Glioblastoma and Radiotherapy: a multi-center AI study for Survival Predictions from MRI (GRASP study)

Authors:

Alysha Chelliah, David A Wood, Liane S Canas, Haris Shuaib, Stuart Currie, Kavi Fatania, Russell Frood, Chris Rowland-Hill, Stefanie Thust, Stephen J Wastling, Sean Tenant, Karen Foweraker, Matthew Williams, Qiquan Wang, Andrei Roman, Carmen Dragos, Mark MacDonald, Yue Hui Lau, Christian A Linares, Ahmed Bassiouny, Aysha Luis, Thomas Young, Juliet Brock, Edward Chandy, Erica Beaumont, Tai-Chung Lam, Liam Welsh, Joanne Lewis, Ryan Mathew, Eric Kerfoot, Richard Brown, Daniel Beasley, Jennifer Glendenning, Lucy Brazil, Angela Swampillai, Keyoumars Ashkan, Sébastien Ourselin, Marc Modat, Thomas C Booth

Affiliations:

1. King's College London, London, United Kingdom (A.C., D.A.W., L.S.C., H.S., A.B., A.L., E.K., R.B., D.B., K.A., S.O., M.M., T.C.B.)

2. Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom (H.S., A.R., M.M., C.A.L., T.Y., D.B., L.B., A.S.)

3. Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom (S.C., K.F., R.F., R.M.)

4. Hull University Teaching Hospitals NHS Trust, England, United Kingdom (C.R-H.)

5. University College London Hospitals NHS Foundation Trust, London, United Kingdom (S.Th., S.J.W.)

6. University College London, London, United Kingdom (S.Th., S.J.W.)

7. Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom (S.Th., K.F.)

8. University of Nottingham, Nottingham, United Kingdom (S.Th.)

9. The Christie NHS Foundation Trust, Withington, Manchester, United Kingdom (S.Te.)

10. Imperial College Healthcare NHS Trust, London, United Kingdom (M.W., Q.W.)

11. Imperial College London, London, United Kingdom (M.W., Q.W.)

12. Oncology Institute Prof. Dr. Ion Chiricuta, Cluj-Napoca, Romania (A.R.)

13. Buckinghamshire Healthcare NHS Trust, Amersham, United Kingdom (C.D.)

14. King's College Hospital NHS Foundation Trust, London, United Kingdom (Y.H.L., A.L., K.A., T.C.B.)

15. Mansoura University, Mansoura, Egypt (A.B.)

16. Brighton and Sussex University Hospitals NHS Trust, England, United Kingdom (J.B., E.C.)

17. Lancashire Teaching Hospitals NHS Foundation Trust, England, United Kingdom (E.B., T-C.L.)

18. The Royal Marsden NHS Foundation Trust, London, United Kingdom (L.W.)

19. Newcastle upon Tyne Hospitals NHS Foundation Trust, England, United Kingdom (J.L.)

20. University of Leeds, Leeds, UK (R.M.)

21. Maidstone and Tunbridge Wells NHS Trust, Kent, United Kingdom (J.G.)

Running title: Predicting glioblastoma survival post-radiotherapy

Corresponding author:

Thomas C Booth

School of Biomedical Engineering & Imaging Sciences, King's College London, St Thomas' Hospital, London SE1 7EH, UK

+447977533022

thomas.booth@kcl.ac.uk

Manuscript word count: 4,006 words

Abstract

Background

The aim was to predict survival of glioblastoma at eight months after radiotherapy (a period allowing for completing a typical course of adjuvant temozolomide), by applying deep learning to the first brain MRI after radiotherapy completion.

Methods

Retrospective and prospective data were collected from 206 consecutive glioblastoma, IDH-wildtype patients diagnosed between March 2014-February 2022 across 11 UK centers. Models were trained on 158 retrospective patients from three centers. Holdout test sets were retrospective (n=19; internal validation), and prospective (n=29; external validation from eight distinct centers).

Neural network branches for T_2-weighted and contrast-enhanced T_1-weighted inputs were concatenated to predict survival. A non-imaging branch (demographics/MGMT/treatment data) was also combined with the imaging model. We investigated the influence of individual MR sequences; non-imaging features; and weighted dense blocks pretrained for abnormality detection.

Results

The imaging model outperformed the non-imaging model in all test sets (area under the receiver-operating characteristic curve, AUC p=0.038) and performed similarly to a combined imaging/non-imaging model (p>0.05). Imaging, non-imaging, and combined models applied to amalgamated test sets gave AUCs of 0.93, 0.79, and 0.91. Initializing the imaging model with pretrained weights from 10,000s of brain MRIs improved performance considerably (amalgamated test sets without pretraining 0.64; p=0.003).

Conclusions

A deep learning model using MRI images after radiotherapy, reliably and accurately determined survival of glioblastoma. The model serves as a prognostic biomarker identifying patients who will not survive beyond a typical course of adjuvant temozolomide, thereby stratifying patients into those who might require early second-line or clinical trial treatment.

Key words: glioblastoma; survival; artificial intelligence; deep learning; magnetic resonance imaging

Key Points:

- A deep learning model predicted post-radiotherapy survival of glioblastoma from MRIs.
- An imaging model was generalizable on internal and prospective external test data.
- Performance was considerably better when initial weights were pretrained on 10,000s of MRIs.

Importance of the Study:

- A deep learning model that used MRI images after radiotherapy, and that was pretrained on 10,000s of brain MRIs, reliably and accurately determined survival of isocitrate dehydrogenase (IDH) wildtype glioblastoma patients after radiotherapy.

Introduction

Glioblastoma is the most aggressive adult primary brain cancer[1]. MRI plays a key role in diagnosis, treatment planning, and treatment response assessment[2]. MRI images can also act as prognostic biomarkers with studies predicting survival from pre-operative MRIs using classical[3] and deep[4,5] machine learning models. However, by the time radiotherapy finishes, considerable intervention potentially confounds survival predictions obtained at the pre-operative time point. Survival predictions from images obtained after radiotherapy could be more accurate. To our knowledge, machine learning has not been applied to the first MRI images after radiotherapy completion to identify patients who will not survive beyond a typical course of adjuvant temozolomide (TMZ). In this scenario, an accurate and generalizable prognostic biomarker would stratify patients into those requiring early second-line treatment or clinical trial enrollment. Additionally, all subsequent tumor boards held during the course of adjuvant TMZ would have an accurate *a priori* survival prediction, therefore improving management decision confidence. This is relevant as often follow-up imaging findings are non-specific and treatment response assessment is not definitive; even when findings are specific, utility is based on low-level evidence[6].

Optimal treatment involves surgical resection, followed by radiotherapy with concomitant TMZ, then adjuvant TMZ[7,8] (see Appendix A for an illustration of treatment and imaging pathways). Modified treatment may be planned for patients who are elderly or have tumors in eloquent areas, or who cannot tolerate optimal treatment[1,2]. This often includes a shorter course and lower dose of radiotherapy, where a longer course of adjuvant chemotherapy may be prescribed. Whilst 99% of US patients ≥66 years undergoing post-surgical treatment receive radiotherapy, just 57% receive TMZ[9]. Only 34% of UK patients between 20-70 years complete optimal treatment[1]. To inform patient management, MRIs are often performed after initial surgery, during radiotherapy planning, and at 2-3 monthly intervals (or if clinically deteriorating) during subsequent follow-up[2,9,10,11]. However imaging studies, including those predicting survival[3-5,12], typically sample patients only from the optimally-treated population limiting biomarker applicability. The unmet need to improve outcomes of patients undergoing modified treatment, highlighted at national strategic level[13,14] (and study stakeholder

feedback; Appendix B), motivated our biomarker design to be applicable to both optimal and modified treatment populations.

This study aimed to apply deep learning to the first brain MRI after radiotherapy, in glioblastoma, IDH-wildtype[15] patients undergoing optimal or modified treatment, to predict survival at eight months after completing radiotherapy (a period allowing for completion of a typical course of adjuvant TMZ). For imaging-based biomarkers to be valuable in the clinic, it is rational that predictions should either be more accurate than those derived from freely available non-imaging information known to be associated with poorer patient survival, or are enhanced when combined. We hypothesized that prediction based on imaging would outperform prediction using only available non-imaging information (demographic, pathological, and treatment-related variables).

Methods

Study reporting followed the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[16]. The UK's Health Research Authority provided ethical approval (ref:18/LO/1873); data were anonymized before analyses.

Patient characteristics

Patient cohort

This study included consecutive retrospective and prospective data from 11 ZGBM (zeugmatography for glioblastoma) consortium centers[17], with diagnoses between March 2014 and February 2022 (a CONSORT diagram displaying the flow of patients included in analyses is presented in Appendix C). The study was pragmatic; imaging regimens were not standardized and were expected to vary over centers and time[18]. Inclusion criteria consisted of adults diagnosed with glioblastoma, IDH-wildtype[15]; who underwent radiotherapy after first surgery; and subsequent MRI with contrast-enhanced T_1-weighted (T1c) and T_2-weighted (T2) sequences; and could be identified as being deceased or not at eight months post-radiotherapy (labeled as short-term or long-term survival, respectively).

Long-term survivors who received second-line or trial treatment within eight months were excluded to prevent confounding from that treatment. As the classifier is designed to help decision-making on expediting early trial or second-line treatment, we excluded those rare patients whose first post-radiotherapy MRI occurred either after second-line treatment started (to prevent confounding), or beyond 24 weeks (arbitrary time threshold). T1c and T2 sequences were selected to maximize the clinical applicability of developed models, as these are acquired in routine clinical settings[18] and were available for all patients in this cohort. It should be noted that other MR sequences such as FLAIR are informative images and are commonly acquired. However, 18.5% (23/124) of patients in the largest retrospective cohort (the KCH cohort) reported here did not have FLAIR imaging during the first post-radiotherapy MRI study.

Of 206 patients included (Table 1), 64 (31.1%) were short-term survivors (<8 months survival). The amalgamated test set consisted of all prospective external data (henceforth *prospective test set*; n=29) and 10.7% of holdout retrospective data (*retrospective test set*; n=19/177). Stratified sampling into training and test sets was performed on retrospective data to avoid bias from imbalances in survival outcome and MRI acquisition dimensionality across sites. We sampled 89.3% of retrospective patients (n=158/177) as the training set, and the remaining were held out for testing. No further variables were stratified due to low patient numbers after controlling for three variables. Description of sample sizes and sampling error associated with survival outcome, acquisition dimension, and variables associated with survival (including age, initial surgery type, and MGMT methylation status) are presented in Appendix D1.

Co-variates

Non-imaging information associated with poorer survival includes patients who are older (>60 years), or have tumors which are unmethylated, have minimal O^6-methylguanine-DNA methyltransferase (MGMT) methylation, are deep-seated (midbrain/thalamus/callosum) or have undergone biopsy alone[19-22]. These, and other demographic, histologic, tumor-related, and prior treatment variables were included in non-imaging models (Table 2). Of available data, the Eastern Cooperative Oncology Group (ECOG)[23] performance status did not differ between short-term and long-term survivors within KCH training patients ($p>0.05$) (median=0; range=0-2); such formal assessments are not regularly administered and, when applied, can be subjective in choice and nature[6]. Performance status was therefore excluded. Mean/mode imputation was used for missing data; labels were added identifying imputed inputs. Numeric attributes were standardized to unit variance using training data. Categoric variables were one-hot encoded. MGMT methylation was handled in two ways. Firstly, a numeric variable identified the MGMT methylation percentage. Secondly, three distinct categoric variables were added identifying if patients had methylated, unmethylated, or unknown (missing) MGMT methylation status. Distributions of non-imaging variables were compared between short-term and long-term survivors using Mann-Whitney U and Chi-squared tests. Significance was set at $p\leq0.05$ for all analyses.

Non-imaging models

Machine learning models (logistic regression, linear and gaussian support vector classifiers (SVC), and decision tree classifiers) were applied to training data with sequential feature selection using scikit-learn[24]. Tuned parameters were logistic regression and SVC regularization parameters, gaussian SVC gamma coefficients, and decision tree gini and entropy criteria. We also applied fully-connected neural networks to non-imaging features alone (Appendix E1).

Imaging and combined models

Whole-brain T1c and T2 images were co-registered and minimally pre-processed using a similar approach to that for a model[25,26] applied for pretraining. MRI inputs were converted from DICOM into NIfTI format. T2 scans were registered to the corresponding T1c image for each patient and MRI study. Images were resampled to common voxel sizes (1 mm^3), and subsequently cropped or padded to a final 3D array of shape 130 x 130 x 130 for inputs to deep learning models. Resampling was performed to address differences in slices thickness and spacing between images. Cropping/padding was performed to preserve aspect ratios of images when resizing to the final shape. Image pre-processing was conducted with niftyreg[27] and MONAI[28].

Network architectures

Model architectures (Fig.1a) were modified from DenseNet121[29] and abnormality detection models[25,26] (Appendix E2 describes an alternative architecture considered). Input images were the final 3D array of shape 130 x 130 x 130. Dense blocks were initialized with weights pretrained on a large dataset containing all neurological abnormalities (10,695 and 50,523 T1c and T2 scans, respectively). The T1c-branch has four pretrained dense blocks. Outputs are flattened to a 1x1920-dimensional vector via pooling, then passed through two linear layers (providing prediction

probabilities). The T2-branch performs the analogous process for T2 inputs. Outputs from the first linear layer per branch are concatenated *(merged branch*); this vector is passed through a linear layer that outputs a 1x2-dimensional vector with prediction probabilities. Since each branch can predict survival separately, distinct loss functions are applied per branch. Outputs from the merged branch were selected as final predictions.

