191, 0.170, and 0.140 respectively ($p < 0.01$).

Zhao et al. [23], used histogram specification-grid search (HS-GS), and Chen et al. [24] used histogram specification with automated selection of reference frames (HSASR), which automatically select the training histogram. Zhao et al. compared the predictive ability of standardised compared to unstandardised images for glioma grading and demonstrated an AUC of 0.956, 27% higher than without standardisation. Using HSASR, Chen et al. achieved 0.9934 AUC for grading (AUC 0.8512 without). These were the highest achieved for glioma grading, although a direct comparison to other ISTs was not presented, therefore limiting the ability to draw comparisons between approaches.

2.4.5 Deep learning

Hu et al. [25] describe 'MIL' pre-processing and intensity standardisation that corrects: modality incompleteness (M), uneven intensity distribution (I), and inconsistent layer spacing (L) in mpMRI datasets of T1W, T1CE, T2W and FLAIR sequences. Modality incompleteness is the absence of MRI sequences (referred to as 'modalities'), for example T1CE. Intensity unevenness is MRI signal

intensity variation introduced by variance in acquisition, and inconsistent layer spacing refers to differences in slice thickness between image sets. Effect of MIL normalisation on accuracy of radiomics models for glioma grading and for IDH1 prediction, and on tumour segmentation accuracy was assessed. A cycle-consistent adversarial network (CycleGAN) standardised signal intensities, and a deep learning network synthesised any missing MRI sequences using a encoder (a modified U-net) and separate decoder. Slice thickness was standardised using interpolation software, Statistical Parametric Mapping 12 (SPM12). AUC 0.693 (95% CI 0.613-0.772) was reported for a radiomics model for pathological grade prediction without any MIL normalisation. Accuracy increased following synthesis of missing sequences (AUC 0.838, 0.772-0.904), intensity standardisation (0.704, 0.626-0.783), and layer space normalisation (0.716, 0.639 – 0.793). Combining 3-steps produced the best performing model (0.89, 0.838 – 0.941). Similarly, the IDH1 mutation prediction model increased from AUC 0.701 (0.623 - 0.779) without MIL to 0.908 (0.863 - 0.954) with all three components of MIL normalisation.

2.4.6 Limiting or rescaling signal intensity

Reuze et al. rescaled the signal intensity between 0 and 32767 per patient and concurrently resampled to $0.5mm^3$ resolution and assessed the impact on feature robustness on images from 11 MRI scanners [14]. From 31 textural features, 11 were found to be robust amongst differing magnetic field strength post-normalisation ($p > 0.05$ on Wilcoxon paired test). Results from intensity standardisation alone were not presented.

Upadhaya et al. assessed the effect of pre-processing steps on the accuracy of a support vector classifier for OS prediction [13]. Patients were divided in to short- or long-term survivors by the median (12 months). Baseline pre-processing steps included bias field correction, skull stripping, and registration, with additional spatial resampling, intensity quantisation (or grey-level discretization), and standardisation. Intensity standardisation ignored any values outside of the range: ($\mu-\sigma$, $\mu+\sigma$), where μ and σ are the mean and standard deviation, respectively of the intensity values within the VOI. If the model utilised additional sequences and pre-processing steps, sensitivity improved from

79% to 93% and specificity from 86% to 93%. The effect on prognostic model performance of only applying intensity standardisation was not presented.

Florez et al. evaluated intensity standardisation on differentiation of tumour volume and oedema in 17 and 20 glioblastoma patients using a logistic regression model with least absolute shrinkage and selection operator penalty [11, 12]. 1% - 99% normalisation, where only intensities within the 1st and 99th centiles of the intensity histogram are included, was compared to no standardisation. Standardised T1CE sequences produced the best classification accuracy with an AUC$>$ 0.97 (0.85 without standardisation) [11]. The performance of standardised T2W images decreased - AUC of 0.85 (standardised) compared to AUC 0.91 (without). In a separate study, utilising only FLAIR, standardisation reduced AUC for discriminating tumour and oedema (AUC without 0.87, AUC with standardisation 0.84) [12].

Vils et al. assessed the impact of linear intensity interpolation on predictive model performance in 118 patients with recurrent glioblastoma [15]. OS, progression free survival and MGMT methylation status were the outcomes for each model. The linear intensity interpolation model used the intensity of ROIs within normal contralateral white matter and the vitreous body:

$$Intensity_{standardised} = Intensity_{original} \frac{500}{Intensity_{whitematter} - Intensity_{eye}} + 800 - \frac{500 Intensity_{whitematter}}{Intensity_{whitematter} - Intensity_{eye}}$$

A radiomic model for prediction of MGMT promotor methylation following standardisation achieved an AUC of 0.673 (95% CI 0.4837 – 0.8618) on the validation set, whereas the model without standardisation could not be validated (AUC 0.660 in training data).

Orlhac et al. assessed the impact of hybrid WhiteStripe standardisation on the distribution of features from normal white matter and tumours in 18 patients with diffuse glioma that had been scanned and rescanned at different field strengths [4]. WhiteStripe subtracts the mean and divides by the standard deviation of normal appearing white matter intensity [26]. WhiteStripe reduced the number of significantly different features in normal white matter (88 to 69%) and tumour (98

to 60%) between the images from different field strengths, demonstrating some improvement but considerable residual scanner dependency.

2.4.7 Comparison of techniques

Carré et al. [3] and Hoebel et al. [16] both used histogram-matching and z-score. Z-score normalisation subtracts the mean signal intensity from each voxel and divides by the standard deviation of the ROI [3]. Carré et al. also used WhiteStripe.

Hoebel et al. assessed the repeatability, using the intraclass correlation coefficient (ICC), of RFs extracted from a set of scan-rescan T1CE and FLAIR images of 48 patients diagnosed with glioblastoma [16]. Z-score and histogram matching improved repeatability of intensity features on FLAIR but not T1CE. Histogram matching improved repeatability of texture features on FLAIR ($p = 0.003$), whereas Z-score did not and neither technique improved the repeatability of texture features on T1CE.

Carré et al. [3], assessed the impact of intensity standardisation on feature robustness and the prediction of glioma grading. Using a scan-rescan dataset of 20 patients with low-grade glioma, histogram matching was found to produce the highest number of robust first-order features on both T1CE and FLAIR images (ICC > 0.80, 16 and 8 features out of 18 respectively). Regarding glioma pathological grade prediction using T1CE images, and only robust features from the first scan-rescan experiment, the average balanced accuracy increased from 0.73 to 0.81, 0.79, and 0.81 for histogram, WhiteStripe, and z-score respectively.

2.5 Discussion

The aim of this review was to evaluate the published literature and compare the efficacy of different ISTs prior to the extraction of RFs in the setting of adult-type diffuse glioma. A multitude of IST were used in the 12 studies reviewed, however there were significant differences in image processing

Table 2.3: Limitations of literature and opportunities for the future

Limitation	Opportunity
1. Assessing the effect of multiple preprocessing steps simultaneously	Effects of preprocessing steps presented independently of others so their effect on the result can be determined.
2. Investigating the effect of only one intensity standardisation technique	Impact of more than one standardisation method on a predictive model or feature robustness could be evaluated.
3. Lack of scan-rescan data used to test the repeatability of radiomic features	Increased availability of datasets that have rescanned a patient with a diffuse glioma within a short time interval (i.e. days) in public databases.
4. Single-centre studies used to assess standardisation techniques	Use of multi-centre datasets in assessing the efficacy of standardisation techniques and repeatability of radiomic features.

and methodology that did not allow quantitative analysis. Although on face value, studies using histogram matching produced the highest AUC values within the included studies [23, 24], these studies lacked a comparison method and therefore we could not confidently produce a consensus on the optimal IST for this context.

To be clinically useful, radiomics needs to be validated [19], and there are unique challenges to evaluating radiomics-based predictive models [27]. For MRI radiomics, a key challenge to assessing repeatability and reproducibility, is to remove the scanner-dependent signal intensity changes [1]. This study confirms that intensity standardisation improves RF repeatability and improves most predictive models, and therefore that there needs to be awareness of this crucial step in any radiomics predictive modelling study. Variation in methodology precluded the direct comparison of results across studies and this review has highlighted potential areas of improvement, which may improve translation of radiomic models into the clinical setting (**Table 2.3**).

In two studies[13, 14], the effects of intensity standardisation were difficult to differentiate from other pre-processing, and authors could have reported separately the impact of different pre-processing steps on feature robustness or model performance. Hu et al. [25] presented all possible combinations of pre-processing steps, with separate AUC results, so the impact of each step was identifiable.

Only two studies [3, 16] performed a comparison of one or more IST. Given the number of methods

and lack of consensus, more studies that directly compare techniques are required as it makes the interpretation of results such as those of the two histogram-specification studies [23, 24] difficult. The AUC for glioma grading was the highest reported out of all classifiers for this outcome across all the included studies, however it is unclear how if other ISTs would have produced similar results on the same data and with the same model building approach. Li et al. [6] compared multiple ISTs and post-feature extraction correction with ComBat, a statistical model for batch-effect correction in genomics that has been applied successfully to RFs in retrospective analyses [1, 4]. Intensity standardisation was insufficient to remove scanner-dependency, but ComBat could remove scanner-dependent information from extracted features [6], similar to the findings of Orlhac et al [4].

Three studies used scan-rescan data, where patients were scanned multiple times with a short delay (hours or days) and either using the same or varied acquisition parameters, thereby providing the opportunity to assess RF reproducibility and repeatability. Although a tumour may change microscopically within several days, these radiomic studies assume that if the imaging phenotype remains identical, then the RFs ought to as well [3, 4, 16]. Test-retest data, along with phantom studies [6], and comparison of RFs extracted from normal structures provides a useful paradigm to compare ISTs. Open access to such data in a public repository may help further validate different ISTs.

Limitations to this review include not being able to retrieve full-text articles for two conference abstracts. Based on the abstracts, it is unlikely they would have been included. Their potential omission will have had a limited impact as a narrative synthesis would still have been required. QUADAS-2 is not specifically designed for assessing the efficacy of MRI ISTs but was considered the only option given the absence of a more specific alternative. The scope of this review was to assess MRI intensity standardisation in the context of diffuse glioma, which will have led to inevitable omission of studies of other organs, brain pathologies and healthy volunteers.

2.6 Conclusion

No clear consensus emerged as to which approach is the most reliable IST. In order to translate radiomics to the clinic, more studies that assess the effects of multiple ISTs on predictive model performance or RF robustness are required. The impact of any intensity standardisation step should be clearly reported and determined independently from the effects of other image manipulation. Collation and sharing of scan-rescan datasets would facilitate production of radiomic models in diffuse glioma and greatly improve the development of clinically translatable models.

References

1. Da-Ano, R., Visvikis, D. & Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine & Biology* **65,** 24TR02 (Dec. 2020).

2. Yang, F., Dogan, N., Stoyanova, R. & Ford, J. C. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth. *Physica Medica* **50,** 26–36 (2018).

3. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10,** 12340 (July 2020).

4. Orlhac, F. *et al.* How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European Radiology* **31,** 2272–2280 (Apr. 2021).

5. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology Biology Physics* **102,** 1143–1158 (2018).

6. Li, Y., Ammari, S., Balleyguier, C., Lassau, N. & Chouzenoux, E. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features. *Cancers* **13,** 1–22 (2021).

7. Baeßler, B., Weiss, K. & Santos, D. P. D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Investigative Radiology* **54,** 221–228 (2019).

8. Pandey, U., Saini, J., Kumar, M., Gupta, R. & Ingalhalikar, M. Normative Baseline for Radiomics in Brain MRI: Evaluating the Robustness, Regional Variations, and Reproducibility on FLAIR Images. *Journal of Magnetic Resonance Imaging* **53,** 394–407 (2021).

9. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372,** n71 (Mar. 2021).

10. Whiting, P. F. *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. eng. *Annals of internal medicine* **155,** 529–536 (Oct. 2011).

11. Florez, E. *et al.* Multiparametric Magnetic Resonance Imaging in the Assessment of Primary Brain Tumors Through Radiomic Features: A Metric for Guided Radiation Treatment Planning. *Cureus* **10,** e3426 (Oct. 2018).

12. Florez, E., Nichols, T. A., Lirette, S. T., Howard, C. M. & Fatemi, A. Developing a Texture Analysis Technique using Fluid-Attenuated Inversion Recovery (FLAIR) to Differentiate Tumor from Edema for Contouring Primary Intracranial Tumors. **4,** 1023 (2018).

13. Upadhaya, T., Morvan, Y., Stindel, E., Le Reste, P.-J. & Hatt, M. Prognosis classification in glioblastoma multiforme using multimodal MRI derived heterogeneity textural features: impact of pre-processing choices. *Medical Imaging 2016: Computer-Aided Diagnosis* **9785,** 97850W (2016).

14. Reuzé, S. *et al. A preliminary MRI harmonization method allowing large scale radiomics analysis in glioblastoma* 2018.

15. Vils, A. *et al.* Radiomic Analysis to Predict Outcome in Recurrent Glioblastoma Based on Multi-Center MR Imaging From the Prospective DIRECTOR Trial. *Frontiers in Oncology* **11,** 1–9 (2021).

16. Hoebel, K. V. *et al.* Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence* **3,** e190199 (2021).

17. Clark, K. *et al.* The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging* **26,** 1045–1057 (2013).

18. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

19. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology* **14,** 169–186 (Mar. 2017).

20. Um, H. *et al.* Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Physics in Medicine & Biology* **64,** 165011 (Aug. 2019).

21. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* **19,** 143–150 (2000).

22. Shah, M. *et al.* Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis* **15,** 267–282 (2011).

23. Zhao, G. *et al.* HS–GS: A method for multicenter MR image standardization. *IEEE Access* **8,** 158512–158522 (2020).

24. Chen, X. *et al.* Automatic Histogram Specification for Glioma Grading Using Multicenter Data. *Journal of Healthcare Engineering* **2019,** 1–12 (Dec. 2019).

25. Hu, Z. *et al.* MIL normalization – prerequisites for accurate MRI radiomics analysis. English. *Computers in Biology and Medicine* **133,** 104403 (2021).

26. Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9–19 (2014).

27. Halligan, S., Menu, Y. & Mallett, S. Why did European Radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting. *European Radiology* **31,** 9361–9368 (2021).

2.7 Search protocols and PRISMA checklists

The following search strategy was used to search Medline and then EMBASE, both of which were accessed via OVID on 05/10/2021. No limits or filters were applied.

1. MRI.ab,sh,ti.
2. magnetic resonance imaging.ab,sh,ti.
3. 1 or 2
4. AI.ab,ti.
5. Machine learning.ab,sh,ti.
6. Neural network.ab,sh,ti.
7. “Neural network*“.ab,ti.
8. “Radiomic*“.ab,ti.
9. “radiogenomic*“.ab,ti.
10. deep learning.ab,ti.
11. Advanced neuroimaging.ab,ti.
12. Artificial intelligence.ab,sh,ti.
13. 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14. “intensity standard*“.ab,ti.
15. “intensity harmon*“.ab,ti.
16. (intensity adj10 standard*).ab,ti.
17. (intensity adj10 harmon*).ab,ti.
18. “image preprocess*“.ab,ti.
19. feature extraction.ab,ti.
20. extracted feature.ab,ti.
21. radiomic feature.ab,ti.
22. texture feature extraction.ab,ti.
23. “harmon*“.ab,ti.
24. “standard*“.ab,ti.

25. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24
26. Glioma.ab,sh,ti.
27. Glioblastoma.ab,sh,ti.
28. glioblastoma multiforme.ab,sh,ti.
29. GBM.ab,ti.
30. Low grade.ab,ti.
31. Grade II.ab,ti.
32. 30 or 31
33. 26 and 32
34. High grade.ab,ti.
35. Grade III.ab,ti.
36. Grade IV.ab,ti.
37. 34 or 35 or 36
38. 26 and 37
39. “Glial cell tumo*“.ab,ti.
40. “Astrocytoma*“.ab,ti.
41. “Oligodendroglioma*“.ab,ti.
42. “Oligoastrocytoma*“.ab,ti.
43. brain cancer.ab,ti.
44. Neuro-oncology.ab,ti.
45. 26 or 27 or 28 or 29 or 33 or 38 or 39 or 40 or 41 or 42 or 43 or 44
46. “normali*ation”.ab,ti.
47. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 46
48. 3 and 13 and 45 and 47

The following strategy was used to search Scopus on 05/10/21. No limits or filters were applied.

(TITLE-ABS (mri) OR TITLE-ABS (magnetic AND resonance AND imaging))

AND

(TITLE-ABS (artificial AND intelligence) OR TITLE-ABS (ai) OR TITLE-ABS (machine AND learning) OR TITLE-ABS (neural AND network) OR TITLE-ABS (radiomic*) OR TITLE-ABS (radiogenomic*) OR TITLE-ABS (deep AND learning) OR TITLE-ABS (advanced AND neuro-imaging)) AND (TITLE-ABS-KEY (intensity AND standard*) OR TITLE-ABS-KEY (intensity AND harmon*) OR TITLE-ABS (intensity AND adj10 AND standard*) OR TITLE-ABS (intensity AND adj10 AND harmon*) OR TITLE-ABS-KEY (image AND preprocess*) OR TITLE-ABS (feature AND extraction) OR TITLE-ABS (extracted AND feature) OR TITLE-ABS (radiomic AND feature) OR TITLE-ABS (texture AND feature AND extraction) OR TITLE-ABS-KEY (harmon*) OR TITLE-ABS (standard*) OR TITLE-ABS-KEY (normali*ation))

AND

(TITLE-ABS (glioma) OR TITLE-ABS (glioblastoma) OR TITLE-ABS (glioblastoma AND multiforme) OR TITLE-ABS (gbm) OR TITLE-ABS (low AND grade AND glioma) OR TITLE-ABS (grade AND ii AND glioma) OR TITLE-ABS (high AND grade AND glioma) OR TITLE-ABS (grade AND iii AND glioma) OR TITLE-ABS (grade AND iv AND glioma) OR TITLE-ABS (glial AND cell AND tumo*) OR TITLE-ABS (astrocytoma*) OR TITLE-ABS (oligodendroglioma*) OR TITLE-ABS (brain AND cancer) OR TITLE-ABS (neuro-oncology))

Table S2.1: Abstract compliance with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines

Section	Item	Description	Reported
TITLE			
Title	1	Identify the report as a systematic review.	Yes
BACKGROUND			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review.	In main text
Information sources	4	Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched.	Yes
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies.	Yes
Synthesis of results	6	Specify the methods used to present and synthesise results.	In results
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarise relevant characteristics of studies.	Yes
Synthesis of results	8	Present results for main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured).	Yes
DISCUSSION			

Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision).	In main text
Interpretation	10	Provide a general interpretation of the results and important implications.	Yes
OTHER			
Funding	11	Specify the primary source of funding for the review.	Main text
Registration	12	Provide the register name and registration number.	N/a

Table S2.2: Study compliance with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines

Section	Item	Description	Reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	See other checklist
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Full protocol in appendix
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods

Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	NA
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Table 2
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	methods

	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	NA
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Methods
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 2
Study characteristics	17	Cite each included study and present its characteristics.	Results, Table 2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Table 1, results
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Table 2, narrative synthesis
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Table 1, results
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	NA
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	NA
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	NA

Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	NA
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
	23b	Discuss any limitations of the evidence included in the review.	Discussion
	23c	Discuss any limitations of the review processes used.	Discussion
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion, Table 3
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Search protocol in appendix
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding statements
Competing interests	26	Declare any competing interests of review authors.	None
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Not publicly available. Can be provided on request

CHAPTER

THREE

TUMOUR SIZE AND OVERALL SURVIVAL IN A COHORT OF PATIENTS WITH UNIFOCAL GLIOBLASTOMA: A UNI- AND MULTIVARIABLE PROGNOSTIC MODELLING AND RESAMPLING STUDY

3.1 Abstract

3.1.1 Background

Published models inconsistently associate tumour size with survival in patients with glioblastoma. Rather than build the best prognostic model, the purpose was to investigate the prognostic effect of tumour size in a large cohort of patients diagnosed with glioblastoma and interrogate how choice of sample size and consideration of non-linear transformations may impact on the likelihood of finding a prognostic effect using univariable and multivariable analysis and data resampling.

3.1.2 Methods

279 patients (mean age 61 years, 39% female) with IDH-wildtype unifocal WHO grade 4 glioblastoma and pre-operative MRI between 2014-2020 from a retrospective cohort were included. Tumours were segmented using deep-learning with manual correction and manual diameter measurements. Uni- and multivariable association between core volume, whole volume (CV, WV) and diameter with OS was assessed with (1) Cox proportional hazard models +/- log transformation and (2) resampling with 1,000,000 repetitions and varying sample size to identify the percentage of models, which showed a significant effect of tumour size.

3.1.3 Results

Diameter or volume models adjusted for operation-type were significant, and diameter adjusted for any clinical variable remained significant ($p = 0.03$). Multivariable resampling increased significant effects ($p < 0.05$) of all size variables as sample size increased. Log-transformation also had a large effect on chances of prognostic effect of WV. For models adjusted for operation-type, 19.5% of WV vs 26.3% log-WV ($n = 50$) and 69.9% WV and 89.9% log-WV ($n = 279$) were significant.

3.1.4 Conclusion

In this large, well-curated cohort, multivariable modelling and resampling suggests tumour volume is prognostic at larger sample sizes and with log-transformation for WV.

3.2 Introduction

Many proposed prognostic models for OS prediction in glioblastoma include IBs [1] as MRI is usedd to assess patients throughout their treatment pathway and captures the entire tumour volume. Radiomic features derived from MRI have gained popularity in published models [2], and can

include features that assess tumour heterogeneity and intensity but also 'simple' features such as diameter and volume.

The diameter of the tumour core, which is commonly defined as the enhancing and necrotic portion of the tumour [3], is also routinely evaluated in clinical practice and is amongst the most common of the imaging predictors of OS to be investigated [4]. T1CE enhancing and T2W high signal areas are also routinely evaluated as markers of prognosis [4]. There are now several methods available for automated or semi-automated segmentation of the different tumour regions of glioblastoma that are apparent on imaging [5], such as the enhancing and necrotic tumour and peritumoural 'oedema', and it has therefore become more feasible to integrate various definitions of 'tumour volume' into prognostic models, including in larger institutional datasets [4, 6, 7].

Intuitively, pre-treatment tumour size is expected to impact on patient outcome as it likely reflects the number of tumour clonogens that require ablation by conventional cytotoxic treatments [8, 9]. Published modelling studies have, however, yielded inconsistent data regarding the prognostic effect of tumour diameter and volume [4, 10–13]. Some of this may be due to sample size; large cohort studies of glioblastoma have demonstrated a weak prognostic effect of tumour diameter [13], but these larger cohort studies could not feasibly assess volume. Studies that have assessed tumour volume [4, 14, 15] have not suggested a definite prognostic relationship to OS. As well as variation in sample size, some of the inconsistency may reflect variations in handling continuous variables during statistical modelling, for example leading to dichotomisation [16], assumptions of a linear relationship to outcome [4] and use of univariable model significance to select predictors [17]. All these choices are known to impact upon the modelling process and may impact on the ability to determine accurate prognostic effects of the candidate predictors [18, 19].

There are a number of ways to select predictors for multivariable prognostic modelling and to evaluate the uncertainty or instability that might arise from choosing predictors in small samples or using univariable significance [18, 19]. This includes the use of internal validation strategies such as data resampling (ie. bootstrapping) to estimate uncertainty in effect size and predictor selection, however has been infrequently assessed in the prognostic modelling of glioblastoma survival despite

its importance [1].

The hypothesis is that inconsistencies in the literature are secondary to varying sample size, predictor selection strategies in multivariable modelling and consideration of data transformation. Rather than build the best prognostic model, the purpose of this study was to investigate the prognostic effect of tumour size in a large cohort of patients diagnosed with glioblastoma and interrogate how choice of sample size and consideration of non-linear transformations may impact on the likelihood of finding a prognostic effect using univariable and multivariable analysis and data resampling.

3.3 Materials and Methods

3.3.1 Ethical approval

This was a retrospective study and therefore informed patient consent was not feasible. Ethical approval and institutional data access was approved via local ethical review committee (REC ref: 19/YH/0300, IRAS project ID: 255585, see section 5).

3.3.2 Patient selection and characteristics

All consecutive patients (16 years and over) with histologically proven glioblastoma according to 2021 WHO classification of central nervous system tumours treated at a single tertiary referral centre between 2014-2020 were identified retrospectively from neuro-oncology multidisciplinary team (MDT) records. The catchment area includes 3.9 – 4.4 million adults, and over this period, 3046 new primary brain neoplasms were reviewed at MDT, with approximately 20% diagnosed with glioblastoma or malignant glioma. Inclusion criteria were: MRI performed prior to any surgery, unifocal tumour (as determined by consultant neuroradiologist with $>$ 10 years experience), and all four of the following MRI sequences acquired: T1W, T2W, FLAIR and T1CE sequences. Exclusion criteria were: absence of pre-operative MRI, significant degradation of imaging due to artefact, or tumours that were multifocal at presentation, documented IDH mutation on immunohistochemistry

or cytogenetic testing. Patients with giant cell glioblastoma were excluded due to the small number of patients affected and the much longer OS associated with this diagnosis [20].

Demographic, clinical and cytogenetic data were obtained from the electronic health records using in-house software. Data included patient age, sex and type of operation. Histopathological and cytogenetic data included histology, IDH1 and 2 mutation and MGMT promoter methylation. Extent of resection was estimated by the same consultant neuroradiologist using the immediate (48-72 hour) post-resection MRI, and grouped based upon the amount of contrast enhancing and necrotic tumour resected – (i) 100%, (ii) $\geq 90\%$ or (iii) $< 90\%$. Adjuvant treatment was categorized as (i) full Stupp protocol – 60 Gy in 30 fractions radiotherapy with concomitant and 6 cycles adjuvant temozolomide; (ii) partial Stupp – 60 Gy in 30 fractions radiotherapy but temozolomide discontinued during either concomitant or adjuvant treatment phase; (iii) non-Stupp – any other treatment protocol. Other clinical information such as performance status, eligibility or entry into clinical trials, socioeconomic status were not widely available due to the retrospective nature of the data collection.

3.3.3 Data preparation

A summary of the data preparation and numbers excluded with reasons is shown in **Figure 3.1**. DICOM image preparation was performed in python 3.9 [21]. DICOM images were retrieved from the institutional picture archive and communication system and pseudonymised, and the image acquisition parameters are summarised in **Table S3.1**. Images were converted to Neuroimaging Informatics Technology Initiative file format using the dicom2nifti (v2.3.4) package [22].

3.3.4 Image pre-processing and tumour segmentation

Semi-automated tumour segmentations were produced using the Federated Tumor Segmentation (FeTS) software, an open-source platform available for processing and segmentation of MRIs for patients with glioblastoma [23]. A detailed description of the software, including packages and

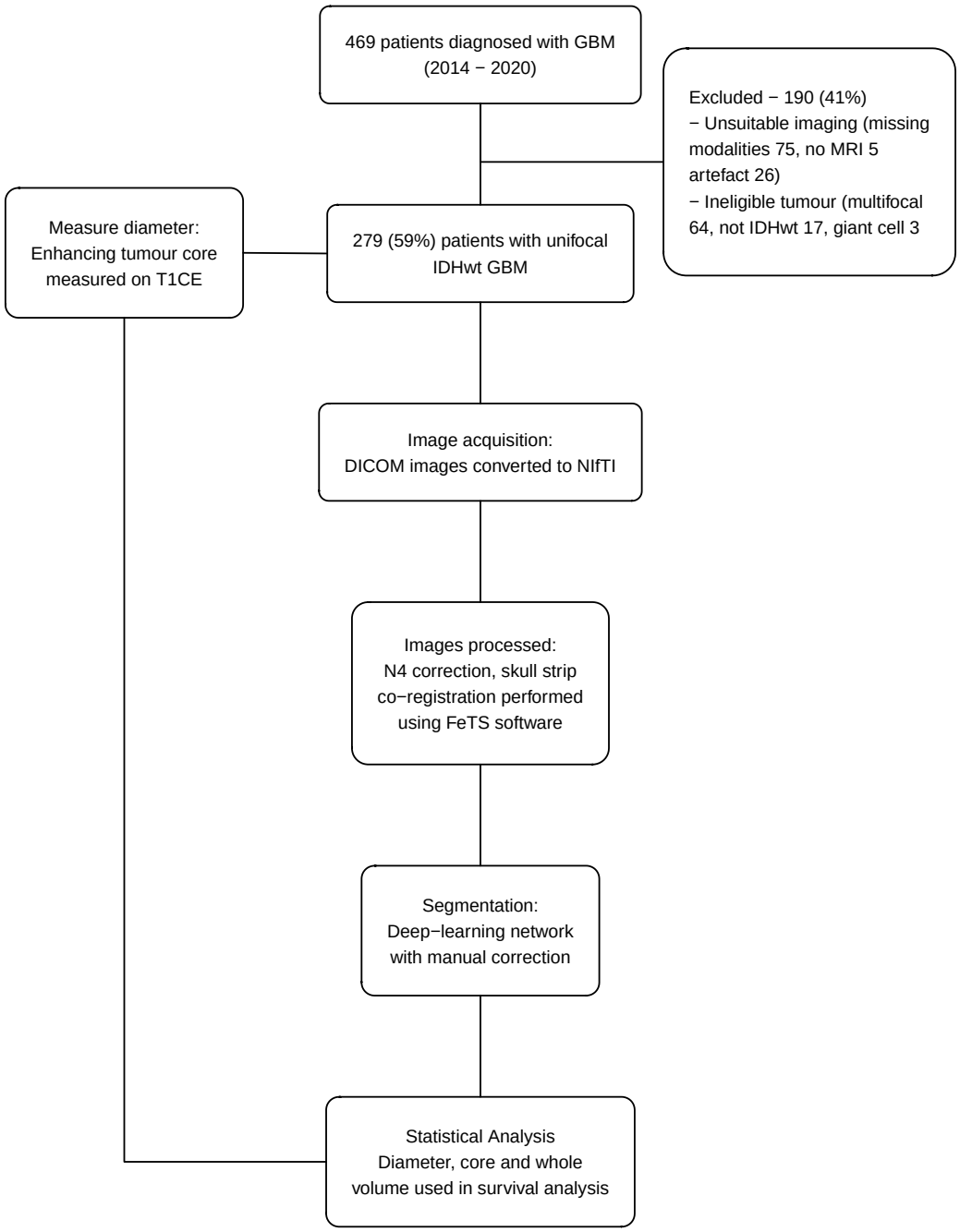

Figure 3.1: CV = Core Volume, DICOM = Digital Imaging and Communications in Medicine, FeTS = Federated Tumor Segmentation software, GBM = glioblastoma, IDHwt = isocitrate dehydrogenase wild-type, NifTI = Neuroimaging Informatics Technology Initiative, PACS = picture archive and communication system, T1CE = gadolinium contrast-enhanced T1-weighted imaging, WV = Whole Volume

libraries used in FeTS is available elsewhere [23] and utilises the same pre-processing steps used in the BraTs challenge [5], and the open-source software Cancer Imaging Phenomics Toolkit (CaPTk) [24]. The key features are outlined below.

The T2W, T1CE and FLAIR sequences were rigidly co-registered first to the T1W sequence, then to the SRI24 brain atlas [25] and also spatially resampled to 1 x 1 x 1mm voxel resolution using the Greedy registration framework [26]. Images were then skull-stripped [27] and tumour segmentation was performed with the 'nnU-net' deep-learning network and pretrained model weights [28]. Tumours were automatically segmented into three VOIs. The three VOIs were defined as: i) necrotic tumour – fluid signal intensity showing very high T2W signal and reduced T1CE signal compared to the same area on T1W images; ii) enhancing tumour – increased signal on T1CE compared to the same area on T1W images and also increased T1CE signal compared to normal white matter regions on T1CE images; iii) peritumoural oedema – high FLAIR and T2W signal of the entire tumour, minus the necrotic and enhancing regions and not including ventricles or extra-axial cerebrospinal fluid spaces [5]. Tumour masks were used to produce two tumour volumes per patient: 1) core volume (CV, cm^3) – combination of necrotic and enhancing components; 2) whole volume (WV, cm^3) – CV combined with the peritumoural oedema (**Figure 3.2**). Although both WV and CV will have included the necrotic portions of the tumour, the decision was made to include the necrosis as it reduces the amount of manual correction of tumour segmentations, the enhancing and necrotic portion of the tumour is often treated surgically as a whole target for debulking, and when measuring tumour diameter, the enhancing and necrotic portion of the tumour is included in the measurement.

The segmentations were checked manually and corrected using FeTS. All segmentations were checked by a neuroradiology fellow (5 years radiology experience). Independently, 50 segmentations were also checked by a consultant neuroradiologist ($>$ 10 years consultant neuroradiology experience), and the inter-rater concordance compared using the dice similarity coefficient [29].

Tumour diameter was defined as the maximum axial or cranio-caudal diameter of the enhancing tumour core and was measured using the T1CE sequence within imaging viewing software (Impax

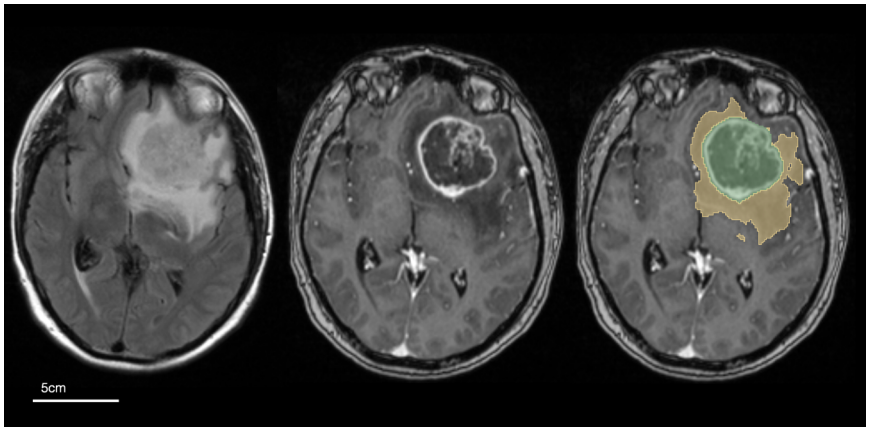

Figure 3.2: Selected MRI axial images: left – fluid-attenuated inversion recovery (FLAIR); middle – gadolinium contrast-enhanced T1-weighted (T1CE); right – T1CE image with overlay of core tumour segmentation (in green) and peritumoural oedema (yellow). Core volume (CV) defined as the enhancing and necrotic component of the tumour (green), and whole volume (WV) defined as the combination of core and peritumour oedema segmentation (green + yellow).

Version 6.5.3.3009, Agfa Healthcare) using in-built calipers on a submillimetre scale (mm - converted to cm) by two radiology trainees (1 and 2 years radiology experience) and corrected by a neuroradiology fellow (5 years radiology experience). All manual correction and measurement was performed without knowledge of individual patient outcome.

3.3.5 Statistical analysis

All statistical analysis was performed in R version 4.2.2 (2022-10-31) and overseen by a highly-experienced career statistician. Univariable association between CV, WV or tumour diameter with OS was investigated using Cox regression modelling. HRs, concordance indices (C-indices) and p-values for each model were used to assess performance.

Possible non-linear relationships between OS and size (volume or diameter) were explored using both logarithmic transformation and penalised spline functions, the latter implemented using penalised spline function within the 'survival' package (v3.6-4) [30]. Penalised spline functions were used to assess for any trends in the data that might not be seen with a linear fit, as splines allow a smooth

curve to be fit to data [31]. Overfitting to the data points is discouraged by the inclusion of a roughness penalty, and the implementation does not require any pre-specification of the number of internal boundaries or knots. Model fits were assessed by plotting each tumour size parameter against the log-HR.

Multivariable association of CV, WV or diameter to OS was also evaluated by (i) adjusting each size variable for either age, sex, type of surgery, MGMT promoter methylation status or adjuvant oncological treatment (i.e. size variable + one clinical variable in turn) and (ii) adjusting size for all clinical parameters. As the aim was to assess the prognostic effect of tumour size, in multivariable models this was assessed using the HR for each size parameter and the Wald test p-value for the size variable's coefficient rather than the overall model p-value. A post-hoc bonferroni correction was considered but not applied, as the aim of the study was to examine the effect of the adjustment on the significance of the Wald test, rather than to definitively conclude that size was a prognostic variable. It is acknowledged that this increases the risk of type 1 error.

To assess the impact that either log-transformation and/or sample size could have on detecting a prognostic effect of tumour size on OS, a resampling study was conducted. Using different sample sizes (50, 100, 150, 200, 250, 258 or 279), bootstrapped samples were generated from the original dataset with replacement. For any multivariable models that were adjusted for MGMT methylation status, the maximum sample size was 258 (not 279) due to the number of cases with a known result. Bootstrapping was carried out for 1,000,000 repetitions at each sample size and for each of the tumour size variables and a Cox regression model for each tumour size variable, both with and without log-transformation, was created. For univariable models, the percentage of models in which the overall model Wald test p-value < 0.05, < 0.01 and < 0.001 was calculated across the 1,000,000 repetitions per sample size. For multivariable models, the percentage of models in which the Wald test p-value for the coefficient of tumour size < 0.05 was calculated (rather than the overall model Wald test significance) across the 1,000,000 repetitions per sample size. Effect of sample size was assessed with two-sided Kolmogorov-Smirnov (KS) tests to compare the p-value distributions from resampling at varying sample sizes.

3.4 Results

3.4.1 Demographics of the study population

279 patients were included - 39% (108/279) of patients were female and the median age was 62 years (interquartile range 31 – 85 years) (**Table 3.1**). 236 deaths occurred before the censor date of 31/10/2020. Median OS was 12 months (95% CI 11 – 14 months), median follow-up time was 45 months (maximum 70 months) and 26% (72/279) patients had a surgical biopsy of their glioblastoma. 20% (57/279) of patients had 100% resection of tumour core and 21% (58/279) completed the full Stupp protocol of adjuvant treatment. Median (IQR) CV was 28.1cm^3 (12.6 – 50.3), WV was 103.3cm^3 (45.6 – 160.1) and tumour diameter was 4.4cm (3.3-5.4cm). Histograms of tumour size (**Figure S3.1**) showed that distributions of CV and WV were slightly positively skewed and tumour diameter was normally distributed prior to any transformation. These data confirm that this population is representative of patients diagnosed with glioblastoma in other typical neurosciences centres [32].

3.4.2 Segmentations and Univariable Cox models of tumour size

The segmentation process explained above yielded 50 patients with two independent sets of contourd, which were the product of manual correction by a fellow and independently a consultant neuroradiologist, and these two segmentations per patient were compared for spatial overlap using dice similarity coefficient (DSC). Accepting that there are myriad of segmentation comparison metrics [33], DSC was chosen as it is the only metric provided for a similar task in this context, that of the BRATS segmentation dataset, in which multiple experts raters segment the same glioblastoma images [5]. Hence the results of this experiment could be directly compared to the literature. The mean (± standard deviation) DSC for the two independent segmentations for the core and oedema regions was 0.94 ± 0.05 and 0.97 ± 0.03 respectively, which are equivalent to values published in the Brats data [5].

Table 3.1: Summary of patient demographics and treatment ($n = 279$)

Demographic	Value
Age, years - median (IQR)	62 (55-68)
Gender - no. female (%)	108 (39%)
Surgical treatment – no. (%)[a]	
Biopsy	71 (25%)
100% resected	57 (20%)
$\geq$ 90% resected	86 (31%)
$<$ 90% resected	65 (23%)
Adjuvant oncology treatment – no. (%)	
No Stupp	150 (54%)
Full Stupp[b]	58 (21%)
Partial Stupp[c]	71 (25%)
MGMT methylation – no. (% of known)[d]	103 (40%)
Overall survival, months – median (95% CI)	12 (11-14)
Maximum tumour diameter, cm – median (IQR)	4.4 (3.3-5.4)
Core volume, cm^3 - median (IQR)	28.1 (12.6-50.3)
Whole volume, cm^3 - median (IQR)	103 (45.6-160)

IQR = interquartile range, MGMT = O6-methylguanine-DNA methyltransferase, CI = Confidence Interval. [a] Percentage of contrast enhancing and necrotic tumour core removed
[b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and 6 cycles adjuvant temozolomide [c] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide [d] 258 cases with result known

Table 3.2: Cox proportional hazards models for each tumour size parameter - whole volume, core volume and diameter - for predicting overall survival ($n = 279$) without adjustment for other clinical predictors. Each size parameter has been tested with and without log-transformation.

	WV	log(WV)	CV	log(CV)	diameter	log(diameter)
C (95% CI)	0.5 (0.46-0.54)	0.5 (0.46-0.54)	0.5 (0.46-0.54)	0.5 (0.46-0.54)	0.5 (0.46-0.54)	0.5 (0.46-0.54)
HR (95% CI)	1 (1-1)	1.1 (0.81-1.6)	1 (1-1)	0.95 (0.71-1.3)	1 (0.93-1.1)	0.94 (0.43-2)
p	0.784	0.475	0.539	0.704	0.745	0.875

C = Concordance index, CV = Core Volume, HR = Hazard Ratio, WV = Whole Volume.

Table 3.2 summarises the univariable Cox regression models for CV, WV and diameter, with and without log-transformation. The results of the models derived from the institutional glioblastoma images show limited evidence for a univariable prognostic relationship between tumour volume or diameter and OS. C-indices for all models were 0.5, and all HRs crossed 1.

In **Figures S3.2 - S3.4**, each tumour size parameter (with and without log-transformation) was plotted against the log-HR. The fit of a linear function to the data was compared with the use of splines, and these suggest that there was limited evidence that the tumour size parameters had a univariable prognostic relationship – the model closely followed the reference line for linear and non-linear functions. These results show that within this cohort, there was no evidence to support a univariable linear or non-linear prognostic relationship between OS and size.

In multivariable analyses, however, there was evidence of a prognostic association between size and OS when adjusting for clinical variables. A summary of the association of size variables in multivariable models with OS is shown in **Table 3.3**. CV, WV, log(WV) and diameter adjusted for type of surgery showed a statistically significant association with OS. Although not significant at the 0.05 level, the CIs for the HRs of log(CV) and log(diameter) were relatively wide, especially the latter indicating uncertainty in the HR estimate. Similarly, for the model adjusted for all clinical variables, only diameter remained statistically significant at the 0.05 threshold, however the HR for log(CV), log(WV) and log(diameter) suggested a potentially prognostic effect with relatively wider confidence intervals (and less certainty) for HR estimate of the latter two variables. The univariable and multivariable prognostic associations of each clinical variable to OS is provided in **Tables S3.2-S3.3**. Data from this cohort of glioblastoma patients therefore suggests that size was

associated with OS in multivariable models and whilst several related parameters did not achieve statistical significance there was supportive evidence of a potential prognostic relationship.