A separate *combined* model adds a non-imaging branch with 1x27-dimensional inputs alongside the T1c and T2 branches (Fig.1b); the non-imaging branch of this combined model additionally included the duration between radiotherapy completion and imaging (Table 2). The merged prediction is obtained by concatenating T1c, T2, and non-imaging vectors.

Final (hyper-)parameters of model training and tuning (Appendix E3) were selected by mean validation area under the receiver-operating characteristic curve (AUC) across training folds. All models incorporating imaging were developed with both PyTorch[30] and the PyTorch-based MONAI[28] framework.

Test set analysis

Five-fold cross-validation was used on training data (stratified by outcome/dimension/center) (Appendix D1). To determine generalizability, individual imaging, non-imaging, and combined models were trained on all training data and assessed on holdout test data.

To check for dependencies between features and outcomes, a permutation test was performed with test set inputs per patient shuffled before determining model performance. Ablation studies were conducted to investigate the relative importance of individual branches and use of pretrained weights. Model explainability was further pursued using a guided backpropagation approach[25] modified to obtain saliency maps from merged branch weights and multiple sequences. As an overview, guided backpropagation is intended to highlight regions of input images which, if modified slightly, would change predictions obtained from the model. The method returns gradient arrays that match dimensions of the original 3D input images. For visualization purposes of volumetric saliency maps,

axial slices that most contributed to model survival predictions were automatically selected and presented, following the methodology reported by Wood et al. (2022)[25].

The primary outcome measure was AUC. We used DeLong's test to compare model performances (pROC R package)[31]. Subgroup analyses considered retrospective/prospective collection, surgery type, age (>60years), sex, and acquisition dimension. Code is available at https://github.com/lyshc/glioblastoma-survival-classifier.

Results

Patient characteristics

The dataset included 206 consecutive patients (Tables 1 and 2; Appendix C). The mean age was 57.4 (standard deviation: 10.6); 72 patients were female and 134 were male. Missing data for at least one variable (age, MGMT status, MGMT methylation percentage, radiotherapy dose, or TMZ dose) were noted in 57/206 (27.7%) patients. For 13 patients, the MGMT status was known while the exact methylation percentage was missing (methylated, n=7; unmethylated, n=6); the percentage was imputed based on the mean percentage for other patients with the same methylation status.

Longer survival was associated with tumors that have higher MGMT methylation percentage, are not deep-seated, are resected and undergo Stupp dose radiotherapy and TMZ (Table 2), supporting prior research[19-22]. It was also related to having a later post-radiotherapy MRI.

Non-imaging models

Among all non-imaging machine learning models, logistic regression with reduced features was selected as the optimal classifier based on the highest validation AUC. The optimal logistic regression model had regularization parameter (C) set to 1.0 and ten features retained (male sex, methylated MGMT status, unmethylated MGMT status, unknown MGMT status, initial biopsy, initial resection, standard radiotherapy dose, reduced radiotherapy dose, reduced TMZ dose, and no TMZ). These were all one-hot encoded categoric variables (for example, separate variables encoded if a patient had methylated, unmethylated, or unknown MGMT status). The AUCs for retrospective, prospective and amalgamated test sets were 0.76, 0.78 and 0.79, respectively (Table 3); performances did not differ between test sets (all $p>0.05$). To aid with assessments of model performances and generalizability across test sets, Figure 2 shows receiver-operating characteristic (ROC) curves for all models (imaging, combined, and non-imaging) on the amalgamated, retrospective, and prospective test sets.

Imaging and combined models

Parameters used to optimize the imaging model are shown in Appendix F1. Initializing the imaging model with pretrained weights from 10,000s of brain MRIs[25,26] improved performance considerably (with and without pretraining on amalgamated test set gave AUCs of 0.93 and 0.64 respectively; p=0.003). Therefore, performances of imaging (and combined) models initialized with pretrained weights are reported (Table 3). The imaging model AUCs for retrospective, prospective and amalgamated test sets were 0.92, 0.93 and 0.93, respectively, and did not differ in performance between sets (p>0.05) (Figure 2).

For the combined model, AUCs for retrospective, prospective and amalgamated test sets were 0.94, 0.89 and 0.91, respectively; performances did not differ across test sets (p>0.05).

All models applied a survival classification threshold of 0.50; an analysis of decision threshold selection is presented in Appendix F2. Description of the interval between radiotherapy completion and the first post-radiotherapy MRI study for patients in the amalgamated test set is presented in Appendix G1.

Model comparison

One way for imaging-based biomarkers to be valuable in the clinic is that, when compared to freely-available non-imaging biomarkers, there is an incremental increase in predictive accuracy when biomarkers are combined. An incremental increase in performance was not clearly proven for the combined model. We found that whilst there was a trend for enhanced performance in the amalgamated test set (AUC 0.91 vs 0.79, p=0.07), in retrospective and prospective test sets this was less clear (Figure 2) (p=0.11 and p=0.16).

Another, plausibly optimal, way for imaging-based biomarkers to be valuable clinically is that, when compared to freely-available non-imaging biomarkers, the predictive accuracy is higher. The advantage of using an imaging model alone is that it can be applied in isolation, without needing additional information gathering. The imaging model outperformed the non-imaging model in amalgamated and prospective test sets (AUC, $p \leq 0.05$) (Table 3 and Figure 2). However, performances did not significantly differ on the retrospective test set ($p=0.14$); comparison of receiver-operating characteristic curves suggest that this may be related to the smaller retrospective test set size (retrospective test n=19) (Appendix G2). The combined model was not superior to the imaging model in any test set ($p>0.05$), despite the combined model incorporating information on the interval between radiotherapy completion and follow-up imaging (the interval was different in the two groups). To further assess whether the model could complement evaluations made in routine hospital settings, we performed a comparison against expert clinical raters reported in Appendix H.

Imaging model explainability

Based on the findings that available non-imaging features did not improve predictive performances, and that the combined model was not superior to the imaging model, the imaging model was selected over non-imaging and combined counterparts for further analysis. ROC curves showing results from the permutation test and ablation studies are provided in Figure 3. Model performances are plotted separately for sample subgroups (initial surgery type, age group, sex, and T1c acquisition dimension; Figure 3). Further detail on imaging model results from the permutation test and ablation studies, along with performances disaggregated for sample subgroups is provided in Appendix G2. The permutation test AUC of 0.49 indicates that the model was not performing by chance.

Ablation studies showed that test set performance using the merged branch was similar to using the T2 branch alone (comparison of AUCs across amalgamated test set, $p=0.19$), but better than the T1c branch alone ($p=0.048$). Performances were similar when using only one sequence (T1c versus T2 branches, $p=0.41$). Together, this suggests that on the rare occasion that a patient does not receive

gadolinium (for example, due to high-grade renal failure, or patient refusal), predictions may remain accurate with only the T2 sequence. We found that test set performance dropped considerably when not training with transfer learning, where initial weights were pretrained on a brain MRI dataset x100 larger than the training dataset (AUCs with and without pretraining 0.93 and 0.64, respectively; p=0.003). This shows that medical image classifiers with high-dimensional and high-resolution inputs such as brain MRIs may benefit from pretraining on larger datasets.

Saliency maps based on predicted survival outcome from the imaging model are presented in Figure 4. These show examples of short-term and long-term survivors from retrospective and prospective external test sets, along with erroneous predictions of both survival outcomes. Across patients, there appears to be variation in the location, size and number of brain areas that are salient. For example, some maps seemingly display coarse localization of tumor regions, as well as ventricles. It is plausible that it may be more difficult to interpret appearances associated with long-term compared to short-term survival in MRIs and saliency maps (i.e., to identify the absence of expected deterioration). Nonetheless we can make some tentative observations. Patient 2, for example, was correctly predicted to have subsequent long-term survival. In this case the presented slices suggest relatively greater contribution from ventricular areas than the treated tumor region. This suggests that both tumor and non-tumor regions provide informative features for deep learning models, and jointly contribute to survival predictions. Among misclassified patients, it is conceivable that model weights associate ventriculomegaly with short-term survival (for example, patient 5). Further analysis of saliency maps is presented in Appendix I. However, it should be noted that saliency maps alone do not identify features that are easily interpretable to human readers[32].

Discussion

We present the first known model that uses imaging to distinguish short-term and long-term survivors within eight months of completing radiotherapy. Eight months represents the period of time to complete adjuvant chemotherapy. Using a multi-center cohort we built a model with T1c and T2 inputs. The transfer learning approach improved predictions. There was no clear benefit of generating predictions with non-imaging data. Using the T2 scan alone was not inferior to using both sequences. The imaging model seemed to generalize both to retrospective and external, prospective test data.

One strength of this study is providing insight into the extent to which neural networks predicting post-treatment survival generalized across multiple external centers. External, prospective sites showed a higher proportion of 2D scanning and short-term survivors than retrospective data which may have been a potential source of bias. Therefore, we stratified data to allow better evaluation of predictions on short-term and long-term survivors, and both 2D and 3D acquisitions. Based on similar performances across the retrospective and external prospective datasets, the imaging model may be robust to variations in imaging protocols and class imbalances.

Another key contribution is the finding that transfer learning can offer a strong benefit to models with large numbers of parameters and small training samples. This accords with other research evaluating benefits of transfer learning for MRIs of glioma patients. For example, one study combining low-grade and high-grade gliomas found that pre-training improved classification accuracy of a deep learning radiogenomic model[33]. Another study combined classical radiomics features with those extracted from a pre-trained neural network to predict overall survival of glioblastoma[34]. These studies used natural images for pre-training, and predicted outcomes from cropped 2D slices of tumor regions from pre-operative MRIs. In comparison, the model used for pre-training in our study was trained on thousands of brain MRIs and was highly successful at detecting abnormalities[25,26].

Previous research that successfully applied machine learning to predict survival of glioblastoma has largely focused on pre-treatment timepoints. One study used a DenseNet-based network with multiple branches to predict three-year survival from 2D T1c and T2 slices[4]. Another applied a neural network

to quantify the temporalis muscle; this predicted survival in distinct datasets[5]. Several studies with multi-center data extracted radiomics features from pre-operative tumor segmentations and applied machine learning to predict survival[3]. To our knowledge, prior studies have not demonstrated benefits of classical or deep machine learning methods on predicting outcomes from post-treatment timepoints, and with whole-brain inputs requiring minimal pre-processing.

Our imaging model is a contribution towards developing networks that could be applied to aid decision-making in hospitals. The two-year survival rate of glioblastoma is just 18%[35]. Such models could prompt closer MRI surveillance of suspected short-term survivors, compared with patients expected to show initial treatment response. Large prospective studies replicating high predictive performances in clinical settings are now desirable. If validated, studies assessing improvements to patient management are required. Researchers could also investigate extending model applicability using, for example, curated second-line therapy trial datasets.

Our model predicted post-radiotherapy *survival* using imaging as a prognostic biomarker which can be used to stratify patients into those requiring early second-line treatment or trial enrollment. An alternative model might predict tumor *treatment response* using imaging as a monitoring biomarker[6]. While not the focus of our study which incorporates all patients consecutively (including complete response, partial response, stable disease, progression, and pseudoprogression), interpreting post-radiotherapy structural MRIs in clinical settings is typically challenging due to difficulty in distinguishing recurrent disease from treatment-related effects – particularly for pseudoprogression[2,3,6,11,17,36,37,38]. However, labelling progression – and pseudoprogression – requires availability of repeated T1c imaging obtained in a timely manner per patient, accompanied by accurate measurements of bidirectional diameters of contrast-enhancing tumors[37-38]. Prior research has reported that there can be substantial inter-rater variability in these measurements, however, which can confound evaluations of treatment response[39-41]. One reason for measurement variability is the irregular shape at the tumor margin[3], whilst another relates to similarities in signal intensity between tumor and non-tumor if pre-contrast T1c scans are not studied carefully[42]. To rule out factors that potentially confound assessments, data on prescribed steroids and longitudinal patient symptom

profiles are additionally needed. In contrast, the approach presented here uses overall survival as the reference standard, free from inter-rater variability and requirements for RANO-compliant longitudinal data collection. Our study was not designed to identify the first occurrence of true tumor progression (and thereby rule out pseudoprogression, which is expected to be associated with longer survival). However, our approach has the potential to provide all tumor boards monitoring patients at all time periods after radiotherapy with an accurate *a priori* survival prediction gained at the first post-radiotherapy scan, thereby improving management decision confidence, including for example, the challenging scenario of pseudoprogression.

While predictions did not improve when incorporating non-imaging features, we had a limited number of these variables. Combined models with a greater range of tumor-related data might show better performances (e.g., Ki67 percentage, ATRX status, genomic variables). Models could also integrate earlier MRI studies which may contain useful features for improving prognostic predictions, for example pre-surgical and pre-radiotherapy studies. For now, a model that could translate most easily across centers would likely benefit from a pragmatic approach that requires collecting widely available non-imaging features and cross-sectional (rather than longitudinal) imaging.