Table 3.3: Prognostic effect of each tumour size parameter (whole volume, core volume and diameter) within multivariable Cox proportional hazards models predicting overall survival that have been adjusted for selected clinical variables ($n = 279$). Each row of the table presents the results from models that include the tumour size size variable specified by the column name (either diameter, core or whole volume or their log-transformed versions) and the clinical variable indicated in the 'Variable' column. The stated hazard ratios (and 95% confidence intervals) refer to the selected tumour size variable and not the clinical variable indicated in the 'Variable' column. The stated p-values refer to the Wald test for the regression coefficient of the tumour size variable and not the overall multivariable Cox model significance/p-value

	Tumour Diameter				Whole Volume				Core Volume			
	Diameter		log(Diameter)		Whole Volume		log(Whole Volume)		Core Volume		log(Core Volume)	
Variable	HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
Age	1.01 (0.916-1.1)	0.91	0.857 (0.391-1.88)	0.7	1 (0.998-1)	0.9	1.09 (0.781-1.52)	0.61	1 (0.997-1.01)	0.56	0.926 (0.696-1.23)	0.6
Gender	1 (0.915-1.1)	0.93	0.874 (0.403-1.89)	0.73	1 (0.998-1)	0.99	1.09 (0.783-1.52)	0.61	1 (0.996-1.01)	0.71	0.915 (0.688-1.22)	0.54
Type of surgery[a]	1.14 (1.02-1.26)	0.016	2.39 (0.955-5.98)	0.063	1 (1-1)	0.013	1.9 (1.28-2.82)	0.0014	1.01 (1-1.01)	0.018	1.29 (0.927-1.79)	0.13
Adjuvant oncology treatment[b]	0.996 (0.909-1.09)	0.93	0.818 (0.374-1.79)	0.61	1 (0.998-1)	0.99	1.05 (0.754-1.47)	0.76	1 (0.996-1.01)	0.67	0.922 (0.693-1.23)	0.58
MGMT methylation[c]	1.02 (0.926-1.12)	0.7	0.963 (0.426-2.18)	0.93	1 (0.998-1)	0.98	1.1 (0.779-1.54)	0.6	1 (0.996-1.01)	0.71	0.941 (0.704-1.26)	0.68
Age + Gender + Surgery + Oncology + MGMT[c]	1.12 (1.01-1.25)	0.032	2.34 (0.914-6.01)	0.076	1 (0.999-1)	0.24	1.45 (0.985-2.14)	0.06	1 (1-1.01)	0.072	1.24 (0.889-1.73)	0.2

HR = Hazard Ratio, CI = Confidence Interval, CV = Core Volume, WV = Whole Volume, MGMT = O6-methylguanine-DNA methyltransferase. [a] Divided into biopsy only or resection (100, ≥ 90, or < 90 [b] Categorical variable: Full Stupp (completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and 6 cycles adjuvant temozolomide), partial Stupp (as per full Stupp but either did not commence or did not complete adjuvant temozolomide), or other adjuvant therapy [c] $n = 258$, cases with known MGMT result

3.4.3 Resampling study

The results of the resampling experiments using univariable and multivariable models, the latter adjusted for operation type and all clinical variables, are shown in **Table S3.4** and **Table 3.4**, respectively. In univariable models of tumour size, for all size variables, higher percentages of models with $p < 0.05$ were seen as the sample size increased, and for tumour volume (CV or WV), the same was observed after log transformation, although the change was modest. For WV, 5.14 vs 5.60% ($n = 50$ vs $n = 279$), for CV 5.07 vs 8.60% and for diameter 5.43 vs 6.39% models had p-values < 0.05 across all repetitions on non-transformed data. The distributions of p-values (across all repetitions per sample size at $n = 50$ vs $n = 279$) differed significantly on two-sided KS testing (test p-values < 0.0001). **Tables S3.5-S3.6** also show the percentages of models with $p < 0.01$ and $p < 0.001$, and this shows the same overall trend for the tumour size parameters, but with successively lower percentages of models as the p-value threshold was lowered.

Table 3.4: Percentage of multivariable tumour size models with model p-values < 0.05 during the resampling study. The percentages in the table cells represent the percentage of resamples in which Wald test p-value for the regression coefficient of the selected tumour size variable (each column) was < 0.05. The left side of the table shows the results when each tumour size variable was adjusted only for type of operation (i.e. size and operation entered into Cox model), and the right side of the table shows the results when size was adjusted for all clinical variables stated.

Adjusted for Operation Type							Adjusted for Age + Gender + Surgery + Oncology + MGMT						
	Tumour Diameter		Whole Volume		Core Volume			Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)	Sample size	Diameter	log(diameter)	WV	log(WV)	CV	log(CV)
50	19.01	14.78	19.45	26.30	17.15	11.93	50	19.53	16.39	14.95	15.54	14.87	12.93
100	31.24	21.54	32.15	47.22	28.95	16.28	100	26.74	20.39	16.13	20.97	19.15	13.78
150	42.94	28.75	44.56	64.84	40.92	21.16	150	35.14	26.01	18.29	27.80	25.15	16.32
200	53.50	36.03	55.58	77.80	51.63	26.03	200	43.47	32.24	20.47	34.84	31.77	19.38
250	62.47	42.82	65.10	86.42	61.30	30.92	250	51.34	38.30	23.06	41.89	38.19	22.67
258[a]	67.16	46.66	69.87	89.94	66.05	33.61	258[a]	55.67	41.79	24.57	45.72	41.89	24.50

CV = Core Volume, WV = Whole Volume, MGMT = O6-methylguanine-DNA methyltransferase. [a] Maximum sample size limited to 258 due to number of cases with a known MGMT result

In the multivariable resampling experiment, increasing sample size increased the percentages models, either adjusted for operation type or all clinical variables, in which the tumour size variable's Cox regression coefficient had a Wald test p-value < 0.05. The impact of increasing sample size was much greater than compared with univariable modelling. Again, the distributions of p-values (comparing $n = 50$ vs $n = 279$) differed significantly on two-sided KS testing (all test p-values < 0.0001). Log-transformation consistently increased the percentages of multivariable models with WV regression coefficient Wald test p-values < 0.05 (**Table 3.4**).

Figure 3.3 show the distributions of the p-values extracted from models during the univariable models in the resampling experiment. For CV and WV, but not diameter, there was a modest downwards trend as sample size increased, suggesting that this increased the probability of seeing a prognostic effect.

Figure 3.4 shows the distribution of p-values across resamples for the regression coefficients of each tumour size variable within multivariable models, which have either been adjusted for only operation type (**Figure 3.4a, c & e**) or adjusted for all clinical variables (**Figure 3.4b, d & f**), at different sample sizes. These charts showed a much greater downwards trend for all size variables, and the consistent effect of log-transformation in shifting the p-value distribution of WV downwards in multivariable modelling. Overall, results from univariable and multivariable resampling indicated that increased sample size for all size parameters and, in the case of WV, log transformation increased the chances of showing a significant univariable and multivariable association with OS.

3.5 Discussion

This study set out to explore the prognostic effect of tumour volume and diameter in this institutional cohort of patients with glioblastoma, and specifically to examine how choice of sample size and consideration of non-linear transformations may impact on the chances of detecting a prognostic effect. Univariable models did not provide any evidence of a linear or non-linear prognostic

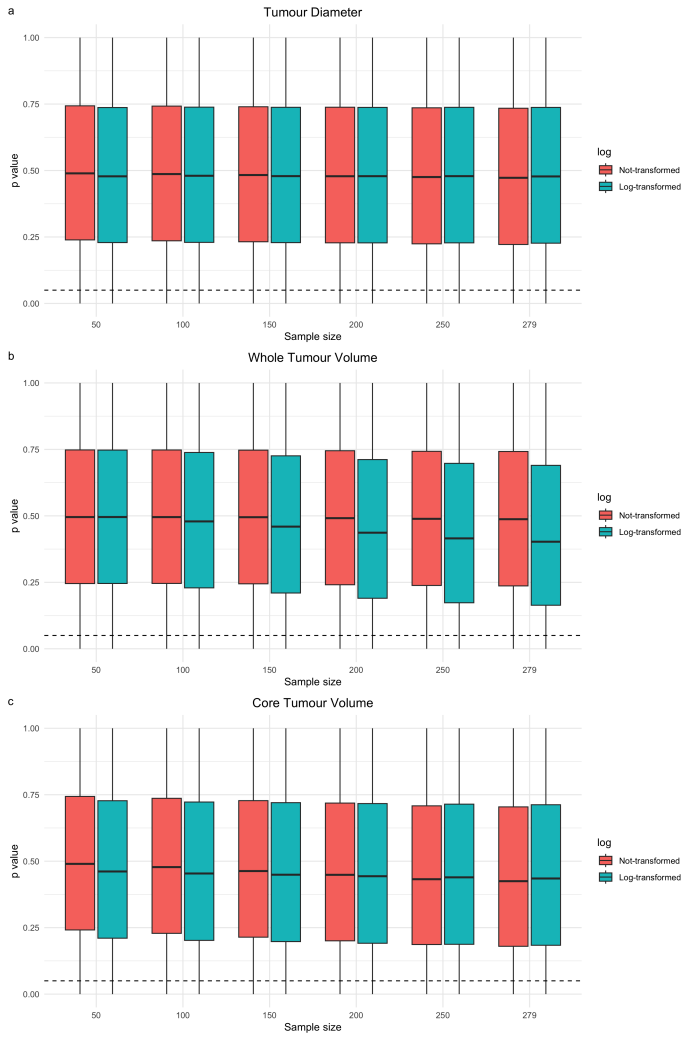

Figure 3.3: Three sets of box plots showing the distribution of p-values (y-axis) extracted from each univariable model for tumour diameter (a), whole volume (b) or core volume (c) vs overall survival created across the 1,000,000 repetitions for each sample size (x-axis). Boxes outline the interquartile range of p-values, with median values indicated by the central, thick black line. Tails represent the range of the distribution. Models with and without log-transformation are shown side by side (see figure legends). The dotted horizontal lines represents the p-value threshold for statistical significance (0.05).

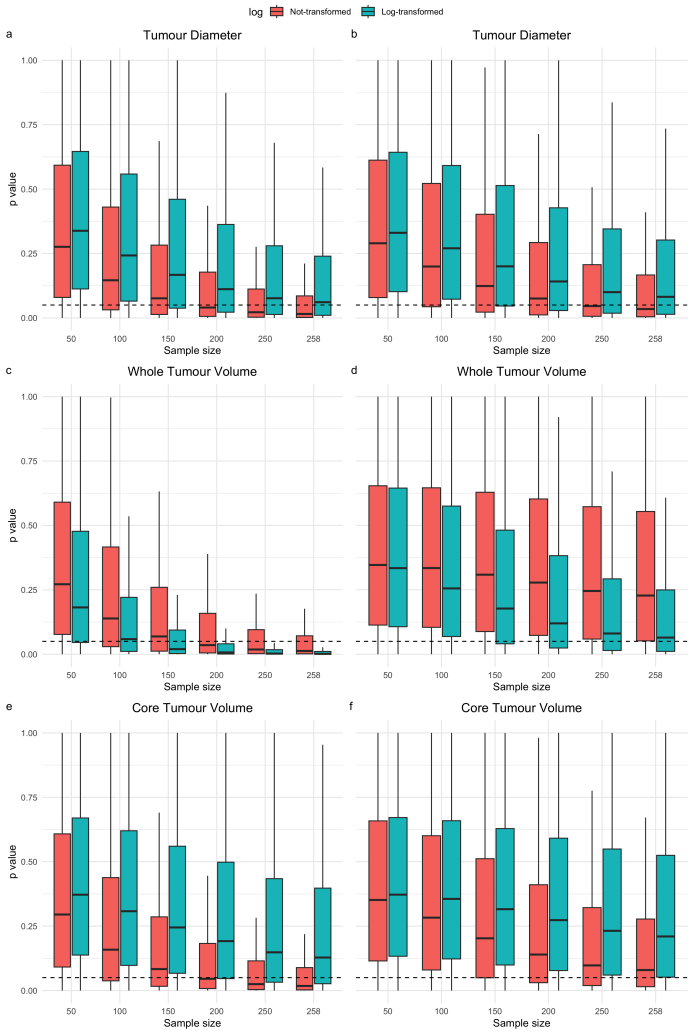

Figure 3.4: Six sets of box plots showing the distribution of p-values (y-axis) for the Wald test of the regression coefficient of tumour diameter (a & b), whole volume (c & d) or core volume (e & f) in a multivariable Cox model vs overall survival created across the 1,000,000 repetitions for each sample size (x-axis). Left column of graphs (a, c & e) show results from models including the selected tumour size variable and operation type only, and right column (b, d & f) show results from models adjusted for all clinical variables. Boxes outline the interquartile range of p-values, with median values indicated by the central, thick black line. Tails represent 1.5× the interquartile range of the distribution (outliers not shown). Models with and without log-transformation are shown side by side (see figure legend). The dotted horizontal lines represents the p-value threshold for statistical significance (0.05).

relationship between size and OS, however the multivariable models and resampling experiments showed that there is a prognostic role for tumour size, and that this is more likely to be detected in larger samples. Tumour diameter was prognostic in multivariable models adjusted for operation type and all clinical variables combined, whereas CV and WV were prognostic for the operation adjusted model and showed evidence of potential prognostic effects in the combined multivariable model as well as the resampling experiments.

For WV, log-transformation could also increase the probability of detecting a statistically significant effect, potentially due to the positively skewed distribution. WV might play a role in prognostication even when adjusted for extent of tumour core resection as illustrated in these multivariable models and resampling experiments and this could be due to WV encompassing more of the infiltrated brain tissue. However, WV is infrequently explored as a candidate prognostic variable in patients with glioblastoma [4, 34–36]. In 65 patients, Iliadis et al. found no significant association between WV and OS using univariable Cox modelling [35], and it is unclear if any log or other transformation was considered. Palpan Flores et al. investigated the equivalent of WV in 44 IDH-wildtype glioblastoma patients and found a significant effect of WV $> 60cm^3$ in univariable and multivariable models (adjusted HR 3.93 95% CI 1.23-10.2, $p = 0.018$) [36]. Other groups have investigated peritumoural oedema alone, rather than WV, and these studies have shown mixed results [16, 34, 37–39]. Fuster-Garcia et al. found no prognostic effect for peritumoural oedema volume in 84 patients [37], whereas Wangaryattawanich et al. showed a statistically significant effect for peritumoural oedema when dichotomising volume using a threshold of $85,000mm^3$ in a cohort of 94 patients [16]. Although the multivariable model adjusted for all clinical parameters in the complete cohort did not show a statistically significant result for log-transformed WV, the confidence interval for its hazard ratio was relatively wide, suggesting a higher degree of uncertainty in the result. Second, the results of the multivariable resampling for the log(WV) full clinical model suggested that at the sample sizes used in the above cited literature there is a lower chance of detecting the potentially prognostic role than in this cohort study.

For tumour diameter and CV, which are more commonly investigated [1, 4], there are several studies

in similarly sized institutional datasets that did not show a prognostic effect for either CV [12, 34, 40] or diameter [17, 41]. However, a small positive effect size has been shown in studies with larger datasets [13, 14]. These findings support the initial multivariable model and experiment resampling findings, which indicates that diameter does have a small multivariable prognostic effect and that CV could potentially have prognostic effects, after adjustment for other clinical parameters and in larger samples. Li et al. for example found that contrast enhancing tumour volume had a small but statistically significant effect in a cohort of 1226 glioblastoma patients (HR 1.004 95% CI 1.002-1.006, $p < 0.001$) [14]. Senders et al. also showed a small (relative survival rate per centimetre increase in diameter - 0.99 95% CI 0.99-1.00) but significant effect of tumour diameter in 16,656 patients [13]. Whilst it could be argued that such a small effect size is not clinically significant, the aim of this study was not to produce a prognostic model for clinical use but to identify the barriers to detecting potentially significant effects in glioblastoma prognostic models, and suggest that resampling and data transformation can have a role in highlighting uncertainty of predictor selection in relatively small datasets. Future studies would benefit from leveraging multi-institutional networks [23] or online imaging repositories to further increase statistical power for studying clinically relevant size parameters including volumetric assessments.

An important consideration in regression modelling is the inclusion of any non-linear transformation of variables [42], although this is not routinely documented in prognostic modelling studies in glioblastoma [1, 4]. The advantages include more flexible modelling of continuous variables that might not have a simple linear relationship to outcome, but this comes with the drawback of potentially overfitting a model to the development dataset. In this univariable resampling study, logarithmic transformation led to a modestly higher rate of significant models for WV, and this effect was much greater in multivariable models. The results of this resampling study point to a possible explanation as to why some prognostic models that assume linear relationships between volume and outcome return non-significant results, particularly in the case of WV. In the present study, the WV shows a small positive skew and large range that might explain why log-transformation increased the chances of detecting a potential prognostic relationship for WV.

Bootstrapping (resampling with replacement) as a method for exploring model uncertainty has been described elsewhere in the statistics literature [18, 19, 42] and was used in this resampling study to demonstrate the variability in the prognostic effect of tumour size. By resampling a dataset multiple times, researchers can identify the variability in multiple aspects of the model building process, such as feature selection, internal validation of model accuracy and model stability [18]. There are other, less computationally intensive, methods available for calculating 95% CIs such as using the standard error of the statistical test, but this assumes a normal distribution of data, whereas bootstrapping allows multiple steps of the modelling process to be included in the uncertainty estimate [19]. This study suggests that when selecting one of these size variables in glioblastoma prognostic models based on univariable model significance, there could be up to 5-10% uncertainty in whether they might be statistically significant, and therefore included in a multivariable model if using this as a selection criteria. The instability is shown to be even greater in the multivariable resampling and there could be a large range in the certainty as to whether a variable is prognostic based on a limited sample size. This is one of the reasons that this approach of univariable screening of candidate predictors is generally not recommended for multivariable model building, and also why a focus on p-values in multivariable modelling may lose some of the important information in estimating prognostic effects [18, 19]. Preferred methods are to preselect variables for inclusion using expert knowledge or using an unsupervised method, such as clustering or principle component analysis, which do not rely on a significance test.

There are several limitations of this study. The MRI acquisition parameters were heterogeneous, especially slice thickness, and this could have impacted upon the accuracy of volume measurements. However, the spatial resampling of images to an isotropic $1mm^3$ voxel resolution should have reduced the impact of acquisition heterogeneity. Furthermore, the dataset represents a retrospective real-world clinical dataset, which in this institution's routine practice is likely to include different imaging acquisitions due to patients being referred from other centres, with their own (varying) MRI protocols. A proportion of the patients had to be excluded due to lack of the necessary MRI sequences for the deep-learning segmentation algorithm. The efficiency of a semi-automated segmentation approach outweighed the potential limitation of a reduced sample size. Only three size

variables were investigated but there are many others described in the literature. Whilst this could be deemed a limitation of this approach, the aim was not to provide a comprehensive study of the prognostic role of all possible tumour size parameters in glioblastoma, but to investigate some of the methodological issues affecting this question, and that could be applied to any of the other continuous measures of tumour size in glioblastoma. Important prognostic factors for OS prediction were not available in this dataset, such as performance status, second-line treatment or trial treatments and it would have been useful to adjust for this in the modelling process given its importance to clinical practice. However, the aim of the study was to illustrate important methodological factors in prognostic factor modelling for tumour size and lack of a complete dataset does not detract from the conclusions.

3.6 Conclusion

Univariable models derived from this large, well curated institutional dataset of patients with glioblastoma showed limited evidence to support a linear or non-linear prognostic association between size and patient outcome, however the multivariable models did support a prognostic role for tumour size. Diameter showed a significant multivariable association with survival, whereas CV and log(WV) showed significant effects when adjusted for operation type and potential for an effect in the full clinical model. Importantly, the resampling experiments demonstrate the impact that increasing sample size and for WV, log-transformation, has in increasing the ability to detect prognostic relationships in univariable and multivariable models.

References

1. Tewarie, I. A. *et al.* 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 (2021).

2. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. English. *Radiology* **278,** 563–577 (Feb. 2016).

3. Currie, S. *et al.* A Comprehensive Clinical Review of Adult-Type Diffuse Glioma Incorporating the 2021 World Health Organization Classification. *Neurographics* **12,** 43–70 (Apr. 2022).

4. Henker, C., Kriesen, T., Glass, Ä., Schneider, B. & Piek, J. Volumetric quantification of glioblastoma: experiences with different measurement techniques and impact on survival. *Journal of Neuro-Oncology* **135,** 391–402 (2017).

5. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

6. Rudie, J. D., Rauschecker, A. M., Bryan, R. N., Davatzikos, C. & Mohan, S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology* **290,** 607–618 (Mar. 2019).

7. Forghani, R. Precision Digital Oncology: Emerging Role of Radiomics-based Biomarkers and Artificial Intelligence for Advanced Imaging and Characterization of Brain Tumors. *Radiology: Imaging Cancer* **2,** e190047 (2020).

8. Sanai, N., Alvarez-Buylla, A. & Berger, M. S. Neural Stem Cells and the Origin of Gliomas. *New England Journal of Medicine* **353,** 811–822 (2005).

9. Laug, D., Glasgow, S. M. & Deneen, B. A glial blueprint for gliomagenesis. *Nature Reviews Neuroscience* **19,** 393–403 (2018).

10. Peeken, J. C. *et al.* Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients. *Strahlentherapie und Onkologie* **194,** 580–590 (2018).

11. Lacroix, M. *et al.* A multivariate analysis of 416 patients with glioblastoma multiforme: Prognosis, extent of resection, and survival. *Journal of Neurosurgery* **95,** 190–198 (2001).

12. Compter, I. *et al.* Deciphering the glioblastoma phenotype by computed tomography radiomics. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **160,** 132–139 (2021).

13. Senders, J. T. *et al.* An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Clinical Neurosurgery* **86,** E184–E192 (2020).

14. Li, Y. M., Suki, D., Hess, K. & Sawaya, R. The influence of maximum safe resection of glioblastoma on survival in 1229 patients: Can we do better than gross-total resection? *Journal of Neurosurgery* **124,** 977–988 (2016).

15. Molinaro, A. M. *et al.* Association of Maximal Extent of Resection of Contrast-Enhanced and Non-Contrast-Enhanced Tumor with Survival Within Molecular Subgroups of Patients with Newly Diagnosed Glioblastoma. *JAMA Oncology* **6,** 495–503 (2020).

16. Wangaryattawanich, P. *et al.* Multicenter imaging outcomes study of The Cancer Genome Atlas glioblastoma patient cohort: Imaging predictors of overall and progression-free survival. *Neuro-Oncology* **17,** 1525–1537 (2015).

17. Park, M. *et al.* Elderly patients with newly diagnosed glioblastoma: can preoperative imaging descriptors improve the predictive power of a survival model? *Journal of Neuro-Oncology* **134,** 423–431 (2017).

18. Harrell, F. E. *Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis* Second edi (Springer, Cham, 2015).

19. Riley, R. D., van der Windt, D., Croft, P. & Moons, K. G. *Prognosis Research in Healthcare: Concepts, Methods, and Impact* (eds Riley, R. D., van der Windt, D. A., Croft, P. & Moons, K. G.) (Oxford University Press, Oxford, UK, 2019).

20. Shinojima, N. *et al.* The influence of sex and the presence of giant cells on postoperative long-term survival in adult patients with supratentorial glioblastoma multiforme. *Journal of neurosurgery* **101,** 219–226 (2004).

21. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

22. Icometrix. *dicom2nifti*

23. Pati, S. *et al.* Federated Learning Enables Big Data for Rare Cancer Boundary Detection (2022).

24. Davatzikos, C. *et al.* Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of Medical Imaging* **5,** 1 (Jan. 2018).

25. Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping* **31,** 798–819 (Dec. 2009).

26. Yushkevich, P. A. *et al.* IC-P-174: Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. *Alzheimer's & Dementia* **12,** 126–127 (July 2016).

27. Thakur, S. *et al.* Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage* **220,** 117081 (Oct. 2020).

28. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18,** 203–211 (Feb. 2021).

29. Zou, K. H. *et al.* Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology* **11,** 178–189 (2004).

30. Therneau, T. M. *A Package for Survival Analysis in R* 2023.

31. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. & Schmid, M. A review of spline function procedures in R. *BMC Medical Research Methodology* **19,** 1–16 (2019).

32. Verduin, M. *et al.* Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma. *Cancers* **13,** 1–20 (2021).

33. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging* **15** (2015).

34. Henker, C. *et al.* Volumetric assessment of glioblastoma and its predictive value for survival. *Acta Neurochirurgica* **161,** 1723–1732 (2019).

35. Iliadis, G. *et al.* Volumetric and MGMT parameters in glioblastoma patients: Survival analysis. *BMC Cancer* **12** (2012).

36. Palpan Flores, A. *et al.* Assessment of Pre-operative Measurements of Tumor Size by MRI Methods as Survival Predictors in Wild Type IDH Glioblastoma. *Frontiers in Oncology* **10,** 1–12 (Sept. 2020).

37. Fuster-Garcia, E. *et al.* Improving the estimation of prognosis for glioblastoma patients by MR based hemodynamic tissue signatures. *NMR in Biomedicine* **31,** 1–10 (2018).

38. Zhang, Z. *et al.* Identifying the survival subtypes of glioblastoma by quantitative volumetric analysis of MRI. *Journal of Neuro-Oncology* **119,** 207–214 (Aug. 2014).

39. Choi, Y. *et al.* Analysis of heterogeneity of peritumoral T2 hyperintensity in patients with pretreatment glioblastoma: Prognostic value of MRI-based radiomics. *European Journal of Radiology* **120,** 108642 (2019).

40. Kickingereder, P. *et al.* Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-Oncology* **20,** 848–857 (2018).

41. Lutterbach, J., Sauerbrei, W. & Guttenberger, R. Multivariate analysis of prognostic factors in patients with glioblastoma. *Strahlentherapie und Onkologie* **179,** 8–15 (2003).

42. Steyerberg, E. W. *Clinical prediction models : a practical approach to development, validation, and updating* (Springer, New York ; 2009).

3.7 Supplementary Materials

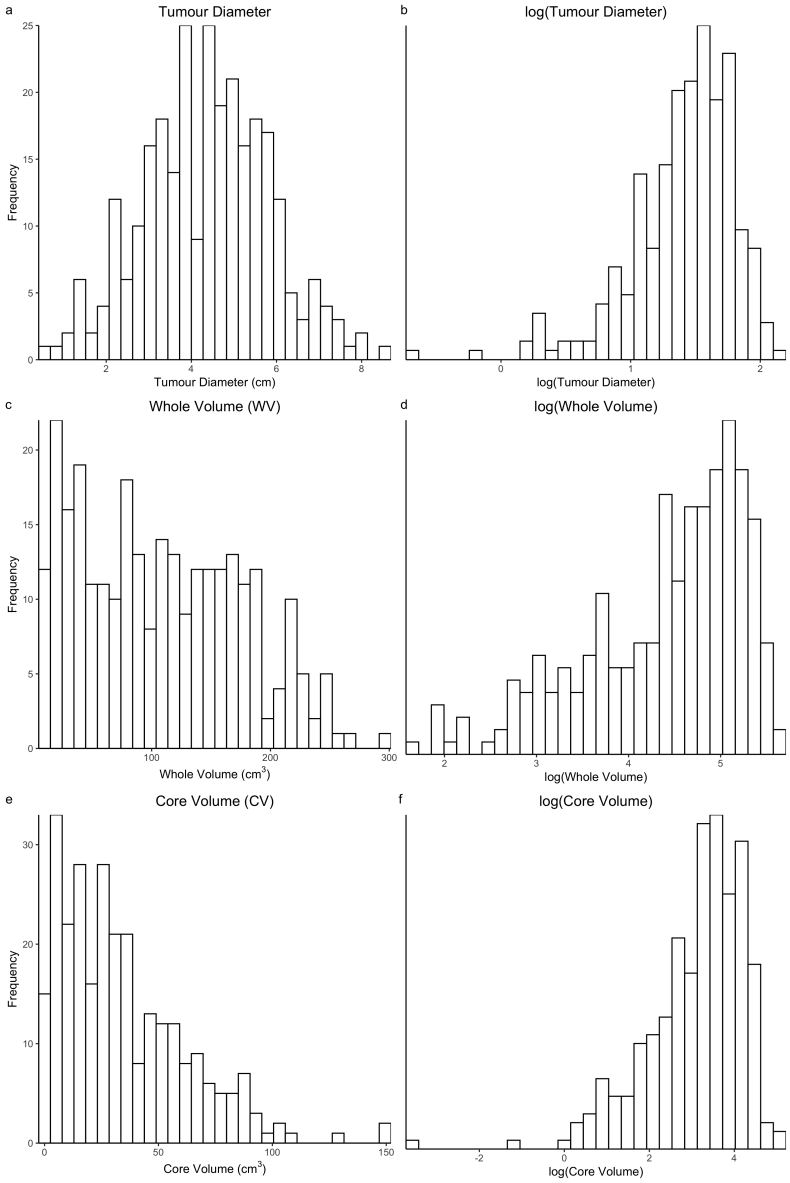

Figure S3.1: Panels a,c and e show histograms of tumour diameter, whole volume and core volume, respectively without any transformation. Panels b,d and f show histograms of these parameters after log transformation.

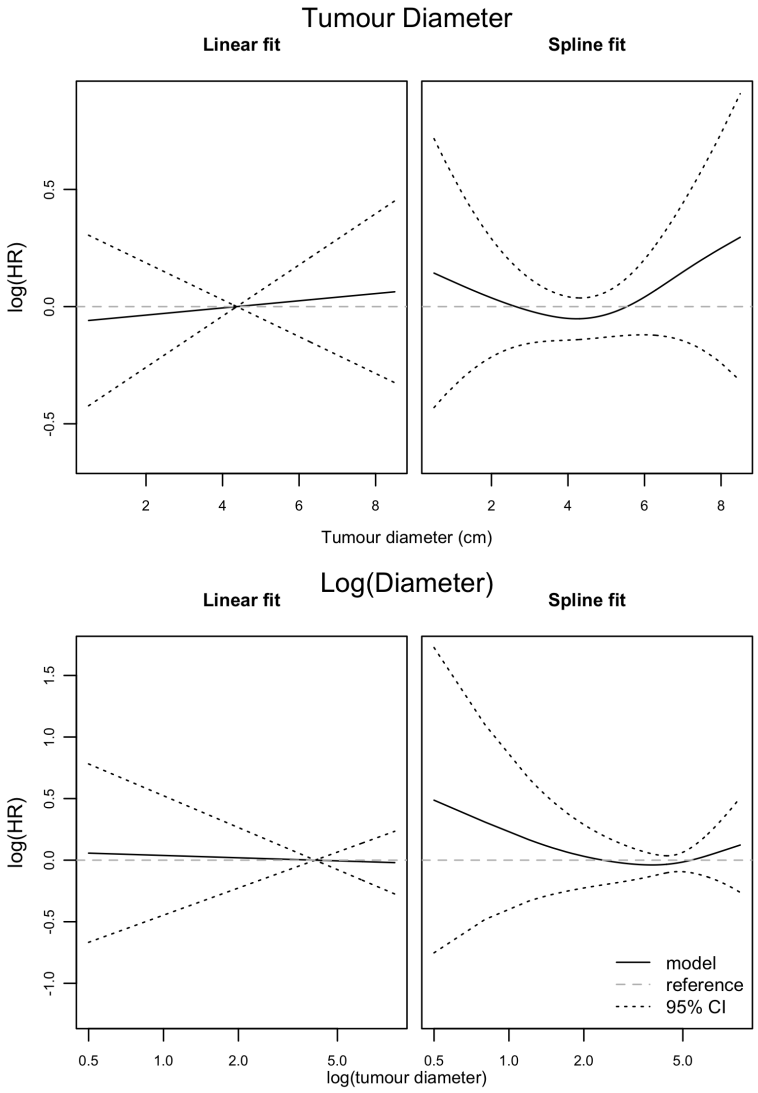

Figure S3.2: Top row shows models before and bottom row after log-transformation. The left column illustrates a linear and the right shows a non-linear fit to the data points, the latter with penalised splines; HR = hazard ratio.

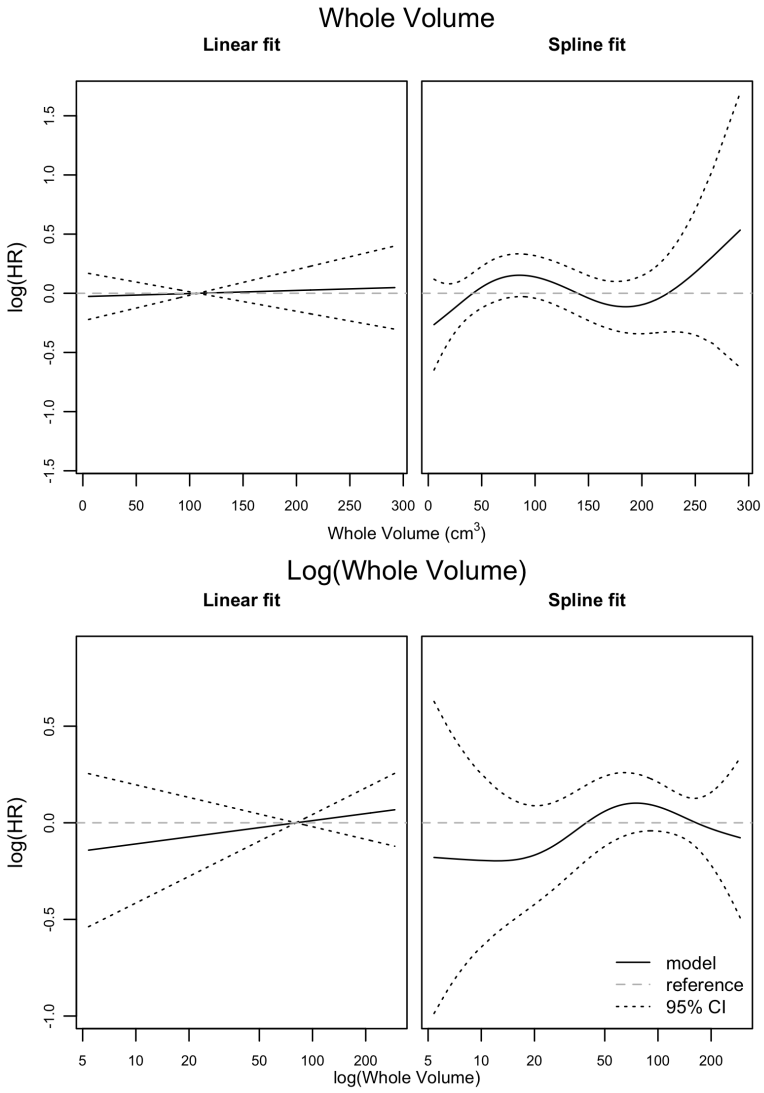

Figure S3.3: Top row shows models before and bottom row after log-transformation. The left column illustrates a linear and the right shows a non-linear fit to the data points, the latter with penalised splines; HR = hazard ratio.

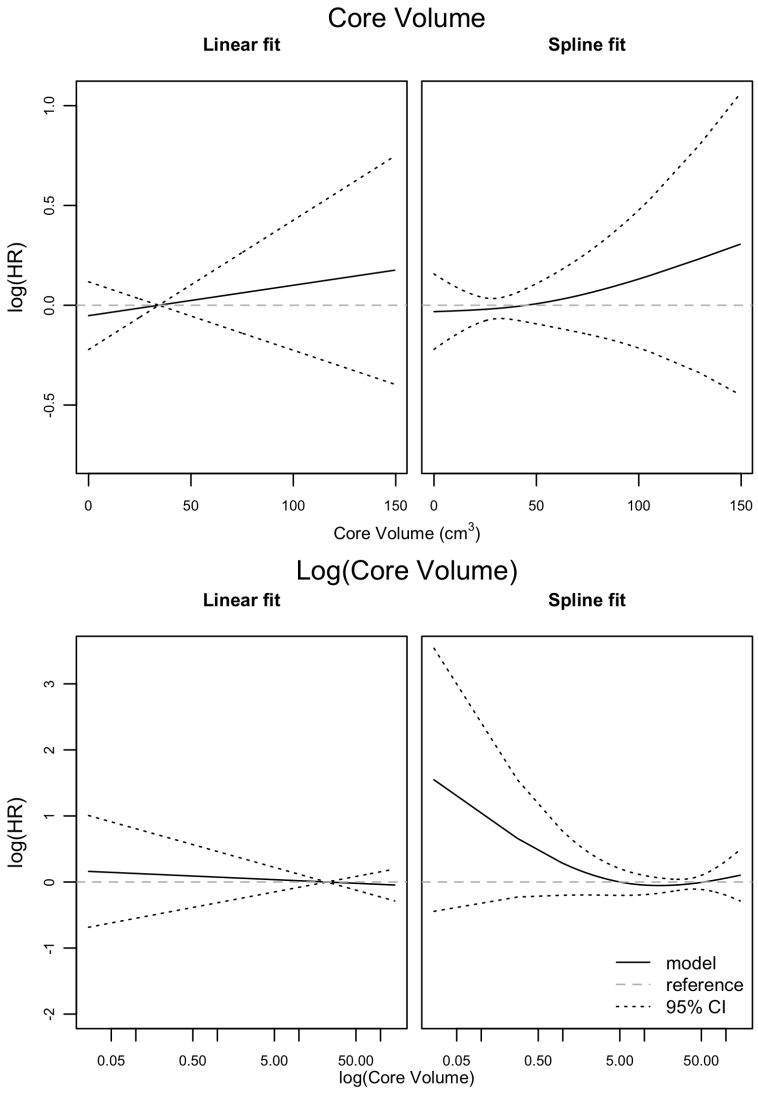

Figure S3.4: Top row shows models before and bottom row after log-transformation. The left column illustrates a linear and the right shows a non-linear fit to the data points, the latter with penalised splines; HR = hazard ratio.

Table S3.1: MRI acquisition parameters per sequence

Sequence	Parameter Summary
T1	Slice thickness $5mm$ (0 – $5mm$), spacing between slices $5.5mm$ (0.6 – $7mm$), echo time $7.8ms$ (3.0 – $58ms$), repetition time $550ms$ (7.0 – $3200ms$), field strength $1.5T$ (1.5 – 3.0 T), flip angle 90° (8 – 150°)
T2	Slice thickness $5mm$ (1.2 – $7mm$), spacing between slices $5.5mm$ (0.6 – $7.7mm$), echo time $98ms$ (25 – $171ms$), repetition time $5268.6ms$ (660 – $6600ms$), field strength $1.5T$ (1.5 – $3.0T$), flip angle 150° (20 – 180°)
FLAIR	Slice thickness $5mm$ (0.7 – $5mm$), spacing between slices $5.5mm$ (0.6 – $7mm$), echo time $109ms$ (82 – $474ms$), inversion time $2500ms$ (1660 – $2880ms$), repetition time $9000ms$ (4610 – $14788ms$), field strength $1.5T$ (1.5 – $3.0T$), flip angle 150° (90 – 180°)
T1CE	Slice thickness $1.1mm$ (0 – $7mm$), spacing between slices $5.9mm$ (0.5 – $7.7mm$), echo time $3.9ms$ (2.3 – $46ms$), repetition time $700ms$ (7.5 – $3200ms$), field strength $1.5T$ (1.5 – $3.0T$), flip angle 15° (8 – 150°)

Values for acquisition parameters are presented as median (range). FLAIR = Fluid attenuated inversion recovery, T1CE = post-gadolinium contraste-enhanced T1-weighted imaging.

Table S3.2: Univariable association between clinical variables and overall survival

Clinical Predictor	n	Event n	HR	95% CI	p-value
Age	279	236	1.01	1.00 - 1.03	0.06
Sex	279	236			
Female			—	—	
Male			1.26	0.96 - 1.64	0.091
Operation[a]	279	236			
Biopsy			—	—	
100% resected			0.34	0.23 - 0.50	<0.001
$\geq$ 90% resected			0.38	0.27 - 0.54	<0.001
$<$ 90% resected			0.5	0.35 - 0.71	<0.001
Stupp	279	236			
No Stupp			—	—	
Full Stupp[b]			0.29	0.20 - 0.41	<0.001
Partial Stupp[c]			0.49	0.36 - 0.67	<0.001
MGMT	258[d]	219			
Unmethylated			—	—	
Methylated			0.56	0.42 - 0.74	<0.001

HR = Hazard Ratio; CI = Confidence Interval. [a] Percentage of contrast enhancing and necrotic tumour core removed
[b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and 6 cycles adjuvant temozolomide
[c] Completed 60Gy in 30 fractions radiotherapy but stopped temozolomide either during radiotherapy or adjuvant course
[d] Number of cases with known result

Table S3.3: Adjusted prognostic effect of clinical variables

Clinical Predictor	HR	95% CI	p-value
Age	1	0.98 - 1.02	0.98
Sex			
Female	—	—	
Male	1.31	0.99 - 1.75	0.061
Operation[a]			
Biopsy	—	—	
100% resected	0.38	0.25 - 0.56	<0.001
$\geq$ 90% resected	0.36	0.25 - 0.52	<0.001
$<$ 90% resected	0.43	0.29 - 0.63	<0.001
Stupp			
No Stupp	—	—	
Full Stupp[b]	0.34	0.23 - 0.50	<0.001
Partial Stupp[c]	0.56	0.39 - 0.79	<0.001
MGMT			
Unmethylated	—	—	
Methylated	0.67	0.50 - 0.90	<0.001

HR = Hazard Ratio; CI = Confidence Interval. Multivariable model included age, sex, operation type, Stupp status, and MGMT methylation. $n = 258$ (219 events) = cases with complete results for all clinical variables.

[a] Percentage of contrast enhancing and necrotic tumour core removed [b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and 6 cycles adjuvant temozolomide

[c] Completed 60Gy in 30 fractions radiotherapy but stopped temozolomide either during radiotherapy or adjuvant course

Table S3.4: Percentage of total number of resamples (1,000,000), in which the univariable tumour size models had a model p-value < 0.05.