A potential limitation is that we did not consider other MRI sequences that may provide insights into tumor recurrence (e.g., diffusion or perfusion imaging)[43]. However, our models used T1c and T2 sequences to maximize clinical utility and translation across hospitals. These sequences were consistently acquired at all centers; conversely, more advanced MRIs are less commonly available[17,18]. Incorporating other anatomical sequences desirable for brain tumor imaging, such as FLAIR sequences, was not also pursued as it would have reduced the patient cohort in this UK-based study where FLAIR imaging was not always performed. A downstream constraint of building a model without the most common MRI sequences is that it reduces the potential for clinical translation. Nonetheless, future models could investigate the extent to which models built with alternative imaging protocols (for example, advanced imaging as well as FLAIR) can predict post-treatment survival.

Another limitation is that we used a small dataset whereas DenseNet[29] is a large model and whole-brain images provide many inputs per patient. Beyond pretraining, future research could use smaller inputs, e.g., bounding boxes cropped to initial tumor sites. This was not pursued because: (i) extracranial information is linked with overall survival[5]; (ii) contrast-enhancing masses remote to the initial site signal recurrence (and shorter survival); (iii) data pre-processing that aligned with pretraining pre-processing was favored[25]; and (iv) whole-brain images require minimal pre-processing (plausibly reducing barriers to translation).

In this multi-center study, we developed a model that predicts survival within eight months of completing radiotherapy. The model is intended for use for patients undergoing optimal treatment as well as the under-studied cohort of patients undergoing modified treatments. A neural network with T1c and T2 branches showed generalizable classification on both retrospective and external, prospective test cohorts. If validated in large prospective studies, such approaches could be used to distinguish patients who show initial response to radiotherapy from those requiring closer image-based monitoring and second-line treatments (or termination of ineffective treatment).

Funding

A.C. is supported by the UK Medical Research Council (MR/N013700/1) and King's College London, MRC Doctoral Training Partnership in Biomedical Sciences. L.S.C. is supported by the Wellcome Trust (215010/Z/18/Z). S.C. is supported by the Leeds Hospitals Charity and Cancer Research UK RadNet. K.Fa. is supported by the Wellcome Trust (203914/Z/16/Z). S.Th. is supported by Cancer Research UK and the Medical Research Council / BeiGene. M.W. is supported by the National Institute for Health and Care Research Imperial Biomedical Research Centre, the Brain Tumour Charity, Macmillan Cancer Care, and Novocure. T-C.L is supported by the Innovation and Technology Commission – Partnership Research Program (PRP/067/20Fx) with Roche, Hong Kong. R.M. is supported by the National Institute for Health and Care Research, Yorkshire's Brain Tumour Charity, and Candlelighters. T.C.B., S.O. and D.W. are supported by the UK Medical Research Council (MR/W021684/1). Development of deep learning networks was enabled by the JADE2-HPC cluster, supported by the EPSRC (EP/T022205/1). This work was also supported by the Wellcome EPSRC Centre for Medical Engineering at King's College London (203148/Z/16/Z) (including authors E.K., R.B., S.O., and T.C.B.).

Conflict of Interest

There is no conflict of interest for all authors as a consortium. A.C. - None declared. D.A.W - None declared. L.S.C. - None declared. H.S. - None declared. S.C. - None declared. K.Fa. - None declared. R.F. - None declared. C.R-H. - None declared. S.Th. - None declared. S.J.W. - None declared. S.Te. - None declared. K.Fo. - None declared. M.W. - stock and other interests: PearBio. Q.W. - None declared. A.R. - None declared. C.D. - None declared. M.M. - None declared. Y.H.L. - None declared. C.A.L. - None declared. A.B. - None declared. A.L. - None declared. T.Y. - None declared. J.B. - None declared. E.C. - None declared. E.B. - None declared. T-C.L. - None declared. L.W. - None declared. J.L. - None declared. R.M. – consultancy: Brainlab, Stryker; payment/honoraria: Baxter, Roswell Comprehensive Cancer Centre, Zeiss; support for attending meetings/travel:

Brainlab, Roswell Comprehensive Cancer Centre, Zeiss; patents: UK patent office; unpaid leadership/fiduciary role: Oscar's Paediatric Brain Tumour Charity, TJBCM-BTR NTA; shareholding: Opto Biosystems, RBM Healthcare, Assemblify; clinical advisor: MHRA. E.K. - None declared. R.B. - None declared. D.B. - None declared. J.G. - None declared. L.B. - None declared. A.S. - None declared. K.A. - None declared. S.O. - consultancy: Proximie, Avatera Medical; stock: Hypervision Surgical Ltd. M.M. - None declared. T.C.B. – consultancy: Microvention; payment/honoraria for education lectures: Siemens Healthineers Speakers Bureau, Medtronic Speakers Bureau; support for attending meetings/travel: Balt.

Authorship

Conception and design: A.C., M.M, T.C.B. Data acquisition and preparation: A.C., H.S., S.C., K.F., R.F., C.R-H., S.T., S.J.W., S.T., K.F., M.W., Q.W., A.R., C.D., M.D., Y.H.L., C.A.L., A.B., A.L., T.Y., J.B., E.C., E.B., T-C.L., L.W., J.L., R.M., D.B., J.G., L.B., A.S., K.A., S.O., T.C.B. Data analysis: A.C., D.A.W., L.S.C., E.K., R.B., M.M., T.C.B. Manuscript drafting: A.C., D.A.W., L.S.C., M.M., T.C.B. Data interpretation, critical review of the work and manuscript, final approval of manuscript, accountability for all aspects of the work: all authors.

Data Availability

Data generated or analyzed during the study are available from the corresponding author by request.

Acknowledgements

We thank Giusi Manfredi, Dijana Vilic, Sharaf Ayinla, Bernice Akpinar, and Rachel Daniel for their contributions to this study.

References

1. Brodbelt A, Greenberg D, Winters T, Williams M, Vernon S, and Collins VP. Glioblastoma in England: 2007–2011. European Journal of Cancer 2015;51:533–542. doi: 10.1016/j.ejca.2014.12.014

2. Weller M, Van Den Bent M, Tonn JC, Stupp R, Preusser M, Cohen-Jonathan-Moyal E, Henriksson R, Le Rhun E, Balana C, Chinot O, Bendszus M. European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas. The Lancet Oncology 2017;18:e315–e329. doi: 10.1016/S1470-2045(17)30194-8

3. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H. Machine learning and glioma imaging biomarkers. Clinical radiology. 2020;75:20-32. doi: 10.1016/j.crad.2019.07.001

4. Fu X, Chen C, Li D. Survival prediction of patients suffering from glioblastoma based on two-branch DenseNet using multi-channel features. International Journal of Computer Assisted Radiology and Surgery. 2021;16:207-17. doi: 10.1007/s11548-021-02313-4

5. Mi E, Mauricaite R, Pakzad-Shahabi L, Chen J, Ho A, Williams M. Deep learning-based quantification of temporalis muscle has prognostic value in patients with glioblastoma. British Journal of Cancer. 2022;126(2):196-203. doi: 10.1038/s41416-021-01590-9

6. Booth TC, Thompson G, Bulbeck H, Boele F, Buckley C, Cardoso J, Dos Santos Canas L, Jenkinson D, Ashkan K, Kreindler J, Huskens N. A position statement on the utility of interval imaging in standard of care brain tumour management: defining the evidence gap and opportunities for future research. Frontiers in oncology 2021;11:620070. doi: 10.3389/fonc.2021.620070

7. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine. 2005;352:987-96. doi: 10.1056/NEJMoa043330

8. Stupp R, Hegi ME, Mason WP, Van Den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P. 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. 2009;10:459-66. doi: 10.1016/S1470-2045(09)70025-7

9. Davies J, Reyes-Rivera I, Pattipaka T, Skirboll S, Ugiliweneza B, Woo S, Boakye M, Abrey L, Garcia J, Burton E. Survival in elderly glioblastoma patients treated with bevacizumab-based regimens in the United States. Neuro-Oncology Practice. 2018;5(4):251-61. doi: 10.1093/nop/npy001

10. Bates A, Gonzalez-Viana E, Cruickshank G, Roques T. Primary and metastatic brain tumours in adults: summary of NICE guidance. BMJ. 2018;362. doi: 10.1136/bmj.k2924

11. Stupp R, Brada M, Van Den Bent MJ, Tonn JC, Pentheroudakis GE. High-grade glioma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2014;25:iii93-101. doi: 10.1093/annonc/mdu050

12. Brancato V, Nuzzo S, Tramontano L, Condorelli G, Salvatore M, Cavaliere C. Predicting Survival in Glioblastoma Patients Using Diffusion MR Imaging Metrics—A Systematic Review. Cancers. 2020; 12(10):2858. doi: 10.3390/cancers12102858

13. National Cancer Research Institute. NCRI Brain Group Strategic Priorities 2021-2024. [Internet]. Available from: https://www.ncri.org.uk/wp-content/uploads/NCRI-Brain-Group-Strategic-Priorities-Document-2021-2024.pdf

14. National Cancer Institute. Advocates In Research Working Group. [Internet]. 2011. Available from: https://deainfo.nci.nih.gov/advisory/ncra/ARWG-recom.pdf

15. WHO Classification of Tumours Editorial Board. (2021). World Health Organization Classification of Tumours of the Central Nervous System (5th ed.). International Agency for Research on Cancer.

16. Mongan J, Moy L, Kahn Jr CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology: Artificial Intelligence. 2020;2(2):e200029. doi: 10.1148/ryai.2020200029

17. Shuaib H, Barker GJ, Sasieni P, De Vita E, Chelliah A, Andrei R, Ashkan K, Beaumont E, Brazil L, Rowland-Hill C, Lau YH, Luis A, Powell J, Swampillai A, Tenant S, Thust SC, Wastling S, Young T, Booth TC. Overcoming challenges of translating deep-learning models for glioblastoma: the ZGBM consortium. The British Journal of Radiology. 2023;96:20220206. doi: 10.1259/bjr.20220206

18. Booth TC, Luis A, Brazil L, Thompson G, Daniel RA, Shuaib H, Ashkan K, Pandey A. Glioblastoma post-operative imaging in neuro-oncology: current UK practice (GIN CUP study). European radiology. 2021;31:2933-43. doi: 10.1007/s00330-020-07387-3

19. Felsberg J, Rapp M, Loeser S, Fimmers R, Stummer W, Goeppert M, Steiger HJ, Friedensdorf B, Reifenberger G, Sabel MC. Prognostic Significance of Molecular Markers and Extent of Resection in Primary Glioblastoma Patients Molecular Markers in Glioblastoma Patients. Clinical Cancer Research. 2009;15(21):6683-93. doi: 10.1158/1078-0432.CCR-08-2801

20. Brown TJ, Brennan MC, Li M, Church EW, Brandmeir NJ, Rakszawski KL, Patel AS, Rizk EB, Suki D, Sawaya R, Glantz M. Association of the extent of resection with survival in glioblastoma: a systematic review and meta-analysis. JAMA oncology. 2016;2(11):1460-9. doi:10.1001/jamaoncol.2016.1373

21. Helseth R, Helseth E, Johannesen TB, Langberg CW, Lote K, Rønning P, Scheie D, Vik A, Meling TR. Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. Acta neurologica scandinavica. 2010;122(3):159-67. doi: 10.1111/j.1600-0404.2010.01350.x

22. Lamborn KR, Chang SM, Prados MD. Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. Neuro-oncology. 2004;6(3):227-35. doi: 10.1215/S1152851703000620

23. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, Carbone PP. Toxicity and response criteria of the Eastern Cooperative Oncology Group. American journal of clinical oncology. 1982;5(6):649-56.

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research. 2011;12:2825-30.

25. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem E, Montvila A, Lynch J, Townend M, Agarwal S, Mazumder A, Barker GJ, Ourselin S. Deep learning models for triaging hospital head MRI examinations. Medical Image Analysis. 2022;78:102391. doi: 10.1016/j.media.2022.102391

26. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem E, Lynch J, Townend M, Montvila A, Siddiqui J, Gadapa N, Benger M, Barker G, Ourselin S, Cole JH, Booth TC. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Proceedings 3; 2020 October 4–8; Lima, Peru. Springer International Publishing; 2020. p. 254-265. doi: 10.1007/978-3-030-61166-8_27

27. Modat M, Cash DM, Daga P, Winston GP, Duncan JS, Ourselin S. Global image registration using a symmetric block-matching approach. Journal of Medical Imaging. 2014;1(2):024003. doi: 10.1117/1.JMI.1.2.024003

28. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, Murray B, Myronenko A, ..., Feng A. MONAI: An open-source framework for deep learning in healthcare. arXiv preprint. 2022. doi: 10.48550/arXiv.2211.02701

29. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;4700-4708.

30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32:8026-8037.