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	5.43	5.95	5.14	5.20	5.07	7.07
100	5.60	5.94	5.12	5.94	5.73	7.57
150	5.84	6.00	5.21	6.93	6.53	8.00
200	6.05	6.04	5.35	8.00	7.32	8.36
250	6.25	6.07	5.53	9.04	8.15	8.68
279	6.39	6.13	5.60	9.70	8.60	8.94

CV = Core Volume; WV = Whole Volume.

Table S3.5: Percentage of total number of resamples (1,000,000), in which the univariable tumour size models had a model p-value < 0.01.

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	1.12	1.31	1.05	1.06	0.91	1.66
100	1.16	1.26	1.03	1.26	1.13	1.91
150	1.26	1.31	1.05	1.57	1.39	2.12
200	1.34	1.33	1.09	1.90	1.66	2.26
250	1.40	1.34	1.15	2.29	1.93	2.39
279	1.44	1.33	1.16	2.51	2.08	2.49

CV = Core Volume; WV = Whole Volume.

Table S3.6: Percentage of total number of resamples (1,000,000), in which the univariable tumour size models had a model p-value < 0.001.

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	0.11	0.15	0.11	0.11	0.08	0.21
100	0.12	0.14	0.10	0.14	0.10	0.27
150	0.14	0.14	0.10	0.18	0.14	0.30
200	0.15	0.15	0.11	0.22	0.18	0.34
250	0.16	0.15	0.11	0.30	0.22	0.36
279	0.17	0.15	0.12	0.33	0.24	0.39

CV = Core Volume; WV = Whole Volume.

Table S3.7: Percentage of resamples in which the multivariable tumour size model adjusted for patient age (ie. Age + tumour size in model) has a tumour size regression coefficient Wald test p-value < 0.05

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	5.61	6.61	5.60	5.00	5.71	7.70
100	5.53	6.75	5.30	5.00	6.30	8.30
150	5.47	6.96	5.22	5.42	7.01	8.87
200	5.44	7.23	5.21	5.90	7.68	9.37
250	5.43	7.49	5.19	6.39	8.43	9.98
258	5.45	7.64	5.25	6.65	8.82	10.28

CV = Core Volume; WV = Whole Volume.

Table S3.8: Percentage of resamples in which the multivariable tumour size model adjusted for patient gender (ie. gender + tumour size in model) has a tumour size regression coefficient Wald test p-value < 0.05

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	5.40	6.63	5.58	4.74	5.33	8.30
100	5.30	6.65	5.28	4.78	5.54	9.08
150	5.27	6.78	5.19	5.10	5.77	9.82
200	5.24	6.95	5.12	5.57	6.07	10.62
250	5.27	7.15	5.07	6.03	6.32	11.39
258	5.20	7.28	5.10	6.33	6.48	11.90

CV = Core Volume; WV = Whole Volume.

Table S3.9: Percentage of resamples in which the multivariable tumour size model adjusted for adjuvant oncology treatment (ie. oncology treatment + tumour size in model) has a tumour size regression coefficient Wald test p-value < 0.05

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	9.50	9.74	7.59	6.94	7.41	10.43
100	9.30	10.10	6.97	6.65	7.68	11.08
150	9.18	10.48	6.72	6.65	7.92	11.73
200	9.08	10.96	6.51	6.67	8.23	12.30
250	9.08	11.42	6.50	6.83	8.53	12.94
258	9.03	11.77	6.42	6.94	8.77	13.22

CV = Core Volume; WV = Whole Volume.

Table S3.10: Percentage of resamples in which the multivariable tumour size model adjusted for MGMT methylation status (ie. MGMT methylation + tumour size in model) has a tumour size regression coefficient Wald test p-value < 0.05

	Tumour Diameter		Whole Volume		Core Volume	
Sample size	Diameter	log(Diameter)	WV	log(WV)	CV	log(CV)
50	6.52	8.29	6.31	5.85	5.89	10.80
100	6.78	8.51	6.03	5.86	6.07	11.44
150	7.05	8.55	5.89	6.26	6.27	11.87
200	7.29	8.57	5.79	6.75	6.53	12.23
250	7.63	8.54	5.75	7.34	6.86	12.58
258	7.80	8.52	5.74	7.63	6.97	12.74

CV = Core Volume; WV = Whole Volume.

CHAPTER

FOUR

IMPACT OF INTENSITY STANDARDISATION AND COMBAT BATCH SIZE ON CLINICAL – RADIOMIC PROGNOSTIC MODELS PERFORMANCE IN A MULTI-CENTRE STUDY OF PATIENTS WITH GLIOBLASTOMA

4.1 Abstract

4.1.1 Background

To assess the effect of different ISTs and ComBat batch sizes on radiomics survival model performance and stability in a heterogeneous, multi-centre cohort of patients with glioblastoma.

4.1.2 Methods

Multi-centre pre-operative MRI acquired between 2014-2020 in patients with IDH-wildtype unifocal WHO grade 4 glioblastoma were retrospectively evaluated. WhiteStripe (WS), Nyul histogram matching (HM) and Z-score (ZS) ISTs were applied before radiomic feature (RF) extraction. RFs were realigned using ComBat and minimum batch size (MBS) of 5, 10 or 15 patients.

Cox proportional hazards models for overall survival (OS) prediction were produced using five different selection strategies and the impact of IST and MBS was evaluated using bootstrapping. Calibration, discrimination, relative explained variation, and model fit were assessed. Instability was evaluated using 95% confidence intervals (95% CIs), feature selection frequency and calibration curves across the bootstrap resamples.

4.1.3 Results

195 patients were included. Median OS = 13 (95% CI 12-14) months. 12-14 unique MRI protocols were used per MRI sequence. HM and WS produced the highest relative increase in model discrimination, explained variation and model fit but IST choice did not greatly impact on stability, nor calibration. Larger ComBat batches improved discrimination, model fit and explained variation but higher MBS (reduced sample size) reduced stability (across all performance metrics) and reduced calibration accuracy.

4.1.4 Conclusion

Heterogeneous, real-world glioblastoma data poses a challenge to the reproducibility of radiomics. ComBat generally improved model performance as MBS increased but reduced stability and calibration (i.e. overfit). HM and WS tended to improve model performance.

4.2 Introduction

Translation of radiomics to clinical practice has been hampered by a lack of reproducibility linked to variables introduced in multi-centre imaging [1–3]. Intensity standardisation (IS) homogenizes the scale and distribution of MRI signal intensity, which is affected by imaging protocol [4], however, there is no consensus on the best IST [5, 6].

Statistical realignment of radiomic features using ComBat, can also reduce the effect of the different imaging acquisitions [7, 8]. ComBat requires sufficient data to estimate these 'batch' effects and the minimum batch size (MBS) must be chosen to ensure accurate results [7, 8]. MBS choice not only affects ComBat performance, but also discards some of the data within heterogeneous, real-world images.

Inconsistent statistical modelling, which in glioblastoma, has tended to focus on prognostic separation ('discrimination') [5, 6] may also play a role in the lack of reproducibility. Model calibration and stability are important but less well evaluated [9]. Calibration compares predictions to observed survival and stability examines the consistency of model performance [10]. To date, the effect of ISTs and ComBat MBS choice have not been thoroughly assessed on model calibration and stability in a multi-centre setting. Thus, the aim of this study was to assess the effect of ISTs and ComBat MBS choice on the calibration, discrimination, relative model fit and explain variation, and stability of prognostic models in a heterogeneous, real-world cohort of patients with glioblastoma, rather than producing the most accurate prognostic model for OS prediction in glioblastoma.

4.3 Materials and Methods

4.3.1 Ethical approval

This was a retrospective study and therefore informed patient consent was not feasible. Ethical approval and institutional data access was approved via local ethical review committee (REC

ref: 19/YH/0300, IRAS project ID: 255585, see section 5). A completed Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [11] is provided in **Table S4.1**.

4.3.2 Patient selection and characteristics

A description of the patient cohort, selection criteria, data collection and image preparation has been previously outlined in section 3.3.2 [12]. Inclusion criteria were: adults ($\geq$ 16) with histologically proven glioblastoma according to 2021 World Health Organisation classification of central nervous system tumours treated between 2014-2020, MRI performed prior to any surgery, unifocal tumour, and all four of the following MRI sequences acquired: T1W, T2W, FLAIR and T1CE sequences. Exclusion criteria were: absence of pre-operative MRI, significant degradation of imaging due to artefact, multifocal tumours at presentation and, IDH mutation.

4.3.3 Clinical predictors

Clinical predictors included patient age, sex and type of operation. Histopathological and cytogenetic data included histology, IDH1 and 2 mutation and MGMT promoter methylation. Maximum axial or cranio-caudal diameter of the enhancing tumour core and was measured using the T1CE sequence. Extent of resection was estimated using the immediate (48-72 hour) post-resection MRI and grouped based upon the amount of contrast enhancing and necrotic tumour resected – (i) 100%, (ii) $\geq$ 90% or (iii) $<$ 90%. Adjuvant treatment was categorized as (i) full Stupp protocol – 60 Gy in 30 fractions radiotherapy with concomitant and 6 cycles adjuvant temozolomide; (ii) partial Stupp – 60 Gy in 30 fractions radiotherapy but temozolomide discontinued during either concomitant or adjuvant treatment phase; (iii) non-Stupp – any other treatment protocol.

4.3.4 Image preparation and tumour segmentation

A graphical illustration of the whole experiment is provided (**Figure 4.1**). As a tertiary referral centre in the UK, it is standard practice for this institution to manage patients with glioblastoma

from the surrounding region (with a catchment of approximately 4 million people), which includes general hospitals ('hub-and-spoke' model). DICOM images were acquired across multiple centres across the region and historically transferred to the local institutional picture archive and communication system (PACS) to facilitate routine patient care (acquisition parameters are summarised in **Tables S4.2-S4.5**). DICOM image preparation was performed in python 3.9 [13]. DICOM data was retrieved from PACS, pseudonymised and converted to Neuroimaging Informatics Technology Initiative (NIfTI) file format.

NIfTI images were processed and segmented using the open-source platform FeTS software, designed for performing these tasks on MRIs from patients with glioblastoma [14]. T2W, T1CE and FLAIR sequences were rigidly co-registered to the T1W sequence, then to the SRI24 brain atlas [15], and spatially resampled to $1 \times 1 \times 1mm$ voxel resolution [16]. Images were skull-stripped [17] and tumours segmented using the 'nnU-net' deep-learning network with pretrained model weights [18]. Core volume (CV) was defined as enhancing and necrotic regions, and whole tumour volume (WTV) defined as CV plus peritumoural high T2 signal. Necrotic portions of tumour were included as there could be important quantitative signal captured in the proportion of tumour that has undergone necrosis, whereas this information would be lost if only the enhancing portion were included.

Segmentations were checked manually and corrected by a neuroradiology fellow (5 years radiology experience). Independently, 50 segmentations were also checked by a consultant neuroradiologist ($>$ 10 years consultant neuroradiology experience), and the inter-rater concordance compared using the dice similarity coefficient (DSC) [19]. Corrections were carried out in the FeTS software package.

MR field inhomogeneity correction was performed with the N4ITK algorithm within the simple ITK package (v2.1.1.2) [20] before applying one of three MRI intensity standardisation techniques (ISTs).

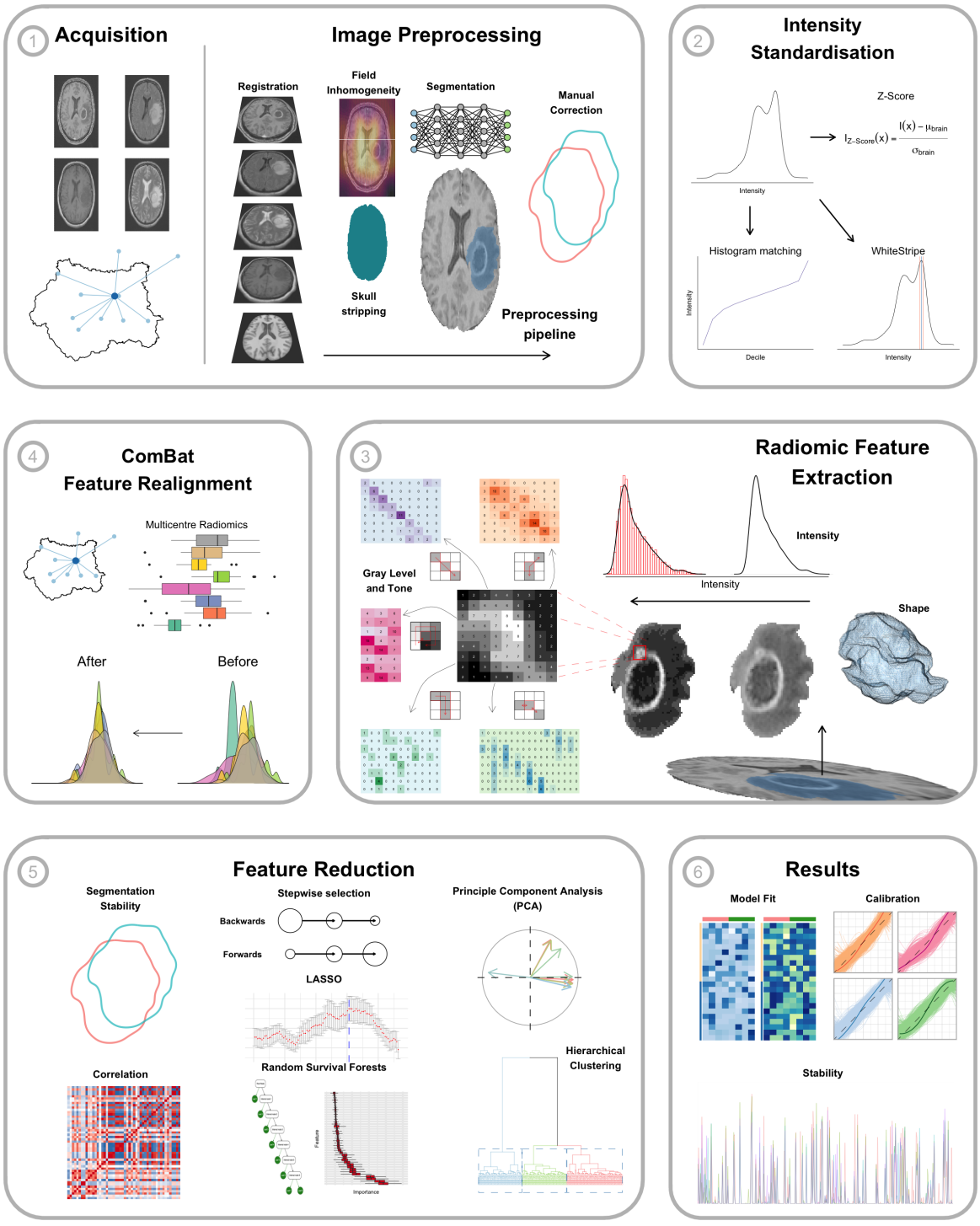

Figure 4.1: Panels 1-6 outline the main steps of the experiment: 1) The acquisition of MRI across multiple sites within the region and pre-processing including registration, skull stripping and field inhomogeneity correction; 2) Intensity Standardisation of MRI signal intensities; 3) Radiomic feature extraction, including calculation of shape, intensity and higher level features; 4) post-extraction realignment of multi-centre radiomics using ComBat; 5) Application of multiple feature reduction techniques to reduce the dimensionality of the data; 6) Calculation of results and data analysis. LASSO = Least Absolute Shrinkage and Selection Operator, PCA = Principle component analysis.

4.3.5 Intensity standardisation

Three ISTs that are commonly used in patients with glioblastoma [6] are WhiteStripe (WS) [21], Nyul histogram matching (HM) [22, 23] and Z-score (ZS). Full details for each technique and its implementation can be found online: https://github.com/jcreinhold/intensity-normalization.

Each IST was applied independent of the other, resulting in four separate images per sequence per patient (**Figure 4.2**) - one per IST, plus the non-standardised images that served as control (referred to a 'RAW' images hereafter).

4.3.5.1 Z-score (ZS)

For each standardised voxel ($I_{Z-Score}(x)$), the initial intensity ($I(x)$) is standardised by subtracting the mean intensity of all brain voxels (μ_{brain}), and then dividing it by the standard deviation of all the brain voxels' intensity values (σ_{brain}).

$$I_{Z-Score}(x) = \frac{I(x) - \mu_{brain}}{\sigma_{brain}}$$

4.3.5.2 WhiteStripe(WS)

WS standardises each voxel ($I_{WhiteStripe}(x)$) by subtracting the mean intensity of normal appearing white matter (NAWM, μ_{NAWM}) and then dividing by the standard deviation of the intensity of NAWM (σ_{NAWM}) [21].

$$I_{WhiteStripe}(x) = \frac{I(x) - \mu_{NAWM}}{\sigma_{NAWM}}$$

4.3.5.3 Nyul histogram matching (HM)

Nyul's piecewise linear HM process [22] requires a standard histogram scale to be produced by averaging the intensity values from a subset of scans, using pre-defined intensity histogram landmarks

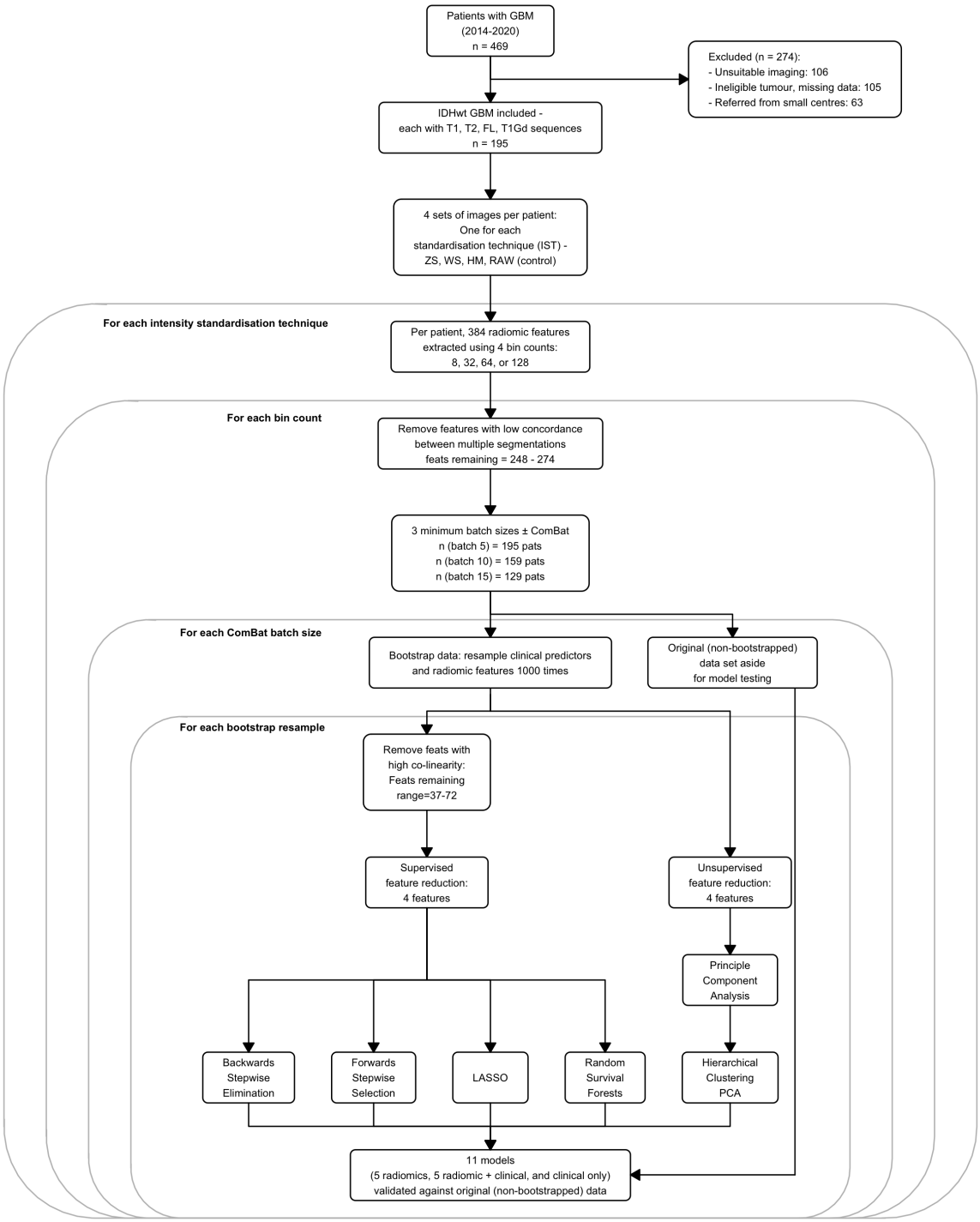

Figure 4.2: FLAIR = Fluid Attenuated Inversion Recovery, GBM = Glioblastoma, HM = Histogram Matching, IDH = Isocitrate dehydrogenase, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = Principle Component Analysis, RAW = no intensity standardisation applied to images (control), T1 = T1-weighted, T1CE = T1-weighted, contrast-enhanced, T2 = T2-weighted, WS = WhiteStripe, ZS = Z-Score.

(step-one). The landmarks are defined as percentiles, ranging from 1-99% of the intensity range (the default values are 1, 10, 20, 30, 40, 50, 60, 70, 80, 90 & 99%), such that outlier values are ignored [23].

All scans are then standardised by dividing the histogram of the new image into deciles and per decile, all voxels that fall into that range of intensities are linearly mapped using the standard scale produced in step-one [22, 23].

4.3.6 Radiomics Feature Extraction

Pyradiomics (v3.0.1) [24] was used to extract RFs from the WTV (**Figure 4.1**). 384 features were extracted from each image set (four sets, one per IST), including 18 first order, 24 gray-level co-occurrence matrix, 16 gray-level run length matrix, 16 gray-level size zone matrix, 14 gray-level dependence matrix and five neighbouring gray-tone difference matrix features from each MR sequence, and 12 shape features extracted from the T1CE sequence.

Features were extracted in 3-dimensions ($3D$), using a voxel size of $1mm^3$. Four bin numbers (8, 32, 64, 128) were used to extract four unique sets of RF per image to determine if ISTs were dependent on the bin number. A fixed bin number (FBN) was used rather than fixed bin size for intensity discretization prior to texture feature calculation. Previous work in diffuse glioma has suggested that a FBN may reduce the need for IST as it has a normalising effect. FBNs rescale the signal intensity within the volume of interest [4] and the IBSI suggests this approach for MRI [25].

Pyradiomics is mostly compliant with the IBSI feature definitions. The definition of bin boundaries when using fixed bin sizes is different (not applicable to this study), Pyradiomics aligns its resampling grid to the origin voxel (rather than to the centre), gray values are not rounded in Pyradiomics and kurtosis is calculated as +3 compared to IBSI [25, 26]. Neither Pyradiomic's nor IBSI's definition of features is necessarily better than the others'. It is important to acknowledge the differences as IBSI is trying to standardize the extraction of features but there is no inherently correct way to do this.

4.3.7 Radiomic feature reproducibility

The 50 patients that had their whole tumour volume segmented independently by two observers were used to measure the reproducibility of radiomic features. It is important to assess IB reproducibility as features that vary significantly due to small changes in segmentations, as might be seen if two independent readers segment the tumour, ought not to be used in the model building process [27]. The two-way random effects ICC for each RF was measured by constructing a linear two-way random effects model with patient and segmentor selected as random effects. Reliable measurement of ICC assumes that the features are normally distributed; that features and model residuals follow a gaussian distribution. If the residuals for a particular RF's model did not follow a normal distribution, the RF was power-transformed using the Box-Cox (or Yeo-Johnson if they contained negative values), to attempt to shift their model residual distribution to a gaussian one. RFs were excluded if, after a power transformation, the two-way random effects model residuals remained non-gaussian. A list of all transformed features, the lambda used for power-transformation and whether the transformation resulted in acceptance of the feature can be found in **Table S4.7**. It is important to note that non-normal features are still retained and only a few features that could not be coerced to a normal distribution using a power transformation are removed. It would not be possible to be confident that these features had a high level of reproducibility between segmentations.

4.3.8 ComBat realigment of multi-centre radiomics

ComBat is a statistical realignment process that aims to estimate the batch effects imparted onto radiomic features by imaging acquisition and variability in patient demographics or clinical variables between sites. To decide which biological co-variates to include in the ComBat model, all continuous predictors were tested with one-way ANOVA and categorical predictors were tested with Fisher's exact test for significant differences ($p < 0.05$) across batches. This was a sub-optimal approach as ideally, all clinical predictors would have been added to the ComBat model as biological co-variates but this increases the sample size requirements for estimation of the batch effects [8], so a more

pragmatic approach was adopted and only those with $p < 0.05$ on significance testing (patient age) were included. Use of significance tests to variable is not generally advised, as discussed in earlier chapters but a pragmatic approach was taken. ComBat realignment was performed per MRI sequence, defining each batch not only on geographical site but also by homogeneity of scan acquisition per site (batch definition and acquisition parameters provided in **Tables S4.2 - S4.5**).

Selecting the minimum batch size (MBS) represents a trade-off between increased performance of ComBat realignment, providing a larger sample to estimate the ComBat model coefficients, against discarding too much data. A minimum of five patients has been previously identified as the lower limit for the MBS [7, 28]. Three MBS values were chosen: five, 10 or 15. Patients in smaller batches were excluded (**Figure 4.2**) so 15 was the maximum chosen as this avoided excessive data loss. Radiomic features without ComBat realignment were also included as a baseline assessment of the effects of IST alone.

4.3.9 Statistical analysis and experimental settings

All statistical analysis was performed in R version 4.2.2 (2022-10-31) and overseen by a career statistician. A summary of the statistical analysis is shown in **Figure 4.2**. Cox proportional hazards (CPH) models for OS prediction (time from surgery to death, censor date 10/10/22) were built. 96 different combinations of 'experimental settings' (**Figure 4.2**) were investigated; with and without ComBat, four ISTs, each with four bin counts and three MBS for ComBat.

4.3.10 Model building and feature selection

Five feature selection (FS) methods were used to reduce dimensionality; four RFs were considered for entry into the radiomics model based on sample size calculations, which are detailed in section 4.7.1. Each FS method was applied within each of the 1000 bootstrap resamples (**Figure 4.2**). Correspondingly, five sets of RFs were selected per bootstrap resample. More detail about the steps involved in the bootstrapping and modelling process is provided in **Table S4.6**.

Unsupervised selection used PCA and hierarchical clustering using the package 'FactoMineR' and default settings [29]. PCA is a linear data reduction technique that describes the variation in the data using linear combinations of non-correlated features (principle components). Hierarchical clustering classified the PCA results so that three to 10 clusters were formed. Four RFs that explain the greatest variation between clusters were selected [29].

Prior to the four supervised feature selection strategies, highly colinear features were removed to reduce redundancy. Features with an absolute Spearman correlation coefficient below 0.8 were kept (i.e. > -0.8 and $< +0.8$). CPH models using backwards or forwards stepwise feature selection and a p-value threshold of 0.1 was used until four features were included using the 'stepAIC' function [30]. CPH model with a LASSO penalty [31] and the smallest value of lambda that would select only four features was used. The optimal value of lambda was selected following 10-fold cross validation applied within each bootstrap resample.

Random survival forests (RSFs) were used to select features using the package 'RandomForestSRC' [32]. Tuning of the optimum number of features to split at each node and the minimum size of the terminal node, as well as calculating the importance of features to the models was done using the in-built package functions. The four most important RFs were selected using this approach [32].

In all, three models were produced. Each set of RFs were used to train a radiomics-only model. A clinical-only model was also trained as a baseline for results comparison using age, gender, MGMT promoter methylation, extent of surgical resection, oncological adjuvant treatment, tumour diameter and log-transformed WTV (prior analysis, chapter 3, indicated log-transformation was an effective non-linear transformation of WTV [12]). Clinical and radiomics features were then combined to produce a clinical-radiomics model.

4.3.11 Model performance

Evaluation of any proposed prognostic model requires assessment across (at least) four domains: discrimination, calibration, relative model fit and relative explained variance.

Calibration shows how closely predictions match observed events; if an individual is predicted to have low risk of death, is the observed death rate also low for similar patients? This was assessed with the calibration slope and calibration plots for 1 year survival prediction. The variability of calibration of these predictions was evaluated; a model that produces highly variable predictions after small alterations to training data indicates that the model or model building process is highly unstable.

Stability of model calibration was assessed using calibration plots. To make the calibration plots, the bootstrap resampling process outlined in **Table S4.6** was followed to produce 1000 predictions for patient survival probabilities at 1 year. Since the data is censored, observed survival times would be potentially misleading and therefore estimated survival times were produced so that a smoothed calibration plot could be drawn using the package 'pseudo' (v1.4.3) [33]. 200 curves were randomly selected (from the 1000 possible) to enhance visualisation.

Discrimination informs how good the model is at dividing patients into high- or low-risk groups. Harrell's C-index (C) is the proportion of all pairs of individuals that can be ordered, in which the person with higher predicted risk has a lower survival. A C of 1 is perfect discrimination and 0.5 indicates no discrimination. It has been suggested that comparing models using C is not very informative despite its popularity in medical literature [9]. C does not give any weighting to the magnitude of correct predictions and is simply a measure of ranking accuracy like Wilcoxon's rank sum test. Royston and Sauerbrei's D-statistic (D), which starts at 0, with no upper bound was also calculated – a higher value is better. Royston and Sauerbrei's D-statistic (D) is the log-hazard ratio for two equal sized groups that are split using the average (median) prognostic risk score. This is more informative than C as the strength of correct predictions is rewarded.

Relative model fit provides an insight into which model (from all models built from this data), contains more information relating to outcome whilst using the fewest variables to do so; a more parsimonious model is generally a better one. Fit was measured with Akaike's information criterion (AIC) - lower values suggest better model fit [34].

Relative explained variation indicates which model best explains the variation in survival times for

glioblastoma patients for competing models. Two measures of explained variation were calculated - Royston and Sauerbrei's R^2 (R^2_D) and Nagelkerke's R^2 (R^2_N) - higher values indicate better performance [35].

Mean and 95% confidence intervals (95% CIs) were calculated across all 1000 bootstrap resamples (**Figure 4.2, Table S4.6**). Bootstrapping rather than a random train-test split was used for optimism adjustment as it is widely recommended in the statistical modelling literature [10, 36].

Heatmaps were created to graphically illustrate the impact of ISTs and MBS. The heatmaps of discrimination, fit and explained variation were centered on the clinical-only model and scaled to the standard deviation of models for each experimental setting to highlight the change in model performance relative to the clinical-only model and allow comparison across settings [35]. For example, results for WS standardised images, bin count of 64 and MBS 10 can be compared fairly to ZS images, bin count 32 and MBS 15.

The impact of IST and MBS on feature selection stability was also assessed by measuring the percentage of times across bootstrap resamples that the same four features were selected together (feature co-occurrence).

4.4 Results

4.4.1 Study population

Cohort demographics are shown in **Table 4.1** and are comparable to those in the scientific literature [37, 38]. Median age was 61 years (IQR 55-68 years), 37% of patients were female, 23% had a biopsy, and 22% had complete resection of their enhancing and necrotic tumour. 48% patients underwent either full or partial Stupp protocol adjuvant treatment, 36% of patients had tumours with methylated MGMT promotors and median survival was 13 months (95% CI 12-14 months) following surgery, with 167 deaths (86%) occurring before the censor date.

Figure 4.3 shows the proportion of eligible data per ComBat MBS that was used for modelling

Table 4.1: Patient demographics and treatment summary for radiomics modelling ($n = 195$)

Demographic	Value
Age, years - median (IQR)	61 (55-68)
Gender - no. female (%)	72 (37%)
Surgical treatment – no. (%)[a]	
Biopsy	44 (23%)
100% resected	42 (22%)
$\geq$ 90% resected	62 (32%)
$<$ 90% resected	47 (24%)
Adjuvant oncology treatment – no. (%)	
No Stupp	102 (52%)
Full Stupp[b]	44 (23%)
Partial Stupp[c]	49 (25%)
MGMT methylation – no. (% of known)	70 (36%)
Overall survival, months – median (95% CI)	13 (12-14)
Maximum tumour diameter, cm – median (IQR)	4.4 (3.35-5.35)
Core volume, cm^3 - median (IQR)[d]	1 (1-1)
Whole volume, cm^3 - median (IQR)[e]	1 (1-1.01)

IQR = interquartile range, MGMT = O6-methylguanine-DNA methyltransferase, CI = Confidence Interval. [a] Percentage of contrast enhancing and necrotic tumour core removed.
[b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and 6 cycles adjuvant temozolomide.
[c] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide. [d] Core Volume includes all enhancing and necrotic tumour. [e] Whole Volume includes the tumour core plus all peritumoural high T2 signal.

and the number of unique batches per MRI sequence (in-depth charts for each MBS are provided in **Figures S4.1-S4.3**). 76% of eligible data was retained when MBS=5 compared to 50% when MBS=15. The bar charts also illustrate the diversity of imaging that had to be harmonised to build radiomic models in this real-world dataset.

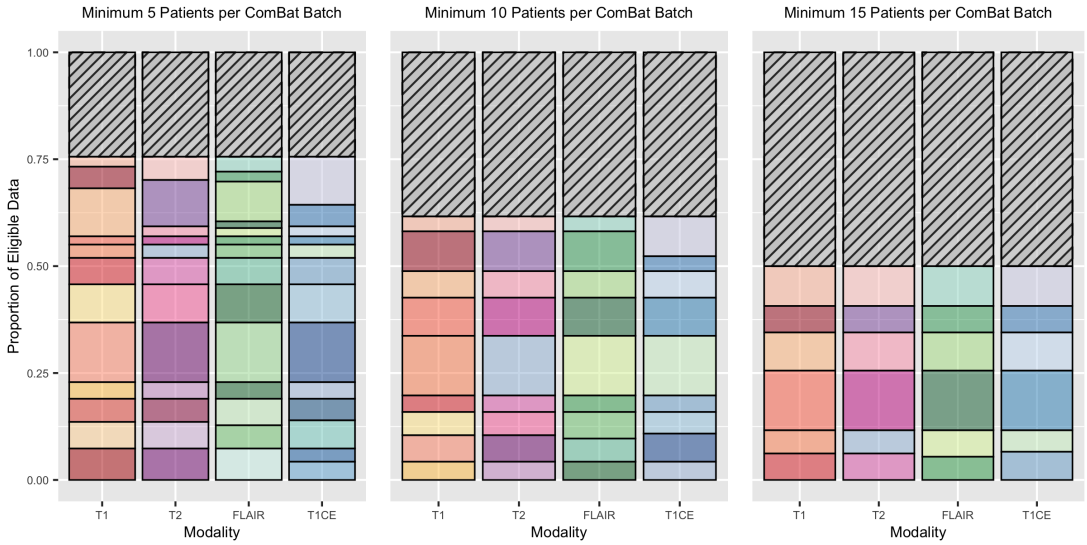

Figure 4.3: Stacked bar chart in which each bar represents one MRI sequence, and the different colours within each bar indicate a unique batch. A unique batch could, for example, indicate a different geographic location or a different set of acquisition parameters within the same location. The shaded regions indicate the proportion of imaging data that had to be excluded to meet the minimum batch size. FLAIR = Fluid Attenuated Inversion Recovery, T1 = T1-weighted, T1CE = T1-weighted, contrast-enhanced, T2 = T2-weighted.

4.4.2 Feature reduction

The mean ($\pm$ standard deviation) DSC for WTV segmentations was 0.96 ± 0.03, which is equivalent to values published in the BRATS segmentation dataset, in which multiple experts raters segment the same glioblastoma images, and the segmentation concordance was therefore within the expected variation of inter-rater agreement [39]. ICC values of radiomic features from the two independent segmentations in the patients with two sets of tumour masks resulted in removal of between 110-136 features (range across all bin counts and ISTs). For supervised feature selection, a range of 32-72

features remained following the removal of those with high co-linearity based on absolute Spearman correlation coefficient.

4.4.3 Model performance - effect of ISTs and ComBat batch size

A summary of the model performance for all experimental settings is shown in **Figure 4.4** and **Tables S4.8-S4.15**.

Figure 4.4 shows that as MBS increased, the average calibration slope range decreased successively from 1 and there was little influence of adding in ComBat. The heatmaps also demonstrated that the results for average calibration slope for all ISTs compared to no standardisation (labelled 'RAW' in **Figure 4.4**) were similar.

Both IST and MBS had a small but appreciable effect on discrimination. MBS 10 and 15 increased the range of z-score adjusted values compared to MBS 5, regardless of the use of ComBat. ComBat had only a minor impact on values. At MBS 15 the range of z-score adjusted values $(0 - 0.36)$ with ComBat increased compared to without ComBat $(0 - 0.34)$, which is reflected in the slight upwards shift of the range of 95% CIs in the absolute values of Royston's D. The 95% CIs increase from $0.83 - 1.2$ to $0.86 - 1.5$ without ComBat and up to $0.88 - 1.5$ with ComBat. The greatest relative improvement in discrimination was seen with LASSO or forwards stepwise feature reduction, HM standardisation, 8 bins and MBS 10 patients (0.41), with or without the use of ComBat. Although not strictly observed, overall, HM and WS standardised images tended to produce the highest relative increase in discrimination compared with ZS.

4.4.4 Relative explained variance and model fit

The additional benefit of ComBat was best seen with MBS 10 or 15. For example, with MBS 10, the max scaled increase was 0.29 with and 0.19 without ComBat. At MBS 15, the score was 0.33 with and 0.28 without ComBat. For MBS 5, the addition of ComBat degraded the scores. The

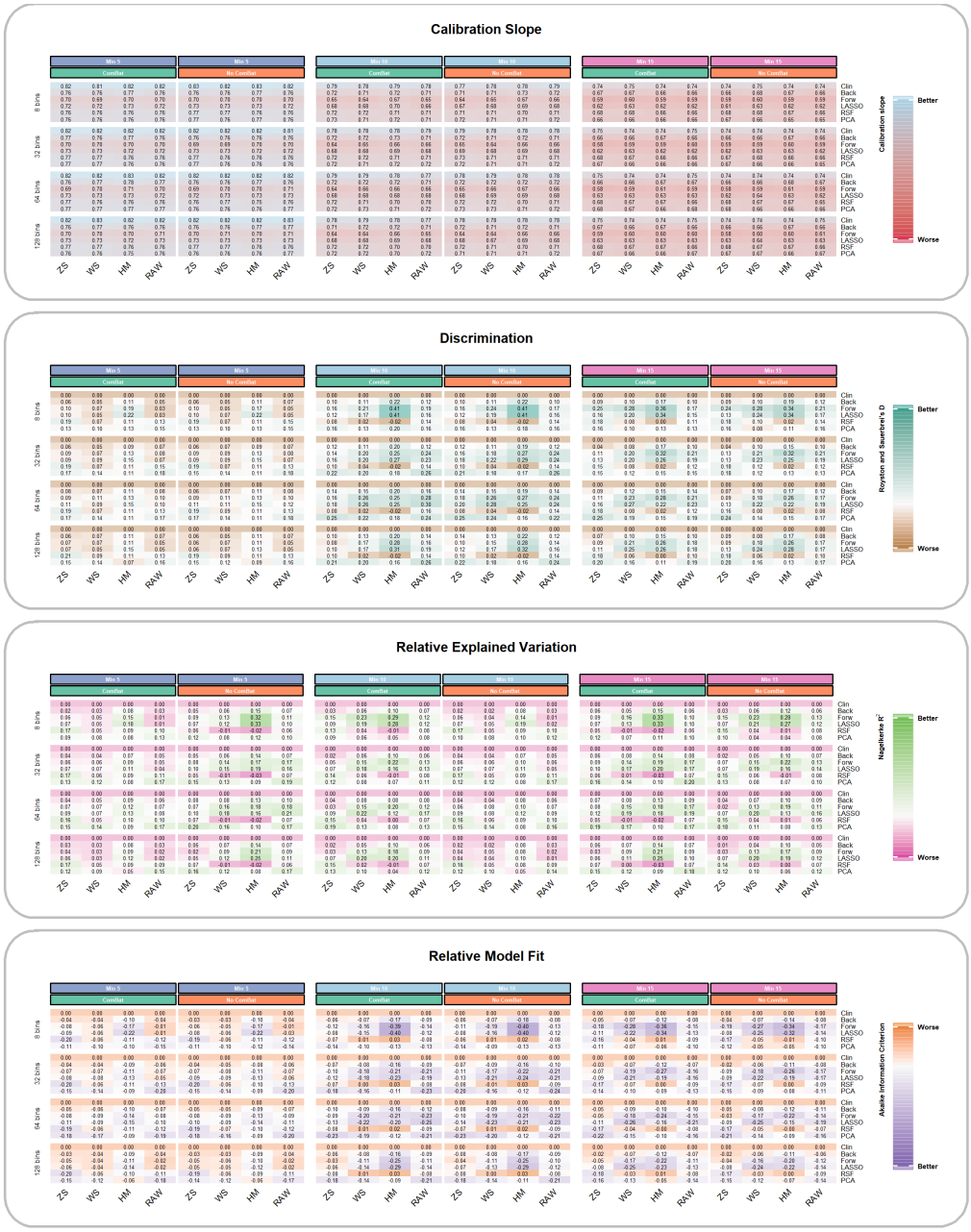

Figure 4.4: Heatmaps show the mean result per model performance statistic (measured in the test data, averaged across the 1000 bootstrap resamples) for the clinical and the combined radiomic and clinical models across different selection procedures for all the experimental settings. The data for discrimination, relative explained variance and model fit statistics have been centred on the mean clinical value and scaled to the standard deviation across all models for that particular experimental setting (ie. for each choice of minimum ComBat batch size, bin count and intensity standardisation) so that it represents change relative to the clinical only model and allows more meaningful comparisons between different experiment settings. Lower values for Akaike's Information Criterion (AIC) show improved model performance, hence the colour bar shows better performance at the bottom (the reverse of the other three colour bars). CmB = ComBat, HM = Histogram Matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = hierarchical clustering of principle component results, RAW = no intensity standardisation applied to images (control), RSF = Random Survival Forests, WS = WhiteStripe, ZS = Z-Score.

largest increase in scores were seen for HM (range $-0.03-0.26$) and WS ($-0.01-0.22$) standardised images and 8 bins.

Model fit showed similar findings; the greatest improvements relative to the clinical model with the addition of ComBat compared to without it, was observed with MBS 15 (max scaled decrease -0.36 and -0.34 respectively). A lower score indicates improved relative model fit.

At other MBS there was less benefit from ComBat realignment. RFs extracted with 8 bins, LASSO or forwards feature selection and HM standardisation produced the largest improvements in model fit (lowest AIC) and explained variation (highest R^2). WS standardisation also performed well across most bin counts. As observed for discrimination, this was not a strictly observed finding and the result also depended upon which feature selection strategy was selected.

4.4.5 Model stability

The size of 95% CIs for model performance measures (**Tables S4.8-S4.15**), the frequency with which the same radiomic features were selected (**Table 4.2**) and the 1-year event prediction calibration plots (**Figure 4.5**) all showed a trend towards reduced stability with increased MBS. All ISTs produced similar findings, as did models with and without ComBat realignment.

Stability of calibration plots for 1-year event prediction using PCA feature selection and a bin count of 32 across all ISTs and ComBat batch sizes is illustrated in **Figure 4.5** (calibration plots using other feature selection methods are shown in **Figures S4.4-S4.7** for bin count 32 - other bin counts not included but illustrated similar findings). As the MBS increased, the stability of predictions decreased, as evidenced by greater spread from the null line of the bootstrapped results (shown in the paler colour). This was observed for all models and ISTs, with no IST clearly outperforming any other.

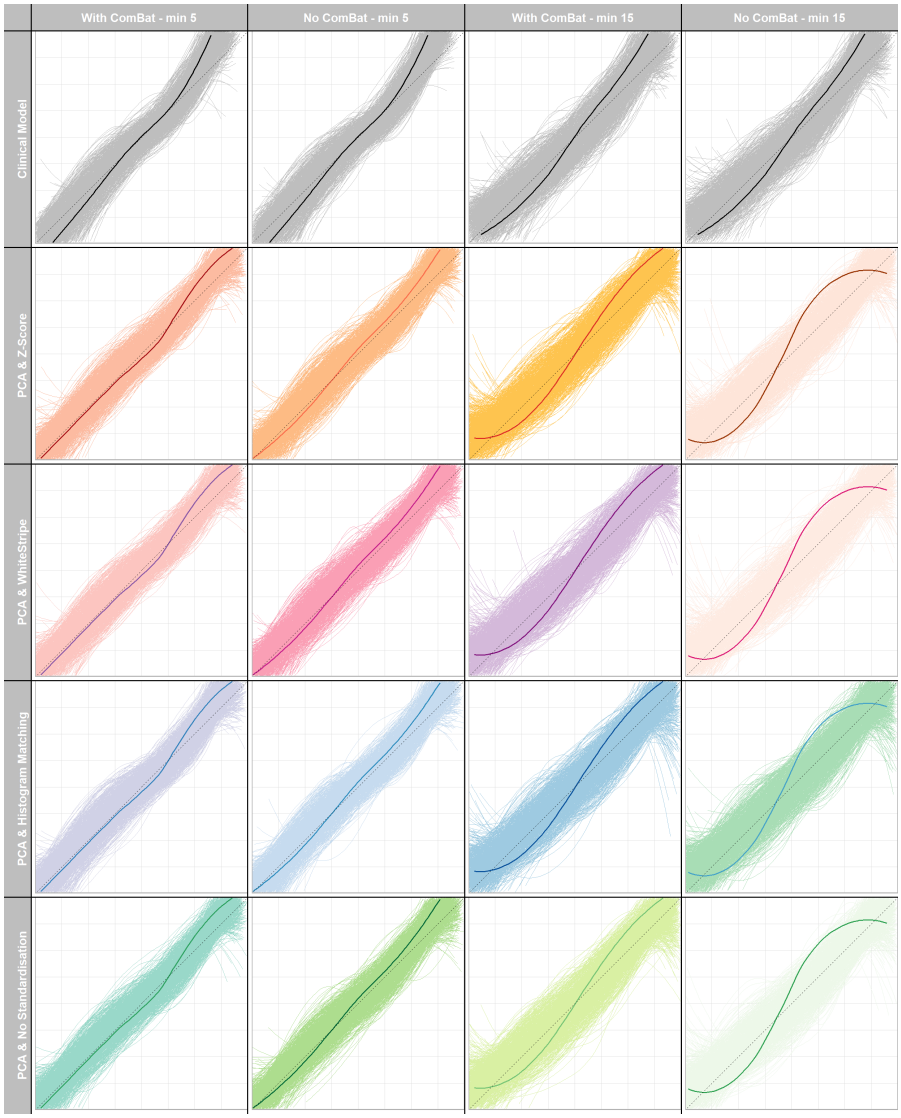

Figure 4.5: Calibration instability plots show the impact of different experimental settings on 1-year survival predictions. ComBat was applied for columns 1 and 3, and without ComBat in columns 2 and 4. Different minimum ComBat batch size were used (5 - columns 1 and 2; 15 - columns 3 and 4) and different intensity standardisation techniques applied per row. Results show predictions across the bootstrap resamples for models built using principle component analysis and bin count 32. x-axes represent predicted and y-axes the observed survival at 1-year. The thin curves represent the predictions from one bootstrap sample and the thicker curve, predictions based on the original, non-bootstrapped data. Only 200, randomly selected, bootstrap results are shown in each calibration plot. The grey dashed line represents the null line, with greater deviation from this indicating worse calibration. Increased spread of the thin curves indicates lower stability of that model building process. The calibration plots resulting from combined clinical and radiomics models, with features selected using hierarchical clustering of principle component analysis results (rows 2,3 and 4) are compared against the clinical only models (grey, top row). PCA = hierarchical clustering of principle component results.

Table 4.2: Percentage of bootstrap resamples in which the same four radiomic features were selected for entry into the final model. Results are shown for radiomic features with ComBat realignment, at all minimum ComBat batch sizes, bin counts and intensity standardisation techniques and all five feature selection techniques. If one intensity standardisation technique performed better than others, the result for that experimental setting is highlighted bold.

	Percentage of resamples in which the same 4 radiomics features were selected											
	Minimum batch size for ComBat realignment											
	Minimum = 5				Minimum = 10				Minimum = 15			
Feat Select[a]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
Bin count 8												
Backwards	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
Forwards	**2**	1	1	<1	1	1	1	<1	<1	1	1	<1
LASSO	1	1	1	<1	<1	1	<1	<1	<1	<1	<1	<1
RSF	77	**78**	16	24	**75**	72	14	21	**72**	46	63	27
PCA	7	7	6	6	5	5	**8**	5	6	7	6	7
Bin count 32												
Backwards	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
Forwards	1	1	<1	1	1	1	<1	1	<1	<1	<1	1
LASSO	1	1	<1	1	<1	1	<1	<1	<1	<1	<1	<1
RSF	27	**78**	16	20	25	**73**	16	19	**83**	47	67	25
PCA	7	10	6	**12**	6	8	6	**10**	5	**8**	4	6
Bin count 64												
Backwards	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
Forwards	1	1	1	1	<1	1	<1	1	<1	<1	1	<1
LASSO	1	1	1	1	1	1	<1	1	<1	1	<1	<1
RSF	33	**80**	18	26	30	**77**	61	22	27	50	**61**	25
PCA	10	14	8	14	8	**10**	7	8	8	**11**	5	7
Bin count 128												
Backwards	<1	<1	1	1	<1	<1	1	1	<1	<1	<1	1
Forwards	1	1	1	1	1	1	1	1	1	1	1	**2**
LASSO	1	1	1	1	1	**2**	1	1	1	1	1	1
RSF	**86**	79	66	22	**80**	74	61	18	**81**	46	63	22
PCA	7	10	**16**	12	7	9	**10**	9	9	**12**	6	9

Minimum batch size refers to the smallest number of patients that had to be scanned in each batch to be included in the ComBat realignment process (other data was excluded). HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis, RAW = No intensity standardisation prior to radiomic extraction, RSF = Random Survival Forests, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation.

[a] Radiomic feature selection method. A maximum of four radiomic features was selected with the chosen method within each bootstrap resample.

Similarly, the 95% CIs for model results (**Tables S4.8-S4.15**) showed a trend towards increased confidence interval size, and hence lower stability, as the MBS was increased. Feature co-occurrence (**Table 4.2**) did broadly show that ZS and WS resulted in higher feature co-occurrence for RSF and PCA based selection methods but that MBS did not have a great impact on feature selection stability.

4.5 Discussion

The aim of this project was to assess the effect of MRI intensity standardisation technique and ComBat minimum batch size on prognostic model performance including calibration and stability in a real-world, multi-centre glioblastoma patient cohort. Results demonstrated worse calibration and model stability as MBS increased, and hence sample size decreased, however discrimination, explained variation, and model fit improved. HM and WS ISTs, overall, improved discrimination, explained variation and model fit, which tended to occur at higher MBS, whereas choice of IST did not impact upon calibration or stability. The relative improvement of ComBat was mostly demonstrated at MBS 10 and 15, whereas there was little difference or even deterioration at lower MBS in some domains. By comparing across multiple domains of performance a more thorough assessment of ISTs and ComBat MBS was produced.

Previous studies that have compared the effect of ISTs on radiomics models [6] often show improved OS prediction [5, 40] or accuracy in differentiating grades of diffuse glioma [4, 41, 42]. Based on discrimination, relative fit or explained variation, performance improved through the choice of IST and, consistent with other studies [5, 6, 43], the current results show that HM and WS produced the highest relative improvements. However, for model calibration accuracy and model stability, IST did not affect results following the application of ComBat to realign features.

Adding ComBat slightly improved performance only at MBS 15 for discrimination and model fit, and at 10 and 15 for explained variation. This is explained by the likely increased accuracy of ComBat model coefficients estimation [8]. Increments of 5 patients were enough to improve model performance. The application of ComBat to real-world datasets, however, poses a challenge due to the wide range of acquisitions and locations [7]. Previous studies have suggested that the MBS for ComBat could be as low as five [7, 28], however others have suggested 20-30 minimum [8]. This study used a compromise to minimize data loss. Reducing the available sample size, by using MBS 10 or 15, improved certain aspects of model performance but this also made them less stable, regardless of whether ComBat feature realignment was performed. No other published studies have examined this impact. For real-world datasets, where scanner protocols are difficult to standardise across a

broad geographical range and many centres, restricting the sample size for ComBat may not be a feasible option as it ignores heterogeneity of imaging data, and more importantly, prediction models developed in this manner may not then be generalizable to sites with fewer patients. In this study, results without ComBat were similar to those with realignment, and a more practical solution may be to use fixed bin number discretization and IST without ComBat in such data. Unsupervised clustering has been used to increase batch sizes [7, 44], grouping patients with similar RFs into clusters for ComBat realignment batches. However, the clustering results were not validated, and this approach would be difficult to validate with the sample size in this study.

This study demonstrated a mixed picture regarding the effects of ISTs and ComBat batch sizes when considering multiple domains of model performance and model stability. A systematic review of prognostic models in patients with glioblastoma reported that 10 of 11 time-to-event models reported just the C-index [45]. A recent comparison of multiple ISTs in radiomics models in patients with 'primary' and recurrent high-grade glioma reported on discrimination, using C-index, and relative model fit (AIC), but did not comment on calibration or ComBat MBS [5]. The present study also included a more in-depth assessment of model stability using bootstrapping, including calibration instability plots [10], which was a useful way to identify consistency of model predictions. Stability is important as it provides information on how well a model performs following variations in input data, and not just how it performs on average [10].

This study has several limitations. Acquisition parameters were heterogeneous, including several centres with relatively few patients scanned, which impacted on the ability to test larger batch sizes for ComBat. This is a real-world dataset, and the restriction of larger batches would have meant too few patients were included. The comparison of the relative impact of different ISTs could still be assessed, and this represents a case where good intensity standardisation is required. Public data could have been used to supplement institutional data, however the aim was to assess the performance of combined clinical-radiomic models, and hence well-curated data on clinical predictors were necessary. Future work could build on these results with additional public data. Only three out of many ISTs available were chosen for evaluation, however these had previously

been identified as the most popular choices in prior studies [12]. The supervised feature selection strategies considered far more than the four radiomic features suggested as the maximum by event per predictor calculation, however they are popular within the literature and the decision will not have impacted upon the assessment of relative model performance due to IST and ComBat batch size. Finally, measurement of IST impact on feature repeatability was not assessed, however to the best of the author's knowledge, a preoperative glioblastoma dataset with test-retest data is not available publicly.

4.6 Conclusions

ISTs and ComBat MBS affected survival model performance in a heterogeneous multi-centre glioblastoma cohort. HM and WS, overall, improved discrimination, relative explained variation, and model fit, as did ComBat at higher MBS. However, calibration and model stability deteriorated as MBS increased and resulted in more data being discarded from modelling. This has clinical implications as referral systems such as the hub-and-spoke model in this study are hampered by varied image acquisitions, and therefore require robust methods for harmonizing heterogeneous datasets without compromising the model performance. Future work to demonstrate methods of improving radiomic model performance in real-world datasets that also preserve model stability is warranted.

References

1. O'Connor, J. P. B. *et al.* Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology* **14,** 169–186 (Mar. 2017).
2. Huisman, M. & Akinci D'Antonoli, T. What a Radiologist Needs to Know About Radiomics, Standardization, and Reproducibility. *Radiology* **310,** e232459 (Feb. 2024).

3. Huang, E. P. *et al.* Criteria for the translation of radiomics into clinically useful tests. *Nature Reviews Clinical Oncology* **20** (2022).

4. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10,** 12340 (July 2020).

5. Salome, P. *et al.* MR Intensity Normalization Methods Impact Sequence Specific Radiomics Prognostic Model Performance in Primary and Recurrent High-Grade Glioma. *Cancers* **15** (2023).

6. Fatania, K. *et al.* Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review. *European Radiology* **32,** 7014–7025 (Oct. 2022).

7. Carré, A. *et al.* AutoComBat: a generic method for harmonizing MRI-based radiomic features. *Scientific Reports* **12,** 1–17 (2022).

8. Orlhac, F. *et al.* A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine* **63,** 172–179 (Feb. 2022).

9. Harrell, F. *Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements* 2023.

10. Riley, R. D. & Collins, G. S. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal* **65,** 1–22 (2023).

11. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence and Medical Imaging (Claim). *Radiology: Artificial Intelligence* (2020).

12. Fatania, K. *et al.* Tumour Size and Overall Survival in a Cohort of Patients with Unifocal Glioblastoma: A Uni- and Multivariable Prognostic Modelling and Resampling Study. *Cancers* **16,** 1301 (Mar. 2024).

13. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

14. Pati, S. *et al.* Federated Learning Enables Big Data for Rare Cancer Boundary Detection (2022).

15. Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping* **31,** 798–819 (Dec. 2009).

16. Yushkevich, P. A. *et al.* IC-P-174: Fast Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. *Alzheimer's & Dementia* **12,** 126–127 (July 2016).

17. Thakur, S. *et al.* Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage* **220,** 117081 (Oct. 2020).

18. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18,** 203–211 (Feb. 2021).

19. Zou, K. H. *et al.* Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology* **11,** 178–189 (2004).

20. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of simpleITK. *Frontiers in Neuroinformatics* **7,** 1–14 (2013).

21. Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9–19 (2014).

22. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* **19,** 143–150 (2000).

23. Shah, M. *et al.* Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis* **15,** 267–282 (2011).

24. Van Griethuysen, J. J. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* **77,** e104–e107 (Nov. 2017).

25. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative (2016).

26. Pyradiomics. *Pyradiomics Frequently Asked Questions*

27. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14,** 749–762 (Oct. 2017).

28. Fortin, J.-P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* **167,** 104–120 (Feb. 2018).

29. Le, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25,** 1–18 (2008).

30. Venables, W. & Ripley, B. *Modern Applied Statistics with S* Fourth (Springer, New York, 2002).

31. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1–22 (2010).

32. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Annals of Applied Statistics* **2,** 841–860 (2008).

33. Klein, J. P., Gerster, M., Andersen, P. K., Tarima, S. & Perme, M. P. SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine* **89,** 289–300 (2008).

34. Sauerbrei, W. The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **48,** 313–329 (1999).

35. Austin, P. C., Pencinca, M. J. & Steyerberg, E. W. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Statistical Methods in Medical Research* **26,** 1053–1077 (2017).

36. Steyerberg, E. W. *Clinical prediction models : a practical approach to development, validation, and updating* (Springer, New York ; 2009).

37. Verduin, M. *et al.* Prognostic and predictive value of integrated qualitative and quantitative magnetic resonance imaging analysis in glioblastoma. *Cancers* **13,** 1–20 (2021).

38. Li, Y. M., Suki, D., Hess, K. & Sawaya, R. The influence of maximum safe resection of glioblastoma on survival in 1229 patients: Can we do better than gross-total resection? *Journal of Neurosurgery* **124,** 977–988 (2016).

39. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

40. Saltybaeva, N. *et al.* Robustness of radiomic features in magnetic resonance imaging for patients with glioblastoma: Multi-center study. *Physics and Imaging in Radiation Oncology* **22,** 131–136 (2022).

41. Li, Y. *et al.* Radiomics-Based Method for Predicting the Glioma Subtype as Defined by Tumor Grade, IDH Mutation, and 1p/19q Codeletion. *Cancers* **14,** 1778 (Mar. 2022).

42. Ubaldi, L., Saponaro, S., Giuliano, A., Talamonti, C. & Retico, A. Deriving quantitative information from multiparametric MRI via Radiomics: Evaluation of the robustness and predictive value of radiomic features in the discrimination of low-grade versus high-grade gliomas with machine learning. *Physica Medica* **107,** 102538 (2023).

43. Foltyn-Dumitru, M. *et al.* Advancing noninvasive glioma classification with diffusion radiomics: Exploring the impact of signal intensity normalization. *Neuro-Oncology Advances* **6,** 1–9 (Jan. 2024).

44. Da-Ano, R., Visvikis, D. & Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine & Biology* **65,** 24TR02 (Dec. 2020).

45. Tewarie, I. A. *et al.* 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 (2021).

4.7 Supplementary Materials

4.7.1 Calculation of sample size and event per predictor

Given that this was an exploratory analysis, comparing ISTs and ComBat batch sizes on model performance, rather than an exercise in producing the best prognostic model, all available data was used rather than calculating sample size a priori. However, as the number of candidate radiomic predictors was high, resulting in a low event per predictor parameter (EPP) rate, a number of feature reduction strategies were adopted and it was useful to know the minimum EPP available for modelling and this was calculated using previously published methodology [1, 2]. Of the glioblastoma prognostic models identified in a recent systematic review of glioblastoma prognostic models [3], none had published the Cox-Snell R^2 if the model had been applied in new data (adjusted Cox-Snell R^2, $R^2_{CS\_adj}$), which is the ideal parameter required for sample size calculation in time-to-event models [1, 2]. Therefore, this had to be estimated from the minimum C-index (C) of models identified in the systematic review - 0.66 [3, 4]. The steps to estimating $R^2_{CS\_adj}$ are outlined below and further detail is found in the study from Riley et al. [1]. The calculations below resulted in a minimum EPP of 22 for the present dataset, which meant that four radiomic features were retained in the final candidate models, with suggested minimum sample size of 175.

First, Royston's D can be estimated from C:

$$D = 5.50(C - 0.5) + 10.26(C - 0.5)^2$$

Having estimated D from the reported C, $R^2_{D\_app}$ can be derived:

$$R^2_{D\_app} = \frac{\frac{\pi}{8}D^2}{\frac{\pi^2}{6} + \frac{\pi}{8}D^2}$$

$R^2_{D\_app}$ is used as a proxy for $R^2_{Royston\_app}$ to derive $R^2_{O'Quigley\_app}$:

$$R^2_{O'Quigley\_app} = \frac{-\frac{\pi^2}{6} R^2_{Royston\_app}}{(1-\frac{\pi^2}{6})R^2_{Royston\_app} - 1}$$

From $R^2_{O'Quigley\_app}$, the total number of events (E) used to derive the model (995)[4], the likelihood ratio (LR) of the model can be estimated:

$$LR = -E \ln(1 - R^2_{O'Quigley\_app})$$

The apparent Cox-Snell R^2 ($R^2_{CS\_app}$) can then be derived, where n is the sample size (1354)[4]:

$$R^2_{CS\_app} = 1 - \exp(\frac{-LR}{n})$$

Next, the Van Houwelingen and Le Cessie shrinkage factor (S_{VH}) can be derived, where p is the number of candidate predictor parameters (ie. all predictors that were tested for inclusion):

$$S_{VH} = 1 + \frac{p}{n \ln(1 - R^2_{CS\_app})}$$

Finally, the adjusted Cox-Snell R^2 ($R^2_{CS\_adj}$) can be derived:

$$R^2_{CS\_adj} = S_{VH} R^2_{CS\_app}$$

From the $R^2_{CS\_adj}$, the R-package 'pmsampsize' [5] was used to calculate the minimum sample size and event per predictor parameter (EPP), using an event rate of 0.5 events/year based on median survival of patients with glioblastoma being 12 months [3, 6] and a timepoint of 1 year.

Supplementary references

1. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* **38,** 1276–1296 (Mar. 2019).
2. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368,** m441 (Mar. 2020).
3. Tewarie, I. A. *et al.* 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 (2021).
4. Gittleman, H. *et al.* An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: NRG Oncology RTOG 0525 and 0825. *Neuro-Oncology* **19,** 669–677 (2017).
5. Ensor, J. *pmsampsize: Sample Size for Development of a Prediction Model* 2023.
6. Fatania, K. *et al.* Tumour Size and Overall Survival in a Cohort of Patients with Unifocal Glioblastoma: A Uni- and Multivariable Prognostic Modelling and Resampling Study. *Cancers* **16,** 1301 (Mar. 2024).

4.7.2 Supplementary Figures and Tables

Table S4.1: Study compliance with Checklist for Artificial Intelligence in Medical Imaging (CLAIM)

Section	Number	Item	Present
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	Y

	2	Structured summary of study design, methods, results, and conclusions	Y
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	Y
	4	Study objectives and hypotheses	Y
METHODS			
Study Design	5	Prospective or retrospective study	Y
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	Y
Data	7	Data sources	Y
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	Y
	9	Data pre-processing steps	Y
	10	Selection of data subsets, if applicable	Y
	11	Definitions of data elements, with references to Common Data Elements	Y
	12	De-identification methods	Y
	13	How missing data were handled	Y
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	Y
	15	Rationale for choosing the reference standard (if alternatives exist)	Y
	16	Source of ground-truth annotations; qualifications and preparation of annotators	Y
	17	Annotation tools	Y

	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	Y
Data Partitions	19	Intended sample size and how it was determined	Y
	20	How data were assigned to partitions; specify proportions	Y
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	Y
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections	Y
	23	Software libraries, frameworks, and packages	Y
	24	Initialization of model parameters (e.g., randomization, transfer learning)	Y
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained	Y
	26	Method of selecting the final model	Y
	27	Ensembling techniques, if applicable	Y
Evaluation	28	Metrics of model performance	Y
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	Y
	30	Robustness or sensitivity analysis	Y
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	Y
	32	Validation or testing on external data	N/A

RESULTS			
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	Y
	34	Demographic and clinical characteristics of cases in each partition	Y
Model performance	35	Performance metrics for optimal model(s) on all data partitions	N/A
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Y

	37	Failure analysis of incorrectly classified cases	N/A
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	Y
	39	Implications for practice, including the intended use and/or clinical role	Y
OTHER IN-FORMATION			
	40	Registration number and name of registry	N/A
	41	Where the full study protocol can be accessed	N/A
	42	Sources of funding and other support; role of funders	Y

Table S4.2: Summary of patient, tumour and MRI acquisition characteristics for T1-weighted images. Values for age, diameter and volume represent medians (range), and values for biopsy, gross resection, stupp and MGMT represent percentages of patients per batch. Values for acquisition parameters represent mean (range) parameters per batch - ranges not stated for parameters that did not vary within the batch.

		Patient and tumour characteristics								Acquisition parameters																
Cluster	Count	Gender	Age	Biopsy	Gross resection[a]	Stupp[b]	MGMT[c]	Diameter (cm)	Volume (cm^3)[d]	Series	Location[e]	Manufacturer	Model	Machine ID[f]	Field (T)	Pixel size (mm)	Slice thick-ness (mm)	Slice spacing (mm)	Orientation	Rows	Columns	Bandwidth (Hz)	Echo time (ms)	Echo train length	Inversion time (ms)	Repetition time (ms)	Flip angle (°)
Batch 1	28	57	65 (51-85)	20	29	61	43	4.3 (1.4-8)	120 (17-230)	3D Fast Spin Echo	Site 6	Philips	Intera	Random ID 710	1.5	0.96 (0.9-0.96)	0.96 (0-1)	1	AX	260 (260-290)	260 (260-290)	240	3.6 (3.5-3.7)	200	-	7.7 (7.6-8)	8
Batch 2	39	64	59 (35-78)	30	13	41	41	4.6 (2-7.9)	100 (5.4-250)	2D Spin Echo	Site 8	Philips	Achieva	Random ID 209	1.5	0.72	5	6	AX	320	320	140 (110-160)	13 (12-15)	1	-	600 (570-730)	64
Batch 3	12	50	60 (49-74)	40	25	50	25	4.4 (1.3-6.6)	110 (8.9-220)	3D Fast Spin Echo	Site 5	Philips	Ingenia	Random ID 294	1.5	0.89	0.92 (0.9-1.1)	0.92 (0.9-1.1)	AX	290	290	220	3.5 (3.4-3.6)	220	-	7.8 (7.5-7.9)	8
Batch 4	16	44	66 (48-74)	10	19	38	38	4.4 (2.3-7.3)	98 (7.3-250)	2D Spin Echo	Site 2	Siemens	Aera	Random ID 361	1.5	0.6	5	5.5	AX	380	340 (300-360)	130	7.7	1	-	530 (400-700)	86 (52-90)
Batch 5	21	52	61 (44-81)	10	24	57	52	4.4 (2.7-7.5)	120 (19-280)	PROPELLOR	Site 3	Siemens	Aera	Random ID 940	1.5	0.92 (0.9-1.1)	4.2 (4-5)	5.4 (5.2-6.5)	AX	260	260	360	46	15	1100 (710-1300)	2700 (1600-3200)	150 (140-150)
Batch 6	7	71	63 (41-80)	40	0	29	14	3.8 (3.3-8.5)	61 (46-160)	3D Fast Spin Echo	Site 4	GE	Discovery MR450	Random ID 544	1.5	0.47	1.2	0.6	COR	510	510	240	12	24	-	600	90
Batch 7	13	85	71 (47-77)	8	31	31	54	4.9 (1.2-7.4)	140 (17-230)	2D Spin Echo	Site 7	Siemens	Aera	Random ID 679	1.5	0.72 (0.72-0.75)	5	6	AX	320	270 (250-320)	150	8.9	1	-	500 (410-550)	90
Batch 8	17	65	65 (45-75)	40	18	41	35	4.3 (0.5-5.9)	89 (6.5-240)	3D Fast Spin Echo	Site 4	Siemens	Avanto Fit	Random ID 383	1.5	1 (0.98-1.1)	1	-	COR	260 (260-320)	200 (180-260)	750	11	63 (49-65)	-	690 (600-700)	120
Batch 9	6	83	63 (55-68)	50	0	50	33	5.2 (2.4-6.7)	95 (16-130)	2D Spin Echo	Site 2	Siemens	Avanto	Random ID 118	1.5	0.6	5	5.5	AX	380	350 (310-360)	130	8.1 (7.8-9.4)	1	-	580 (500-660)	80 (58-90)
Batch 10	15	67	56 (45-69)	7	40	67	33	5.7 (1.8-7.8)	150 (13-250)	2D Spin Echo	Site 1	Siemens	Avanto	Random ID 933	1.5	0.6 (0.6-0.62)	5.1 (5-7)	5.6 (5.5-7.7)	AX	380	350 (290-350)	130	7.8	1	-	510 (450-620)	86 (64-90)
Batch 11	31	55	57 (34-81)	20	26	55	39	4.8 (1.4-7.4)	120 (8.4-220)	2D Spin Echo	Site 1	Siemens	Avanto	Random ID 534	1.5	0.61 (0.6-0.9)	5	5.5	AX	380 (260-380)	340 (260-350)	130	7.9 (7.8-12)	1	-	550 (450-680)	85 (57-90)
Batch 12	20	70	59 (31-77)	20	25	40	20	3.9 (1.7-6.1)	85 (9-240)	2D Spin Echo	Site 2	Siemens	Aera	Random ID 78	1.5	0.6 (0.6-0.65)	5	5.5	AX	380	350 (310-360)	130	7.7	1	-	520 (400-660)	89 (82-90)

AX = axial, COR = coronal, GE = General Electric, MGMT = O6-methylguanine-DNA methyltransferase, PROPELLER = Periodically Rotated Overlapping ParallEL Lines with Enhanced Reconstruction, - = Not applicable or missing from DICOM header. [a] 100 [b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide. [c] Percentage of patients per batch with methylation of the MGMT promoter. [d] Whole tumour volume (includes enhancement, necrosis and peritumoural high T2 signal). [e] Site identifiers such as scanner location and machine identifier were anonymised.

Table S4.3: Summary of patient, tumour and MRI acquisition characteristics for T2-weighted images. Values for age, diameter and volume represent medians (range), and values for biopsy, gross resection, stupp and MGMT represent percentages of patients per batch. Values for acquisition parameters represent mean (range) parameters per batch - ranges not stated for parameters that did not vary within the batch.

		Patient and tumour characteristics								Acquisition parameters																
Cluster	Count	Gender	Age	Biopsy	Gross resection[a]	Stupp[b]	MGMT[c]	Diameter (cm)	Volume (cm^3)[d]	Series	Location[e]	Manufacturer	Model	Machine ID[f]	Field (T)	Pixel size (mm)	Slice thickness (mm)	Slice spacing (mm)	Orientation	Rows	Columns	Bandwidth (Hz)	Echo time (ms)	Echo train length	Inversion time (ms)	Repetition time (ms)	Flip angle (°)
Batch 1	17	59	57 (45-76)	6	35	71	24	5 (1.8-7.8)	110 (13-250)	2D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 255	1.5	0.51	5	5.5	AX	450	390 (340-390)	100	96	11	-	5400 (4900-5800)	150
Batch 2	28	57	65 (51-85)	20	29	61	43	4.3 (1.4-8)	120 (17-230)	2D Fast Spin Echo	Site 6	Philips	Intera	Random ID 140	1.5	0.45 (0.41-0.57)	5	6.0	AX	530 (400-640)	530 (400-640)	200 (120-210)	110 (110-120)	20 (15-30)	-	5600 (4000-6300)	90
Batch 3	37	65	59 (35-78)	30	14	43	41	4.6 (2-7.9)	100 (5.4-250)	2D Fast Spin Echo	Site 8	Philips	Achieva	Random ID 882	1.5	0.45	5	6.0	AX	510 (260-510)	510 (260-510)	140 (100-220)	100 (100-110)	15 (12-23)	-	5100 (4900-5500)	90
Batch 4	10	50	60 (49-74)	40	20	50	20	4.4 (2.9-6.6)	110 (8.9-220)	2D Fast Spin Echo	Site 5	Philips	Ingenia	Random ID 430	1.5	0.41 (0.4-0.41)	5	6.0	AX	560 (560-580)	560 (560-580)	160 (150-180)	100	15	-	4900 (4500-5200)	90
Batch 5	16	44	66 (48-74)	10	19	38	38	4.4 (2.3-7.3)	98 (7.3-250)	2D Fast Spin Echo	Site 2	Siemens	Aera	Random ID 969	1.5	0.51	5	5.5	AX	450	370 (340-390)	100	95	11	-	5600 (5400-6100)	150
Batch 6	21	52	61 (44-81)	10	24	57	52	4.4 (2.7-7.5)	120 (19-280)	2D Fast Spin Echo	Site 3	Siemens	Aera	Random ID 943	1.5	0.55 (0.51-0.6)	4	5.2	AX	450	330 (290-360)	170	80	19	-	4000 (3400-6600)	150
Batch 7	28	57	58 (34-81)	20	29	50	43	4.6 (1.4-7.4)	130 (8.4-220)	2D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 233	1.5	0.51 (0.51-0.54)	5	5.5	AX	450	390 (340-450)	100 (100-130)	96 (90-110)	11 (11-15)	-	5400 (4100-6200)	150 (120-150)
Batch 8	7	71	65 (55-68)	40	0	57	29	5.3 (2.4-6.7)	110 (16-160)	2D Fast Spin Echo	Site 2	Siemens	Avanto	Random ID 947	1.5	0.51	5	5.5	AX	450	380 (350-390)	100	96	11	-	5500 (5400-5800)	150
Batch 9	5	100	67 (41-80)	40	0	20	20	3.8 (3.3-8.5)	61 (46-160)	2D Fast Spin Echo	Site 4	GE	Discovery MR450	Random ID 721	1.5	0.45 (0.43-0.47)	5	6.0	AX	510	510	200	98 (97-98)	24	-	5500 (5300-5700)	160
Batch 10	8	88	72 (48-77)	0	25	25	25	4.7 (1.2-5.8)	110 (17-230)	2D Fast Spin Echo	Site 7	Siemens	Aera	Random ID 846	1.5	0.54 (0.39-0.56)	3	3.3	SAG	450	450	190	92 (83-110)	17	-	4300 (3800-4700)	150
Batch 11	17	65	65 (45-75)	40	18	41	35	4.3 (0.5-5.9)	89 (6.5-240)	2D Fast Spin Echo	Site 4	Siemens	Avanto Fit	Random ID 669	1.5	0.57 (0.29-0.65)	5	6.5	AX	430 (380-770)	370 (290-670)	130	120	13	-	5000 (4600-5600)	150
Batch 12	22	68	59 (31-77)	20	23	45	23	4 (1.7-6.1)	85 (9-270)	2D Fast Spin Echo	Site 2	Siemens	Aera	Random ID 901	1.5	0.51	5	5.5	AX	450	380 (350-410)	100 (100-190)	95 (95-96)	11 (11-17)	-	5600 (5300-6100)	150

AX = axial, COR = coronal, GE = General Electric, MGMT = O6-methylguanine-DNA methyltransferase, PROPELLER = Periodically Rotated Overlapping ParallEL Lines with Enhanced Reconstruction, - = Not applicable or missing from DICOM header. [a] 100 [b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide. [c] Percentage of patients per batch with methylation of the MGMT promoter. [d] Whole tumour volume (includes enhancement, necrosis and peritumoural high T2 signal). [e] Site identifiers such as scanner location and machine identifier were anonymised.

Table S4.4: Summary of patient, tumour and MRI acquisition characteristics for fluid-attenuated inversion recovery (FLAIR) T2-weighted images. Values for age, diameter and volume represent medians (range), and values for biopsy, gross resection, stupp and MGMT represent percentages of patients per batch. Values for acquisition parameters represent mean (range) parameters per batch - ranges not stated for parameters that did not vary within the batch.

		Patient and tumour characteristics								Acquisition parameters																
Cluster	Count	Gender	Age	Biopsy	Gross resection[a]	Stupp[b]	MGMT[c]	Diameter (cm)	Volume (cm^3)[d]	Series	Location[e]	Manufacturer	Model	Machine ID[f]	Field (T)	Pixel size (mm)	Slice thickness (mm)	Slice spacing (mm)	Orientation	Rows	Columns	Bandwidth (Hz)	Echo time (ms)	Echo train length	Inversion time (ms)	Repetition time (ms)	Flip angle (°)
Batch 1	10	80	56 (46-69)	10	30	60	10	5.9 (2.9-7.8)	180 (29-250)	2D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 732	1.5	0.45 (0.45-0.47)	5.0	5.5	AX	510	450	130	110	21	2500	9000	150
Batch 2	24	62	64 (51-85)	20	25	58	42	4.3 (1.4-8)	130 (17-230)	2D Fast Spin Echo	Site 6	Philips	Intera	Random ID 941	1.5	0.46 (0.41-0.8)	5.0	6	AX	530 (290-640)	530 (290-640)	220 (200-320)	140	28 (27-50)	2800	10000 (10000-11000)	90
Batch 3	38	66	59 (35-78)	30	13	39	39	4.6 (2-7.9)	100 (5.4-250)	2D Fast Spin Echo	Site 8	Philips	Achieva	Random ID 480	1.5	0.9	3.5	4.5	COR	260	260	370 (350-370)	120	47	2800	11000	90
Batch 4	12	50	60 (49-74)	40	25	50	25	4.4 (1.3-6.6)	110 (8.9-220)	2D Fast Spin Echo	Site 5	Philips	Ingenia	Random ID 362	1.5	0.65	5.0	6	AX	350	350	390 (370-420)	130	53	2800	11000	90
Batch 5	18	78	59 (31-72)	20	22	44	17	3.8 (1.7-6.1)	91 (9-270)	2D Fast Spin Echo	Site 2	Siemens	Aera	Random ID 242	1.5	0.87 (0.45-0.9)	5.0	5.5	AX	270 (260-510)	230 (210-420)	130	110	21	2500	9000	150
Batch 6	15	47	66 (48-74)	10	20	33	40	4.4 (2.3-7.3)	100 (7.3-250)	2D Fast Spin Echo	Site 2	Siemens	Aera	Random ID 210	1.5	0.9	5.0	5.5	AX	260	220 (190-260)	150 (130-360)	110 (98-110)	22 (21-35)	2500	9000	150
Batch 7	21	52	61 (44-81)	10	24	57	52	4.4 (2.7-7.5)	120 (19-280)	2D Fast Spin Echo	Site 3	Siemens	Aera	Random ID 685	1.5	0.77 (0.72-0.84)	4.0	5.2	AX	320	230 (200-260)	180	84	17	1900 (1700-2000)	5300 (4600-5700)	150
Batch 8	5	40	57 (45-76)	20	60	60	40	2.8 (1.8-7)	25 (13-190)	3D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 732	1.5	1.1	1.1	-	AX	260	190	440	470	110	1800	5000	120
Batch 9	7	71	63 (41-80)	40	0	29	14	3.8 (3.3-8.5)	61 (46-160)	2D Spin Echo	Site 4	GE	Discovery MR450	Random ID 652	1.5	0.44 (0.43-0.47)	5.0	6	AX	510	510	120	120 (120-130)	1	2000	8000	160
Batch 10	13	85	71 (47-77)	8	31	31	54	4.9 (1.2-7.4)	140 (17-230)	2D Fast Spin Echo	Site 7	Siemens	Aera	Random ID 217	1.5	0.72 (0.72-0.75)	5.0	6	AX	320	250 (240-260)	190	82	19 (16-22)	2300 (1800-2400)	7600 (5000-8000)	150
Batch 11	16	62	66 (45-75)	40	19	38	38	4.4 (0.5-5.9)	100 (6.5-240)	2D Fast Spin Echo	Site 4	Siemens	Avanto Fit	Random ID 503	1.5	0.45 (0.43-0.49)	5.0	6.5	AX	510	420 (380-450)	130	110	21	2500	8800 (8000-9000)	150
Batch 12	7	71	65 (55-68)	40	0	57	29	5.3 (2.4-6.7)	110 (16-160)	2D Fast Spin Echo	Site 2	Siemens	Avanto	Random ID 889	1.5	0.45	5.0	5.5	AX	510	430 (390-450)	130	110	21	2500	8800 (7700-9000)	150
Batch 13	25	60	59 (37-81)	20	24	44	40	4.8 (1.8-7.4)	140 (8.4-220)	2D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 534	1.5	0.45 (0.45-0.47)	5.0	5.5	AX	510	440 (390-450)	130	110	21	2500	8900 (8000-9000)	150 (120-150)

Batch 14	5	20	54 (34-64)	20	40	100	40	3 (1.4-5.4)	21 (14-120)	3D Fast Spin Echo	Site 1	Siemens	Avanto	Random ID 534	1.5	1.1	1.1	-	AX	260	190	440	470	110	1800	5000	120

AX = axial, COR = coronal, GE = General Electric, MGMT = O6-methylguanine-DNA methyltransferase, PROPELLER = Periodically Rotated Overlapping ParallEL Lines with Enhanced Reconstruction, - = Not applicable or missing from DICOM header. [a] 100 [b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide. [c] Percentage of patients per batch with methylation of the MGMT promoter. [d] Whole tumour volume (includes enhancement, necrosis and peritumoural high T2 signal). [e] Site identifiers such as scanner location and machine identifier were anonymised.

Table S4.5: Summary of patient, tumour and MRI acquisition characteristics for contrast-enhanced T1-weighted images. Values for age, diameter and volume represent medians (range), and values for biopsy, gross resection, stupp and MGMT represent percentages of patients per batch. Values for acquisition parameters represent mean (range) parameters per batch - ranges not stated for parameters that did not vary within the batch.