31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;7:77. doi: 10.1186/1471-2105-12-77

32. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Advances in Neural Information Processing Systems. 2018;31

33. Cluceru J, Interian Y, Phillips JJ, Molinaro AM, Luks TL, Alcaide-Leon P, Olson MP, Nair D, LaFontaine M, Shai A, Chunduru P. Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. Neuro-Oncology. 2022;24(4);639-52. doi: 10.1093/neuonc/noab238

34. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, Zhai G. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Scientific Reports. 2017;7(1):10353. doi: 10.1038/s41598-017-10649-8

35. Poon MT, Sudlow CL, Figueroa JD, Brennan PM. Longer-term (≥ 2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis. Scientific Reports. 2020;10(1);1-10. doi: 10.1038/s41598-020-68011-4

36. Booth TC, Larkin TJ, Yuan Y, Kettunen MI, Dawson SN, Scoffings D, Canuto HC, Vowler SL, Kirschenlohr H, Hobson MP, Markowetz F, Jefferies S, Brindle KM. Analysis of heterogeneity in T2-weighted MR images can differentiate pseudoprogression from progression in glioblastoma. PLoS One. 2017;12(5):e0176528. doi: 10.1371/journal.pone.0176528

37. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, DeGroot J, Wick W, Gilbert MR, Lassman AB, Tsien C. Updated response assessment criteria for high-grade gliomas:

response assessment in neuro-oncology working group. Journal of Clinical Oncology. 2010;10;28(11):1963-72

38. Wen PY, van den Bent M, Youssef G, Cloughesy TF, Ellingson BM, Weller M, Galanis E, Barboriak DP, de Groot J, Gilbert MR, Huang R, Lassman AB, Mehta M, Molinaro AM, Preusser M, Rahman R, Shankar LK, Stupp R, Villanueva-Meyer JE, Wick W, Macdonald DR, Reardon DA, Vogelbaum MA, Chang SM. RANO 2.0: Update to the Response Assessment in Neuro-Oncology Criteria for High-and Low-Grade Gliomas in Adults. Journal of Clinical Oncology. 2023;41(33); 5187-5199. doi: 10.1200/JCO.23.01059

39. Vos MJ, Uitdehaag BM, Barkhof F, Heimans JJ, Baayen HC, Boogerd W, Castelijns JA, Elkhuizen PH, Postma TJ. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. Neurology. 2003;60(5):826-30. doi: 10.1212/01.WNL.0000049467.54667.92

40. Deeley MA, Chen A, Datteri R, Noble JH, Cmelak AJ, Donnelly EF, Malcolm AW, Moretti L, Jaboin J, Niermann K, Yang ES, Yu DS, Yei F, Koyama T, Ding GX, Dawant BM. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Physics in Medicine & Biology. 2011;56(14):4557. doi: 10.1088/0031-9155/56/14/021

41. Boxerman JL, Zhang Z, Safriel Y, Larvie M, Snyder BS, Jain R, Chi TL, Sorensen AG, Gilbert MR, Barboriak DP. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. Neuro-oncology. 2013;15(7):945-54. doi: 10.1093/neuonc/not049

42. Chang K, Beers AL, Bai HX, Brown JM, Ly KI, Li X, Senders JT, Kavouridis VK, Boaro A, Su C, Bi WL. Rapalino O, Liao W, Shen Q, Zhou H, Xiao B, Wang Y, Zhang PJ, Pinho MC, Wen PY, Batchelor TT, Boxerman JL, Arnaout O, Rosen, BR, Gerstner ER, Yang L, Huang RY, Kalpathy-Cramer J. Automatic assessment of glioma burden: a deep learning algorithm for fully automated

volumetric and bidimensional measurement. Neuro-oncology. 2019;21(11):1412-22. doi: 10.1093/neuonc/noz106

43. Blasel S, Zagorcic A, Jurcoane A, Bähr O, Wagner M, Harter PN, Hattingen E. Perfusion MRI in the evaluation of suspected glioblastoma recurrence. Journal of Neuroimaging. 2016;26(1);116-23. doi: 10.1111/jon.12247

Figures

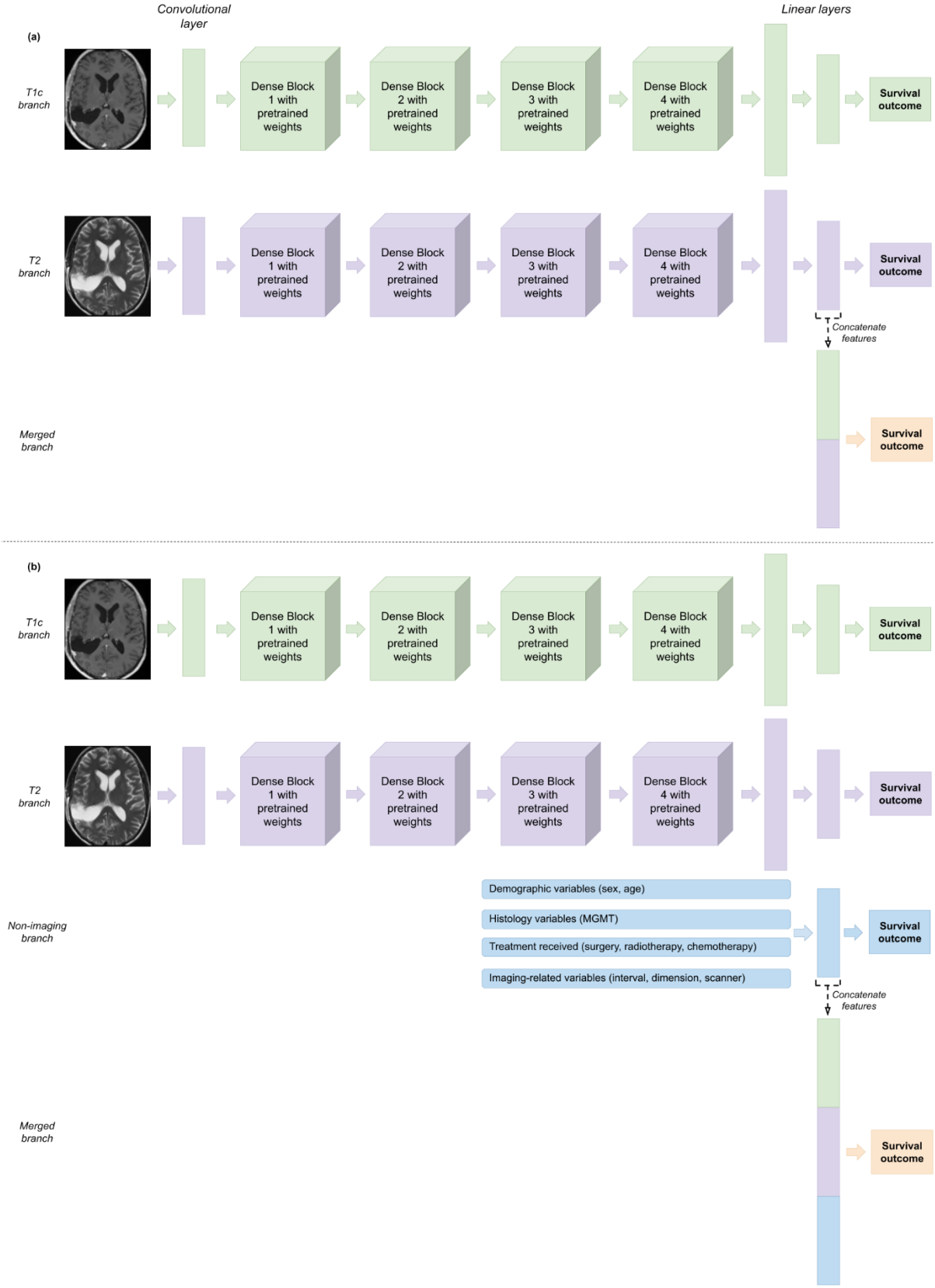

Convolutional layer
Linear layers
(a)
T1c branch
Dense Block 1 with pretrained weights
Dense Block 2 with pretrained weights
Dense Block 3 with pretrained weights
Dense Block 4 with pretrained weights
Survival outcome
T2 branch
Dense Block 1 with pretrained weights
Dense Block 2 with pretrained weights
Dense Block 3 with pretrained weights
Dense Block 4 with pretrained weights
Survival outcome
Concatenate features
Merged branch
Survival outcome
(b)
T1c branch
Dense Block 1 with pretrained weights
Dense Block 2 with pretrained weights
Dense Block 3 with pretrained weights
Dense Block 4 with pretrained weights
Survival outcome
T2 branch
Dense Block 1 with pretrained weights
Dense Block 2 with pretrained weights
Dense Block 3 with pretrained weights
Dense Block 4 with pretrained weights
Survival outcome
Non-imaging branch
Demographic variables (sex, age)
Histology variables (MGMT)
Treatment received (surgery, radiotherapy, chemotherapy)
Imaging-related variables (interval, dimension, scanner)
Survival outcome
Concatenate features
Merged branch
Survival outcome

Figure 1. Architectures for dense neural networks. **(a)** Imaging model: The model inputs whole brain contrast-enhanced T_1-weighted sequences, and T_2-weighted sequences as separate branches (T1c and T2 branches). These are passed through dense blocks with pretrained weights. Outputs are flattened and reduced before feature concatenation. Predictions are obtained from the merged linear layer (concatenating vectors from T1c and T2 branches). **(b)** *Combined model*: Modified version of the architecture with an additional branch consisting of non-imaging inputs and linear layers.
For illustrative purposes, 3D MR volumes are shown as 2D images and 4D dense blocks as 3D representations.

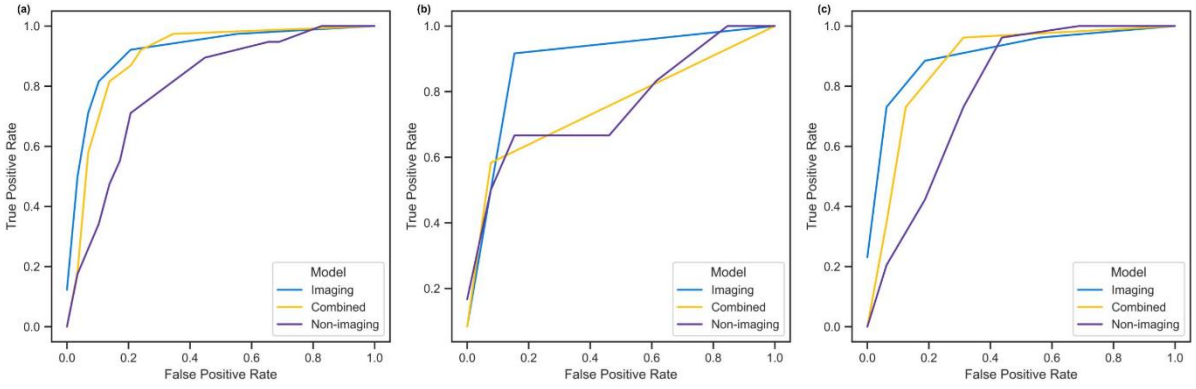


Figure 2. Receiver-operating characteristic curves for imaging, combined, and non-imaging models on holdout test data. **(a)** Model performances on the amalgamated test set. AUCs were 0.93, 0.91 and 0.79 for the imaging, combined, and non-imaging models respectively. **(b)** Model performances on the retrospective test set. AUCs were 0.92, 0.94 and 0.76 for the imaging, combined, and non-imaging models respectively. **(c)** Model performances on the external, prospective test set. AUCs were 0.93, 0.89 and 0.78 for the imaging, combined, and non-imaging models respectively.
AUC: area under the receiver-operating characteristic curves.

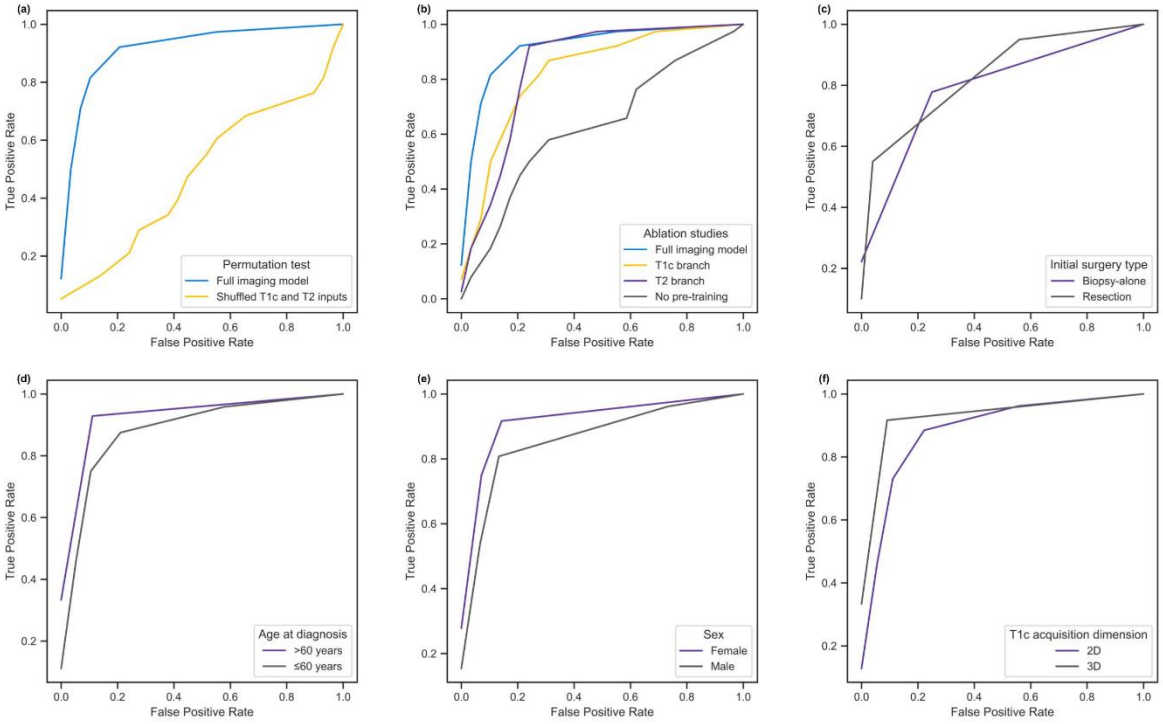


Figure 3. Receiver-operating characteristic curves displaying imaging model performances for additional analyses run on the amalgamated test set. **(a)** Permutation test results (full imaging model, AUC = 0.93; permutation test, AUC = 0.49*****). **(b)** Results from ablation studies (full imaging model, AUC = 0.93; predictions from T1c branch, AUC = 0.83*****; predictions from T2 branch, AUC = 0.85; trained model initializing random weights – i.e., with no pre-training, AUC = 0.64*****). Panels (c) to (f) show imaging model results disaggregated for sample subgroups. **(c)** Performances based on the initial surgery type (biopsy-alone, AUC = 0.89; resection, AUC = 0.87). **(d)** Curves plotted separately for age at first diagnosis (> 60 years, AUC = 0.98; ≤ 60 years, AUC = 0.89). **(e)** Performances based on sex (female, AUC = 0.96; male = 0.89). **(f)** Performances split by the acquisition dimension of the input T1c MRI (2D, AUC = 0.90; 3D, AUC = 0.98).

AUC: area under the receiver-operating characteristic curves. *T1c*: contrast-enhanced T1-weighted MRI. *T2*: T2-weighted MRI.

*: significantly different AUC compared to the full imaging model using DeLong's test with a threshold of $p \leq 0.05$.

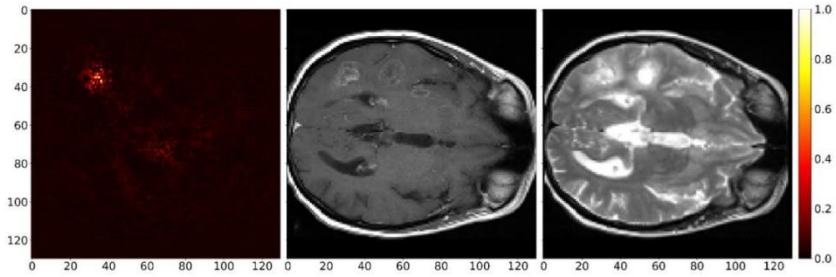
Patient 1

Test cohort: *retrospective*

True outcome: *short-term*

Predicted outcome: *short-term*

Slice number: *60*



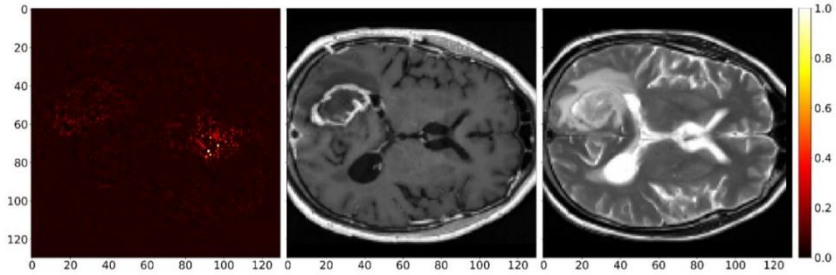
Patient 2

Test cohort: *retrospective*

True outcome: *long-term*

Predicted outcome: *long-term*

Slice number: *73*



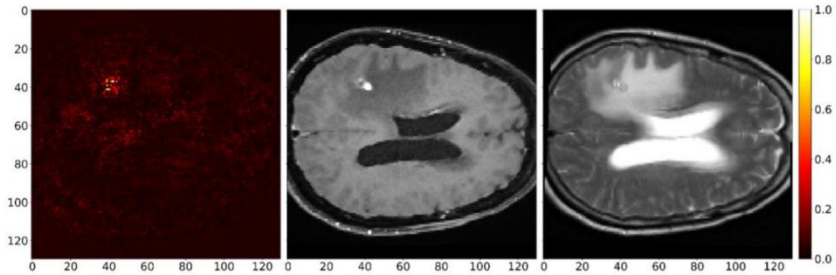
Patient 3

Test cohort: *prospective*

True outcome: *short-term*

Predicted outcome: *short-term*

Slice number: *84*



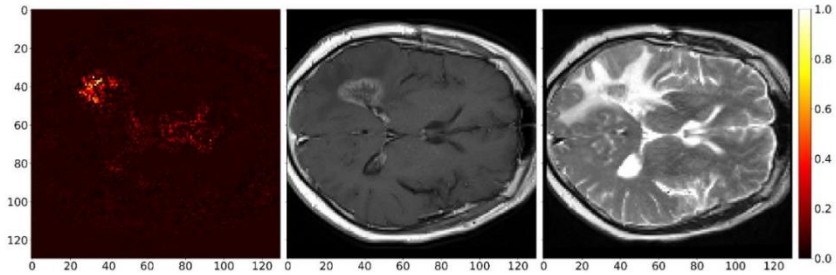
Patient 4

Test cohort: *prospective*

True outcome: *long-term*

Predicted outcome: *long-term*

Slice number: *64*



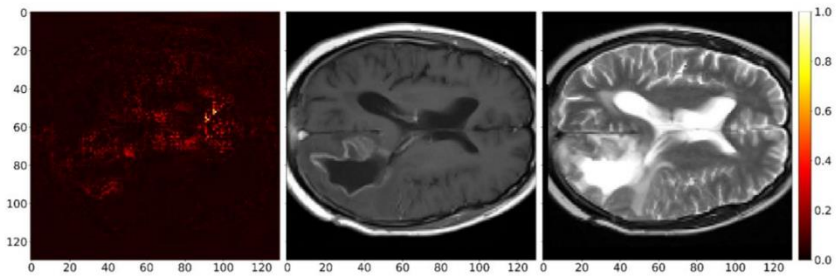
Patient 5

Test cohort: *prospective*

True outcome: *long-term*

Predicted outcome: *short-term*

Slice number: *68*



Patient 6

Test cohort: *prospective*

True outcome: *short-term*

Predicted outcome: *long-term*

Slice number: *73*

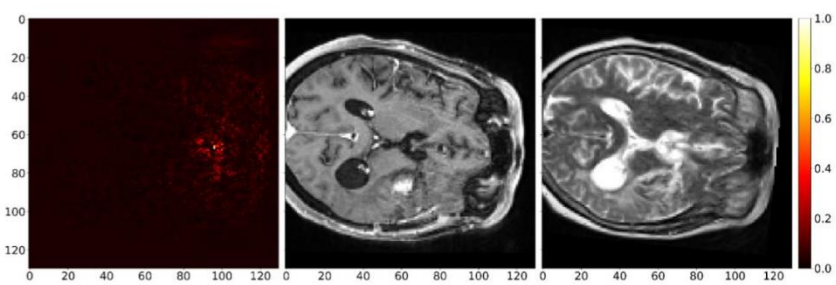


Figure 4. Saliency maps from guided backpropagation on the merged branch of imaging models using T1c and T2 inputs. Patients from retrospective and prospective test sets were selected including erroneous classification predictions (patients 5 and 6).

T1c: contrast-enhanced T_1-weighted MR sequence. *T2*: T_2–weighted MR sequence.

Tables

Table 1. Patient cohort described by center, data collection period (retrospective/prospective), outcome (short/long-term survival), and MRI acquisition dimension (2D/3D). The amalgamated holdout test set consists of a *prospective test set* (all patients from eight prospective centers, n=29) and a *retrospective test set* (n=19 patients from two centers; of which KCH n=13, LTHT n=6).

		Survival Outcome		**T1c Acquisition Dimension**	
Centre	**N total (% of dataset)**	**Short-term N (% of center)**	**Long-term N (% of center)**	**2D N (% of center)**	**3D N (% of center)**
Retrospective data collection					
KCH	124 (70.1%)	35 (28.2%)	89 (71.8%)	39 (31.5%)	85 (68.5%)
LTHT	47 (26.6%)	14 (29.8%)	33 (70.2%)	41 (87.2%)	6 (12.8%)
UCLH	6 (3.4%)	2 (33.3%)	4 (66.7%)	1 (16.7%)	5 (83.3%)
Total	*177 (85.9%)*	*51 (28.8%)*	*126 (71.2%)*	*81 (45.8%)*	*96 (54.2%)*
Prospective data collection					
BSUH	2 (1.0%)	1 (50.0%)	1 (50.0%)	2 (100.0%)	0 (0.0%)
Christie	7 (3.4%)	1 (14.3%)	6 (85.7%)	7 (100.0%)	0 (0.0%)
HEY	8 (3.9%)	6 (75.0%)	2 (25.0%)	5 (62.5%)	3 (37.5%)
ICHT	4 (1.9%)	1 (25.0%)	3 (75.0%)	1 (25.0%)	3 (75.0%)
LTHTR	2 (1.0%)	1 (50.0%)	1 (50.0%)	2 (100.0%)	0 (0.0%)
Marsden	1 (0.5%)	0 (0.0%)	1 (100.0%)	0 (0.0%)	1 (100.0%)
NUH	1 (0.5%)	0 (0.0%)	1 (100.0%)	1 (100.0%)	0 (0.0%)
NUTH	4 (1.9%)	3 (75.0%)	1 (25.0%)	4 (100.0%)	0 (0.0%)
Total	*29 (14.1%)*	*13 (44.8%)*	*16 (55.2%)*	*22 (75.9%)*	*7 (24.1%)*
Total	*206 (100%)*	*64 (31.1%)*	*142 (68.9%)*	*103 (50%)*	*103 (50%)*

KCH: King's College Hospital NHS Foundation Trust; patients were treated across KCH, Guy's and St Thomas' NHS Foundation Trust, and the Kent Oncology Centre. *LTHT*: Leeds Teaching Hospitals NHS Trust. *UCLH*: University College London Hospitals NHS Foundation Trust. *BSUH*: Brighton and Sussex University Hospitals NHS Trust. *Christie*: The Christie NHS Foundation Trust. *HEY*: Hull University Teaching Hospitals NHS Trust. *ICHT*: Imperial College Healthcare NHS Trust. *LTHTR***:** Lancashire Teaching Hospitals NHS Foundation Trust. *Marsden*: The Royal Marsden NHS Foundation Trust. *NUH*: Nottingham University Hospitals NHS Trust. *NUTH*: Newcastle upon Tyne Hospitals NHS Foundation Trust.

Table 2. Patient characteristics described overall (all patients), and by survival outcome (short-term or long-term survivors defined as ≤ or > eight months survival from end of radiotherapy, respectively).

Variable	All patients (n=206)	Short-term survivors (n=64)	Long-term survivors (n=142)	*P value* [a]
Survival				
Deceased date, n (%)				-
Known	183 (88.8%)	64 (100.0%)	119 (57.8%)	
Unknown	23 (11.2%)	0 (0.0%)	23 (11.2%)[b]	
Survival time from end of radiotherapy, in weeks				-
Mean (SE[c])	73.1 (4.1)	21.9 (1.1)	96.2 (4.8)	
Demographic variables				
Sex, n (%)				0.12
Female	72 (35.0%)	17 (26.6)	55 (38.7%)	
Male	134 (65.0%)	47 (73.4%)	87 (61.3%)	
Age at first diagnosis, in years				0.28
Mean (SE)	57.4 (0.7)	59.0 (1.13)	56.7 (0.9)	
Unknown, n (%)	1 (0.5%)	0 (0.0%)	1 (0.7%)	
Histologic variables				
MGMT[d] status, n (%)				0.13
Methylated	87 (42.2%)	21 (32.8%)	66 (46.5%)	
Unmethylated	114 (55.3%)	42 (65.6%)	72 (50.7%)	
Unknown	5 (2.4%)	1 (1.6%)	4 (2.8%)	
MGMT methylation percentage				0.04
Mean (SE)	16.4 (1.4)	10.9 (1.9)	18.7 (1.8)	
Unknown, n (%)	26 (12.6%)	10 (26.6%)	16 (12.0%)	
Tumor location				
Deep-seated location[e], n (%)				0.21
Deep-seated	25 (12.1%)	11 (17.2%)	14 (9.9%)	
Not deep-seated	181 (87.9%)	53 (82.8%)	128 (90.1%)	
Treatment variables				
Surgery type, n (%)				<0.001
Biopsy-only	48 (23.3%)	25 (39.1%)	23 (16.2%)	
Resection	158 (76.7%)	39 (60.9%)	119 (83.8%)	
Radiotherapy dose, n (%)				0.03
Stupp dose[f]	160 (77.7%)	43 (67.2%)	117 (82.4%)	
Reduced dose	36 (17.5%)	18 (28.1%)	18 (12.7%)	
Not documented	10 (4.9%)	3 (4.7%)	7 (4.9%)	
Concomitant temozolomide dose, n (%)				0.001
Stupp dose	126 (61.2%)	29 (45.3%)	97 (68.3%)	
Reduced dose	26 (12.6%)	15 (23.4%)	11 (7.7%)	
No temozolomide	23 (11.2%)	11 (17.2%)	12 (8.5%)	
Not documented	31 (15.0%)	9 (14.1%)	22 (15.5%)	
Imaging-related variables				
Duration between radiotherapy and input MRI, in weeks[g]				0.02
Mean (SE)	8.7 (0.3)	7.5 (0.6)	9.2 (0.4)	
Scanner manufacturer, n (%)				0.43
General Electric	55 (26.7%)	14 (21.9%)	41 (28.9%)	
Mirada	1 (0.5%)	0 (0.0%)	1 (0.7%)	
Philips	8 (3.9%)	3 (4.7%)	5 (3.5%)	
Siemens	141 (68.4%)	46 (71.9%)	95 (66.9%)	
Toshiba	1 (0.5%)	1 (1.6%)	0 (0.0%)	

T1c dimension, n (%)				0.45
2D	103 (50.0%)	35 (54.7%)	68 (47.9%)	
3D	103 (50.0%)	29 (45.3%)	74 (52.1%)	

[a] *P* values reflect statistical significance of distributions for demographic, histologic, tumour location, treatment-related, and imaging-related variables between short-term and long-term survivors, calculated with Mann-Whitney U and Chi-squared tests.