		Patient and tumour characteristics								Acquisition parameters																	
Cluster	Count	Gender	Age	Biopsy	Gross resection[a]	Stupp[b]	MGMT[c]	Diameter (cm)	Volume (cm^3)[d]	Series	Location[e]	Manufacturer	Model	Machine ID[f]	Field (T)	Pixel size (mm)	Slice thick-ness (mm)	Slice spacing (mm)	Orientation	Rows	Columns	Bandwidth (Hz)	Echo time (ms)	Echo train length	Inversion time (ms)	Repetition time (ms)	Flip angle (°)
Batch 1	27	59	65 (51-85)	20	30	63	44	4.3 (1.4-8)	110 (17-230)	3D Fast Spin Echo	Site 6	Philips	Intera	Random ID 590	1.5	0.96 (0.9-0.96)	0.98 (0.5-1)	0.98 (0.5-1)	AX	260 (260-290)	260 (260-290)	240	3.6 (3.5-3.7)	200	-	7.7 (7.6-8)	8
Batch 2	39	64	59 (35-78)	30	13	41	41	4.6 (2-7.9)	100 (5.4-250)	2D Spin Echo	Site 8	Philips	Achieva	Random ID 321	1.5	0.72	5	6	AX	320	320	160 (110-160)	12 (12-15)	1	-	570 (540-730)	64
Batch 3	11	45	61 (49-74)	50	18	45	27	4.4 (1.3-6.6)	110 (8.9-220)	3D Fast Spin Echo	Site 5	Philips	Ingenia	Random ID 358	1.5	0.89	0.92 (0.9-1.1)	0.92 (0.9-1.1)	AX	290	290	220	3.5 (3.4-3.6)	220	-	7.8 (7.5-7.9)	8
Batch 4	16	69	62 (31-77)	20	19	38	19	3.8 (1.7-7.1)	85 (9-230)	3D MP-RAGE	Site 2	Siemens	Aera	Random ID 560	1.5	0.55	1.1	-	AX	510	380	250	2.3	1	1100	1900	15
Batch 5	5	40	73 (45-76)	20	20	40	60	4.3 (1.5-6.2)	90 (32-190)	3D FLAIR	Unknownh	Siemens	Aera	Random ID 180	1.5	1 (0.98-1.1)	1.1 (1-1.2)	-	AX	250 (230-260)	240 (190-260)	170 (150-170)	4.4 (3.2-4.8)	1	900	410 (9.6-2000)	18 (8-20)
Batch 6	8	50	62 (50-74)	10	12	50	38	4.2 (3.5-5.3)	150 (63-180)	PROPELLER	Site 3	Siemens	Aera	Random ID 957	1.5	0.92 (0.9-0.98)	4.9 (4.5-5)	6.3 (5.8-6.5)	AX	260	260	360	46	15	840 (820-900)	2000 (1900-2100)	150
Batch 7	31	58	58 (34-81)	20	26	52	42	4.5 (1.4-7.4)	120 (8.4-220)	3D MP-RAGE	Site 1	Siemens	Avanto	Random ID 488	1.5	0.55	1.1	-	AX	510 (480-510)	380 (340-380)	250	2.3	1	1100	1900 (1600-1900)	15
Batch 8	7	71	63 (41-80)	40	0	29	14	3.8 (3.3-8.5)	61 (46-160)	3D Fast Spin Echo	Site 4	GE	Discovery MR450	Random ID 402	1.5	0.47	1.2	0.6	COR	510	510	240	12	24	-	570 (400-600)	90
Batch 9	13	85	71 (47-77)	8	31	31	54	4.9 (1.2-7.4)	140 (17-230)	2D Spin Echo	Site 7	Siemens	Aera	Random ID 270	1.5	0.72 (0.72-0.75)	5	6	AX	320	280 (250-320)	150	9 (8.9-10)	1	-	500 (410-550)	90
Batch 10	17	65	65 (45-75)	40	18	41	35	4.3 (0.5-5.9)	89 (6.5-240)	3D Fast Spin Echo	Site 4	Siemens	Avanto Fit	Random ID 24	1.5	1 (0.98-1.1)	1	-	COR	260 (250-260)	200 (180-260)	750	11	63 (49-65)	-	690 (600-700)	120
Batch 11	15	67	57 (45-69)	10	40	67	27	5.7 (1.8-7.8)	150 (13-250)	3D MP-RAGE	Site 1	Siemens	Avanto	Random ID 989	1.5	0.55	1.1	-	AX	510 (450-510)	370 (280-380)	250	2.3	1	1100	1900 (1600-1900)	15
Batch 12	6	83	63 (55-68)	50	0	50	33	5.2 (2.4-6.7)	95 (16-130)	3D MP-RAGE	Site 2	Siemens	Avanto	Random ID 975	1.5	0.55	1.1	-	AX	510	380	250	2.3	1	1100	1900	15
Batch 13	18	56	62 (46-73)	10	17	39	33	4.4 (2.3-7.3)	110 (7.3-240)	3D MP-RAGE	Site 2	Siemens	Aera	Random ID 388	1.5	0.55	1.1	-	AX	510	380	250	2.3	1	1100	1900 (1600-1900)	15

Batch 14	11	45	58 (44-71)	20	36	73	64	4.4 (2.7-7.5)	86 (19-280)	PROPELLER	Site 3	Siemens	Aera	Random ID 957	1.5	0.94 (0.9-1.1)	4	5.2	AX	260	260	360	46	15	1200 (910-1300)	2900 (2100-3200)	150 (140-150)

AX = axial, COR = coronal, GE = General Electric, MGMT = O6-methylguanine-DNA methyltransferase, PROPELLER = Periodically Rotated Overlapping ParallEL Lines with Enhanced Reconstruction, - = Not applicable or missing from DICOM header. [a] 100 [b] Completed 60Gy in 30 fractions radiotherapy with concomitant temozolomide and began adjuvant temozolomide. [c] Percentage of patients per batch with methylation of the MGMT promoter. [d] Whole tumour volume (includes enhancement, necrosis and peritumoural high T2 signal). [e] Site identifiers such as scanner location and machine identifier were anonymised.

Table S4.7: List of radiomic features that were power-transformed using Box Cox transformation, the lambda value used and whether the feature was retained after transformation.

Bin Count	Intensity Standardisation	Lambda Used for Transformation	Feature Kept after Transformation	MRI sequence	Feature Class	Feature Name
8	Z-Score	-0.5	Yes	FLAIR	gldm	Small Dependence Low Gray Level Emphasis
8	WhiteStripe	-0.5	Yes	FLAIR	gldm	Small Dependence Low Gray Level Emphasis
32	Z-Score	-0.7	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
32	Z-Score	-0.8	Yes	FLAIR	gldm	Low Gray Level Emphasis
64	Z-Score	-0.6	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
64	Z-Score	-0.6	Yes	FLAIR	gldm	Low Gray Level Emphasis
64	Z-Score	-0.7	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	Z-Score	-0.4	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
128	Z-Score	-0.4	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
128	Z-Score	-0.5	Yes	FLAIR	gldm	Low Gray Level Emphasis
128	Z-Score	-0.5	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
128	Z-Score	-0.1	Yes	FLAIR	glszm	Large Area Low Gray Level Emphasis
128	Z-Score	-0.3	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
128	Z-Score	-0.3	Yes	T1	gldm	Low Gray Level Emphasis
128	Z-Score	-0.4	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
128	Z-Score	-0.2	Yes	T1CE	gldm	Large Dependence Low Gray Level Emphasis
32	WhiteStripe	-0.7	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
32	WhiteStripe	-0.8	Yes	FLAIR	gldm	Low Gray Level Emphasis
32	WhiteStripe	-0.9	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	WhiteStripe	-0.6	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
64	WhiteStripe	-0.6	Yes	FLAIR	gldm	Low Gray Level Emphasis
64	WhiteStripe	-0.7	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	WhiteStripe	-0.4	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
128	WhiteStripe	-0.4	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
128	WhiteStripe	-0.5	Yes	FLAIR	gldm	Low Gray Level Emphasis
128	WhiteStripe	-0.5	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
128	WhiteStripe	-0.1	Yes	FLAIR	glszm	Large Area Low Gray Level Emphasis
128	WhiteStripe	-0.3	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis

128	WhiteStripe	-0.3	Yes	T1	gldm	Low Gray Level Emphasis
128	WhiteStripe	-0.4	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
128	WhiteStripe	-0.2	Yes	T1CE	gldm	Large Dependence Low Gray Level Emphasis
32	Histogram Matching	-0.6	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
32	Histogram Matching	-0.7	Yes	FLAIR	gldm	Low Gray Level Emphasis
32	Histogram Matching	-0.7	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	Histogram Matching	-0.5	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
64	Histogram Matching	-0.7	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	Histogram Matching	-0.5	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
64	Histogram Matching	-0.4	Yes	T1	gldm	Low Gray Level Emphasis
64	Histogram Matching	-0.6	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
64	Histogram Matching	-0.1	Yes	T1CE	gldm	Large Dependence Low Gray Level Emphasis
128	Histogram Matching	-0.4	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
128	Histogram Matching	-0.4	Yes	FLAIR	gldm	Low Gray Level Emphasis
128	Histogram Matching	-0.5	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
128	Histogram Matching	0.0	Yes	FLAIR	glszm	Large Area Low Gray Level Emphasis
128	Histogram Matching	-0.4	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
128	Histogram Matching	-0.4	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
128	Histogram Matching	-0.2	Yes	T1	glszm	Large Area Low Gray Level Emphasis
128	Histogram Matching	-0.2	Yes	T1CE	gldm	Large Dependence Low Gray Level Emphasis
128	Histogram Matching	-0.3	Yes	T1CE	gldm	Low Gray Level Emphasis
128	Histogram Matching	-0.4	Yes	T1CE	glrlm	Long Run Low Gray Level Emphasis
32	No standardisation	-0.7	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
32	No standardisation	-0.8	Yes	FLAIR	gldm	Low Gray Level Emphasis
32	No standardisation	-0.9	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	No standardisation	-0.6	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
64	No standardisation	-0.6	Yes	FLAIR	gldm	Low Gray Level Emphasis
64	No standardisation	-0.7	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
64	No standardisation	-0.4	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
64	No standardisation	-0.5	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
128	No standardisation	-0.4	Yes	FLAIR	gldm	Large Dependence Low Gray Level Emphasis
128	No standardisation	-0.5	Yes	FLAIR	gldm	Low Gray Level Emphasis
128	No standardisation	-0.5	Yes	FLAIR	glrlm	Long Run Low Gray Level Emphasis
128	No standardisation	-0.1	Yes	FLAIR	glszm	Large Area Low Gray Level Emphasis

128	No standardisation	-0.3	Yes	T1	gldm	Large Dependence Low Gray Level Emphasis
128	No standardisation	-0.4	Yes	T1	glrlm	Long Run Low Gray Level Emphasis
128	No standardisation	-0.1	Yes	T1	glszm	Large Area Low Gray Level Emphasis
128	No standardisation	-0.2	Yes	T1CE	gldm	Large Dependence Low Gray Level Emphasis

glcm = gray level co-occurrence matrix, gldm = gray level dependence matrix, glrlm = gray level run length matrix, glszm = gray level size zone matrix, FLAIR = Fluid Attenuated Inversion Recovery image, T1 = T1-weighted image, T1CE = T1-weighted post-gadolinium contrast-enhanced image, T2 = T2-weighted image.

Table S4.6: A step-by-step guide to how the bootstrap resampling process was conducted for the study. Model performance statistics that were calculated during step 5 included measures of concordance, model fit, explained variation and calibration.

Steps	Description
1	Randomly sample patients (b) from the original development dataset (O), allowing duplicates to be selected, until the sample equals the size of original (sample size, n).
2	Apply one of five feature selection strategies until the required number of radiomic features (four) is selected.
3	Create Cox proportional hazards models (i) using the selected radiomic features in sample b (radiomics only model, Rad_b), (ii) using only clinical variables in sample b (clinical only model, $Clin_b$), and (iii) a 'combined' model using the values of the selected radiomic features and clinical variables in sample b ($Comb_b$).
4	Use the models produced in bootstrap sample b (step 3) to make survival predictions ($\hat{p}$) for each patient in the original data at a given time-point (1 year in this study). Hence, for each bootstrap sample,b, n survival predictions ($\hat{p}$) will be produced ($\hat{p}_b$, where b is each bootstrap sample).
5	Measure the 'test' performance of the models produced in the bootstrap sample b, by supplying the original dataset O to the bootstrap models and calculate performance statistics S_b.
6	Repeat steps 1-5 for a large number of repetitions (B, $B = 1000$ for this study).
7	Steps 1-6 will result in a set (of size between 1 and B) of survival predictions and model performance statistics for each feature selection process. This set is used to calculate the mean and 95% confidence interval of each model performance statistic.
8	Each of the predicted survival probabilities ($\hat{p}_b$) is plot against the survival that is actually observed in the original data O at 1-year (or another time point) to produce a calibration plot. Each bootstrap resample (between 1 and B) will result in a unique calibration plot, and by overlaying these onto the same plotting region (up to B lines on one plot), a calibration instability plot can be drawn.

Table S4.8: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 8 bin count, with ComBat feature realignment and all minimum ComBat batch sizes. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
Clin	C	0.82 (0.61 - 1.1)	0.82 (0.59 - 1.1)	0.81 (0.58 - 1.1)	0.83 (0.59 - 1.1)	0.27 (0.22 - 0.29)	0.27 (0.22 - 0.29)	0.27 (0.22 - 0.29)	0.27 (0.23 - 0.29)	0.21 (0.17 - 0.24)	0.21 (0.16 - 0.24)	0.21 (0.17 - 0.24)	0.21 (0.17 - 0.25)	1445 (1438 - 1456)	1445 (1438 - 1456)	1445 (1438 - 1456)	1444 (1438 - 1455)	1.1 (0.92 - 1.2)	1.1 (0.91 - 1.2)	1.1 (0.91 - 1.2)	1.1 (0.92 - 1.2)	0.72 (0.69 - 0.73)	0.71 (0.69 - 0.73)	0.72 (0.69 - 0.73)	0.72 (0.69 - 0.73)
Back	CR	0.76 (0.55 - 1)	0.76 (0.54 - 1)	0.76 (0.53 - 1)	0.77 (0.56 - 1)	0.27 (0.22 - 0.32)	0.27 (0.22 - 0.31)	0.27 (0.22 - 0.32)	0.28 (0.22 - 0.33)	0.23 (0.17 - 0.27)	0.22 (0.17 - 0.27)	0.22 (0.16 - 0.27)	0.23 (0.17 - 0.28)	1443 (1431 - 1457)	1443 (1432 - 1457)	1443 (1432 - 1458)	1441 (1428 - 1456)	1.1 (0.91 - 1.3)	1.1 (0.91 - 1.2)	1.1 (0.89 - 1.2)	1.1 (0.94 - 1.3)	0.72 (0.69 - 0.74)	0.71 (0.69 - 0.73)	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)
Forw	CR	0.7 (0.49 - 0.91)	0.7 (0.49 - 0.91)	0.7 (0.49 - 0.9)	0.7 (0.48 - 0.91)	0.27 (0.2 - 0.32)	0.28 (0.21 - 0.33)	0.28 (0.21 - 0.33)	0.29 (0.22 - 0.35)	0.22 (0.15 - 0.28)	0.23 (0.16 - 0.29)	0.23 (0.16 - 0.28)	0.25 (0.17 - 0.31)	1444 (1430 - 1463)	1442 (1426 - 1460)	1442 (1429 - 1460)	1438 (1421 - 1457)	1.1 (0.86 - 1.3)	1.1 (0.89 - 1.3)	1.1 (0.89 - 1.3)	1.2 (0.92 - 1.4)	0.71 (0.68 - 0.74)	0.71 (0.68 - 0.74)	0.71 (0.68 - 0.74)	0.72 (0.69 - 0.74)
LASSO	CR	0.72 (0.52 - 0.94)	0.72 (0.52 - 0.94)	0.72 (0.51 - 0.92)	0.73 (0.53 - 0.95)	0.27 (0.21 - 0.32)	0.28 (0.21 - 0.33)	0.28 (0.21 - 0.32)	0.3 (0.23 - 0.36)	0.22 (0.16 - 0.27)	0.23 (0.17 - 0.29)	0.23 (0.17 - 0.27)	0.25 (0.18 - 0.32)	1444 (1431 - 1460)	1441 (1428 - 1458)	1442 (1430 - 1459)	1436 (1419 - 1454)	1.1 (0.88 - 1.2)	1.1 (0.92 - 1.3)	1.1 (0.91 - 1.2)	1.2 (0.96 - 1.4)	0.71 (0.68 - 0.73)	0.72 (0.69 - 0.74)	0.71 (0.68 - 0.73)	0.72 (0.69 - 0.74)
RSF	CR	0.76 (0.56 - 0.99)	0.76 (0.55 - 1)	0.76 (0.54 - 1)	0.76 (0.53 - 1)	0.29 (0.23 - 0.32)	0.29 (0.24 - 0.33)	0.27 (0.22 - 0.31)	0.28 (0.23 - 0.32)	0.24 (0.18 - 0.28)	0.25 (0.2 - 0.28)	0.23 (0.17 - 0.27)	0.23 (0.17 - 0.27)	1440 (1430 - 1455)	1437 (1428 - 1451)	1443 (1432 - 1456)	1440 (1429 - 1454)	1.2 (0.96 - 1.3)	1.2 (1 - 1.3)	1.1 (0.92 - 1.2)	1.1 (0.94 - 1.2)	0.72 (0.7 - 0.74)	0.72 (0.7 - 0.74)	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)
PCA	CR	0.76 (0.55 - 1)	0.76 (0.55 - 1)	0.75 (0.53 - 1)	0.76 (0.56 - 1)	0.29 (0.23 - 0.34)	0.28 (0.22 - 0.33)	0.28 (0.22 - 0.33)	0.28 (0.22 - 0.32)	0.24 (0.18 - 0.29)	0.24 (0.17 - 0.28)	0.23 (0.16 - 0.28)	0.24 (0.18 - 0.28)	1439 (1425 - 1455)	1440 (1428 - 1456)	1441 (1428 - 1458)	1440 (1430 - 1456)	1.2 (0.95 - 1.3)	1.1 (0.93 - 1.3)	1.1 (0.91 - 1.3)	1.1 (0.94 - 1.3)	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)	0.72 (0.69 - 0.74)
10 patients per batch																									

	C	0.78	0.79	0.78	0.79	0.28	0.28	0.28	0.28	0.21	0.21	0.21	0.22	1139	1139	1138	1138	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.53	(0.54	(0.53	(0.53	(0.23	(0.23	(0.23	(0.22	(0.16	(0.15	(0.15	(0.15	(1132	(1132	(1132	(1132	(0.88	(0.85	(0.87	(0.87	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1149)	1150)	1149)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.72	0.71	0.72	0.29	0.28	0.29	0.3	0.24	0.23	0.23	0.25	1136	1137	1136	1133	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.47	(0.47	(0.47	(0.48	(0.22	(0.21	(0.22	(0.22	(0.15	(0.15	(0.15	(0.16	(1122	(1124	(1124	(1119	(0.87	(0.85	(0.86	(0.91	(0.69	(0.69	(0.69	(0.69
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)		0.97)	0.98)	0.35)	0.34)	0.34)	0.36)	0.31)	0.3)	0.3)	0.32)	1150)	1152)	1150)	1150)	1.4)	1.3)	1.3)	1.4)	0.75)	0.74)	0.74)	0.75)
	CR	0.65	0.65	0.65	0.67	0.3	0.29	0.3	0.33	0.25	0.24	0.25	0.28	1134	1135	1133	1126	1.2	1.2	1.2	1.3	0.72	0.72	0.72	0.73
Forw		(0.44	(0.42	(0.43	(0.45	(0.21	(0.2	(0.21	(0.24	(0.16	(0.15	(0.15	(0.18	(1119	(1118	(1117	(1109	(0.89	(0.86	(0.85	(0.96	(0.69	(0.69	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.89)	0.87)	0.9)	0.36)	0.36)	0.37)	0.4)	0.33)	0.33)	0.34)	0.36)	1152)	1155)	1153)	1146)	1.4)	1.4)	1.5)	1.5)	0.75)	0.75)	0.75)	0.76)
	CR	0.66	0.68	0.68	0.7	0.29	0.29	0.3	0.33	0.24	0.23	0.24	0.28	1135	1136	1134	1126	1.2	1.1	1.2	1.3	0.72	0.72	0.73	0.73
LASSO		(0.45	(0.46	(0.47	(0.48	(0.22	(0.21	(0.22	(0.23	(0.15	(0.15	(0.15	(0.18	(1121	(1120	(1119	(1109	(0.87	(0.88	(0.87	(0.97	(0.69	(0.69	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.9)	0.91)	0.91)	0.94)	0.35)	0.36)	0.36)	0.4)	0.32)	0.32)	0.33)	0.36)	1151)	1153)	1150)	1149)	1.4)	1.4)	1.4)	1.5)	0.75)	0.74)	0.75)	0.76)
	CR	0.71	0.72	0.72	0.71	0.29	0.29	0.28	0.27	0.24	0.23	0.22	0.21	1136	1137	1139	1139	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.48	(0.49	(0.49	(0.47	(0.22	(0.22	(0.21	(0.2	(0.17	(0.16	(0.15	(0.15	(1126	(1128	(1129	(1130	(0.93	(0.91	(0.87	(0.84	(0.7	(0.69	(0.69	(0.68
		-	- 1)	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.98)		0.33)	0.33)	0.32)	0.31)	0.3)	0.27)	0.27)	0.26)	1150)	1150)	1152)	1155)	1.3)	1.2)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.71	0.73	0.71	0.72	0.3	0.3	0.29	0.3	0.24	0.24	0.24	0.25	1134	1134	1135	1134	1.2	1.2	1.1	1.2	0.73	0.72	0.72	0.73
PCA		(0.47	(0.48	(0.46	(0.47	(0.23	(0.23	(0.22	(0.22	(0.16	(0.16	(0.16	(0.16	(1120	(1121	(1120	(1121	(0.88	(0.9	(0.88	(0.91	(0.69	(0.69	(0.69	(0.7
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.98)	0.98)	0.36)	0.35)	0.36)	0.35)	0.32)	0.31)	0.32)	0.31)	1150)	1149)	1151)	1150)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.74	0.74	0.75	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884.1	884.2	884.1	884.2	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.48	(0.47	(0.48	(0.47	(0.21	(0.21	(0.21	(0.21	(0.14	(0.14	(0.15	(0.14	(877.5	(877.6	(877.6	(877.5	(0.82	(0.82	(0.85	(0.83	(0.68	(0.68	(0.68	(0.68
		- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.26)	0.26)	895.2)	895.5)	895.1)	894.5)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.67	0.67	0.66	0.28	0.28	0.28	0.29	0.23	0.22	0.23	0.24	882.1	883.1	882.5	881.3	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.42	(0.41	(0.42	(0.42	(0.2	(0.2	(0.2	(0.21	(0.13	(0.14	(0.13	(0.15	(870.4	(871.6	(870.1	(868.6	(0.8	(0.81	(0.8	(0.84	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.96)	0.96)	0.96)	0.35)	0.34)	0.35)	0.36)	0.31)	0.3)	0.31)	0.32)	896)	897)	897.1)	895.2)	1.4)	1.3)	1.4)	1.4)	0.75)	0.74)	0.75)	0.75)

	CR	0.59	0.59	0.6	0.59	0.29	0.3	0.31	0.32	0.24	0.24	0.25	0.27	880.4	879.9	877.4	875.6	1.2	1.2	1.2	1.2	0.72	0.72	0.73	0.73
Forw		(0.37	(0.35	(0.37	(0.37	(0.18	(0.18	(0.2	(0.21	(0.13	(0.13	(0.14	(0.16	(864.4	(862.5	(860	(859.7	(0.8	(0.78	(0.82	(0.88	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.85)	0.84)	0.84)	0.85)	0.38)	0.39)	0.4)	0.4)	0.34)	0.36)	0.36)	0.36)	899.4)	899.2)	896.7)	895.3)	1.5)	1.5)	1.5)	1.5)	0.75)	0.75)	0.76)	0.77)
	CR	0.62	0.62	0.63	0.62	0.29	0.29	0.3	0.32	0.24	0.23	0.24	0.26	880.7	881.7	878.7	876.1	1.1	1.1	1.2	1.2	0.72	0.72	0.73	0.73
LASSO		(0.39	(0.38	(0.4	(0.39	(0.2	(0.19	(0.22	(0.22	(0.14	(0.13	(0.16	(0.17	(867.9	(865.9	(864	(861	(0.83	(0.8	(0.88	(0.92	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.87)	0.9)	0.89)	0.88)	0.36)	0.37)	0.38)	0.39)	0.33)	0.33)	0.34)	0.35)	896)	897.8)	894)	893.8)	1.4)	1.4)	1.5)	1.5)	0.75)	0.75)	0.75)	0.77)
	CR	0.66	0.68	0.68	0.66	0.29	0.29	0.28	0.27	0.23	0.23	0.22	0.21	881.7	880.5	883.1	884.5	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.71
RSF		(0.42	(0.43	(0.42	(0.41	(0.2	(0.21	(0.2	(0.19	(0.14	(0.16	(0.14	(0.13	(872.5	(870.9	(872.8	(874.8	(0.82	(0.89	(0.83	(0.79	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.95)	0.95)	0.95)	0.34)	0.34)	0.34)	0.32)	0.29)	0.29)	0.29)	0.27)	896.1)	894.4)	897.2)	898)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.66	0.66	0.29	0.29	0.28	0.28	0.23	0.23	0.23	0.23	881.5	881.7	882.5	882.7	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
PCA		(0.4	(0.39	(0.4	(0.39	(0.19	(0.2	(0.19	(0.18	(0.13	(0.14	(0.13	(0.13	(869.6	(870.4	(871.6	(870.6	(0.81	(0.82	(0.8	(0.8	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.95)	0.96)	0.98)	0.35)	0.35)	0.34)	0.35)	0.31)	0.31)	0.3)	0.31)	898.2)	897.1)	897.6)	898.9)	1.4)	1.4)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.9: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 8 bin count and all minimum ComBat batch sizes but without using ComBat feature realignment. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.82	0.83	0.82	0.83	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	1445	1445	1444	1445	1.1	1.1	1.1	1.1	0.71	0.72	0.72	0.72
Clin		(0.59	(0.6	(0.61	(0.6	(0.22	(0.23	(0.23	(0.23	(0.16	(0.17	(0.17	(0.17	(1438	(1438	(1438	(1438	(0.91	(0.92	(0.92	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.25)	0.25)	0.25)	0.24)	1456)	1455)	1456)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.76	0.76	0.76	0.77	0.27	0.27	0.27	0.28	0.22	0.22	0.22	0.23	1443	1443	1443	1441	1.1	1.1	1.1	1.1	0.72	0.71	0.71	0.72
Back		(0.53	(0.55	(0.55	(0.56	(0.22	(0.22	(0.22	(0.23	(0.16	(0.16	(0.16	(0.17	(1431	(1433	(1432	(1428	(0.91	(0.9	(0.89	(0.94	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.31)	0.31)	0.33)	0.27)	0.27)	0.27)	0.28)	1457)	1458)	1457)	1456)	1.3)	1.2)	1.2)	1.3)	0.74)	0.73)	0.73)	0.74)
	CR	0.7	0.7	0.7	0.7	0.27	0.28	0.28	0.29	0.22	0.23	0.23	0.24	1444	1442	1442	1438	1.1	1.1	1.1	1.2	0.71	0.71	0.71	0.72
Forw		(0.49	(0.46	(0.5	(0.49	(0.2	(0.2	(0.21	(0.22	(0.15	(0.15	(0.16	(0.17	(1430	(1426	(1430	(1422	(0.86	(0.85	(0.89	(0.92	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.92)	0.92)	0.9)	0.32)	0.33)	0.32)	0.35)	0.27)	0.29)	0.27)	0.31)	1462)	1462)	1459)	1457)	1.3)	1.3)	1.3)	1.4)	0.74)	0.74)	0.74)	0.75)
	CR	0.72	0.73	0.73	0.73	0.27	0.28	0.28	0.3	0.22	0.23	0.23	0.25	1444	1442	1442	1436	1.1	1.1	1.1	1.2	0.71	0.72	0.71	0.72
LASSO		(0.51	(0.52	(0.52	(0.55	(0.2	(0.21	(0.22	(0.23	(0.15	(0.17	(0.16	(0.19	(1432	(1428	(1430	(1422	(0.85	(0.93	(0.89	(1 -	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.4)	-	-	-	-
		0.95)	0.96)	0.96)	0.97)	0.31)	0.33)	0.32)	0.35)	0.27)	0.29)	0.27)	0.31)	1461)	1458)	1458)	1453)	1.2)	1.3)	1.3)		0.73)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.77	0.77	0.28	0.29	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1440	1.1	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.54	(0.56	(0.55	(0.54	(0.23	(0.24	(0.22	(0.23	(0.18	(0.19	(0.16	(0.18	(1430	(1428	(1431	(1430	(0.95	(1 -	(0.91	(0.96	(0.69	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	1.3)	-	-	-	-	-	-
						0.32)	0.33)	0.32)	0.32)	0.28)	0.28)	0.27)	0.27)	1455)	1451)	1456)	1454)	1.3)		1.2)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.76	0.77	0.29	0.28	0.28	0.28	0.24	0.24	0.23	0.24	1439	1440	1441	1440	1.2	1.1	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.54	(0.54	(0.54	(0.54	(0.23	(0.22	(0.22	(0.23	(0.17	(0.17	(0.17	(0.18	(1426	(1428	(1428	(1429	(0.93	(0.92	(0.91	(0.95	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.33)	0.33)	0.33)	0.32)	0.29)	0.28)	0.28)	0.28)	1455)	1457)	1457)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.79	0.77	0.78	0.78	0.28	0.28	0.28	0.28	0.21	0.21	0.21	0.21	1139	1139	1139	1139	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.53	(0.51	(0.52	(0.52	(0.22	(0.22	(0.23	(0.23	(0.15	(0.16	(0.15	(0.15	(1132	(1132	(1132	(1132	(0.88	(0.88	(0.86	(0.87	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1150)	1150)	1150)	1149)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.72	0.71	0.71	0.73	0.29	0.28	0.29	0.3	0.23	0.23	0.23	0.25	1136	1137	1136	1133	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.48	(0.46	(0.48	(0.47	(0.21	(0.21	(0.22	(0.23	(0.15	(0.15	(0.15	(0.16	(1122	(1125	(1124	(1117	(0.85	(0.86	(0.85	(0.9	(0.69	(0.69	(0.69	(0.69
		-	- 1)	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.96)		0.35)	0.34)	0.34)	0.37)	0.32)	0.3)	0.31)	0.33)	1152)	1152)	1151)	1148)	1.4)	1.4)	1.4)	1.4)	0.75)	0.74)	0.75)	0.75)
	CR	0.66	0.64	0.65	0.67	0.3	0.29	0.3	0.33	0.25	0.24	0.25	0.28	1134	1135	1132	1126	1.2	1.1	1.2	1.3	0.72	0.72	0.72	0.73
Forw		(0.43	(0.42	(0.45	(0.45	(0.21	(0.2	(0.22	(0.23	(0.15	(0.15	(0.16	(0.16	(1119	(1120	(1116	(1108	(0.86	(0.86	(0.88	(0.9	(0.69	(0.68	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.89)	0.88)	0.91)	0.36)	0.36)	0.37)	0.4)	0.33)	0.33)	0.35)	0.36)	1153)	1154)	1151)	1149)	1.4)	1.4)	1.5)	1.5)	0.75)	0.75)	0.75)	0.76)
	CR	0.68	0.67	0.68	0.69	0.29	0.29	0.3	0.33	0.24	0.23	0.25	0.28	1134	1136	1133	1126	1.2	1.1	1.2	1.3	0.72	0.72	0.73	0.73
LASSO		(0.46	(0.45	(0.46	(0.47	(0.21	(0.21	(0.22	(0.24	(0.14	(0.15	(0.16	(0.18	(1120	(1121	(1119	(1109	(0.84	(0.87	(0.88	(0.97	(0.69	(0.69	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.92)	0.9)	0.92)	0.36)	0.35)	0.36)	0.4)	0.32)	0.31)	0.32)	0.35)	1153)	1152)	1151)	1146)	1.4)	1.4)	1.4)	1.5)	0.75)	0.75)	0.75)	0.76)
	CR	0.71	0.71	0.71	0.7	0.29	0.28	0.28	0.27	0.24	0.23	0.22	0.21	1136	1137	1139	1139	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.49	(0.47	(0.48	(0.46	(0.22	(0.22	(0.21	(0.21	(0.17	(0.16	(0.15	(0.15	(1126	(1128	(1129	(1130	(0.92	(0.9	(0.85	(0.84	(0.69	(0.68	(0.69	(0.68
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.97)	0.98)	0.33)	0.32)	0.32)	0.31)	0.3)	0.27)	0.27)	0.26)	1151)	1151)	1153)	1154)	1.3)	1.3)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.71	0.71	0.3	0.29	0.29	0.29	0.24	0.24	0.24	0.24	1134	1134	1136	1135	1.2	1.1	1.1	1.2	0.72	0.72	0.72	0.73
PCA		(0.48	(0.45	(0.45	(0.46	(0.22	(0.22	(0.21	(0.23	(0.16	(0.16	(0.15	(0.16	(1118	(1121	(1121	(1122	(0.88	(0.9	(0.85	(0.9	(0.69	(0.69	(0.69	(0.7
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.98)	0.99)	0.37)	0.35)	0.35)	0.35)	0.32)	0.31)	0.31)	0.31)	1150)	1150)	1152)	1149)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.74	0.74	0.75	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884.2	884.4	883.9	884.1	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.48	(0.48	(0.49	(0.49	(0.21	(0.21	(0.21	(0.21	(0.15	(0.14	(0.15	(0.14	(877.7	(877.8	(877.4	(877.5	(0.85	(0.84	(0.84	(0.84	(0.68	(0.68	(0.69	(0.68
		- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.26)	0.25)	0.26)	0.25)	895)	894.8)	894.8)	895.3)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.68	0.67	0.28	0.28	0.28	0.29	0.23	0.22	0.23	0.24	882.2	883.4	882.1	880.8	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.41	(0.42	(0.41	(0.42	(0.2	(0.2	(0.2	(0.2	(0.14	(0.14	(0.13	(0.14	(869.6	(870.8	(869.2	(867.2	(0.82	(0.82	(0.78	(0.84	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.96)	0.96)	0.96)	0.35)	0.34)	0.35)	0.36)	0.31)	0.31)	0.31)	0.33)	896.4)	896.9)	896.8)	896.6)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)

	CR	0.59	0.59	0.6	0.59	0.3	0.29	0.31	0.32	0.24	0.24	0.25	0.26	880	880.1	877.3	875.8	1.2	1.2	1.2	1.2	0.72	0.72	0.73	0.73
Forw		(0.38	(0.35	(0.38	(0.35	(0.18	(0.19	(0.19	(0.2	(0.14	(0.13	(0.14	(0.15	(864.7	(864.1	(859.3	(859.8	(0.81	(0.8	(0.82	(0.86	(0.68	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.82)	0.84)	0.84)	0.85)	0.38)	0.38)	0.4)	0.4)	0.34)	0.35)	0.36)	0.36)	899.1)	898.8)	898.2)	897.1)	1.5)	1.5)	1.5)	1.5)	0.75)	0.75)	0.76)	0.77)
	CR	0.62	0.61	0.63	0.62	0.29	0.28	0.31	0.31	0.24	0.23	0.25	0.26	880.6	882.5	877.9	876.4	1.1	1.1	1.2	1.2	0.72	0.72	0.73	0.73
LASSO		(0.39	(0.39	(0.41	(0.41	(0.19	(0.19	(0.21	(0.22	(0.14	(0.13	(0.15	(0.17	(866.7	(865	(862.5	(861.2	(0.81	(0.8	(0.85	(0.91	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.87)	0.87)	0.87)	0.87)	0.37)	0.37)	0.39)	0.39)	0.33)	0.34)	0.35)	0.35)	898.3)	898.1)	894.5)	893.1)	1.4)	1.5)	1.5)	1.5)	0.75)	0.75)	0.75)	0.77)
	CR	0.66	0.68	0.68	0.67	0.29	0.29	0.28	0.27	0.23	0.23	0.23	0.21	881.6	880.4	882.8	883.9	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.71
RSF		(0.43	(0.45	(0.43	(0.43	(0.21	(0.22	(0.2	(0.19	(0.15	(0.16	(0.15	(0.13	(871.9	(871.5	(873.2	(874.8	(0.87	(0.89	(0.85	(0.8	(0.69	(0.68	(0.68	(0.67
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.97)	0.95)	0.98)	0.34)	0.34)	0.33)	0.32)	0.29)	0.29)	0.29)	0.27)	894.7)	893.7)	896)	897.4)	1.3)	1.3)	1.3)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.65	0.67	0.66	0.65	0.29	0.29	0.28	0.28	0.23	0.23	0.22	0.23	881.7	881.6	882.8	882.9	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.4	(0.42	(0.38	(0.36	(0.2	(0.2	(0.18	(0.19	(0.13	(0.14	(0.13	(0.14	(869.8	(870.5	(871	(871.6	(0.8	(0.82	(0.81	(0.83	(0.69	(0.69	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.92)	0.96)	0.93)	0.95)	0.35)	0.35)	0.34)	0.34)	0.31)	0.3)	0.3)	0.31)	896.9)	896.2)	899.5)	898.3)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.10: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 32 bin count, with ComBat feature realignment and all minimum ComBat batch sizes. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.82	0.83	0.82	0.82	0.27	0.27	0.27	0.27	0.21	0.22	0.21	0.21	1444	1444	1445	1445	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.59	(0.6	(0.6	(0.6	(0.23	(0.22	(0.22	(0.23	(0.17	(0.17	(0.16	(0.17	(1438	(1438	(1438	(1438	(0.92	(0.92	(0.91	(0.93	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.24)	0.24)	0.24)	0.24)	1455)	1457)	1456)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.76	0.77	0.76	0.77	0.28	0.27	0.27	0.28	0.23	0.23	0.23	0.23	1442	1443	1443	1441	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.55	(0.54	(0.56	(0.56	(0.22	(0.21	(0.22	(0.23	(0.17	(0.16	(0.17	(0.17	(1430	(1431	(1431	(1429	(0.91	(0.89	(0.92	(0.94	(0.69	(0.68	(0.68	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.32)	0.32)	0.32)	0.28)	0.27)	0.27)	0.28)	1456)	1459)	1457)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.7	0.7	0.7	0.7	0.28	0.28	0.28	0.28	0.23	0.23	0.23	0.24	1442	1442	1442	1440	1.1	1.1	1.1	1.1	0.72	0.71	0.71	0.72
Forw		(0.51	(0.48	(0.51	(0.5	(0.21	(0.2	(0.21	(0.22	(0.16	(0.16	(0.16	(0.17	(1424	(1427	(1426	(1426	(0.9	(0.89	(0.89	(0.93	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.92)	0.93)	0.91)	0.91)	0.34)	0.33)	0.34)	0.34)	0.3)	0.29)	0.29)	0.29)	1459)	1461)	1460)	1458)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.72	0.73	0.73	0.72	0.27	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1443	1441	1441	1439	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
LASSO		(0.54	(0.52	(0.53	(0.5	(0.21	(0.22	(0.22	(0.22	(0.16	(0.17	(0.17	(0.17	(1430	(1428	(1428	(1426	(0.91	(0.91	(0.93	(0.94	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.96)	0.97)	0.93)	0.32)	0.33)	0.33)	0.33)	0.28)	0.28)	0.28)	0.29)	1458)	1458)	1456)	1457)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.77	0.76	0.29	0.3	0.28	0.28	0.24	0.25	0.23	0.23	1439	1437	1442	1440	1.2	1.2	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.55	(0.56	(0.55	(0.54	(0.24	(0.24	(0.22	(0.23	(0.19	(0.19	(0.16	(0.18	(1430	(1428	(1432	(1430	(0.98	(0.99	(0.91	(0.94	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.33)	0.32)	0.32)	0.28)	0.28)	0.27)	0.27)	1453)	1451)	1457)	1455)	1.3)	1.3)	1.2)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.77	0.76	0.3	0.29	0.29	0.28	0.25	0.24	0.24	0.23	1437	1439	1439	1441	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.55	(0.55	(0.57	(0.55	(0.24	(0.23	(0.23	(0.23	(0.18	(0.18	(0.17	(0.17	(1423	(1426	(1426	(1430	(0.97	(0.95	(0.93	(0.94	(0.7	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.34)	0.34)	0.33)	0.32)	0.31)	0.3)	0.29)	0.28)	1452)	1455)	1455)	1455)	1.4)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.79	0.79	0.78	0.79	0.28	0.28	0.28	0.28	0.21	0.22	0.21	0.21	1139	1139	1139	1138	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.54	(0.54	(0.52	(0.54	(0.22	(0.23	(0.22	(0.22	(0.15	(0.16	(0.15	(0.15	(1132	(1132	(1132	(1132	(0.87	(0.88	(0.87	(0.87	(0.69	(0.69	(0.69	(0.69
		-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)		1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1150)	1149)	1150)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.72	0.72	0.73	0.29	0.29	0.29	0.3	0.24	0.23	0.24	0.25	1136	1136	1136	1133	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.48	(0.48	(0.48	(0.49	(0.22	(0.22	(0.22	(0.23	(0.15	(0.15	(0.16	(0.17	(1123	(1123	(1124	(1119	(0.88	(0.87	(0.89	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)	0.98)	0.98)	0.98)	0.35)	0.35)	0.34)	0.36)	0.31)	0.3)	0.3)	0.32)	1152)	1151)	1150)	1148)	1.4)	1.3)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)
	CR	0.66	0.64	0.65	0.66	0.31	0.29	0.3	0.31	0.26	0.24	0.25	0.26	1132	1135	1133	1132	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.46	(0.43	(0.44	(0.46	(0.21	(0.2	(0.22	(0.21	(0.16	(0.15	(0.16	(0.16	(1114	(1118	(1116	(1114	(0.88	(0.86	(0.88	(0.88	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.87)	0.89)	0.88)	0.38)	0.37)	0.37)	0.38)	0.35)	0.32)	0.33)	0.34)	1153)	1154)	1151)	1152)	1.5)	1.4)	1.4)	1.5)	0.76)	0.75)	0.75)	0.75)
	CR	0.68	0.68	0.68	0.69	0.3	0.29	0.3	0.31	0.25	0.24	0.25	0.26	1133	1135	1133	1131	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.46	(0.45	(0.47	(0.47	(0.22	(0.22	(0.23	(0.22	(0.16	(0.16	(0.16	(0.18	(1116	(1120	(1119	(1117	(0.88	(0.88	(0.89	(0.95	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.94)	0.91)	0.92)	0.37)	0.36)	0.36)	0.37)	0.34)	0.31)	0.31)	0.33)	1151)	1151)	1148)	1150)	1.5)	1.4)	1.4)	1.4)	0.76)	0.75)	0.75)	0.75)
	CR	0.71	0.72	0.72	0.71	0.29	0.29	0.28	0.27	0.24	0.23	0.22	0.21	1136	1136	1139	1139	1.2	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.48	(0.49	(0.48	(0.48	(0.22	(0.23	(0.22	(0.2	(0.17	(0.17	(0.16	(0.14	(1126	(1127	(1129	(1130	(0.92	(0.92	(0.88	(0.83	(0.7	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)	0.98)	0.96)	0.98)	0.33)	0.33)	0.32)	0.32)	0.3)	0.27)	0.27)	0.26)	1150)	1149)	1151)	1154)	1.3)	1.3)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.71	0.72	0.31	0.3	0.3	0.29	0.26	0.25	0.25	0.24	1131	1133	1133	1135	1.2	1.2	1.2	1.2	0.73	0.73	0.73	0.72
PCA		(0.48	(0.48	(0.47	(0.49	(0.22	(0.22	(0.22	(0.22	(0.17	(0.16	(0.16	(0.17	(1114	(1119	(1118	(1124	(0.93	(0.9	(0.9	(0.91	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)	0.99)	0.96)	0.97)	0.38)	0.36)	0.36)	0.34)	0.35)	0.32)	0.33)	0.3)	1150)	1150)	1150)	1150)	1.5)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.75	0.75	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884	884.2	884.2	884.3	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.48	(0.48	(0.47	(0.49	(0.21	(0.21	(0.2	(0.21	(0.14	(0.15	(0.14	(0.14	(877.1	(877.6	(877.7	(877.7	(0.83	(0.85	(0.84	(0.82	(0.68	(0.68	(0.68	(0.68
		-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)		1.1)	0.31)	0.31)	0.31)	0.31)	0.26)	0.25)	0.26)	0.25)	895.2)	894.5)	896.4)	895)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.66	0.67	0.28	0.28	0.28	0.29	0.23	0.22	0.22	0.23	882.1	883.6	882.5	881.5	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.42	(0.43	(0.42	(0.42	(0.21	(0.2	(0.2	(0.2	(0.14	(0.14	(0.14	(0.14	(869.2	(872.7	(871.2	(869	(0.84	(0.82	(0.83	(0.82	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.94)	0.96)	0.96)	0.35)	0.34)	0.34)	0.35)	0.31)	0.29)	0.3)	0.31)	895.5)	896.5)	896.6)	896.6)	1.4)	1.3)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