[b] Albeit known to be alive beyond eight months post-radiotherapy.

[c] *SE*: standard error.

[d] *MGMT*: O6-methylguanine-DNA methyltransferase methylation. Methylated status refers to an MGMT methylation percentage above a 10% cutoff point.

[e] Deep-seated location: tumour infiltrates midbrain, thalamus, or callosum.

[f] Stupp dose: radiotherapy dose of 60 Gy delivered in 30 fractions.

[g] A histogram showing time between radiotherapy and first MRI images after radiotherapy completion is presented in Appendix B2.

Table 3. Holdout test set performances from imaging, combined (imaging/non-imaging), and non-imaging models. The retrospective test set is an internal validation dataset. The prospective test set is an external validation dataset using data from geographically distinct sites. The amalgamated test set refers to the combination of the retrospective and prospective test sets.

Description	AUC[a]	Precision	Recall	F1	Specificity	NPV[b]	BAR[c]	Accuracy
Amalgamated test set (n=48 patients, from 10 centers)								
Imaging model	**0.93 ± 0.07***	**0.77**	**0.89**	**0.83**	**0.83**	**0.92**	**0.86**	**0.85**
Combined model	0.91 ± 0.08	0.63	1.00	0.78	0.62	1.00	0.81	0.77
Non-imaging model	0.79 ± 0.12	0.67	0.32	0.43	0.90	0.67	0.61	0.67
Retrospective test set (n=19 patients, from 2 centers)								
Imaging model	0.92 ± 0.12	0.67	1.00	0.80	0.77	1.00	0.88	0.84
Combined model	**0.94 ± 0.11**	**0.55**	**1.00**	**0.71**	**0.62**	**1.00**	**0.81**	**0.74**
Non-imaging model	0.76 ± 0.19	0.67	0.33	0.44	0.92	0.75	0.62	0.74
Prospective test set (n=29 patients, from 8 centers)								
Imaging model	**0.93 ± 0.09***	**0.85**	**0.85**	**0.85**	**0.88**	**0.88**	**0.86**	**0.86**
Combined model	0.89 ± 0.11	0.68	1.00	0.81	0.63	1.00	0.81	0.79
Non-imaging model	0.78 ± 0.15	0.57	0.31	0.4	0.81	0.59	0.56	0.59

[a]*AUC*: area under the receiver operating characteristic curve. The key results for machine learning models are the generalizability of holdout test set values. We also compute the sample size-based 95% confidence intervals using Bernoulli trials formula ($z \times \sqrt{\frac{AUC \times (1-AUC)}{n}}$).

[b]*NPV*: negative predictive value.

[c]*BAR*: balanced accuracy rate.

Bold rows are those with highest AUC scores.

*: significantly different AUC compared to the non-imaging model using DeLong's test with a threshold of $p \leq 0.05$.

Supplementary Material

Appendix A. Overview of the treatment and imaging pathway for glioblastoma

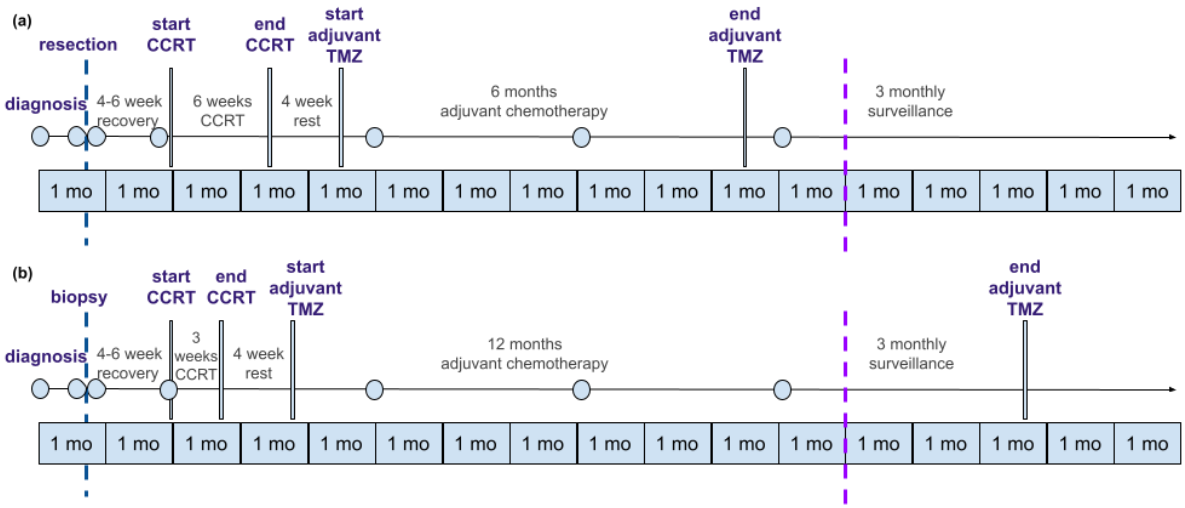


Figure A1. Simplified illustration of the recommended treatment and imaging pathway for glioblastoma. Some patients may not tolerate recommended post-surgical therapies, and therefore undergo an altered course of treatment. Circles are illustrative of recommended interval imaging timepoints. To inform patient management, MRIs are often performed after initial surgery, during radiotherapy planning, and at three-monthly intervals (or if clinically deteriorating) during follow-up after radiotherapy completion[1,2,3,4]. **(a)** Pathway for patients receiving the optimal treatment (i.e., completing the "Stupp" protocol, which consists of surgical resection, followed by radiotherapy with concomitant TMZ, then adjuvant TMZ)[5,6]. **(b)** Pathway for patients receiving a common modified treatment. Modified treatment, as shown in this example, often includes a shorter course of CCRT with a lower dose of radiotherapy; a longer course of adjuvant chemotherapy may be prescribed.

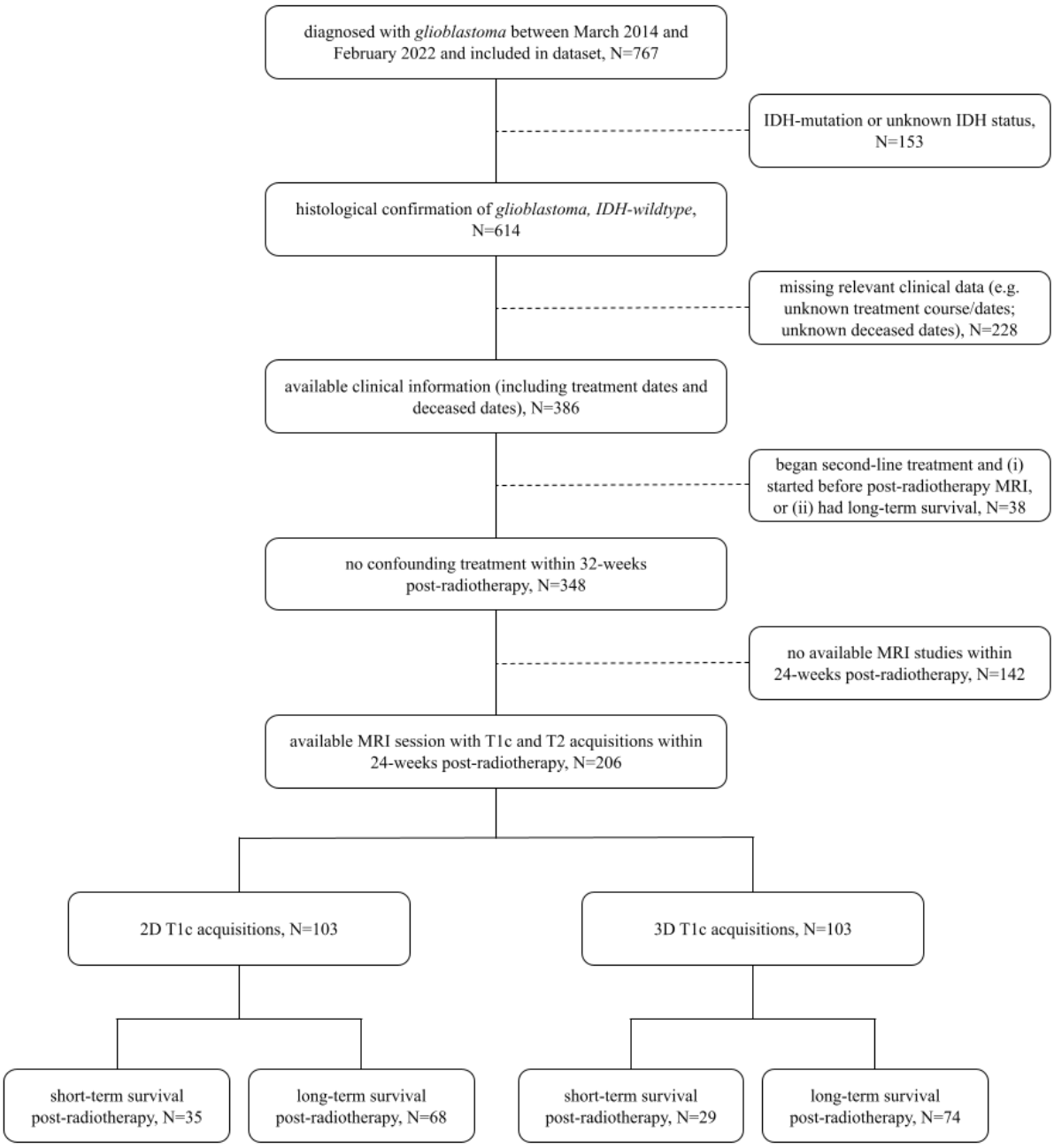
CCRT: radiotherapy and concomitant chemotherapy. *TMZ*: temozolomide.

Appendix B. Patient and Public Feedback

The research proposal was presented to the Next Generation Medical Imaging Advisory Group at King's College London (January 2021), and the Guy's Cancer Group at Guy's and St Thomas' Hospital (February 2021).

Based on feedback, this study considered (i) the influence of non-imaging features on model development, and (ii) in addition to patients undergoing optimal treatment (i.e., completing the "Stupp" protocol, which consists of surgical resection, followed by radiotherapy with concomitant TMZ, then adjuvant TMZ), inclusion of patients without optimal treatment (e.g., those who could not tolerate full-dose of radiotherapy or temozolomide chemotherapy, or who had an initial biopsy without a maximal safe resection). Test evaluation focused on the area under the receiver operating characteristic curve (AUC) metric. Groups commented that multi-center, prospective test data would provide greater reassurance of detecting post-treatment changes in an evidence-based manner.

Appendix C. Patient Cohort



CONSORT diagram displaying flow of patients included in analyses. Also demonstrated is the different contrast-enhanced T_1-weighted (T1c) sequences categorized as 3D or 2D.

Appendix D. Dataset sampling

Appendix D1. Patient characteristics in full and test datasets.

To check for potential bias(es) in the overall dataset sampling strategy, we compared the portion of patients in the test set to the full dataset and measured the sampling error for the following variables: survival outcome, contrast-enhanced T_1-weighted acquisition dimension (2D or 3D), initial surgery type (biopsy or maximal safe resection), MGMT methylation status at first diagnosis, and age at first diagnosis. These findings are presented in Table B1 below. Sampling for validation folds stratified outcome, acquisition dimension, and acquisition center (KCH/LTHT/UCLH) in training data. Surgery type, MGMT methylation status, and age are reported here as these factors are related to survival of glioblastoma[7-10]; these were not stratified during cross-validation due to low patient numbers after stratifying for three variables.

Table D1. Patient characteristics of the full dataset compared to test data described by survival outcome, MRI acquisition dimension, surgery type, MGMT status, and age group. Test set sampling error is relative to the full dataset.

Stratified variable		All patients, N=206	Combined test set, N=48		Retrospective test set, N=19		Prospective test set, N=29	
		N (%)	N (%)	Sampling error	N (%)	Sampling error	N (%)	Sampling error
Survival outcome	Short-term	64 (31.1%)	19 (39.6%)	27.4	6 (31.6%)	1.64	13 (44.8%)	44.29
	Long-term	142 (68.9%)	29 (60.4%)	-12.4	13 (68.4%)	-0.74	16 (55.2%)	-19.96
Acquisition dimension	2D	103 (50.0%)	31 (64.6%)	29.17	9 (47.4%)	-5.26	22 (75.9%)	51.72
	3D	103 (50.0%)	17 (35.4%)	-29.17	10 (52.6%)	5.26	7 (24.1%)	-51.72
Surgery type	Biopsy	48 (23.3%)	13 (27.1%)	16.23	4 (21.1%)	-9.65	9 (31.0%)	33.19
	Resection	158 (76.7%)	35 (72.9%)	-4.93	15 (78.9%)	2.93	20 (69.0%)	-10.08
MGMT status	Methylated	87 (42.2%)	20 (41.6%)	-1.34	8 (42.1%)	-0.30	12 (41.4%)	-2.02
	Unmethylated	114 (55.3%)	28 (58.3%)	5.41	11 (57.9%)	4.62	17 (58.6%)	5.93
	Unknown	5 (2.4%)	0 (0.0%)	-	0 (0.0%)	-	0 (0.0%)	-
Age at first diagnosis	≤ 60 years	89 (43.2%)	32 (66.7%)	17.38	14 (73.7%)	29.73	18 (62.1%)	9.28
	> 60 years	117 (56.8%)	16 (33.3%)	-22.85	5 (26.3%)	-39.09	11 (37.9%)	-12.20

Appendix D2

A visualization of the distribution of time between radiotherapy completion and the first MR study (used as inputs to imaging/combined models) is presented in Figure B1, split by survival outcome. The study was pragmatic, and imaging was carried out in line with local practice, and at a time dictated by clinical protocols, or additional clinical concerns. Detail on local UK imaging protocols is shown in the GIN CUP study[11].