	CR	0.6	0.58	0.59	0.59	0.29	0.28	0.3	0.31	0.25	0.23	0.24	0.26	880	882.5	879.6	878	1.2	1.1	1.2	1.2	0.72	0.72	0.72	0.73
Forw		(0.36	(0.35	(0.36	(0.36	(0.18	(0.18	(0.19	(0.19	(0.13	(0.13	(0.14	(0.14	(863.6	(867.7	(863.7	(859.5	(0.81	(0.79	(0.82	(0.82	(0.68	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.87)	0.83)	0.85)	0.85)	0.38)	0.36)	0.38)	0.4)	0.35)	0.32)	0.34)	0.36)	899.7)	899.9)	897.9)	898.8)	1.5)	1.4)	1.5)	1.5)	0.76)	0.75)	0.75)	0.76)
	CR	0.62	0.62	0.63	0.62	0.3	0.28	0.3	0.3	0.24	0.23	0.24	0.25	880	882.2	879	879.8	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.72
LASSO		(0.4	(0.4	(0.41	(0.42	(0.21	(0.2	(0.2	(0.2	(0.15	(0.14	(0.15	(0.15	(866.9	(870.5	(866	(864.1	(0.86	(0.82	(0.86	(0.87	(0.69	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.89)	0.9)	0.87)	0.36)	0.35)	0.37)	0.38)	0.32)	0.3)	0.32)	0.33)	895.1)	897.2)	895.8)	896.8)	1.4)	1.3)	1.4)	1.5)	0.75)	0.75)	0.75)	0.76)
	CR	0.66	0.67	0.67	0.66	0.29	0.29	0.28	0.27	0.23	0.23	0.23	0.21	881.6	880.3	882.5	884.4	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.71
RSF		(0.43	(0.42	(0.43	(0.42	(0.21	(0.21	(0.2	(0.18	(0.15	(0.16	(0.15	(0.13	(872.1	(870.7	(873.2	(874.5	(0.87	(0.89	(0.84	(0.78	(0.69	(0.69	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.97)	0.96)	0.93)	0.34)	0.35)	0.33)	0.33)	0.29)	0.29)	0.29)	0.27)	894.9)	894.2)	897)	899.5)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.66	0.67	0.66	0.66	0.29	0.29	0.29	0.28	0.24	0.23	0.23	0.23	880.7	880.9	881.7	882.2	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
PCA		(0.42	(0.4	(0.41	(0.42	(0.21	(0.21	(0.2	(0.19	(0.14	(0.15	(0.14	(0.14	(867.7	(870.3	(869.3	(870.4	(0.84	(0.85	(0.83	(0.84	(0.69	(0.69	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.95)	0.95)	0.93)	0.36)	0.35)	0.35)	0.35)	0.31)	0.3)	0.3)	0.31)	895.6)	895.7)	896.8)	898.1)	1.4)	1.3)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.11: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 32 bin count and all minimum ComBat batch sizes but without ComBat feature realignment. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.81	0.82	0.82	0.82	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	1445	1444	1444	1445	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.59	(0.6	(0.59	(0.59	(0.23	(0.23	(0.22	(0.22	(0.17	(0.17	(0.17	(0.17	(1438	(1438	(1438	(1438	(0.93	(0.92	(0.91	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.24)	0.25)	0.25)	0.24)	1454)	1456)	1457)	1457)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.76	0.77	0.76	0.76	0.28	0.27	0.27	0.28	0.23	0.23	0.23	0.23	1442	1443	1443	1441	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.55	(0.56	(0.54	(0.56	(0.22	(0.22	(0.22	(0.22	(0.17	(0.17	(0.16	(0.17	(1430	(1431	(1430	(1429	(0.94	(0.91	(0.9	(0.93	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
					0.99)	0.32)	0.32)	0.32)	0.32)	0.28)	0.27)	0.27)	0.28)	1456)	1457)	1458)	1457)	1.3)	1.3)	1.2)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.7	0.7	0.69	0.7	0.28	0.28	0.28	0.28	0.23	0.23	0.23	0.24	1442	1442	1442	1440	1.1	1.1	1.1	1.1	0.72	0.72	0.71	0.72
Forw		(0.51	(0.49	(0.5	(0.49	(0.21	(0.21	(0.21	(0.22	(0.16	(0.16	(0.16	(0.17	(1425	(1427	(1425	(1426	(0.89	(0.91	(0.88	(0.91	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.9)	0.9)	0.92)	0.9)	0.34)	0.33)	0.34)	0.33)	0.3)	0.29)	0.29)	0.3)	1460)	1458)	1460)	1458)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.72	0.73	0.73	0.72	0.28	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1442	1441	1441	1440	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
LASSO		(0.53	(0.53	(0.51	(0.51	(0.21	(0.22	(0.21	(0.23	(0.16	(0.17	(0.16	(0.18	(1429	(1429	(1429	(1425	(0.9	(0.94	(0.89	(0.95	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.94)	0.96)	0.94)	0.32)	0.33)	0.32)	0.34)	0.28)	0.28)	0.28)	0.29)	1458)	1457)	1458)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.76	0.76	0.29	0.3	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1441	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.55	(0.55	(0.54	(0.55	(0.23	(0.24	(0.21	(0.23	(0.19	(0.19	(0.17	(0.18	(1430	(1428	(1432	(1430	(0.98	(0.99	(0.91	(0.94	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.33)	0.32)	0.32)	0.28)	0.28)	0.27)	0.27)	1453)	1451)	1459)	1455)	1.3)	1.3)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.76	0.76	0.3	0.29	0.29	0.28	0.25	0.24	0.24	0.23	1437	1439	1439	1441	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.55	(0.55	(0.55	(0.55	(0.24	(0.23	(0.23	(0.22	(0.18	(0.18	(0.17	(0.17	(1422	(1426	(1426	(1431	(0.96	(0.95	(0.93	(0.93	(0.7	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.35)	0.33)	0.33)	0.32)	0.31)	0.29)	0.29)	0.28)	1452)	1455)	1455)	1457)	1.4)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.78	0.79	0.78	0.78	0.28	0.28	0.28	0.28	0.21	0.21	0.21	0.21	1139	1139	1139	1138	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.53	(0.53	(0.53	(0.52	(0.22	(0.23	(0.22	(0.23	(0.15	(0.16	(0.15	(0.16	(1132	(1132	(1132	(1132	(0.85	(0.88	(0.86	(0.89	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1150)	1149)	1150)	1149)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.72	0.71	0.72	0.29	0.29	0.29	0.3	0.24	0.23	0.23	0.25	1136	1136	1136	1133	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.48	(0.49	(0.49	(0.48	(0.22	(0.22	(0.22	(0.23	(0.15	(0.15	(0.16	(0.17	(1124	(1125	(1123	(1119	(0.87	(0.88	(0.89	(0.91	(0.69	(0.69	(0.69	(0.7
		-	-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.97)	0.97)		0.34)	0.34)	0.34)	0.36)	0.31)	0.3)	0.3)	0.32)	1151)	1151)	1151)	1148)	1.4)	1.3)	1.3)	1.4)	0.75)	0.74)	0.75)	0.75)
	CR	0.66	0.65	0.64	0.66	0.31	0.29	0.3	0.31	0.26	0.24	0.25	0.26	1132	1135	1133	1131	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.43	(0.43	(0.44	(0.45	(0.2	(0.2	(0.21	(0.22	(0.15	(0.15	(0.15	(0.17	(1112	(1118	(1115	(1114	(0.86	(0.84	(0.86	(0.91	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.89)	0.9)	0.88)	0.9)	0.39)	0.37)	0.38)	0.38)	0.34)	0.32)	0.33)	0.34)	1154)	1154)	1152)	1150)	1.5)	1.4)	1.4)	1.5)	0.76)	0.75)	0.75)	0.75)
	CR	0.68	0.69	0.68	0.69	0.31	0.29	0.3	0.31	0.26	0.24	0.25	0.26	1132	1135	1132	1131	1.2	1.2	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.46	(0.46	(0.47	(0.46	(0.22	(0.21	(0.23	(0.24	(0.16	(0.16	(0.16	(0.18	(1115	(1121	(1117	(1116	(0.9	(0.9	(0.9	(0.95	(0.7	(0.69	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.94)	0.9)	0.93)	0.38)	0.35)	0.37)	0.37)	0.34)	0.31)	0.32)	0.33)	1151)	1152)	1149)	1147)	1.5)	1.4)	1.4)	1.4)	0.76)	0.75)	0.75)	0.76)
	CR	0.71	0.73	0.71	0.71	0.29	0.29	0.28	0.27	0.24	0.23	0.22	0.21	1136	1136	1139	1139	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.49	(0.49	(0.47	(0.47	(0.22	(0.23	(0.21	(0.2	(0.16	(0.17	(0.15	(0.14	(1126	(1128	(1129	(1130	(0.9	(0.92	(0.87	(0.84	(0.69	(0.69	(0.68	(0.68
		-	-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.99)	0.96)		0.33)	0.33)	0.32)	0.31)	0.3)	0.27)	0.27)	0.26)	1151)	1149)	1152)	1154)	1.3)	1.3)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.73	0.71	0.71	0.31	0.3	0.3	0.29	0.26	0.25	0.25	0.24	1131	1132	1134	1135	1.2	1.2	1.2	1.2	0.73	0.73	0.73	0.73
PCA		(0.49	(0.49	(0.48	(0.48	(0.23	(0.23	(0.22	(0.22	(0.17	(0.16	(0.16	(0.16	(1115	(1119	(1119	(1123	(0.92	(0.89	(0.88	(0.9	(0.7	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)	0.98)	0.97)	0.99)	0.38)	0.36)	0.36)	0.34)	0.35)	0.32)	0.33)	0.31)	1148)	1148)	1151)	1150)	1.5)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.74	0.74	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884.2	884.4	884.3	884.1	1.1	1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.48	(0.47	(0.48	(0.48	(0.21	(0.2	(0.21	(0.2	(0.14	(0.14	(0.15	(0.14	(877.6	(877.5	(877.5	(877.4	(0.82	(0.82	(0.85	(0.83	(0.68	(0.68	(0.68	(0.68
		- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
				1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.26)	0.26)	0.25)	0.26)	894.6)	896.1)	894.2)	895.9)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.66	0.67	0.28	0.27	0.28	0.29	0.23	0.22	0.22	0.23	882.1	883.9	882.8	881.4	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.41	(0.41	(0.42	(0.42	(0.2	(0.18	(0.2	(0.2	(0.14	(0.13	(0.14	(0.14	(869.9	(872.7	(871.1	(869.6	(0.82	(0.78	(0.82	(0.84	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.91)	0.95)	0.97)	0.35)	0.34)	0.34)	0.35)	0.3)	0.29)	0.3)	0.31)	895.7)	899.1)	896.6)	896.3)	1.4)	1.3)	1.3)	1.4)	0.75)	0.74)	0.75)	0.75)

	CR	0.59	0.59	0.59	0.6	0.3	0.28	0.3	0.31	0.25	0.23	0.24	0.26	880	882.3	879.9	877.7	1.2	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Forw		(0.35	(0.35	(0.35	(0.36	(0.18	(0.18	(0.19	(0.19	(0.13	(0.13	(0.13	(0.15	(863.8	(867.4	(862.6	(860.1	(0.8	(0.79	(0.8	(0.85	(0.68	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.84)	0.85)	0.83)	0.84)	0.38)	0.36)	0.39)	0.4)	0.34)	0.32)	0.34)	0.36)	899)	899.1)	897.7)	898.6)	1.5)	1.4)	1.5)	1.5)	0.76)	0.75)	0.75)	0.76)
	CR	0.62	0.62	0.63	0.63	0.3	0.28	0.3	0.3	0.24	0.23	0.24	0.25	880	882.4	879.1	879.6	1.2	1.1	1.2	1.2	0.73	0.72	0.72	0.72
LASSO		(0.38	(0.38	(0.41	(0.41	(0.21	(0.2	(0.21	(0.2	(0.15	(0.14	(0.14	(0.14	(866.6	(871.2	(866.5	(864.5	(0.85	(0.83	(0.84	(0.83	(0.69	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.87)	0.88)	0.87)	0.37)	0.34)	0.37)	0.38)	0.32)	0.29)	0.32)	0.33)	895.7)	897.3)	895.3)	896.5)	1.4)	1.3)	1.4)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.66	0.67	0.68	0.66	0.29	0.29	0.28	0.27	0.23	0.23	0.23	0.21	881.8	880.5	882.5	884.2	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.71
RSF		(0.43	(0.42	(0.44	(0.42	(0.21	(0.21	(0.21	(0.19	(0.15	(0.15	(0.14	(0.13	(872.2	(871.3	(872.5	(875.1	(0.86	(0.87	(0.82	(0.8	(0.69	(0.68	(0.68	(0.67
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.94)	0.94)	0.94)	0.34)	0.34)	0.34)	0.32)	0.29)	0.29)	0.29)	0.27)	895)	894.7)	895.3)	898)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.65	0.67	0.66	0.66	0.29	0.29	0.28	0.28	0.23	0.23	0.23	0.23	881.3	881	882.2	882.2	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.41	(0.4	(0.4	(0.41	(0.2	(0.2	(0.19	(0.2	(0.14	(0.14	(0.13	(0.15	(868.7	(870	(870.2	(869.9	(0.81	(0.82	(0.8	(0.85	(0.68	(0.69	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.92)	0.94)	0.95)	0.96)	0.36)	0.35)	0.35)	0.35)	0.31)	0.3)	0.3)	0.31)	896.9)	896.2)	897.7)	897.2)	1.4)	1.3)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.12: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 64 bin count, with ComBat feature realignment and all minimum ComBat batch sizes. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.82	0.82	0.82	0.83	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	1445	1444	1445	1445	1.1	1.1	1.1	1.1	0.72	0.72	0.71	0.72
Clin		(0.6	(0.6	(0.59	(0.59	(0.23	(0.23	(0.22	(0.23	(0.16	(0.17	(0.17	(0.16	(1438	(1438	(1438	(1438	(0.91	(0.92	(0.91	(0.91	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.25)	0.25)	0.25)	0.24)	1455)	1455)	1456)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.77	0.76	0.77	0.78	0.28	0.27	0.28	0.28	0.23	0.23	0.23	0.23	1442	1443	1442	1441	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.56	(0.55	(0.55	(0.55	(0.22	(0.22	(0.22	(0.23	(0.17	(0.17	(0.17	(0.17	(1429	(1431	(1430	(1428	(0.91	(0.91	(0.92	(0.93	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.32)	0.32)	0.33)	0.28)	0.28)	0.27)	0.28)	1457)	1457)	1457)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.7	0.69	0.7	0.71	0.28	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1441	1441	1441	1439	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Forw		(0.5	(0.46	(0.48	(0.5	(0.2	(0.21	(0.2	(0.22	(0.15	(0.16	(0.15	(0.16	(1424	(1426	(1426	(1424	(0.88	(0.88	(0.86	(0.89	(0.68	(0.69	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.92)	0.92)	0.94)	0.92)	0.34)	0.33)	0.34)	0.34)	0.3)	0.29)	0.29)	0.3)	1461)	1460)	1461)	1457)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.72	0.73	0.73	0.73	0.28	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1441	1440	1441	1439	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
LASSO		(0.52	(0.51	(0.51	(0.52	(0.21	(0.23	(0.21	(0.23	(0.16	(0.17	(0.16	(0.17	(1427	(1427	(1428	(1425	(0.9	(0.93	(0.9	(0.94	(0.68	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.95)	0.95)	0.94)	0.33)	0.33)	0.33)	0.34)	0.29)	0.28)	0.28)	0.29)	1460)	1455)	1458)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.75)
	CR	0.76	0.77	0.76	0.76	0.29	0.29	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1440	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.55	(0.56	(0.55	(0.54	(0.23	(0.24	(0.22	(0.23	(0.18	(0.19	(0.16	(0.18	(1430	(1428	(1431	(1430	(0.97	(1 -	(0.91	(0.95	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	1.3)	-	-	-	-	-	-
						0.32)	0.33)	0.32)	0.32)	0.28)	0.28)	0.27)	0.27)	1454)	1451)	1457)	1454)	1.3)		1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.77	0.77	0.77	0.77	0.3	0.29	0.29	0.28	0.25	0.24	0.24	0.23	1437	1438	1438	1441	1.2	1.2	1.2	1.1	0.72	0.72	0.72	0.72
PCA		(0.56	(0.55	(0.55	(0.55	(0.24	(0.23	(0.23	(0.23	(0.19	(0.18	(0.18	(0.18	(1425	(1426	(1427	(1430	(0.98	(0.97	(0.95	(0.95	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.34)	0.33)	0.33)	0.32)	0.3)	0.29)	0.29)	0.28)	1452)	1453)	1454)	1455)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.77	0.79	0.79	0.78	0.28	0.28	0.28	0.28	0.21	0.21	0.22	0.21	1139	1138	1139	1139	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.52	(0.52	(0.54	(0.54	(0.22	(0.23	(0.23	(0.22	(0.15	(0.15	(0.16	(0.15	(1132	(1132	(1132	(1132	(0.85	(0.87	(0.89	(0.86	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1151)	1149)	1149)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.72	0.72	0.72	0.29	0.29	0.29	0.3	0.24	0.24	0.24	0.25	1135	1135	1136	1134	1.2	1.1	1.1	1.2	0.73	0.72	0.72	0.73
Back		(0.48	(0.47	(0.51	(0.48	(0.22	(0.22	(0.22	(0.23	(0.16	(0.16	(0.16	(0.16	(1122	(1122	(1123	(1120	(0.88	(0.88	(0.9	(0.9	(0.69	(0.69	(0.69	(0.69
		-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)	0.99)		0.99)	0.35)	0.35)	0.34)	0.36)	0.31)	0.31)	0.3)	0.32)	1151)	1151)	1150)	1149)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.74)	0.75)
	CR	0.66	0.64	0.66	0.66	0.31	0.29	0.3	0.31	0.26	0.24	0.26	0.26	1131	1136	1132	1132	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.39	(0.32	(0.4	(0.45	(0.2	(0.17	(0.21	(0.21	(0.15	(0.14	(0.15	(0.16	(1112	(1119	(1116	(1115	(0.85	(0.81	(0.85	(0.89	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.91)	0.89)	0.9)	0.9)	0.39)	0.36)	0.37)	0.38)	0.35)	0.32)	0.34)	0.34)	1154)	1160)	1153)	1153)	1.5)	1.4)	1.5)	1.5)	0.76)	0.75)	0.75)	0.75)
	CR	0.68	0.68	0.68	0.68	0.31	0.29	0.31	0.3	0.26	0.24	0.26	0.26	1130	1135	1131	1132	1.2	1.2	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.42	(0.42	(0.43	(0.46	(0.22	(0.21	(0.21	(0.22	(0.16	(0.14	(0.16	(0.17	(1115	(1120	(1118	(1117	(0.89	(0.84	(0.89	(0.91	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.92)	0.93)	0.93)	0.38)	0.36)	0.37)	0.37)	0.34)	0.32)	0.34)	0.33)	1151)	1153)	1152)	1151)	1.5)	1.4)	1.5)	1.4)	0.76)	0.75)	0.75)	0.76)
	CR	0.7	0.72	0.72	0.71	0.29	0.29	0.28	0.27	0.24	0.23	0.22	0.21	1136	1136	1139	1139	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.48	(0.5	(0.49	(0.47	(0.22	(0.23	(0.21	(0.21	(0.17	(0.17	(0.15	(0.15	(1126	(1128	(1129	(1130	(0.93	(0.92	(0.86	(0.85	(0.7	(0.69	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.98)	0.98)	0.97)	0.33)	0.33)	0.32)	0.32)	0.3)	0.27)	0.27)	0.26)	1150)	1149)	1152)	1153)	1.3)	1.2)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.73	0.71	0.31	0.31	0.3	0.29	0.26	0.26	0.25	0.24	1132	1131	1132	1135	1.2	1.2	1.2	1.2	0.73	0.73	0.73	0.72
PCA		(0.48	(0.48	(0.5	(0.48	(0.24	(0.23	(0.23	(0.22	(0.18	(0.18	(0.17	(0.16	(1118	(1120	(1120	(1124	(0.94	(0.95	(0.93	(0.91	(0.69	(0.69	(0.69	(0.69
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)		0.99)	0.98)	0.37)	0.36)	0.36)	0.34)	0.33)	0.32)	0.32)	0.31)	1147)	1148)	1148)	1150)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.75	0.75	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884.2	884.4	884.3	884.1	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.49	(0.48	(0.48	(0.48	(0.21	(0.21	(0.21	(0.21	(0.14	(0.14	(0.14	(0.14	(877.3	(877.7	(877.4	(877.3	(0.84	(0.84	(0.82	(0.83	(0.68	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.26)	0.26)	0.26)	894.6)	894.8)	894.9)	894.8)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.67	0.67	0.66	0.67	0.29	0.28	0.28	0.29	0.23	0.22	0.23	0.23	881.8	883.1	882.2	881.7	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.4	(0.43	(0.42	(0.42	(0.2	(0.2	(0.2	(0.2	(0.14	(0.14	(0.14	(0.14	(869.8	(872.1	(869.4	(869.3	(0.83	(0.83	(0.82	(0.83	(0.68	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.98)	0.98)	0.97)	0.35)	0.34)	0.35)	0.35)	0.3)	0.29)	0.31)	0.31)	896.9)	896)	896.2)	896.4)	1.4)	1.3)	1.4)	1.4)	0.75)	0.75)	0.75)	0.75)

	CR	0.59	0.58	0.59	0.61	0.29	0.28	0.3	0.3	0.24	0.22	0.24	0.25	880.4	883.3	879.9	878.4	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.3	(0.31	(0.31	(0.36	(0.17	(0.17	(0.17	(0.2	(0.13	(0.13	(0.13	(0.15	(865.5	(869.8	(864.5	(862.6	(0.79	(0.78	(0.79	(0.85	(0.68	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.85)	0.86)	0.85)	0.89)	0.37)	0.35)	0.38)	0.39)	0.33)	0.31)	0.34)	0.35)	901.7)	901.6)	901.1)	896.6)	1.4)	1.4)	1.5)	1.5)	0.76)	0.75)	0.75)	0.76)
	CR	0.63	0.63	0.63	0.63	0.3	0.29	0.31	0.29	0.25	0.23	0.25	0.24	879	881.9	878.1	880.3	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.38	(0.38	(0.38	(0.39	(0.2	(0.2	(0.21	(0.18	(0.14	(0.14	(0.15	(0.14	(867	(870.4	(864.1	(865.4	(0.83	(0.84	(0.84	(0.81	(0.69	(0.69	(0.7	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.89)	0.91)	0.87)	0.9)	0.36)	0.35)	0.38)	0.37)	0.32)	0.3)	0.34)	0.33)	896.2)	896.1)	895.2)	899)	1.4)	1.4)	1.5)	1.4)	0.76)	0.75)	0.75)	0.76)
	CR	0.66	0.68	0.67	0.67	0.28	0.29	0.28	0.27	0.23	0.23	0.22	0.22	882.1	880.4	883.2	884	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.71
RSF		(0.43	(0.45	(0.44	(0.42	(0.21	(0.22	(0.2	(0.19	(0.15	(0.16	(0.14	(0.13	(872.4	(871.2	(872.8	(874.4	(0.86	(0.89	(0.83	(0.8	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.97)	0.96)	0.95)	0.34)	0.34)	0.34)	0.33)	0.29)	0.29)	0.29)	0.27)	895.5)	894)	896.3)	897.9)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.67	0.68	0.67	0.66	0.29	0.3	0.29	0.29	0.24	0.24	0.24	0.23	880.2	879.3	880.7	881.7	1.2	1.2	1.1	1.1	0.72	0.73	0.73	0.72
PCA		(0.42	(0.41	(0.39	(0.41	(0.21	(0.22	(0.21	(0.2	(0.15	(0.15	(0.15	(0.14	(868.8	(869.1	(870.2	(869.8	(0.87	(0.87	(0.85	(0.81	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)	0.98)	0.95)	0.95)	0.36)	0.35)	0.35)	0.35)	0.31)	0.31)	0.3)	0.31)	894.3)	893.4)	895.7)	896.9)	1.4)	1.4)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.13: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 64 bin count and all minimum ComBat batch sizes but without ComBat feature realignment. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.82	0.83	0.82	0.83	0.27	0.27	0.27	0.27	0.21	0.21	0.22	0.22	1445	1444	1444	1444	1.1	1.1	1.1	1.1	0.72	0.72	0.71	0.72
Clin		(0.59	(0.6	(0.6	(0.58	(0.22	(0.22	(0.22	(0.23	(0.17	(0.16	(0.17	(0.17	(1438	(1438	(1438	(1438	(0.91	(0.91	(0.91	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.25)	0.24)	0.24)	0.25)	1456)	1456)	1456)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.76	0.77	0.76	0.78	0.28	0.27	0.28	0.28	0.23	0.23	0.23	0.23	1442	1443	1442	1441	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.55	(0.55	(0.55	(0.56	(0.22	(0.22	(0.22	(0.23	(0.17	(0.17	(0.17	(0.17	(1429	(1431	(1429	(1429	(0.92	(0.91	(0.92	(0.94	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.32)	0.32)	0.32)	0.32)	0.28)	0.27)	0.28)	0.28)	1457)	1457)	1456)	1454)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.7	0.7	0.7	0.71	0.28	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1441	1442	1441	1439	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Forw		(0.49	(0.49	(0.51	(0.51	(0.2	(0.2	(0.2	(0.22	(0.15	(0.15	(0.16	(0.17	(1424	(1426	(1426	(1424	(0.86	(0.87	(0.89	(0.93	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.91)	0.95)	0.92)	0.93)	0.34)	0.33)	0.33)	0.34)	0.3)	0.29)	0.29)	0.3)	1461)	1461)	1461)	1457)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.73	0.74	0.73	0.73	0.28	0.28	0.28	0.29	0.23	0.23	0.23	0.24	1441	1441	1441	1439	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
LASSO		(0.52	(0.54	(0.52	(0.52	(0.21	(0.21	(0.21	(0.22	(0.16	(0.16	(0.16	(0.17	(1427	(1427	(1429	(1425	(0.91	(0.89	(0.9	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.99)	0.94)	0.94)	0.33)	0.33)	0.33)	0.34)	0.29)	0.28)	0.28)	0.29)	1459)	1459)	1458)	1457)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.76	0.76	0.29	0.29	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1440	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.54	(0.56	(0.56	(0.53	(0.23	(0.24	(0.22	(0.23	(0.18	(0.19	(0.17	(0.18	(1430	(1428	(1432	(1430	(0.96	(0.98	(0.93	(0.94	(0.7	(0.7	(0.69	(0.69
		-	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)				0.32)	0.33)	0.32)	0.32)	0.29)	0.28)	0.27)	0.27)	1455)	1452)	1458)	1455)	1.3)	1.3)	1.2)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.77	0.78	0.76	0.76	0.3	0.29	0.29	0.28	0.25	0.24	0.24	0.24	1436	1437	1438	1441	1.2	1.2	1.2	1.1	0.72	0.72	0.72	0.72
PCA		(0.55	(0.58	(0.55	(0.55	(0.24	(0.23	(0.23	(0.23	(0.18	(0.18	(0.18	(0.17	(1424	(1426	(1426	(1430	(0.97	(0.96	(0.96	(0.94	(0.7	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.34)	0.34)	0.33)	0.32)	0.3)	0.29)	0.29)	0.28)	1452)	1453)	1454)	1456)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.79	0.78	0.79	0.77	0.28	0.28	0.28	0.28	0.22	0.21	0.21	0.21	1138	1139	1139	1139	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.53	(0.51	(0.52	(0.52	(0.23	(0.22	(0.22	(0.22	(0.16	(0.15	(0.15	(0.16	(1132	(1132	(1132	(1132	(0.88	(0.87	(0.87	(0.88	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1148)	1150)	1150)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.73	0.72	0.3	0.29	0.29	0.3	0.24	0.24	0.24	0.25	1134	1136	1136	1134	1.2	1.1	1.1	1.2	0.73	0.72	0.72	0.73
Back		(0.47	(0.46	(0.5	(0.49	(0.23	(0.22	(0.22	(0.23	(0.17	(0.16	(0.15	(0.17	(1122	(1123	(1122	(1120	(0.91	(0.89	(0.86	(0.91	(0.7	(0.69	(0.69	(0.69
		-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)	0.99)		0.98)	0.35)	0.34)	0.35)	0.36)	0.32)	0.31)	0.31)	0.32)	1148)	1150)	1151)	1148)	1.4)	1.4)	1.4)	1.4)	0.75)	0.74)	0.75)	0.75)
	CR	0.66	0.64	0.66	0.66	0.31	0.29	0.3	0.3	0.26	0.24	0.26	0.26	1131	1136	1132	1132	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.41	(0.35	(0.39	(0.45	(0.2	(0.17	(0.19	(0.22	(0.15	(0.13	(0.15	(0.16	(1111	(1119	(1117	(1116	(0.86	(0.81	(0.85	(0.9	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.9)	0.89)	0.9)	0.89)	0.39)	0.36)	0.37)	0.37)	0.35)	0.32)	0.34)	0.33)	1155)	1160)	1156)	1151)	1.5)	1.4)	1.5)	1.4)	0.76)	0.75)	0.75)	0.76)
	CR	0.69	0.68	0.69	0.68	0.32	0.3	0.31	0.3	0.27	0.25	0.26	0.26	1129	1134	1131	1133	1.2	1.2	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.43	(0.42	(0.45	(0.46	(0.22	(0.19	(0.22	(0.23	(0.16	(0.15	(0.16	(0.17	(1112	(1120	(1117	(1117	(0.89	(0.87	(0.89	(0.92	(0.7	(0.69	(0.7	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.94)	0.93)	0.93)	0.39)	0.36)	0.37)	0.37)	0.35)	0.32)	0.34)	0.33)	1151)	1156)	1151)	1149)	1.5)	1.4)	1.5)	1.4)	0.76)	0.75)	0.75)	0.75)
	CR	0.71	0.71	0.73	0.7	0.29	0.29	0.28	0.27	0.24	0.23	0.22	0.21	1135	1137	1139	1139	1.2	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.49	(0.47	(0.48	(0.47	(0.23	(0.22	(0.21	(0.21	(0.17	(0.16	(0.15	(0.15	(1125	(1128	(1130	(1131	(0.93	(0.9	(0.86	(0.86	(0.7	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)	0.99)	0.98)	0.98)	0.33)	0.32)	0.32)	0.31)	0.3)	0.27)	0.26)	0.26)	1149)	1150)	1152)	1153)	1.4)	1.2)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.73	0.71	0.31	0.31	0.3	0.29	0.26	0.25	0.25	0.24	1132	1132	1133	1135	1.2	1.2	1.2	1.2	0.73	0.73	0.73	0.72
PCA		(0.47	(0.48	(0.48	(0.46	(0.24	(0.24	(0.23	(0.22	(0.18	(0.17	(0.17	(0.16	(1117	(1121	(1120	(1124	(0.95	(0.93	(0.93	(0.89	(0.7	(0.7	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.98)	0.98)	0.97)	0.37)	0.36)	0.36)	0.34)	0.33)	0.32)	0.32)	0.3)	1147)	1147)	1149)	1150)	1.4)	1.4)	1.4)	1.3)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.74	0.74	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	883.9	883.9	884.1	884	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.47	(0.48	(0.48	(0.48	(0.22	(0.21	(0.21	(0.21	(0.15	(0.14	(0.14	(0.14	(877.5	(877	(877.5	(877.3	(0.86	(0.84	(0.84	(0.83	(0.69	(0.69	(0.69	(0.68
		-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)		1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.26)	0.26)	0.26)	0.26)	893.9)	894.2)	894.7)	894.8)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.66	0.67	0.29	0.28	0.28	0.29	0.23	0.22	0.23	0.23	881.3	882.8	882.2	881.6	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.41	(0.41	(0.42	(0.42	(0.2	(0.2	(0.2	(0.19	(0.14	(0.14	(0.15	(0.14	(868.4	(871.5	(870.7	(870.2	(0.83	(0.82	(0.85	(0.83	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.95)	0.96)	0.95)	0.36)	0.34)	0.35)	0.35)	0.31)	0.3)	0.3)	0.31)	897)	897)	896.3)	897.5)	1.4)	1.3)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

	CR	0.59	0.58	0.59	0.6	0.29	0.28	0.3	0.3	0.24	0.23	0.24	0.25	880.5	883	880	878.2	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
Forw		(0.31	(0.29	(0.31	(0.36	(0.17	(0.16	(0.16	(0.19	(0.13	(0.12	(0.13	(0.13	(865.9	(869.2	(863.6	(861.4	(0.81	(0.76	(0.79	(0.8	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.86)	0.87)	0.86)	0.86)	0.37)	0.35)	0.38)	0.39)	0.33)	0.31)	0.34)	0.35)	901.2)	903.5)	902.3)	898.8)	1.4)	1.4)	1.5)	1.5)	0.76)	0.75)	0.75)	0.76)
	CR	0.62	0.63	0.62	0.62	0.3	0.29	0.3	0.29	0.25	0.23	0.25	0.24	879.2	881.5	878.5	880.3	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.36	(0.4	(0.37	(0.4	(0.2	(0.18	(0.2	(0.2	(0.14	(0.14	(0.14	(0.15	(866.8	(870.1	(864.8	(866.2	(0.84	(0.84	(0.82	(0.85	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.91)	0.91)	0.9)	0.89)	0.37)	0.35)	0.38)	0.37)	0.33)	0.3)	0.33)	0.33)	896.5)	899.3)	896.4)	897)	1.4)	1.3)	1.4)	1.4)	0.76)	0.75)	0.75)	0.76)
	CR	0.66	0.68	0.67	0.66	0.28	0.3	0.28	0.27	0.23	0.23	0.22	0.22	882	880.1	882.9	883.9	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.71
RSF		(0.42	(0.43	(0.43	(0.43	(0.2	(0.23	(0.21	(0.19	(0.15	(0.17	(0.15	(0.13	(872.6	(871.4	(873.2	(874.4	(0.87	(0.91	(0.85	(0.8	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.95)	0.97)	0.94)	0.34)	0.34)	0.33)	0.33)	0.29)	0.29)	0.28)	0.28)	896.3)	892.3)	895.2)	898.5)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.67	0.68	0.66	0.66	0.3	0.3	0.29	0.29	0.24	0.24	0.24	0.23	880.1	879.3	880.8	881.8	1.2	1.2	1.1	1.1	0.73	0.73	0.72	0.72
PCA		(0.4	(0.43	(0.41	(0.41	(0.21	(0.22	(0.21	(0.2	(0.15	(0.15	(0.15	(0.15	(869.2	(868.6	(870.6	(870.2	(0.86	(0.87	(0.84	(0.85	(0.69	(0.69	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.97)	0.96)	0.94)	0.35)	0.36)	0.35)	0.35)	0.31)	0.31)	0.3)	0.31)	895.2)	893.7)	895.5)	896.5)	1.4)	1.4)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.14: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 128 bin count, with ComBat realignment and all minimum ComBat batch sizes. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.82	0.82	0.83	0.82	0.27	0.27	0.27	0.27	0.22	0.21	0.22	0.22	1444	1444	1444	1445	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.6	(0.6	(0.58	(0.61	(0.23	(0.23	(0.23	(0.23	(0.17	(0.17	(0.17	(0.17	(1438	(1438	(1438	(1438	(0.93	(0.92	(0.92	(0.92	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.24)	0.25)	0.25)	0.25)	1454)	1455)	1455)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.76	0.76	0.77	0.76	0.27	0.27	0.27	0.28	0.23	0.22	0.23	0.23	1443	1443	1443	1441	1.1	1.1	1.1	1.1	0.72	0.71	0.72	0.72
Back		(0.55	(0.55	(0.55	(0.55	(0.22	(0.22	(0.22	(0.22	(0.17	(0.16	(0.17	(0.17	(1432	(1433	(1433	(1429	(0.93	(0.89	(0.92	(0.93	(0.69	(0.68	(0.69	(0.69
		- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
					0.99)	0.31)	0.31)	0.31)	0.33)	0.27)	0.27)	0.27)	0.28)	1457)	1458)	1456)	1456)	1.3)	1.2)	1.2)	1.3)	0.74)	0.73)	0.74)	0.74)
	CR	0.71	0.7	0.7	0.7	0.27	0.27	0.27	0.28	0.22	0.23	0.23	0.23	1444	1443	1443	1440	1.1	1.1	1.1	1.1	0.71	0.72	0.71	0.72
Forw		(0.5	(0.48	(0.49	(0.48	(0.2	(0.2	(0.2	(0.2	(0.15	(0.15	(0.14	(0.16	(1431	(1429	(1429	(1425	(0.87	(0.88	(0.84	(0.9	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.93)	0.93)	0.93)	0.32)	0.32)	0.33)	0.34)	0.27)	0.28)	0.28)	0.29)	1462)	1462)	1461)	1461)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.73	0.73	0.73	0.72	0.27	0.28	0.27	0.29	0.23	0.23	0.23	0.24	1443	1442	1443	1439	1.1	1.1	1.1	1.1	0.71	0.72	0.71	0.72
LASSO		(0.53	(0.53	(0.53	(0.51	(0.21	(0.22	(0.22	(0.22	(0.16	(0.17	(0.17	(0.17	(1432	(1431	(1432	(1426	(0.91	(0.93	(0.93	(0.94	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.98)	0.97)	0.93)	0.31)	0.32)	0.31)	0.34)	0.27)	0.27)	0.27)	0.29)	1458)	1457)	1457)	1457)	1.2)	1.3)	1.2)	1.3)	0.74)	0.74)	0.73)	0.75)
	CR	0.76	0.77	0.77	0.76	0.28	0.3	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1441	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.55	(0.57	(0.55	(0.54	(0.23	(0.24	(0.22	(0.23	(0.18	(0.2	(0.17	(0.17	(1431	(1428	(1432	(1430	(0.97	(1 -	(0.93	(0.94	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.3)	-	-	-	-	-	-
					0.99)	0.32)	0.33)	0.31)	0.32)	0.28)	0.28)	0.27)	0.27)	1453)	1451)	1456)	1455)	1.3)		1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.76	0.77	0.76	0.75	0.29	0.29	0.29	0.28	0.25	0.24	0.24	0.23	1437	1439	1440	1442	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.56	(0.56	(0.54	(0.55	(0.24	(0.23	(0.23	(0.22	(0.18	(0.18	(0.18	(0.17	(1426	(1427	(1427	(1431	(0.97	(0.95	(0.94	(0.93	(0.7	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
					0.99)	0.34)	0.33)	0.33)	0.32)	0.3)	0.29)	0.29)	0.28)	1452)	1455)	1454)	1456)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.77	0.78	0.79	0.78	0.28	0.28	0.28	0.28	0.21	0.21	0.22	0.21	1139	1139	1138	1139	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.52	(0.53	(0.54	(0.53	(0.22	(0.23	(0.23	(0.22	(0.15	(0.15	(0.16	(0.15	(1132	(1132	(1132	(1132	(0.85	(0.87	(0.89	(0.86	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1151)	1149)	1149)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.71	0.72	0.72	0.29	0.29	0.29	0.3	0.24	0.23	0.24	0.25	1136	1137	1136	1134	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.47	(0.47	(0.5	(0.48	(0.22	(0.22	(0.22	(0.23	(0.15	(0.15	(0.16	(0.16	(1123	(1124	(1124	(1119	(0.84	(0.87	(0.88	(0.89	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.98)	0.98)	0.98)	0.35)	0.34)	0.34)	0.36)	0.31)	0.3)	0.3)	0.32)	1151)	1151)	1151)	1149)	1.4)	1.3)	1.3)	1.4)	0.75)	0.74)	0.74)	0.75)
	CR	0.65	0.64	0.65	0.66	0.29	0.28	0.29	0.31	0.24	0.23	0.24	0.26	1135	1138	1135	1131	1.2	1.1	1.2	1.2	0.72	0.72	0.72	0.73
Forw		(0.42	(0.4	(0.4	(0.44	(0.2	(0.19	(0.2	(0.21	(0.14	(0.14	(0.13	(0.16	(1121	(1124	(1120	(1114	(0.83	(0.83	(0.8	(0.88	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.9)	0.88)	0.89)	0.35)	0.34)	0.36)	0.38)	0.31)	0.3)	0.32)	0.33)	1155)	1156)	1155)	1152)	1.4)	1.3)	1.4)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.68	0.68	0.68	0.69	0.3	0.28	0.3	0.32	0.25	0.23	0.24	0.26	1134	1137	1134	1130	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.44	(0.43	(0.46	(0.46	(0.21	(0.21	(0.22	(0.22	(0.15	(0.15	(0.15	(0.16	(1122	(1126	(1122	(1115	(0.86	(0.86	(0.86	(0.9	(0.69	(0.68	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.92)	0.93)	0.91)	0.91)	0.35)	0.33)	0.35)	0.38)	0.31)	0.29)	0.31)	0.33)	1152)	1153)	1151)	1150)	1.4)	1.3)	1.4)	1.4)	0.75)	0.74)	0.75)	0.76)
	CR	0.7	0.72	0.72	0.7	0.29	0.29	0.27	0.27	0.24	0.23	0.22	0.21	1136	1136	1139	1140	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.48	(0.5	(0.49	(0.47	(0.22	(0.23	(0.21	(0.21	(0.16	(0.17	(0.15	(0.15	(1126	(1127	(1129	(1130	(0.91	(0.93	(0.86	(0.85	(0.69	(0.69	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.99)	0.99)	0.97)	0.33)	0.33)	0.32)	0.32)	0.3)	0.27)	0.27)	0.26)	1150)	1149)	1153)	1153)	1.3)	1.2)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.72	0.72	0.7	0.31	0.3	0.3	0.29	0.26	0.25	0.25	0.24	1132	1133	1134	1136	1.2	1.2	1.2	1.1	0.73	0.73	0.72	0.72
PCA		(0.48	(0.49	(0.48	(0.47	(0.23	(0.23	(0.23	(0.21	(0.17	(0.17	(0.17	(0.16	(1117	(1121	(1119	(1125	(0.93	(0.93	(0.92	(0.88	(0.7	(0.7	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.99)	0.99)	0.96)	0.37)	0.35)	0.36)	0.34)	0.33)	0.32)	0.33)	0.3)	1148)	1148)	1149)	1152)	1.4)	1.4)	1.4)	1.3)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.75	0.75	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884.3	884.4	884.4	884.2	1.1	1.1	1	1.1	0.72	0.72	0.72	0.72
Clin		(0.48	(0.49	(0.49	(0.49	(0.21	(0.21	(0.21	(0.21	(0.14	(0.15	(0.14	(0.14	(877.6	(877.7	(877.4	(877.5	(0.84	(0.87	(0.82	(0.83	(0.69	(0.68	(0.68	(0.68
		-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)		1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.26)	0.26)	0.26)	895.4)	895)	895.7)	894.9)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.67	0.66	0.67	0.28	0.27	0.28	0.29	0.23	0.22	0.22	0.23	882.6	883.8	882.8	881.4	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.42	(0.43	(0.43	(0.41	(0.2	(0.19	(0.2	(0.2	(0.14	(0.14	(0.14	(0.14	(871	(873.1	(871.2	(869.1	(0.83	(0.82	(0.83	(0.84	(0.68	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.97)	0.96)	0.95)	0.34)	0.33)	0.34)	0.35)	0.3)	0.3)	0.3)	0.31)	897)	897.3)	896.6)	896.8)	1.3)	1.3)	1.3)	1.4)	0.75)	0.75)	0.75)	0.75)

	CR	0.6	0.59	0.6	0.6	0.29	0.28	0.29	0.3	0.24	0.22	0.24	0.25	881.5	883.2	880.3	879	1.1	1.1	1.2	1.2	0.72	0.72	0.72	0.73
Forw		(0.35	(0.35	(0.36	(0.37	(0.18	(0.17	(0.2	(0.19	(0.13	(0.11	(0.14	(0.14	(868	(868.4	(866	(864.7	(0.8	(0.74	(0.81	(0.82	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.88)	0.87)	0.86)	0.89)	0.36)	0.36)	0.37)	0.38)	0.32)	0.31)	0.33)	0.33)	899.4)	900.9)	896.9)	898.8)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.63	0.63	0.63	0.63	0.29	0.28	0.3	0.3	0.24	0.23	0.25	0.25	881	882.5	878.4	878.7	1.1	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.39	(0.39	(0.4	(0.4	(0.2	(0.19	(0.22	(0.21	(0.14	(0.13	(0.15	(0.16	(869.1	(871.1	(865.9	(866.4	(0.84	(0.8	(0.87	(0.89	(0.69	(0.68	(0.69	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.91)	0.9)	0.91)	0.9)	0.35)	0.34)	0.37)	0.37)	0.32)	0.3)	0.33)	0.32)	897)	897.9)	893.8)	894.5)	1.4)	1.4)	1.4)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.66	0.68	0.67	0.67	0.28	0.29	0.28	0.27	0.23	0.24	0.22	0.21	882.4	880.3	883.7	884.4	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.71
RSF		(0.42	(0.45	(0.43	(0.41	(0.2	(0.22	(0.2	(0.19	(0.15	(0.16	(0.14	(0.13	(873	(871.2	(873.9	(874.3	(0.85	(0.89	(0.82	(0.79	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.96)	0.94)	0.95)	0.33)	0.34)	0.33)	0.33)	0.29)	0.29)	0.28)	0.27)	896.1)	893.5)	897)	898.8)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.67	0.67	0.66	0.66	0.29	0.29	0.29	0.28	0.24	0.24	0.23	0.23	880.8	880.7	881.4	883	1.2	1.1	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.4	(0.42	(0.4	(0.39	(0.21	(0.21	(0.21	(0.19	(0.15	(0.15	(0.14	(0.14	(869.7	(869.8	(870.3	(871.3	(0.85	(0.86	(0.83	(0.82	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.95)	0.93)	0.94)	0.35)	0.35)	0.35)	0.34)	0.31)	0.31)	0.3)	0.3)	895.2)	895.5)	895.1)	897.8)	1.4)	1.4)	1.4)	1.3)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

Table S4.15: Mean and 95% confidence intervals of the performance measures are derived across the 1000 bootstrap repetitions in the 'test' sample (ie. data withheld from bootstrap resample, and not used to build initial/training model). Results are shown for 128 bin count and all minimum ComBat batch sizes but without ComBat feature realignment. The models shown here are the clinical only and combined radiomics + clinical models, built using five different feature selection processes to select the radiomic features

		Calibration				Relative Explained Variation								Relative Model Fit				Discrimination							
		Calibration Slope				Nagelkerke's R^2				Royston Sauerbrei's R^2				Akaike's Information Criterion				Royston Sauerbrei's D				Concordance Index			
Feats[a]	Model[b]	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW	ZS	WS	HM	RAW
5 patients per batch																									
	C	0.83	0.82	0.82	0.82	0.27	0.27	0.27	0.27	0.21	0.21	0.22	0.21	1445	1444	1444	1444	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.6	(0.59	(0.6	(0.59	(0.23	(0.22	(0.22	(0.23	(0.17	(0.17	(0.17	(0.17	(1438	(1438	(1438	(1438	(0.91	(0.92	(0.92	(0.91	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.29)	0.29)	0.29)	0.29)	0.24)	0.24)	0.25)	0.25)	1455)	1456)	1456)	1455)	1.2)	1.2)	1.2)	1.2)	0.73)	0.73)	0.73)	0.73)
	CR	0.77	0.76	0.76	0.77	0.27	0.27	0.27	0.28	0.23	0.22	0.22	0.23	1443	1443	1443	1441	1.1	1.1	1.1	1.1	0.72	0.71	0.72	0.72
Back		(0.55	(0.54	(0.55	(0.55	(0.22	(0.22	(0.22	(0.23	(0.16	(0.16	(0.17	(0.18	(1432	(1433	(1433	(1429	(0.91	(0.9	(0.92	(0.94	(0.69	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.31)	0.31)	0.31)	0.32)	0.27)	0.27)	0.27)	0.28)	1457)	1458)	1457)	1455)	1.3)	1.2)	1.2)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.7	0.71	0.7	0.27	0.27	0.28	0.28	0.22	0.23	0.23	0.23	1444	1443	1442	1441	1.1	1.1	1.1	1.1	0.71	0.72	0.71	0.72
Forw		(0.5	(0.49	(0.5	(0.47	(0.2	(0.2	(0.21	(0.2	(0.15	(0.15	(0.16	(0.16	(1430	(1429	(1427	(1426	(0.87	(0.87	(0.89	(0.9	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.92)	0.94)	0.91)	0.32)	0.32)	0.33)	0.34)	0.28)	0.28)	0.28)	0.29)	1462)	1461)	1460)	1461)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.74	0.73	0.73	0.73	0.27	0.27	0.28	0.29	0.22	0.23	0.23	0.24	1444	1443	1442	1440	1.1	1.1	1.1	1.1	0.72	0.72	0.71	0.72
LASSO		(0.52	(0.5	(0.52	(0.51	(0.21	(0.2	(0.21	(0.22	(0.16	(0.16	(0.16	(0.17	(1432	(1431	(1431	(1426	(0.88	(0.89	(0.9	(0.94	(0.69	(0.69	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.98)	0.99)	0.95)	0.94)	0.31)	0.32)	0.32)	0.33)	0.27)	0.27)	0.27)	0.28)	1459)	1461)	1459)	1456)	1.2)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
	CR	0.77	0.77	0.77	0.76	0.28	0.29	0.28	0.28	0.24	0.25	0.23	0.23	1440	1437	1442	1441	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
RSF		(0.57	(0.56	(0.56	(0.54	(0.23	(0.24	(0.22	(0.23	(0.18	(0.19	(0.17	(0.18	(1430	(1428	(1432	(1431	(0.97	(1 -	(0.93	(0.95	(0.7	(0.7	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	1.3)	-	-	-	-	-	-
						0.32)	0.33)	0.32)	0.32)	0.29)	0.28)	0.27)	0.27)	1454)	1451)	1457)	1454)	1.3)		1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.77	0.76	0.76	0.76	0.29	0.29	0.29	0.28	0.25	0.24	0.24	0.23	1438	1439	1440	1442	1.2	1.2	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.56	(0.53	(0.55	(0.53	(0.24	(0.22	(0.23	(0.22	(0.18	(0.17	(0.17	(0.17	(1426	(1428	(1428	(1430	(0.97	(0.94	(0.93	(0.94	(0.7	(0.69	(0.69	(0.69
		- 1)	- 1)	- 1)	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
						0.33)	0.33)	0.33)	0.32)	0.3)	0.29)	0.29)	0.28)	1453)	1457)	1455)	1457)	1.3)	1.3)	1.3)	1.3)	0.74)	0.74)	0.74)	0.74)
10 patients per batch																									

	C	0.78	0.78	0.78	0.78	0.28	0.27	0.28	0.28	0.21	0.21	0.21	0.21	1138	1139	1139	1139	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.53	(0.52	(0.53	(0.52	(0.23	(0.22	(0.23	(0.22	(0.15	(0.14	(0.15	(0.15	(1132	(1132	(1132	(1132	(0.88	(0.84	(0.87	(0.87	(0.69	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		1.1)	1.1)	1.1)	1.1)	0.31)	0.31)	0.31)	0.31)	0.25)	0.25)	0.25)	0.25)	1149)	1151)	1149)	1150)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.71	0.72	0.71	0.72	0.29	0.28	0.29	0.3	0.24	0.23	0.24	0.25	1135	1137	1136	1133	1.1	1.1	1.1	1.2	0.72	0.72	0.72	0.73
Back		(0.48	(0.48	(0.49	(0.48	(0.22	(0.22	(0.22	(0.23	(0.16	(0.16	(0.15	(0.17	(1124	(1126	(1125	(1119	(0.88	(0.88	(0.87	(0.92	(0.69	(0.69	(0.69	(0.7
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.98)	0.99)	0.98)	0.34)	0.33)	0.34)	0.36)	0.3)	0.3)	0.3)	0.32)	1151)	1151)	1150)	1149)	1.3)	1.3)	1.3)	1.4)	0.75)	0.74)	0.74)	0.75)
	CR	0.66	0.64	0.64	0.66	0.29	0.28	0.29	0.31	0.24	0.23	0.24	0.26	1135	1138	1135	1131	1.2	1.1	1.2	1.2	0.72	0.72	0.72	0.73
Forw		(0.41	(0.4	(0.43	(0.44	(0.2	(0.18	(0.2	(0.21	(0.14	(0.14	(0.14	(0.16	(1120	(1123	(1120	(1115	(0.82	(0.81	(0.84	(0.91	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.91)	0.88)	0.88)	0.89)	0.36)	0.34)	0.36)	0.38)	0.31)	0.29)	0.32)	0.34)	1155)	1158)	1154)	1152)	1.4)	1.3)	1.4)	1.5)	0.75)	0.74)	0.75)	0.76)
	CR	0.68	0.68	0.67	0.69	0.3	0.28	0.29	0.32	0.24	0.23	0.24	0.27	1134	1137	1134	1130	1.2	1.1	1.2	1.2	0.73	0.72	0.72	0.73
LASSO		(0.45	(0.45	(0.46	(0.48	(0.22	(0.2	(0.22	(0.22	(0.15	(0.15	(0.16	(0.18	(1122	(1126	(1123	(1115	(0.88	(0.85	(0.88	(0.96	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.93)	0.92)	0.9)	0.94)	0.35)	0.33)	0.35)	0.38)	0.31)	0.29)	0.3)	0.33)	1151)	1154)	1151)	1150)	1.4)	1.3)	1.4)	1.4)	0.75)	0.74)	0.75)	0.76)
	CR	0.71	0.72	0.71	0.7	0.29	0.28	0.28	0.27	0.24	0.23	0.22	0.21	1136	1137	1139	1140	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.72
RSF		(0.48	(0.48	(0.5	(0.47	(0.22	(0.22	(0.21	(0.21	(0.17	(0.16	(0.15	(0.14	(1125	(1127	(1129	(1130	(0.92	(0.89	(0.86	(0.83	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.97)	0.98)	0.97)	0.99)	0.34)	0.33)	0.32)	0.31)	0.3)	0.27)	0.27)	0.26)	1150)	1151)	1152)	1153)	1.3)	1.2)	1.2)	1.2)	0.75)	0.74)	0.74)	0.74)
	CR	0.72	0.71	0.71	0.71	0.31	0.3	0.3	0.29	0.26	0.25	0.25	0.24	1131	1134	1134	1135	1.2	1.2	1.2	1.2	0.73	0.72	0.73	0.72
PCA		(0.49	(0.49	(0.48	(0.48	(0.23	(0.22	(0.22	(0.22	(0.17	(0.16	(0.16	(0.16	(1117	(1120	(1121	(1125	(0.93	(0.88	(0.89	(0.9	(0.7	(0.69	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.99)	0.96)	0.96)	0.98)	0.37)	0.36)	0.35)	0.34)	0.33)	0.32)	0.32)	0.3)	1149)	1150)	1150)	1150)	1.4)	1.4)	1.4)	1.3)	0.75)	0.75)	0.75)	0.75)
15 patients per batch																									
	C	0.75	0.74	0.74	0.74	0.27	0.27	0.27	0.27	0.21	0.21	0.21	0.21	884	884.4	883.9	884.2	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Clin		(0.49	(0.47	(0.48	(0.48	(0.21	(0.21	(0.21	(0.21	(0.14	(0.14	(0.14	(0.15	(877.4	(877.6	(877.6	(877.4	(0.84	(0.83	(0.84	(0.84	(0.68	(0.68	(0.69	(0.69
		- 1)	-	-	- 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			1.1)	1.1)		0.31)	0.31)	0.31)	0.31)	0.26)	0.26)	0.25)	0.26)	894.4)	895.7)	894.5)	894.6)	1.2)	1.2)	1.2)	1.2)	0.74)	0.74)	0.74)	0.74)
	CR	0.66	0.66	0.66	0.67	0.28	0.27	0.28	0.29	0.23	0.22	0.23	0.23	882.5	883.8	882.6	881.6	1.1	1.1	1.1	1.1	0.72	0.72	0.72	0.72
Back		(0.41	(0.42	(0.43	(0.41	(0.19	(0.18	(0.2	(0.2	(0.14	(0.13	(0.14	(0.14	(870.5	(873.2	(871.1	(869.5	(0.83	(0.8	(0.81	(0.83	(0.68	(0.68	(0.68	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.94)	0.94)	0.96)	0.93)	0.35)	0.33)	0.34)	0.35)	0.3)	0.29)	0.3)	0.31)	897.3)	899.1)	896.1)	896.9)	1.4)	1.3)	1.3)	1.4)	0.75)	0.74)	0.75)	0.75)

	CR	0.61	0.58	0.6	0.6	0.29	0.28	0.3	0.3	0.24	0.22	0.24	0.25	881.1	883.4	880	879.5	1.1	1.1	1.2	1.2	0.72	0.72	0.72	0.73
Forw		(0.37	(0.34	(0.38	(0.35	(0.18	(0.16	(0.2	(0.19	(0.13	(0.12	(0.13	(0.14	(866.2	(868.9	(865.9	(865	(0.79	(0.75	(0.8	(0.82	(0.68	(0.67	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.87)	0.87)	0.85)	0.86)	0.37)	0.36)	0.37)	0.37)	0.33)	0.31)	0.33)	0.33)	899.7)	902.8)	897.3)	897.4)	1.4)	1.4)	1.5)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.63	0.63	0.64	0.63	0.29	0.28	0.31	0.3	0.24	0.23	0.25	0.25	880.5	882.4	878.2	879	1.2	1.1	1.2	1.2	0.73	0.72	0.73	0.73
LASSO		(0.39	(0.39	(0.39	(0.41	(0.2	(0.19	(0.21	(0.2	(0.15	(0.14	(0.15	(0.15	(868.3	(871	(866	(866.2	(0.86	(0.81	(0.84	(0.86	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.9)	0.9)	0.92)	0.91)	0.36)	0.34)	0.37)	0.37)	0.32)	0.3)	0.33)	0.32)	897)	897.3)	895.3)	896.9)	1.4)	1.3)	1.5)	1.4)	0.75)	0.75)	0.75)	0.76)
	CR	0.66	0.68	0.67	0.67	0.29	0.29	0.28	0.27	0.23	0.23	0.22	0.21	881.9	880.5	883.2	884.3	1.1	1.1	1.1	1.1	0.73	0.72	0.72	0.71
RSF		(0.43	(0.42	(0.41	(0.43	(0.21	(0.21	(0.21	(0.19	(0.16	(0.15	(0.14	(0.13	(872	(871.1	(873.6	(874.8	(0.88	(0.88	(0.83	(0.79	(0.69	(0.68	(0.68	(0.68
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.95)	0.96)	0.95)	0.92)	0.34)	0.34)	0.33)	0.32)	0.29)	0.29)	0.28)	0.27)	895.6)	895.3)	895.5)	897.9)	1.3)	1.3)	1.3)	1.3)	0.75)	0.74)	0.74)	0.74)
	CR	0.67	0.67	0.67	0.66	0.29	0.29	0.29	0.28	0.24	0.24	0.24	0.23	880.5	881	881.1	882.6	1.2	1.1	1.1	1.1	0.72	0.72	0.72	0.72
PCA		(0.41	(0.42	(0.4	(0.4	(0.2	(0.2	(0.21	(0.19	(0.14	(0.15	(0.15	(0.14	(868.8	(869.7	(870.7	(871	(0.82	(0.85	(0.85	(0.82	(0.69	(0.68	(0.69	(0.69
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		0.96)	0.97)	0.95)	0.93)	0.36)	0.35)	0.35)	0.34)	0.32)	0.31)	0.31)	0.3)	896.5)	895.9)	894.5)	897.9)	1.4)	1.4)	1.4)	1.3)	0.75)	0.75)	0.75)	0.75)

Calibration slopes closer to 1 indicates better calibration. Relative explained variation ranges from 0 to 1; higher values are better. Lower relative model fit indicates a better performing model. Royston and Sauerbrei's *D* statistic indicates better discrimination as the value moves away from 0. C-index ranges from 0.5 to 1; 0.5 indicating no and 1 indicating perfect discrimination. Back = backwards stepwise feature elimination, Clin = clinical features only, CI = confindence interval, Forw = forwards stepwise feature selection, HM = histogram matching, LASSO = Least Absolute Shrinkage and Selection Operator, PCA = principle component analysis (with clustering of results), RSF = random survival forests, RAW = no intensity standardisation prior to radiomic extraction, WS = WhiteStripe standardisation, ZS = z-score intensity standardisation

[a] Maximum of four radiomic features selected with the chosen method [b] Clinical features only or a combination of both clinical and radiomic features in the Cox proportional hazards model. C = Clinical only, CR = combined clinical-radiomics model.

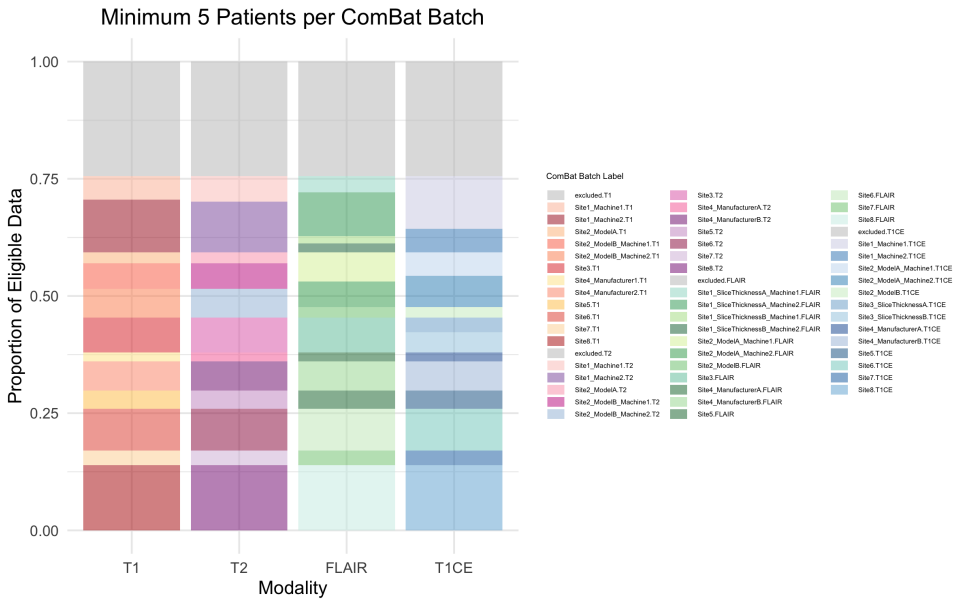

Figure S4.1: Stacked bar charts demonstrating the different ComBat batch labels per MRI sequence, for minimum batch size = 5. Each bar represents a different MRI sequence. Each segment of a bar represents a unique batch label (see key for details).

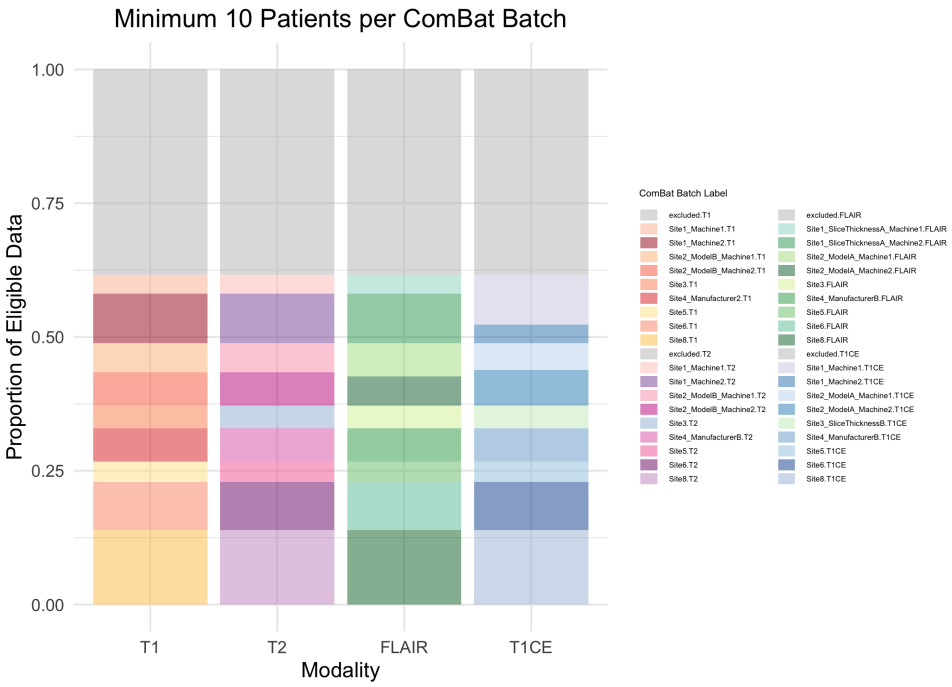

Figure S4.2: Stacked bar charts demonstrating the different ComBat batch labels per MRI sequence, for minimum batch size = 10. Each bar represents a different MRI sequence. Each segment of a bar represents a unique batch label (see key for details).

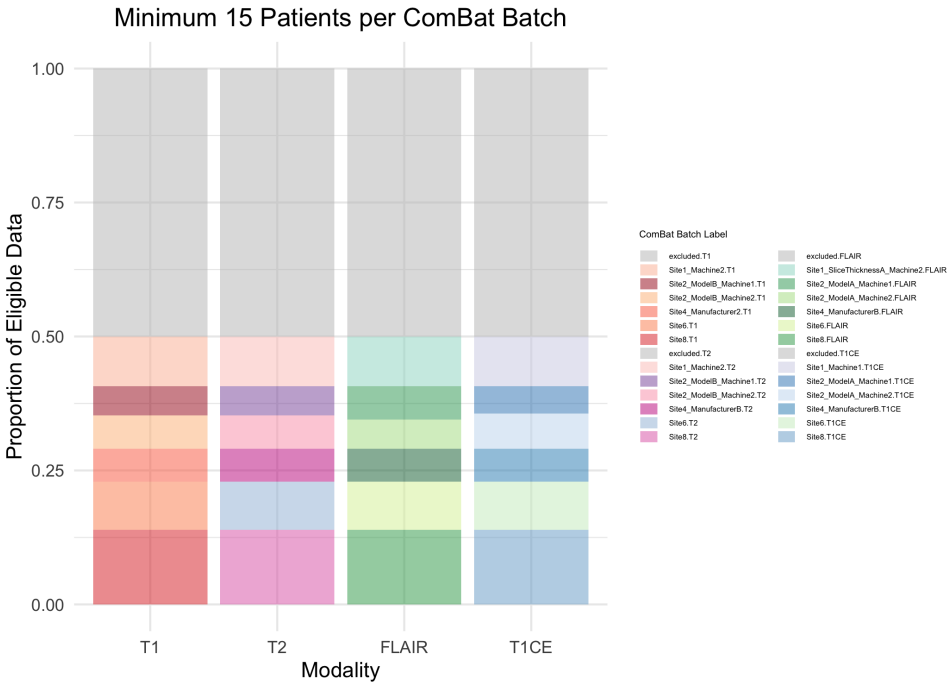

Figure S4.3: Stacked bar charts demonstrating the different ComBat batch labels per MRI sequence, for minimum batch size = 15. Each bar represents a different MRI sequence. Each segment of a bar represents a unique batch label (see key for details).

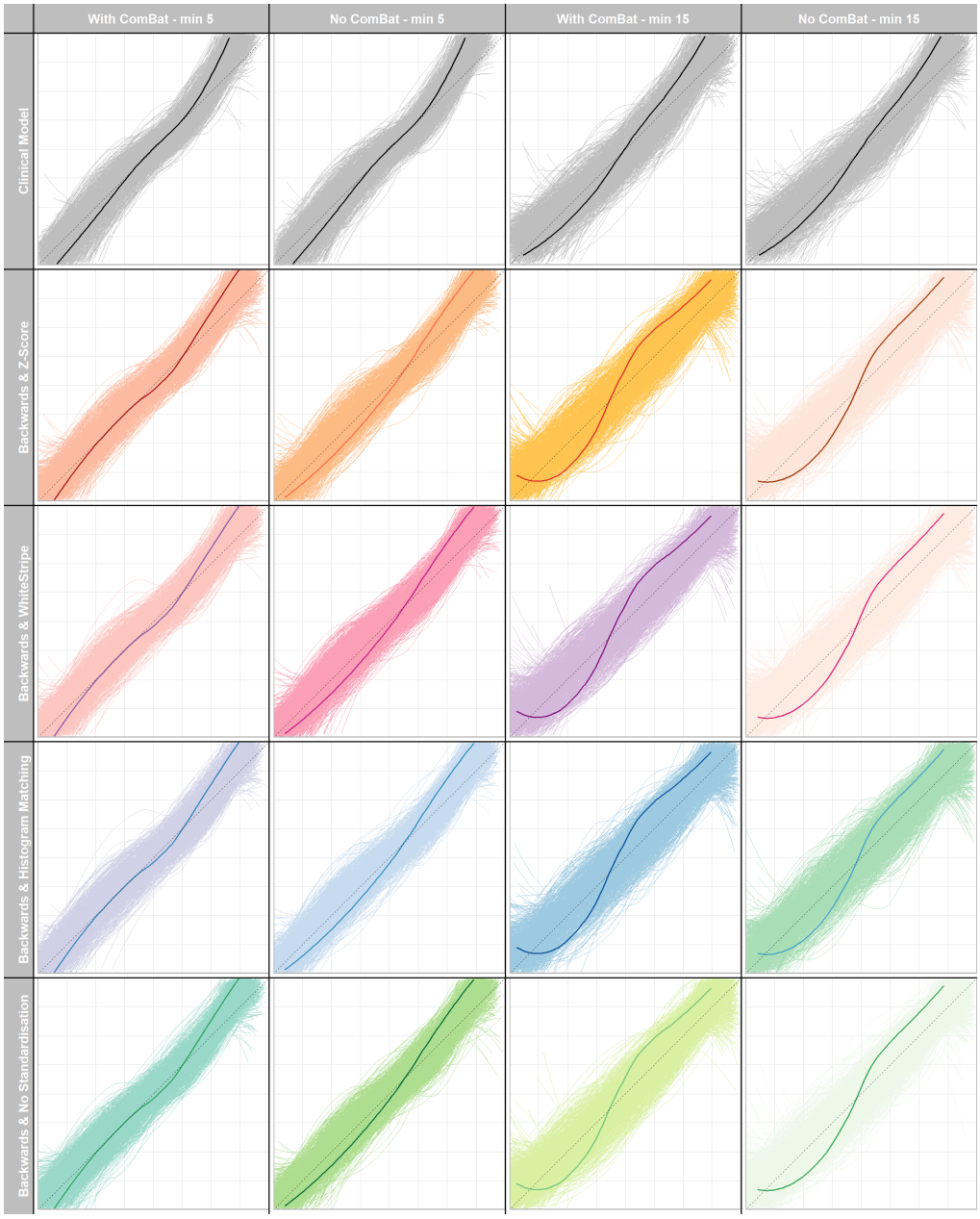

Figure S4.4: Calibration instability plots show the impact of different experimental settings on 1-year survival predictions. ComBat was applied for columns 1 and 3, and without ComBat in columns 2 and 4. Different minimum ComBat batch size were used (5 - columns 1 and 2; 15 - columns 3 and 4) and different intensity standardisation techniques applied per row. Results show individual survival predictions at 1 year, across the bootstrap resamples for models built using backwards feature elimination and bin count 32. x-axes represent predicted and y-axes the observed survival at 1-year. The thin curves represent the predictions from one bootstrap sample and the thicker curve, predictions based on the original, non-bootstrapped data. Only 200, randomly selected, bootstrap results are shown in each calibration plot. The grey dashed line represents the null line, with greater deviation from this indicating worse calibration. Increased spread of the thin curves indicates lower stability of that model building process. The calibration plots resulting from combined clinical and radiomics models, with features selected using hierarchical clustering of principle component analysis results (rows 2,3 and 4) are compared against the clinical only models (grey, top row).

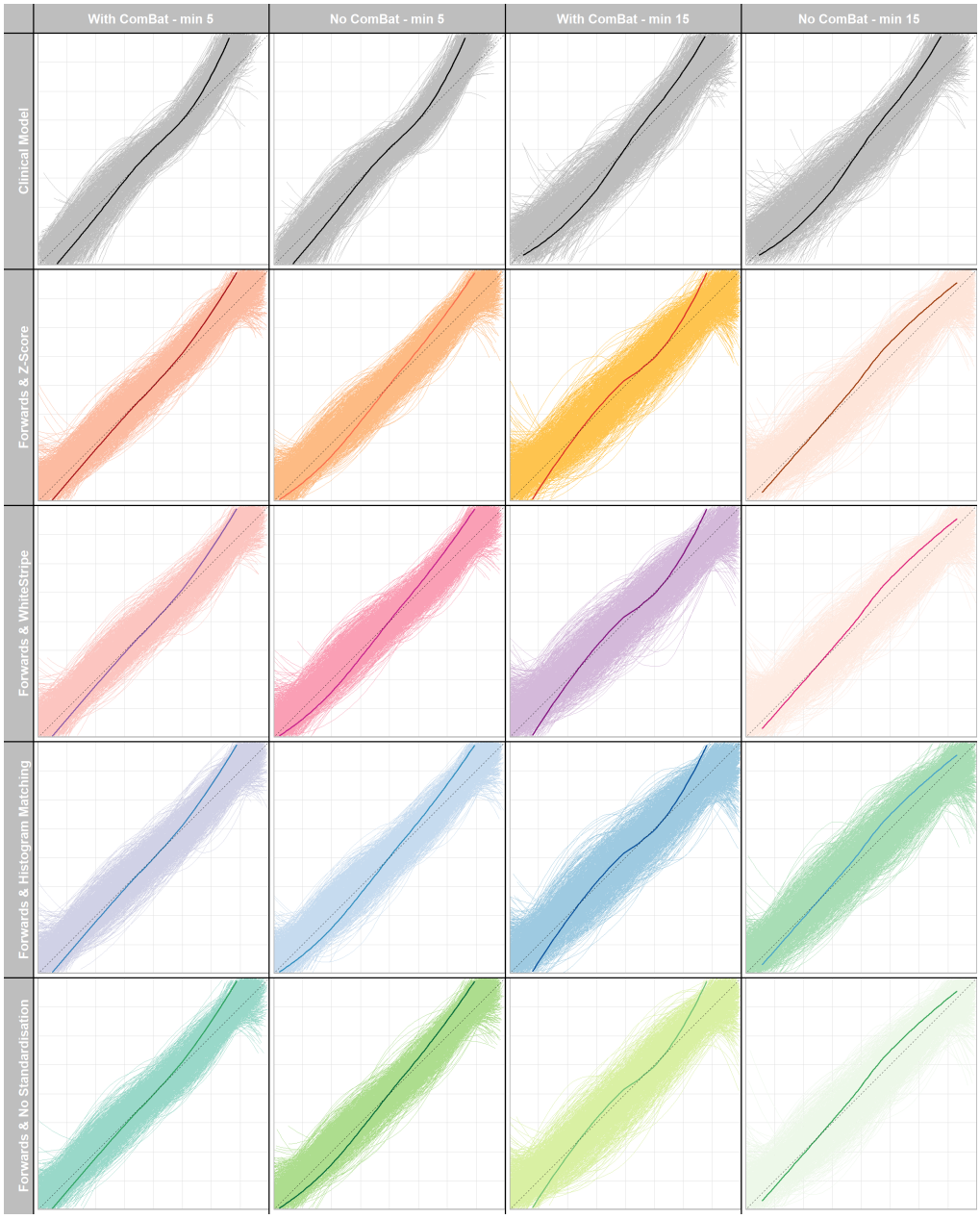

Figure S4.5: Calibration instability plots show the impact of different experimental settings on 1-year survival predictions. ComBat was applied for columns 1 and 3, and without ComBat in columns 2 and 4. Different minimum ComBat batch size were used (5 - columns 1 and 2; 15 - columns 3 and 4) and different intensity standardisation techniques applied per row. Results show individual survival predictions at 1 year, across the bootstrap resamples for models built using forwards stepwise selection and bin count 32. x-axes represent predicted and y-axes the observed survival at 1-year. The thin curves represent the predictions from one bootstrap sample and the thicker curve, predictions based on the original, non-bootstrapped data. Only 200, randomly selected, bootstrap results are shown in each calibration plot. The grey dashed line represents the null line, with greater deviation from this indicating worse calibration. Increased spread of the thin curves indicates lower stability of that model building process. The calibration plots resulting from combined clinical and radiomics models, with features selected using hierarchical clustering of principle component analysis results (rows 2,3 and 4) are compared against the clinical only models (grey, top row).

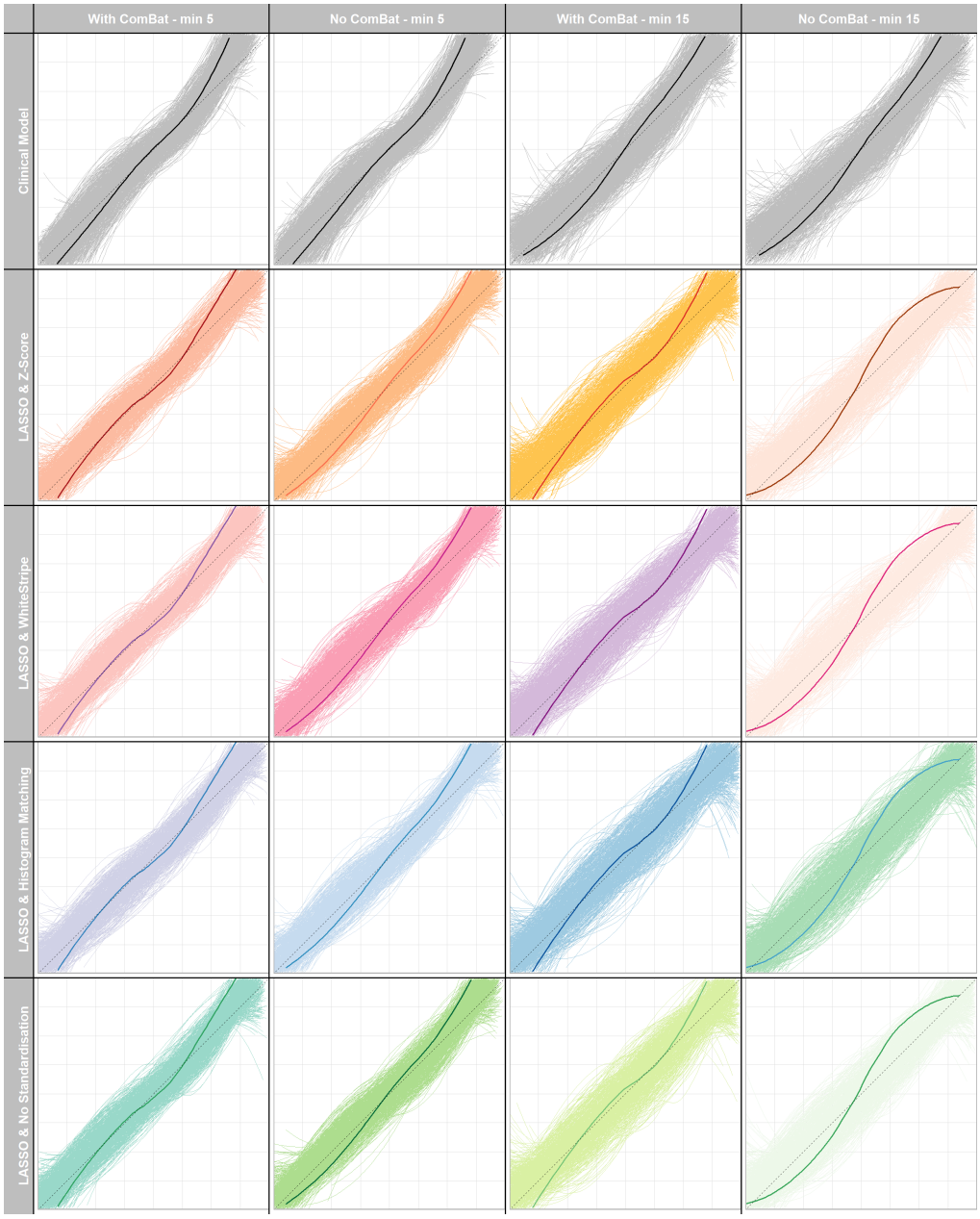

Figure S4.6: Calibration instability plots show the impact of different experimental settings on 1-year survival predictions. ComBat was applied for columns 1 and 3, and without ComBat in columns 2 and 4. Different minimum ComBat batch size were used (5 - columns 1 and 2; 15 - columns 3 and 4) and different intensity standardisation techniques applied per row. Results show individual survival predictions at 1 year, across the bootstrap resamples for models built using Least Absolute Shrinkage and Selection Operator (LASSO) selection and bin count 32. x-axes represent predicted and y-axes the observed survival at 1-year. The thin curves represent the predictions from one bootstrap sample and the thicker curve, predictions based on the original, non-bootstrapped data. Only 200, randomly selected, bootstrap results are shown in each calibration plot. The grey dashed line represents the null line, with greater deviation from this indicating worse calibration. Increased spread of the thin curves indicates lower stability of that model building process. The calibration plots resulting from combined clinical and radiomics models, with features selected using hierarchical clustering of principle component analysis results (rows 2,3 and 4) are compared against the clinical only models (grey, top row).

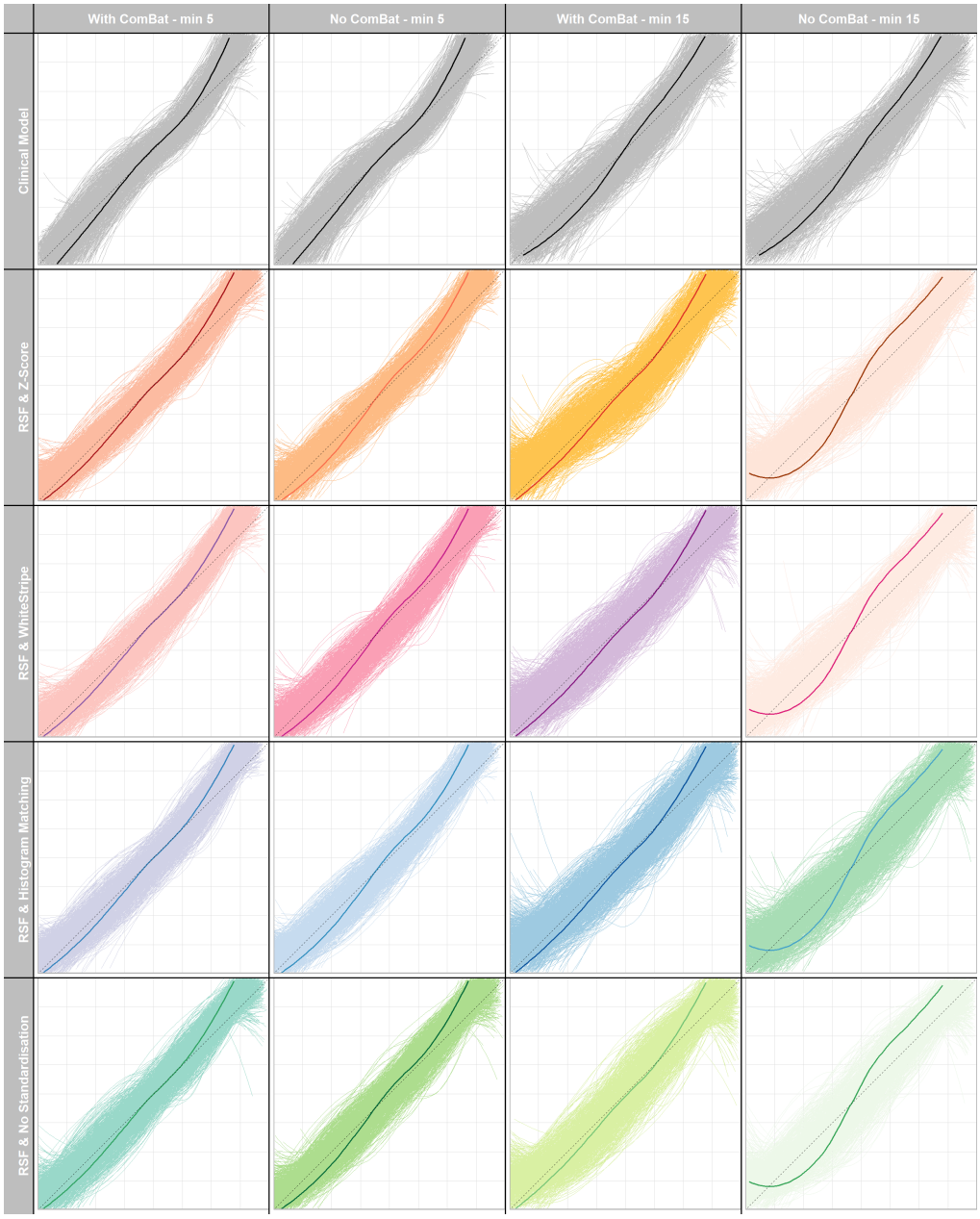

Figure S4.7: Calibration instability plots show the impact of different experimental settings on 1-year survival predictions. ComBat was applied for columns 1 and 3, and without ComBat in columns 2 and 4. Different minimum ComBat batch size were used (5 - columns 1 and 2; 15 - columns 3 and 4) and different intensity standardisation techniques applied per row. Results show individual survival predictions at 1 year, across the bootstrap resamples for models built using random survival forests (RSF) selection and bin count 32. x-axes represent predicted and y-axes the observed survival at 1-year. The thin curves represent the predictions from one bootstrap sample and the thicker curve, predictions based on the original, non-bootstrapped data. Only 200, randomly selected, bootstrap results are shown in each calibration plot. The grey dashed line represents the null line, with greater deviation from this indicating worse calibration. Increased spread of the thin curves indicates lower stability of that model building process. The calibration plots resulting from combined clinical and radiomics models, with features selected using hierarchical clustering of principle component analysis results (rows 2,3 and 4) are compared against the clinical only models (grey, top row).