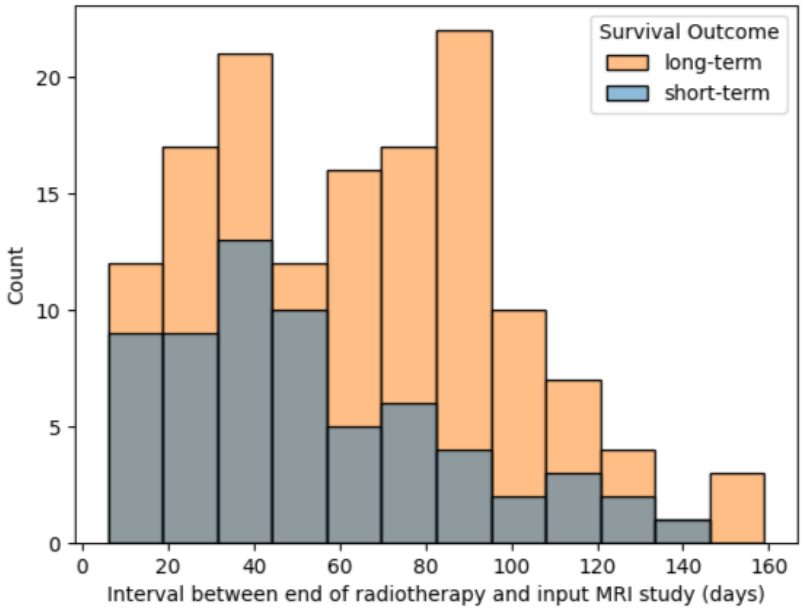


Figure D1. Histogram showing time between end of radiotherapy and the first MRI examination for all patients, stratified by survival outcome. The first MRI images after radiotherapy completion are used as inputs for the imaging model, as well as the duration between end of radiotherapy and the scan.

Appendix E. Further information on survival classifiers

Appendix E1. Description of non-imaging models

Description of non-imaging classical machine learning models

Four types of machine learning models were applied to training data with parameter tuning. Logistic regression models, linear support vector classifiers (SVCs) and gaussian SVCs were run with tuning of the regularization parameter (C) (range of C: 0.1 to 1000). For the gaussian SVC, gamma coefficients were additionally tuned using grid search (range of gamma: 0.1 to 100). Finally, decision tree classifiers were developed with selection between gini and entropy criteria for evaluating partitions. Based on validation performances, backward sequential feature selection was applied until the area under the receiver-operating characteristic curve (AUC) decreased.

Numeric variables were standardized to unit variance using training data, and categoric variables were one-hot encoded. Where data were missing, three approaches were used: (i) mean/mode imputation was used with labels added identifying imputed inputs, (ii) patients with missing data were excluded, (iii) variables with missing data were excluded.

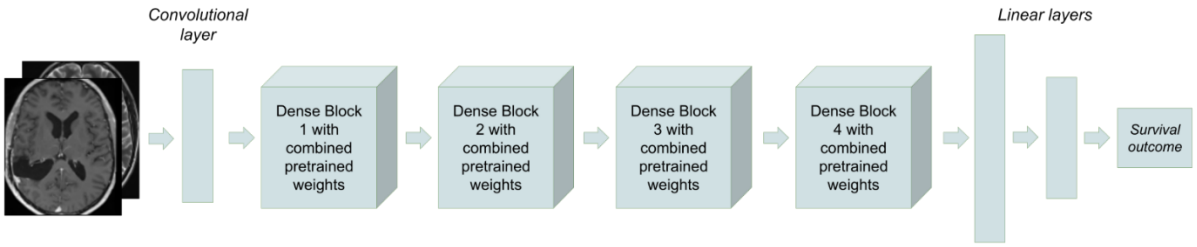
Description of non-imaging fully-connected neural networks

As an additional baseline measure, we trained shallow neural networks to predict survival outcomes from available non-imaging features alone (demographic, histologic, tumor-related, and prior treatment variables). These non-imaging features were passed through either one or two fully-connected linear layers before providing the binary survival prediction. Tuned parameters were the number of linear layers, learning rate and schedule, and probability of dropout; where there were two linear layers, the size of linear layers was also tuned (range=4-24). The fully-connected neural networks did not have a higher

validation performance than corresponding machine learning models, so were not pursued further. Models were developed with PyTorch[12].

Appendix E2. Description of alternative imaging model.

Instead of splitting contrast-enhanced T1-weighted (T1c) and T2-weighted (T2) MR sequence inputs across branches, we also tested a version of the model with both sequences joined as two channels (one branch). Pretrained weights were combined by either taking the average weight per convolutional layer and block, or by selecting the maximum weight. These models were not pursued further as they displayed poorer performances than counterparts with separate branches per input MR sequence.



Appendix E3. Procedure for training and tuning imaging and combined models.

Models were trained measuring cross-entropy loss weighted by class, and with the Novograd optimizer. Tuned (hyper-)parameters included: learning rate and schedule, linear layer sizes, number frozen/updated pretrained blocks, loss weighting per branch, dropout, pooling, and degree/probability of augmentation.

Appendix F.

Appendix F1. Parameters for the optimal imaging model

After tuning the imaging model, pretrained weights were frozen for the first convolutional layer and two dense blocks per branch. Blocks were flattened via maximum pooling; flattened feature vectors were mapped to a 1x56 then 1x2 vector per branch; the SeLU activation function was used. The merged branch therefore inputs a 1x112 vector. Losses were weighted for T1c, T2, and merged branches at a ratio of 1:1:3 respectively. The probability for each augmentation was 0.65. Applied augmentations included random left-right flipping, zooming, shearing, translation, rotation, adjusting intensity, adjusting contrast, adding Gaussian noise, and adding coarse dropout. A cyclical learning rate was applied (range: 4^{-8}-2^{-5}). The model was trained for 170 epochs before frozen and evaluated on holdout test data.

Appendix F2. Selecting the classification threshold

Description of threshold analysis

All models applied a classification threshold of 0.50 to determine the survival prediction label.

As an additional analysis to investigate the optimal decision threshold, Youden's J statistic was calculated on validation folds for the imaging model; the mean threshold was selected.

Test set results when applying Youden's J threshold

The Youden's J analysis suggested a threshold of 0.35. Applying this decision threshold to test set predictions did not improve imaging model performances (AUC=0.75; balanced accuracy rate=0.74). Therefore, a threshold of 0.50 was retained for all models (imaging/combined/non-imaging).

Appendix G. Additional analyses of test set predictions

Appendix G1. Interval between radiotherapy completion and MRI study used for survival prediction

As a supplementary analysis, we investigated whether survival predictions may be influenced by the interval between radiotherapy completion and the MRI study used as model inputs. For potential clinical translation, it is important to assess if there is a potential source of bias, where patients with a longer duration between radiotherapy completion and the first MRI may be predicted as long-term survivors and vice versa. For example, rather than identifying features related to future survival in neuroimaging, the model may be identifying that patients who are less well are brought in for MRI follow-up sooner than those who are responding to treatment. This possibility remains despite model inputs being limited to those MRIs obtained within 24-weeks of radiotherapy completion. We therefore further investigated patients with erroneous survival predictions.

The amalgamated test set has n=48 patients, of whom 19 were short-term survivors (39.6%). The imaging model made erroneous predictions for 7/48 patients (85.4% accuracy). The interval between radiotherapy completion and the MRI study used as model inputs is shown for the test set, based on the survival outcome:

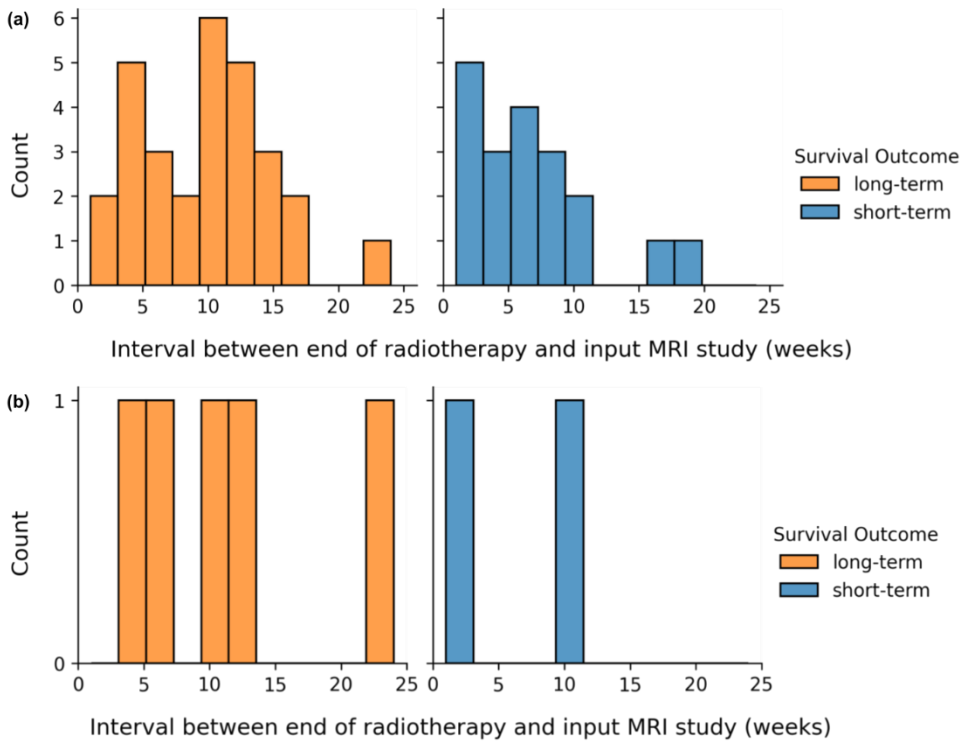


Figure G1. Interval between radiotherapy completion and the MRI study used as imaging (and combined) model inputs. **(a)** The interval for the amalgamated test set, split by survival outcome. **(b)** Intervals for patients with erroneous prognostic predictions (where long-term survivors were misclassified as short-term survivors, and vice versa).

Importantly, we found that the combined model did not perform significantly better than the imaging model, despite receiving this interval as an input variable. Furthermore, based on the overlapping intervals between misclassified short-term and long-term survivors in panel (b) of Figure G1, it is not overtly evident that the imaging model is biased due to variations in the duration between radiotherapy completion and the first post-radiotherapy MRI study. Further research should verify this expectation with a larger test set however, particularly for patients with the first MRI study obtained between 20-24 weeks after radiotherapy completion.

Appendix G2. Imaging model performances for additional analyses run on the amalgamated test set

Performance of the imaging model on the amalgamated test set is further described in Table G2. Additional analyses were performed including permutation testing, ablation studies, and grouping predictions by sample characteristics.

Table G2. Imaging model performances for additional analyses run on the amalgamated test set (the amalgamation of the retrospective and prospective test sets). Performances are shown for the permutation test and ablation studies; they are also reported disaggregated for sample subgroups (surgery type, age (>60years), sex, and acquisition dimension).

Description	AUC[a]	Precision	Recall	F1	Specificity	NPV[b]	BAR[c]	Accuracy
Full imaging model								
Imaging model	0.93	0.77	0.89	0.83	0.83	0.92	0.86	0.85
Permutation test results								
Shuffled T1c[d] and T2[e] inputs	0.49*	0.40	1.00	0.57	0.00	0.00	0.50	0.40
Ablation studies								
Predictions from T1c branch	0.83*	0.65	0.89	0.76	0.69	0.91	0.79	0.77
Predictions from T2 branch	0.85	0.67	0.63	0.65	0.79	0.77	0.71	0.73
Train model initializing random weights	0.64*	0.50	0.21	0.30	0.86	0.63	0.54	0.60
Initial surgery type (n)								
Biopsy-alone (13)	0.89	0.89	0.89	0.89	0.75	0.75	0.82	0.85
Resection (35)	0.87	0.69	0.90	0.78	0.84	0.95	0.87	0.86
Age at diagnosis (n) (missing n=1)								
>60 years (16)	0.98	0.78	1.00	0.88	0.78	1.00	0.89	0.88
≤60 years (31)	0.89	0.77	0.83	0.80	0.84	0.89	0.84	0.84
Sex (n)								
Female (20)	0.96	0.71	0.83	0.77	0.86	0.92	0.85	0.85
Male (28)	0.89	0.80	0.92	0.86	0.80	0.92	0.86	0.86
T1c acquisition dimension (n)								
2D (31)	0.90	0.79	0.85	0.81	0.83	0.88	0.84	0.84
3D (17)	0.98	0.75	1.00	0.86	0.82	1.00	0.91	0.88

[a] *AUC*: area under the receiver operating characteristic curve.