CHAPTER

FIVE

DISCUSSION

5.1 Summary of aims

The thesis set out to first systematically review the literature regarding the use of ISTs in the processing of diffuse glioma and glioblastoma MRI prior to the extraction of RFs, and tried to determine the optimal IST (chapter 2).

Chapter 3 examined the prognostic effect of tumour size, a simple quantitative imaging feature, in a large cohort of patients with glioblastoma and also examined the effect of varying sample size and non-linear transformation on the ability to reproduce the modelling results.

As a result of the preceding components, the final experiment (chapter 4) assessed the impact of ISTs and ComBat realignment, a statistical approach to multi-centre RF harmonisation, on prognostic model performance in patients with unifocal glioblastoma. In particular, the focus was on comprehensive model performance assessment, which included examining calibration, stability, discrimination, relative explained variation and fit.

5.2 Intensity standardisation of MRI prior to radiomic feature extraction for artificial intelligence research in glioma – a systematic review (chapter 2)

5.2.1 Summary

In chapter two, the published literature was systematically reviewed with the aim of comparing different methods of MRI ISTs prior to RF extraction in patients with diffuse glioma, including glioblastoma. MEDLINE, EMBASE, and SCOPUS, were searched for articles that included: MRI radiomic studies where one IST was compared with another or no standardisation, and original research concerning patients diagnosed with diffuse gliomas. After title and abstract screening, PRISMA criteria were used to extract data from potentially eligible studies including number of patients, MRI sequences, validation status, radiomics software, method of segmentation and intensity standardisation. QUADAS-2 was used for quality appraisal.

After duplicate removal, 743 results were returned from database and reference searches and from these, 12 papers were eligible. Due to a lack of common pre-processing and different analyses, a narrative synthesis was sought. Three different approaches to intensity standardisation (IS) have been studied: histogram matching (HM, 5/12), limiting or rescaling signal intensity (8/12), and deep learning (1/12) - only two papers compared different methods. From these studies, on face value, HM produced the highest AUC values but these particular studies failed to include a comparison method and hencde no consensus could be reached on an optimal strategy.

5.2.2 Limitations

As mentioned in chapter two, there were several limitations of this review. Full-text articles for two conference abstracts were not available. Based on the abstracts, it is unlikely they would have been included and their potential omission will have had a limited impact as a narrative synthesis would

still have been required. Similarly, there are several studies that have been published on this topic since the initial searches were carried out. These are discussed in depth in the following section, and they would not have altered the finding and conclusions of the narrative review.

Included studies were evaluated with the QUADAS-2 tool [1], however it is not specifically designed for assessing the efficacy of MRI ISTs, rather it is designed for assessment of primary diagnostic accuracy studies. A more specific alternative was not available and radiomics or artificial intelligence-specific reporting tools such as the Radiomics Quality Score [2] or CLAIM [3], respectively are intended to evaluate the quality of radiomic or AI model development.

The scope of this review was to assess MRI ISTs in the context of diffuse gliomas, which will have led to inevitable omission of studies of other organs, brain pathologies and healthy volunteers. As noted in the introduction, there is evidence from these other studies that supports the use of ISTs such as HM and WS in other contexts. The scope of the review was limited to diffuse glioma and radiomic studies as this would reduce the amount of heterogeneity in the included studies and increase the specificity for neuro-oncology settings. Also, WS is a popular IST and is specific to neuroimaging as it relies on the segmentation of normal appearing white matter (NAWM) [4]. Hence, it would be difficult to compare the results of studies that included this approach with others.

The review was also limited to assessing methods of harmonisation that are applied to images prior to feature extraction, and therefore did not include an assessment of strategies such as ComBat. ComBat is a statistical model for feature realignment that is applied to RFs directly, following their extraction [5]. Studies were still eligible for inclusion if they did use ComBat [6] but they had to specifically assess the impact of ISTs alone. The debate about whether ComBat is a useful tool for multi-centre RF realignment [5, 7–9] is one that would have added further complexity to the assessment of MRI IST. This is a related area but may require a separate review to address.

5.2.3 Future work

Since this systematic review was undertaken, more studies have investigated MRI ISTs in patients with diffuse glioma, however the additional evidence adds further weight to the conclusion that a consensus is difficult to reach and further work is necessary. The studies can be broadly divided into those that chose to use only one type of IST, ZS [7, 9] and those that compared multiple approaches [10–12].

The benefit of using ZS standardisation was demonstrated in studies by Carré et al. [9] and Li et al. [7], although neither compared the results to other ISTs (Carré et al. had previously published a comparative study, which was included in the systematic review [13]). Carré et al. [9] compared ZS standardised images to images without standardisation on 174 patients with glioblastoma and low-grade glioma (LGG) from TCGA multi-centre dataset on TCIA. It was assumed that NAWM RFs should be equal across patients, and any observed differences are due to non-biological, site-dependent effects. Hence, the relative standard deviation (RSD, standard deviation divided by the absolute mean) of NAWM RFs ought to be lower if the site-effects are minimised. MRIs without any pre-processing were compared to those that had been standardised with ZS, as well as spatially resampled ($1mm^3$ resolution) and had N4 bias field inhomogeneity correction (N4) applied. The RSD of NAWM RFs was reduced in 78% of FLAIR images, and 82% T1CE images compared with unprocessed images. They also investigated the effect of pre-processing on the balanced accuracy (arithmetic mean of sensitivity and specificity) of five ML classifiers for differentiation of glioblastoma ($n = 125$) versus LGG ($n = 107$) and results were mixed depending on the sequence. For T1W and T1CE sequences, pre-processed images led to increased accuracy, whereas the reverse was true for FLAIR images. T2W results depended upon the specific ML classifier chosen.

The impact of ZS standardisation compared to no IST was investigated by Li et al. [7] on the prediction accuracy of glioma grade, IDH mutation and 1p/19q co-deletion in 212 patients. Images for patients with glioblastoma ($n = 107$) and LGG ($n = 105$) from TCGA were used to build seven ML classifiers and the AUC was used to evaluate their accuracy. Using ZS standardisation increased the consistency of the histograms of MRI signal intensity and increased the AUC for all

seven classifiers for the three prediction tasks. Carré et al. may have avoided comparison between ZS and other ISTs as they concluded in a prior study that ZS offers the simplest approach compared to HM or WS and the results are comparable to other methods. Li et al. do not justify their choice of IST [7] but it may have been impractical to investigate ISTs as well as ComBat realignment, amongst other factors included in their study. However, given the lack of consensus around this in the literature [8, 14], supplementary results with other ISTs would be beneficial to help the community understand how the results of any radiomics study is impacted by this pre-processing step.

Saltybaeva et al. [10] investigated the impact of HM and ZS IST, as well as different grey-level discretization approaches, on the reproducibility of RFs in a two-stage study. 11 patients with glioblastoma that had a T1CE sequence acquired on two different MRI machines (one internal and another external to the institution) within a specific time interval (range 1-36 days) were included in stage-one. Reproducibility of features between the two scans was low. The percentage of features with a one-way random effects ICC > 0.9 was at most 8.0% in the case of HM (1.9% for no IST, 5.9% ZS). In stage-two, they compared the results of univariable logistic regression models per feature for OS at 18 months in 60 patients (35 deaths) with glioblastoma acquired from the public BraTs collection [15]. Again, the impact of IST was small but noticeable with 7/97 features showing a prognostic relationship with HM (3/97 without IST, ZS not tested). The conclusions that can be drawn from this study are limited by the small sample size and also by the inclusion criteria for the institutional data. The 11 patients with glioblastoma had a large range in intervals between both scans, and although the authors stated that the volume change was $< 30\%$, this still seems substantial when trying to use this as a test-retest dataset for reproducibility analysis. Further studies using larger test-retest datasets are required, with shorter intervals between studies to ensure that there is no macroscopic change between the tumours if they are to be used for RF robustness assessment.

The impact of 15 different ISTs, including ZS, WS and HM, on OS prediction models was also evaluated by Salome et al. [8], in patients with recurrent glioblastoma ($n = 197$, 15 scanners) and

glioblastoma at initial presentation ($n = 144$, 14 scanners). One strength of their study is the real-world dataset used and the range of MRI acquisitions but also, as with the study by Carré et al., they investigated the differential impact of ISTs per MRI sequence. For Cox proportional hazards models built with RFs from T1CE images of glioblastoma at diagnosis, WS performed better than other popular methods with C-index 0.652 (no IST = 0.608, ZS = 0.628, HM = 0.639) and AIC 547 (no IST = 558, ZS = 559, HM = 560). For T2W images, HM had C-index 0.67 (no IST = 0.648, ZS = 0.665, WS = 0.639) and AIC 415 (no IST = 420, ZS = 420, WS = 414). One reason for the lack of consensus in optimal IST could be the nature of each sequence, particularly the range and distribution of signal intensity, which necessitates different approaches to harmonising the signals across patients. However, their results showed minimal absolute differences in model performance metrics and the order of change in C-index or AIC is difficult to interpret [16]. Using multiple ISTs for standardising a MRI dataset also introduces further complexity to the radiomics pipeline, when there is insufficient evidence on the basis of this study to do so.

ZS and WS performed similarly when assessing the impact of different pre-processing techniques on the performance of different ML classifiers for molecular glioma subtype prediction by Foltyn-Dumitru et al. [12]. Using homogeneously acquired internal data ($n = 610$) and heterogeneous external, publicly available data from University of California San Francisco (UCSF) ($n = 410$) and TCGA ($n = 160$) to train nine ML classifiers to predict three classes of glioma subtype (IDH-mutant, 1p/19q intact vs IDH-mutant, 1p/19q co-deleted vs IDH-wild type). They found that there was a large impact on model performance from using IST in the external datasets compared to the more homogeneous internal data. For example, the best classifier in the institutional data hold-out test set (80:20 data partition), had AUC 0.84 (95% CI = 0.75-0.89) without any pre-processing compared to 0.87 (95% CI = 0.81-0.91) with N4 in conjunction with ZS and very similar results for N4/WS. In the external TCGA data, however, AUC increased from 0.45 (95% CI = 0.35-0.54) for unprocessed images to 0.85 (95% CI = 0.76-0.89) with N4/WS and 0.87 (95% CI = 0.80-0.91) with N4/ZS. Similar changes were seen with application to the UCSF data. Overall, they could not detect any large differences between ZS or WS ISTs, which is consistent with other groups [8, 13]. In particular, these prior studies have used a FBN grey-level discretization method for calculating

the texture features, which may provide a way of normalising these RFs without the need for ISTs or limiting their impact, and is recommended by the IBSI for discretization of MRI images [13, 17].

Lastly, Ubaldi et al. [11] compared the impact of ISTs on a random forest classifier for differentiation of low *versus* high-grade glioma in 158 patients taken from the BraTs public dataset. They used three ISTs: i) subtracting the minimum intensity and dividing by the range of intensities in the whole brain (MinMax normalisation); ii) subtracting the median intensity and dividing by the interquartile range of intensities in the whole brain (RobustScaler normalisation); iii) subtracting the median intensity and dividing by the interquartile range of intensities in the brainstem (Brainstem normalisation). They found that RobustScaler and Brainstem normalisation led to greater uniformity in histograms of MRI intensity and increased the AUC of the classifier when considering only intensity-based RFs, but not texture features and this was not dependent on the number of bins used for grey-level discretization when a FBN was used. Again, this supports the view that a FBN may negate the impact of ISTs for texture-based features, but it is clearly important for a multi-centre MRI dataset to produce more harmonisation of the signal intensity distribution across patients.

Despite a number of studies being published following the systematic review of ISTs in diffuse glioma radiomics, there is still little consensus on the optimal approach and this is still an active field of research. Future studies will benefit from better availability of large volumes of scan-rescan data, with short imaging intervals. This should allow for improved assessment of IST impact on repeatability of RFs in a greater range of institutions. Producing such datasets prospectively may be technically challenging and expensive however, and may explain why approaches such as that by Saltybaeva et al. was used [10]. Another area for research to focus upon, relates to radiomics study design. Studies that investigate the impact of ISTs on the results of predictive models could provide the results of alternative ISTs, in particular the more popular choices of HM, ZS or WS, which would represent minimal additional workload when considering the nature of the radiomics pipeline.

Regarding IST impact on survival models in glioma [8, 10, 18–20], the approach to modelling has

been varied and more work is required to ascertain the impact of ISTs on survival models that assess time-to-event models thoroughly. Studies have dichotomised patients into two risk groups [10, 18, 19] or considered only discrimination performance [20], whereas it would be helpful to also assess the impact of ISTs on calibration and stability and focus on continuous event analysis [21]. One of the key areas for future research would be to conduct a study with multi-centre MRI acquisition and produce a radiomics prognostic model that could be thoroughly evaluated using a scan-rescan dataset as well as large volumes of publicly available data on glioblastoma patients, with accompanying clinical predictor data.

There are DL approaches to this problem that will need to be compared with radiomics models. Convolutional neural networks (CNNs) could alleviate the need to find optimal IST strategies for harmonising signal intensity of MRIs in multicentre datasets [22] but they could also produce accurate survival prediction models without the need for radiomic feature extraction [23]. This could produce many potential benefits, including removing the need for laborious tumour segmentation, image pre-processing and devising a ML or statistical strategy for modelling the RFs against a chosen outcome. Another advantage of a DL-approach is that 'transfer learning' can leverage pre-trained CNNs to build models for outcome prediction on relatively small volumes of new data [24]. A CNN that has been previously trained on a large volume of brain MRI data, for example, can be trained on unseen imaging to predict a new outcome such as OS in much smaller samples and perform accurately [23, 24].

One of the main concerns, however, with DL models is the relative lack of explainability of model decisions when compared to traditional prediction modelling based on clinical parameters or IBs that can be measured or derived directly from the images [25]. Trust between the clinicians and any new model is important for its use and translation too [26]. There are now many ways in which DL models can be made more 'explainable' and future work will be necessary to show how a model has come to its decision for OS prediction. For example, saliency maps can be produced using a number of different methods to generate input images with a heatmap of the most important parts of an image used for making a particular decision [25]. Such approaches might help to increase

the trust of clinicians and users in DL approaches, although maps still require some assumptions to be made about how the particular parts of the image have contributed to the decision. Future studies are required to determine which approach to improved OS prediction is most accurate in multicentre data, but also the more explainable and trustworthy for patients and clinicians.

To summarise, the main recommendations for future research are:

- Greater availability of scan-rescan data, which will allow assessment of the repeatability of novel IBs, and greater availability of publicly-available glioma datasets including imaging and comprehensive clinical data.
- When investigating the impact of ISTs, researchers could provide results of alternative ISTs, in particular the more popular choices of HM, ZS or WS.
- Investigate the effect of ISTs on continuous outcome time-to-event models for patients with diffuse glioma, paying attention to metrics such as calibration and stability.

5.3 Tumour size and overall survival in a cohort of patients with unifocal Glioblastoma: a uni- and multivariable prognostic modelling and resampling study (Chapter 3)

5.3.1 Summary

In chapter 3, the prognostic effect of tumour size in a large cohort of patients diagnosed with glioblastoma was investigated. Secondary analyses evaluated the impact of sample size choice and consideration of non-linear transformations on the likelihood of finding a prognostic effect using univariable and multivariable analysis and data resampling. 279 patients with IDH-wildtype unifocal WHO grade 4 glioblastoma and pre-operative MRI between 2014-2020 from a retrospective cohort were investigated. Uni- and multivariable association between core volume, whole volume

(CV, WV) and diameter with OS was assessed with (1) Cox proportional hazard models ± log transformation and (2) resampling with 1,000,000 repetitions and varying sample size to identify the percentage of models, which showed a significant effect of tumour size.

Diameter or volume models adjusted for operation-type were significant, and diameter adjusted for all clinical variables including age, gender, adjuvant therapy, MGMT and operation type remained significant ($p = 0.03$). Multivariable resampling increased significant effects ($p < 0.05$) of all size variables as sample size increased. Log-transformation also had a large effect on chances of prognostic effect of WV. For models adjusted for operation-type, 19.5% of WV vs 26.3% log-WV ($n = 50$) and 69.9% WV and 89.9% log-WV ($n = 279$) were significant. The study suggested that tumour volume is prognostic in multivariable analysis and that this is most likely detected at larger sample sizes and with log-transformation for WV.

5.3.2 Limitations

The MRI acquisition parameters in our dataset were heterogeneous, especially slice thickness, and this could have impacted upon the accuracy of volume measurements. The images were standardised prior to volume assessment by spatially resampling to an isotropic $1mm^3$ voxel resolution, which should have reduced the impact of acquisition heterogeneity. Heterogeneous data acquisition will be encountered in routine clinical practice, and an IB such as tumour volume will have to be robust to variation in acquisition parameters in order to be clinically relevant.

A proportion of our patients had to be excluded due to lack of the necessary MRI sequences for the DL segmentation algorithm. The efficiency of a semi-automated segmentation approach outweighed the potential limitation of a reduced sample size. A manual approach could have been adopted for the patients who did not have all four sequences (T1W, T2W, FLAIR and T1CE) required for the algorithm, however the lack of certain MRI sequences could feasibly impact on even expert manual segmentation, and would have increased the workload of the study.

We investigated only three size variables but there are many others described in the literature.

The experiment did not aim to provide a comprehensive evaluation of the prognostic role of all possible tumour size parameters in glioblastoma, but to investigate some of the methodological issues affecting this question, and that could be applied to any of the other continuous measures of tumour size in glioblastoma.

The resampling study was limited by the maximum sample size, and the results could have been further enhanced with access to larger amounts of patient imaging and corresponding well-curated datasets. Several of the large publicly available repositories such as BraTs [15] or TCGA data do not have as in-depth clinical labels for building multivariable models, which does hamper their use.

The analysis of tumour size as a prognostic factor in glioblastoma patients would have been even more clinically relevant if the dataset had included other key variables such as performance status, trial inclusion and socioeconomic status and to adjust for these in the modelling process. Despite this, it was still possible to highlight methodological issues and common pitfalls in this area of literature.

5.3.3 Future work

Multi-centre prognostic factor studies, with larger sample sizes will be needed to investigate the role of different tumour volumes such as CV and WV with appropriate use of non-linear transformations of volume and other clinical predictors to validate the findings of chapter 3. Ideally, these studies will use optimal statistical methodology [27], and avoid potential pitfalls that can be encountered in glioblastoma prognostication literature [28, 29].

A recent large multi-centre study conducted by Karschnia et al.[28, 29] illustrates how statistical methodology in glioblastoma prognostic factor research can be optimised. The study aimed to evaluate the prognostic impact of the extent of surgical resection in 744 patients with glioblastoma, diagnosed according to 2021 WHO criteria and treated with Stupp protocol adjuvant treatment [30]. In univariable Cox regression, pre-operative enhancing tumour volume was not prognostic (hazard ratio 1.00, 95% CI 1.0-1.0). To improve this study, future work could use this data to investigate

pre-operative non-enhancing tumour volume, or WV for prognostic effect and could also consider whether any continuous variables could be modelled non-linearly. A separate analysis of 98 patients with glioblastoma that showed no contrast enhancement did include non-enhancing pre-operative tumour volume, which was not found to be prognostic in univariable analysis [29].

Their study only included predictors into multivariable models based on univariable screening (included if univariable model $p < 0.05$). Additional analyses could explore whether unsupervised feature selection could demonstrate other prognostic factors and this might avoid some of the drawbacks of univariable feature screening [27]. Riley et al. also suggest that exploratory prognostic factor research should present adjusted hazard ratios; estimates that are adjusted for optimism by performing internal validation [27]. Availability of such datasets in the public domain would greatly enhance prognostic factor research, as well as IST exploration as mentioned in section 5.2.3.

5.4 Impact of intensity standardisation and ComBat batch size on clinical – radiomic prognostic models performance in a multi-centre study of patients with Glioblastoma (Chapter 4)

5.4.1 Summary

The final experiment of the thesis assessed the effect of different ISTs and ComBat batch sizes on radiomics survival model performance and stability in a heterogeneous, multi-centre cohort of patients with glioblastoma. Multi-centre pre-operative MRI acquired between 2014-2020 in patients with IDH-wildtype unifocal WHO grade 4 glioblastoma were retrospectively evaluated. WS, HM and ZS ISTs were applied before RF extraction. RFs were realigned using ComBat and minimum batch size (MBS) of 5, 10 or 15 patients (RFs without ComBat realignment were also used for comparison). Cox proportional hazards models for OS prediction were produced using five different

selection strategies and the impact of IST and MBS was evaluated using bootstrapping. Calibration, discrimination, relative explained variation, and model fit were assessed. Instability was evaluated using 95% CIs, feature selection frequency and calibration curves across the bootstrap resamples.

195 patients were included. Median OS = 13 (95% CI 12-14) months. 12-14 unique MRI protocols were used per MRI sequence. HM and WS produced the highest relative increase in model discrimination, explained variation and model fit but IST choice did not greatly impact on stability, nor calibration. Larger batches improved discrimination, model fit and explained variation and, at MBS 10 and 15, using ComBat feature realignment also led to slight improvements in performance. However, higher MBS (reduced sample size) reduced stability (across all performance metrics) and reduced calibration accuracy. The results showed how heterogeneous, real-world glioblastoma data poses a challenge to the reproducibility of radiomics, and that whilst ComBat generally improved model performance as MBS increased, this required a reduced sample size and therefore reduced stability and calibration. HM and WS tended to improve model performance.

5.4.2 Limitations

As noted in section 5.3.2, the acquisition parameters for the MRI used in this modelling study were heterogeneous, including several centres with relatively few patients scanned, which impacted on our ability to test larger batch sizes for ComBat. Again, given that this is a real-world dataset, encountering this degree of heterogeneity was inevitable, particularly considering the 'hub-and-spoke' system of neuro-oncology referrals to the central neurosciences centre. This impacted upon the batch labelling for ComBat realignment; some peripheral centres scanned only a small number of patients so there were not enough to be included when selecting the MBS for ComBat. For the purposes of this study, this limitation did not impact upon the main findings, as it was still possible to compare the relative impact of different ISTs on model performance, and it was still possible to thoroughly evaluate the models at three different MBSs for ComBat.

Public data could have been used to either supplement or replace institutional data and therefore increase the patients scanned in each batch. However, one strength of this thesis was the availability

of well-curated clinical data, as this allowed comparison between a gold-standard clinical model and clinical-radiomic models in our dataset. Whilst there are larger MRI datasets online, the most popularly cited collections do not contain in-depth clinical parameters such as adjuvant treatment history [31, 32]. Better availability of datasets, or adopting a federated approach, which could reduce the need for in-depth data protection approvals, to such tasks might offer a more suitable approach going forwards [33].

The analysis in this PhD did not attempt to increase the ComBat batch sizes using unsupervised clustering, but this has been used [9, 34] to group patients with similar RFs into clusters to avoid excessive data loss in ComBat realignment. A small number of patients that would ordinarily have been assigned to a separate batch because they were scanned at a small site or with a different set of acquisition parameters to the majority of patients could be incorporated into another batch label if the RFs were found to be similar by clustering algorithms such as hierarchical clustering [34]. Da-Ano et al. [34], for instance, used hierarchical clustering of RFs in 98 patients with laryngeal cancer who had contrast-enhanced computed tomography acquired with 15 different protocols and found that they could cluster patients based on the RFs from their tumours into 2 clusters. Novel batch labels resulted in larger cluster sizes (38 and 60 per cluster) for estimation of ComBat equation coefficients, than using the 15 protocols as batches. Accuracy of ML classifiers for prediction of response to chemotherapy improved following the use of ComBat (balanced accuracy 27% untransformed vs 69% with ComBat).

Several limitations of a clustering approach influenced the decision not to use unsupervised clustering for this PhD. Da-Ano et al. checked that the new batch labels did not differ in terms of the number of events per cluster (16% in cluster 1 vs 15% in cluster 2), and concluded that since these values are similar, clustering based on RFs was not influenced by outcome or biological differences between patients. The distribution of clinical predictors between clusters was not presented, and statistical testing of group differences was not conducted, although it is accepted that the sample size was unlikely to be sufficient to accurately detect cluster-wise differences in all clinical predictors. Alternatively, Da-Ano et al. could have adopted the approach used by Carré et al., using

DICOM metadata and MR quality metrics to cluster patients for ComBat batch labels. These metrics included simple statistical measures of foreground signal intensity (mean, range, variance), as well as more complex measures such as ratio of foreground to background standard deviation in signal intensity [9]. It is much less likely that these metrics would be influenced by the tumour characteristics or clinical predictors.

Another limitation is that unsupervised clustering requires the user to arbitrarily decide where the best cut-off is located and determine into how many clusters the dataset is optimally divided, therefore it is not completely objective. Da-Ano et al. used the silhouette method [34] to try to optimise this decision, which measures the average of the ratios between intra- and nearest-cluster distances; smaller distances between patients within the cluster, than the distance to the centre of the nearest neighbouring clusters suggests a compact and well separated group of patients. A smaller ratio, therefore is optimal. This score can be compared against a number of user-defined clusters, and the point at which the score is no longer decreasing substantially, as the number of clusters increases, determines the optimal cut-point and this can be represented graphically (referred to as the 'elbow' method) [35]. This requires some subjectivity and therefore the results of clustering should ideally be validated in another dataset. However, Da-Ano et al. did not perform any internal validation and attempting to do so in a small dataset would have been difficult. The clustering used by Carré et al. was also not validated on unseen data, however they did note an improved accuracy of ML classifer for high versus LGG prediction with T2W RFs compared to standard labelling for ComBat [9].

Only three out of many ISTs available were chosen for evaluation in chapter 4, however these had previously been identified as the most popular choices in prior studies [36] and the approach to model performance evaluation is still justified. The supervised feature selection strategies considered far more than the four radiomic features suggested as the maximum by event per predictor calculation, and therefore these strategies would not have been optimal if the aim of the study was to identify potential prognostic IBs. However, these selection strategies are popular within the radiomics modelling literature and the potential for overfitting in my selection strategy will not have impacted

upon our assessment of relative model performance due to IST and ComBat batch size. Finally, measurement of IST impact on feature repeatability was not assessed, however to the best of my knowledge, a preoperative glioblastoma dataset with test-retest data is not available publicly. Again, this would have been more imperative had the study aimed to produce an accurate radiomics prognostic model.

5.4.3 Future work

Many of the strategies noted in section 5.2.3 could also be applied to the current work. Thorough assessment of prognostic radiomics models in patients with glioblastoma including the impact on stability and calibration, use of larger and publicly-available multi-centre glioblastoma datasets with comprehensive clinical metadata and scan-rescan data, and evaluation of multiple ISTs on the results of any candidate models would all be ways to extend the work in chapter 4. Ultimately these studies would also need large external validation datasets and DL approaches could also be compared to radiomics based approaches.

Building upon the limitations section, unsupervised clustering could be evaluated in this context but should be evaluated in large, dedicated studies. It would be invaluable to build larger clusters of patients in multi-centre data and avoid the instability introduced by discarding patients when the MBS was increased, particularly if the ideal of 20-30 per cluster, per biological covariate [5] is to be achieved. It is likely that the lack of added benefit of ComBat realignment in the study conducted by Salome et al. [8] was in part due to the small number of patients per batch, and it may have led to the only marginal gains with ComBat in the study performed in this thesis. For example, Salome et al. had 197 patients with recurrent glioblastoma scanned used 15 scanners, and 144 patients with new glioblastoma scanned using 14 machines. They found, for T1CE images in recurrent glioblastoma that the C-index with ComBat and WS was 0.68 (95% CI 0.66-0.69), compared to 0.71 (95% CI 0.69-0.74) with just WS and this trend was replicated for other sequences and the newly diagnosed glioblastoma cohort. If they had used ComBat with clustering to determine larger batches, perhaps a greater improvement would have been seen with ComBat as the coefficients

would have been more accurately estimated [5].

Therefore, it would be useful to conduct a proof-of-concept study in a large multi-centre study to ensure that any clustering algorithm based on RFs separates patients in a way where they do not vary in a statistically significant manner between cluster, avoiding the assumptions made by Da-Ano et al. [34]. Additionally, it would be useful to validate the results with unseen external data to determine its validity, which was not conducted by Carré et al. or Da-Ano et al [9, 34]. This is likely to require large datasets and dedicated sample size calculations.

5.5 Future perspectives and considerations

Further testing of the findings raised in this thesis will need to be explored in new datasets. A collaboration of 10 institutions, Radiomic Signatures for PrecisiON Diagnostics (ReSPOND) consortium on glioblastoma, has collated 3,300 de novo cases of glioblastoma that have undergone Stupp treatment with the aim of developing novel IBs [37]. The models developed in chapter 3 and 4 could be extended and externally validated in these larger datasets. Given the lack of consensus on the optimal IST for images prior to radiomic extraction in neuro-oncology, IBSI's documentation could be updated to reflect the current trends in the literature and issue a position statement on best practice for future research on this topic.

The prospect of DL in medical imaging analysis may make the need for thorough assessment of ISTs and radiomics prognostic models unnecessary going forwards. Particularly given that the advantage of CNNs is that they can learn complex representations of multi-centre data without any need for researchers to spend time deciding the optimal strategy to harmonise the input images. These advances are welcome, however they may need comprehensive evaluation against an optimised radiomics model given the intense research interest and promise in using radiomic IBs to enhance glioblastoma prognostication. Hence, the work presented in this thesis can help to give a framework of how to model any quantitative IBs such as tumour volume, diameter or more complex RFs, as well as how to thoroughly assess putative prognostic models. It may be that DL models outperform

even the most optimised radiomics approaches, however the methodology of building and evaluating the radiomics approaches needs to be improved for an adequate comparison to take place.

This thesis primarily considered IBs from anatomical and structural MRI and some clinical predictors including molecular pathology such as MGMT methylation status, but there are a myriad of constantly improving tests and modalities for assessing the patient with diffuse glioma and it is possible that a multi-modal approach that integrates data from advanced MRI, positron-emission tomograpy (PET), genomics, biochemistry, clinical and socioeconomic information may produce the best explanatory model for OS prognostication [38]. For instance, amino acid based PET tracers can readily cross the blood brain barrier and are taken up by glioma cells via the L-type large amino acid transporters (LAT), which are overexpressed in glioma cells and therefore they generally have excellent tumour-to-background uptake ratios [39]. O-(2-F^{18}-fluoroethyl)-L-tyrosine (FET) is an example of a radiolabelled PET-tracer, which has been shown to provide prognostic information in the context of suspected recurrent glioma [40] and show slight positive correlation with glioma cellularity in pre-operative imaging in glioblastoma patients [41]. Newer advanced MRI sequences such as amide proton transfer-weighted (APTw) imaging can semiquantitatively estimate the concentration of endogenous proteins in a tissue [42] and has been shown to also correlate with cellularity in glioblastoma [41]. It is possible that future studies will demonstrate additional prognostic benefit for using these imaging modalities and sequences in pre-operative settings.

5.6 Conclusions

This PhD has evaluated the current landscape of intensity standardisation techniques prior to radiomic extraction in patients with glioblastoma and shown that there is more research needed into the optimal strategies, and better reporting of IST evaluation on radiomic models. The thesis has also added to the general field of prognostic factor research, specifically in this context by highlighting some areas of improvement in exploratory prognostic research such as careful attention to non-linear modelling, sample size calculations, and thorough time-to-event model evaluation

including model calibration and stability.

Recommendations for future work investigating prognostic modelling in patients with glioblastoma and using radiomics-based imaging biomarkers include:

- Prognostic models using radiomic IBs need to be comprehensively assessed, which includes measuring not only discrimination but also calibration and stability as well as evaluating the impact of popular ISTs on model performance.
- Future radiomic model building will greatly benefit from large volumes of publicly-available and multi-centre imaging data (or federated approaches), including scan-rescan imaging and also accompanying well-curated clinical and molecular pathology information.
- Understanding of prognostic factors for patients with glioblastoma could be improved with greater emphasis on modelling predictors non-linearly and not using univariable screening methods for predictor selection.

References

1. Whiting, P. F. *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. eng. *Annals of internal medicine* **155,** 529–536 (Oct. 2011).
2. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14,** 749–762 (Oct. 2017).
3. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence and Medical Imaging (Claim). *Radiology: Artificial Intelligence* (2020).
4. Shinohara, R. T. *et al.* Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical* **6,** 9–19 (2014).
5. Orlhac, F. *et al.* A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *Journal of Nuclear Medicine* **63,** 172–179 (Feb. 2022).

6. Orlhac, F. *et al.* How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *European Radiology* **31,** 2272–2280 (Apr. 2021).

7. Li, Y. *et al.* Radiomics-Based Method for Predicting the Glioma Subtype as Defined by Tumor Grade, IDH Mutation, and 1p/19q Codeletion. *Cancers* **14,** 1778 (Mar. 2022).

8. Salome, P. *et al.* MR Intensity Normalization Methods Impact Sequence Specific Radiomics Prognostic Model Performance in Primary and Recurrent High-Grade Glioma. *Cancers* **15** (2023).

9. Carré, A. *et al.* AutoComBat: a generic method for harmonizing MRI-based radiomic features. *Scientific Reports* **12,** 1–17 (2022).

10. Saltybaeva, N. *et al.* Robustness of radiomic features in magnetic resonance imaging for patients with glioblastoma: Multi-center study. *Physics and Imaging in Radiation Oncology* **22,** 131–136 (2022).

11. Ubaldi, L., Saponaro, S., Giuliano, A., Talamonti, C. & Retico, A. Deriving quantitative information from multiparametric MRI via Radiomics: Evaluation of the robustness and predictive value of radiomic features in the discrimination of low-grade versus high-grade gliomas with machine learning. *Physica Medica* **107,** 102538 (2023).

12. Foltyn-Dumitru, M. *et al.* Impact of signal intensity normalization of MRI on the generalizability of radiomic-based prediction of molecular glioma subtypes. *European Radiology* **34,** 2782–2790 (2024).

13. Carré, A. *et al.* Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Scientific Reports* **10,** 12340 (July 2020).

14. Fatania, K. *et al.* Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review. *European Radiology* **32,** 7014–7025 (Oct. 2022).

15. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

16. Harrell, F. *Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements* 2023.

17. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative (2016).

18. Upadhaya, T., Morvan, Y., Stindel, E., Le Reste, P.-J. & Hatt, M. Prognosis classification in glioblastoma multiforme using multimodal MRI derived heterogeneity textural features: impact of pre-processing choices. *Medical Imaging 2016: Computer-Aided Diagnosis* **9785,** 97850W (2016).

19. Um, H. *et al.* Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Physics in Medicine & Biology* **64,** 165011 (Aug. 2019).

20. Vils, A. *et al.* Radiomic Analysis to Predict Outcome in Recurrent Glioblastoma Based on Multi-Center MR Imaging From the Prospective DIRECTOR Trial. *Frontiers in Oncology* **11,** 1–9 (2021).

21. Riley, R. D. & Collins, G. S. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal* **65,** 1–22 (2023).

22. Dinsdale, N. K., Jenkinson, M. & Namburete, A. I. L. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 369–378 (2020).

23. Chelliah, A. *et al.* Glioblastoma and radiotherapy: A multicenter AI study for Survival Predictions from MRI (GRASP study). *Neuro-Oncology* **26,** 1138–1151 (June 2024).

24. Kim, H. E. *et al.* Transfer learning for medical image classification: a literature review. *BMC Medical Imaging* **22,** 69 (Dec. 2022).

25. Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine* **140,** 105111 (Jan. 2022).

26. Hallowell, N., Badger, S., Sauerbrei, A., Nellåker, C. & Kerasidou, A. "I don't think people are ready to trust these algorithms at face value": trust and the use of machine learning algorithms in the diagnosis of rare disease. *BMC Medical Ethics* **23,** 112 (Nov. 2022).

27. Riley, R. D., van der Windt, D., Croft, P. & Moons, K. G. *Prognosis Research in Healthcare: Concepts, Methods, and Impact* (eds Riley, R. D., van der Windt, D. A., Croft, P. & Moons, K. G.) (Oxford University Press, Oxford, UK, 2019).

28. Karschnia, P. *et al.* Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO resect group. *Neuro-Oncology* **25,** 940–954 (May 2023).

29. Karschnia, P. *et al.* Surgical management and outcome of newly diagnosed glioblastoma without contrast enhancement (low-grade appearance): a report of the RANO resect group. *Neuro-Oncology,* 1–12 (2023).

30. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23,** 1231–1251 (Aug. 2021).

31. Calabrese, E. *et al.* The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. *Radiology: Artificial Intelligence* **4,** 2–6 (Nov. 2022).

32. Bakas, S. *et al.* The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Scientific Data* **9,** 453 (July 2022).

33. Pati, S. *et al.* Federated Learning Enables Big Data for Rare Cancer Boundary Detection (2022).

34. Da-Ano, R., Visvikis, D. & Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Physics in Medicine & Biology* **65,** 24TR02 (Dec. 2020).

35. Traverso, A., Dankers, F. J. W. M., Osong, B., Wee, L. & van Kuijk, S. M. J. in (eds Kubben, P., Dumontier, M. & Dekker, A.) 121–133 (Springer International Publishing, Cham, 2019).

36. Fatania, K. *et al.* Tumour Size and Overall Survival in a Cohort of Patients with Unifocal Glioblastoma: A Uni- and Multivariable Prognostic Modelling and Resampling Study. *Cancers* **16,** 1301 (Mar. 2024).

37. Davatzikos, C. *et al.* AI-based prognostic imaging biomarkers for precision neuro-oncology: the ReSPOND consortium. English. *Neuro-Oncology* **22,** 886–888 (June 2020).

38. De Godoy, L. L., Chawla, S., Brem, S. & Mohan, S. Taming Glioblastoma in "Real Time": Integrating Multimodal Advanced Neuroimaging/AI Tools Towards Creating a Robust and Therapy Agnostic Model for Response Assessment in Neuro-Oncology. *Clinical Cancer Research* **29,** 2588–2592 (July 2023).

39. Galldiks, N., Lohmann, P., Fink, G. R. & Langen, K.-J. Amino Acid PET in Neurooncology. *Journal of Nuclear Medicine* **64,** 693–700 (May 2023).

40. Celli, M. *et al.* Diagnostic and Prognostic Potential of 18F-FET PET in the Differential Diagnosis of Glioma Recurrence and Treatment-Induced Changes After Chemoradiation Therapy. *Frontiers in Oncology* **11,** 1–10 (2021).

41. Schön, S. *et al.* Imaging glioma biology: spatial comparison of amino acid PET, amide proton transfer, and perfusion-weighted MRI in newly diagnosed gliomas. *European Journal of Nuclear Medicine and Molecular Imaging* **47,** 1468–1475 (2020).

42. Yan, K. *et al.* Assessing Amide Proton Transfer (APT) MRI Contrast Origins in 9 L Gliosarcoma in the Rat Brain Using Proteomic Analysis. *Molecular Imaging and Biology* **17,** 479–487 (2015).

Research Database: Leeds Cancer Centre
Computer Aided Theragnostics (LeedsCAT)
v1.0
REC reference: 19/YH/0300
IRAS project ID: 255585

Radiotherapy Research Department
Level 4 Offices, Bexley Wing
Leeds Teaching Hospitals NHS Trust
Leeds
LS97TF

DATE 20/11/2021

Dear Kavi and Stuart

RE project:
1) HarMOnAE: Harmonisation of MR imaging for Oncology with style-blind AutoEncoders
2) INTRIGUE - INvestigating Treatment Response In Glioblastoma Using radiomic Evaluation

Your project has been considered by the LeedsCAT Governance board on 18/11/2021. The LeedsCAT Governance board consists of representatives from Research and Innovation, Information Governance, PPI and experts in Radiotherapy.

A favourable decision was made and we can confirm that we are able to approve your project within the scope of the LeedsCAT research database ethical approval.

As you indicated that no patient data is to leave LTHT no further Information Governance will be needed. In addition approval by the LeedsCAT Governance Board means there is no requirement to have HRA approval.

With respect to the use of data we expect that you will comply with GDPR, Caldicott guidance, Information Governance procedures and all Trust policies. If there are any significant changes to the project, including change in the list of people who will access project data, you will need to notify the LeedsCAT project manager. You will be routinely asked every 6 months to provide a project update, including changes to the project form and any outputs from the project.

Regards,

John Lilley
on behalf of the LeedsCAT Governance Board