[b] *NPV*: negative predictive value.
[c] *BAR*: balanced accuracy rate.
[d] *T1c*: contrast-enhanced T1-weighted MRI.
[e] *T2*: T2-weighted MRI.
* : significantly different AUC compared to the full imaging model using DeLong's test with a threshold of $p \leq 0.05$

Appendix H. Comparative predictions from expert clinical raters

In the main manuscript, we provided a comparison of predictions from artificial intelligence models based on imaging-alone, combined information, and non-imaging clinical variables alone. The imaging model was selected for further analysis, based on the observation that using non-imaging features did not improve model performances. Readers may also be interested in how the selected model performs in comparison to physicians' interpretation of the same imaging data. It is acknowledged that such a prediction of long- or short-term survival is not expected when reporting on neuroimaging in clinical settings. Nonetheless, such a comparison might help to determine whether the model could be complementary in routine hospital practice.

To provide this comparison, we conducted a blinded study with predictions obtained from expert clinicians reviewing patient imaging. Three senior neuroradiologists (UK consultant grade; US attending equivalent) who present the imaging at the joint neuro-oncology meeting (UK multi-disciplinary meeting; US tumor board) at three UK neuro-oncology centers, made the equivalent survival predictions as the image-based model presented here, using the same T1 post-contrast (T1c) and T2 MRIs.

Predictive performance of each rater on the amalgamated test set is presented in Table H1, in addition to performance based on inter-rater consensus (i.e., the mode/majority vote). Since raters were predicting the binary survival outcome rather than providing prediction probabilities per class, no receiver-operating characteristic curves are presented (equivalently, the area under the receiver-operating characteristic curve metric was not calculated). The precision and recall of the imaging model were 0.77 and 0.89 on the amalgamated test set respectively. In comparison, we found that predictions made by consensus had a precision of 0.79 (range across raters: 0.70-0.85) and recall of 0.79 (range: 0.58-0.84). A Fleiss Kappa score of 0.74 was obtained for inter-rater agreement.

We acknowledge that the three senior neuroradiologists had as much time as required to make the decision. We also acknowledge that consensus readings cannot be obtained routinely in the clinic and therefore our clinical comparator is an optimal scenario which may not be reflected in a routine clinical setting. Nonetheless, based on the consensus predictions and range in inter-rater predictive performances, the presented model (which returns predictions immediately and only requires images alone) performs at least similarly to the consensus of three experts given the same imaging. We therefore expect that the proposed deep learning model could provide relevant information for routine clinical practice, by distinguishing those patients who are and are not expected to survive the eight-month window after radiotherapy completion. Studies validating imaging model performance in clinical settings are required to test this possibility.

Table H1. Comparison of imaging model performance to those obtained by expert clinical raters on the amalgamated test set (the amalgamation of the retrospective and prospective test sets). Performances are shown for predictions made by each reader based on the same images used as inputs to the reported imaging model. Performance of predictions based on consensus agreement across raters is also provided (based on mode/majority vote).

Description	Precision	Recall	F1	Specificity	NPV[a]	BAR[b]	Accuracy
Imaging model	0.77	0.89	0.83	0.83	0.92	0.86	0.85
Predictions from clinical experts, based on T1c and T2 post-radiotherapy imaging							
Reader 1	0.85	0.58	0.69	0.93	0.77	0.75	0.79
Reader 2	0.70	0.84	0.76	0.76	0.88	0.80	0.79
Reader 3	0.70	0.74	0.72	0.79	0.82	0.76	0.77
Consensus vote	0.79	0.79	0.79	0.86	0.86	0.83	0.83

[a] *NPV*: negative predictive value.

[b] *BAR*: balanced accuracy rate.

[c] *T1c*: contrast-enhanced T1-weighted MRI.

[d] *T2*: T2-weighted MRI.

Appendix I. Further analysis of saliency maps

Appendix I1. Analysis of saliency maps in relation to model predictions and tumor appearances

As a further analysis, we investigated the relationship between imaging model predictions, axially-selected slices of 3D saliency maps for the predicted outcome, and tumor-related regions in input MRI scans (including treated/resected tumour areas).

In Table I1 below, we show amalgamated test set patient numbers grouped by (a) prediction accuracy of survival outcome from the imaging model (accurately versus inaccurately classified) and (b) whether any visible heatmap points in axially-selected slices of volumetric saliency maps overlay tumor-related regions in T1c/T2 slices (intersection of (i) tumor regions in MRI scans, and (ii) visible "hot" heatmap area after thresholding to remove the lowest 10% of values).

		Intersection between salient heatmap points in automatically selected axial slices and tumor region(s) in T1c/T2 MRIs	
		Intersection	**No Intersection**
Predictive accuracy (imaging model)	**True positive (short-term survival)**	15	2
	True negative (long-term survival)	13	11
	False positive	4	1
	False negative	1	1

As noted in the main manuscript, saliency maps alone should not be interpreted as showing task-related features that are easily interpretable to human readers[13]. In this test set, 29/48 patients were long-term survivors (60.4%). It is conceivable, for example, that anatomical appearances reflecting long-term survival may be difficult to interpret. Indicators of longer survival may relate to the absence of features that signify more marked disease progression. Indeed, the tumor region in long term survivors (true negatives) appears less likely to contribute to the decision making (a relatively smaller portion of patients have MRI scans where the "hot" heatmap points intersects with the tumor).

Appendix I2. Saliency maps for patients misclassified as short-term survivors

We further reviewed all cases where patients with long-term survival after radiotherapy were mistakenly predicted as short-term survivors by the imaging model. There were five patients in the amalgamated test set in this category. All cases are presented in Figure I2 below to explore the possibility that these patients show signs of pseudoprogression, where the model mistakenly predicted short-term survival given treatment-related effects in MRIs. Since the maps do not consistently suggest that contrast-enhancing tumor regions contributed greatly to predictions, it is not evident that the mistaken predictions of short-term survival could relate to pseudoprogression in all cases. Further comments on the challenging and clinically important scenario of pseudoprogression are presented in the Discussion section of the main manuscript.

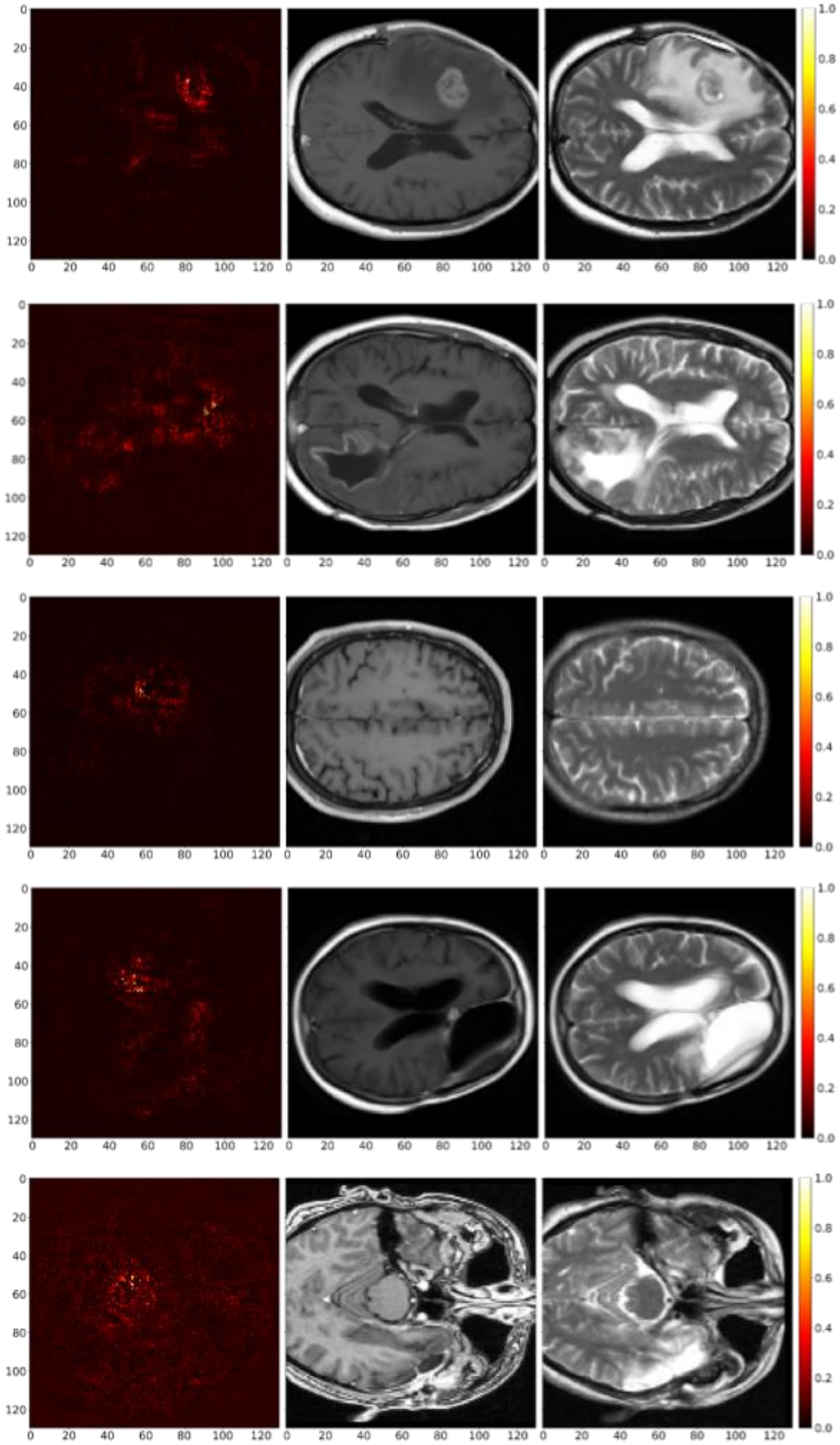


Figure I2. Visualization of saliency map, T1c, and T2 axial slices for all cases where patients with long-term survival were mistakenly predicted as short-term survivors.
T1c: contrast-enhanced. T_1-weighted MR sequence. *T2*: T_2–weighted MR sequence.

References for Supplemental Material

1. Weller M, Van Den Bent M, Tonn JC, Stupp R, Preusser M, Cohen-Jonathan-Moyal E, Henriksson R, Le Rhun E, Balana C, Chinot O, Bendszus M. European Association for Neuro-Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and oligodendroglial gliomas. The Lancet Oncology 2017;18:e315–e329. doi: 10.1016/S1470-2045(17)30194-8

2. Davies J, Reyes-Rivera I, Pattipaka T, Skirboll S, Ugiliweneza B, Woo S, Boakye M, Abrey L, Garcia J, Burton E. Survival in elderly glioblastoma patients treated with bevacizumab-based regimens in the United States. Neuro-Oncology Practice. 2018;5(4):251-61. doi: 10.1093/nop/npy001

3. Bates A, Gonzalez-Viana E, Cruickshank G, Roques T. Primary and metastatic brain tumours in adults: summary of NICE guidance. BMJ. 2018;362. doi: 10.1136/bmj.k2924

4. Stupp R, Brada M, Van Den Bent MJ, Tonn JC, Pentheroudakis GE. High-grade glioma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2014;25:iii93-101. doi: 10.1093/annonc/mdu050

5. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New Engadebland Journal of Medicine. 2005;352:987-96. doi: 10.1056/NEJMoa043330

6. Stupp R, Hegi ME, Mason WP, Van Den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P. 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. 2009;10:459-66. doi: 10.1016/S1470-2045(09)70025-7

7. Felsberg J, Rapp M, Loeser S, Fimmers R, Stummer W, Goeppert M, Steiger HJ, Friedensdorf B, Reifenberger G, Sabel MC. Prognostic Significance of Molecular Markers and Extent of Resection in Primary Glioblastoma Patients Molecular Markers in Glioblastoma Patients. Clinical Cancer Research. 2009;15(21):6683-93. doi: 10.1158/1078-0432.CCR-08-2801

8. Brown TJ, Brennan MC, Li M, Church EW, Brandmeir NJ, Rakszawski KL, Patel AS, Rizk EB, Suki D, Sawaya R, Glantz M. Association of the extent of resection with survival in glioblastoma: a systematic review and meta-analysis. JAMA oncology. 2016;2(11):1460-9. doi:10.1001/jamaoncol.2016.1373

9. Helseth R, Helseth E, Johannesen TB, Langberg CW, Lote K, Rønning P, Scheie D, Vik A, Meling TR. Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. Acta neurologica scandinavica. 2010;122(3):159-67. doi: 10.1111/j.1600-0404.2010.01350.x

10. Lamborn KR, Chang SM, Prados MD. Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. Neuro-oncology. 2004;6(3):227-35. doi: 10.1215/S1152851703000620

11. Booth TC, Luis A, Brazil L, Thompson G, Daniel RA, Shuaib H, Ashkan K, Pandey A. Glioblastoma post-operative imaging in neuro-oncology: current UK practice (GIN CUP study). European radiology. 2021;31:2933-43. doi: 10.1007/s00330-020-07387-3

12. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32:8026-8037

13. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Advances in Neural Information Processing Systems. 2018;31