Detection in Progress - A Multimodal Segmentation-based Approach for Predicting Glioblastoma Recurrence

Anonymous Author(s)
Affiliation
Address
email

Abstract

Radiation therapy planning for patients with glioblastoma requires defining the clinical-target-volume by delineating the tumor and including a margin of healthy tissue to account for microscopic tumor spread post radiation therapy. The current standard-of-care practice for defining the clinical-target-volume still employs an isotropic 1–2 cm expansion of the identified T2-hyperintensity lesion. As a consequence, normal-appearing brain tissue is overtreated, and it also ends up missing progression regions as it overlooks the heterogeneous infiltrative nature of these tumors. We propose incorporating anatomical, metabolic, and diffusion-weighted imaging acquired before surgical resection or between surgery and radiation therapy, using a lesion-size-aware segmentation objective to improve clinical-target-volume definition. The results in multiple metrics demonstrate better prediction of tumor progression than standard of care, as indicated by contrast enhancement and T2-hyperintensity at recurrence. Overall, our approach minimizes treatment of normal-appearing brain and captures progressed voxels beyond the 2 cm expansion.

1 Introduction

The current standard-of-care (SOC) for highly infiltrative glioblastoma (GBM) begins with maximal safe surgical resection, followed by external beam radiotherapy (RT) in conjunction with temozolomide chemotherapy [1]. Despite decades of clinical trials incorporating novel systemic and targeted agents and radiation dosing schemes, the only change to SOC treatment of GBM is the inclusion of Tumor-Treating-Fields upon completion of RT, which has resulted in minimal improvements in outcome beyond the typical 12–15 month dismal prognosis [2, 3]. This is in part due to the difficulty in identifying and treating the full extent of these highly infiltrative tumors with RT, while also sparing critical brain tissue to preserve normal brain function [4].

Although recent advances in RT delivery can provide millimeter-scale precision and dose modulation, current RT treatment planning protocols are still based on a uniform 1–2 cm geometric expansion of the gross-tumor volume defined on conventional post-contrast T1-weighted and T2-weighted FLAIR MRI, without considering spatial heterogeneity. This has the unintended consequences of undertreating subclinical disease not yet visible on anatomical MRI, as well as unnecessarily irradiating normal brain tissue, adversely affecting clinical outcome and increasing toxicity. While most tumor progression occurs locally within the 2 cm expansion of the hyperintensity lesion from T2-weighted images, tumor progression can occur beyond the 2 cm expansion target volume for about 10-37% of patients [5, 6, 7, 8, 9], with more recent studies reporting upwards of 25% [10]. Up to 60% of the irradiated tissue in the high-dose field can be normal-appearing brain [11], which can cause neurotoxicity. This, in turn, can negatively affect a patient's cognitive function, quality-of-life, and overall survival (OS)[12, 13]. The introduction of anti-angiogenic agents also alters the pattern of

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

tumor recurrence, with non-enhancing tumor progression becoming more prevalent than previously observed, further complicating target planning.

In this work, we propose a novel multimodal approach for defining RT clinical target volumes (CTVs) utilizing deep learning-driven predictions of GBM recurrence patterns from either pre-RT or pre-surgery anatomical, diffusion, and metabolic MR images. We compare the resulting predicted volumes to the SOC 1–2 cm uniform expansion of anatomical lesion CTVs for their ability to cover the extent of the lesion at the time of recurrence. Our results demonstrate that this comprehensive personalized strategy, either with pre-surgery or pre-RT scans, produces a more biologically relevant definition of RT target volumes based on the true extent of infiltrating tumor, which more closely overlaps with the region of progression, while minimizing the dose to the normal brain, thus encouraging future work in this area.

Our contributions include: 1) the formulation of planning the CTV for GBM RT as a modified segmentation task that predicts tumor progression, improving upon SOC definition; 2) a novel loss function and tumor segmentation metric that takes into account lesion size; 3) experimental results demonstrating the flexibility of the approach using MR imaging obtained at either pre-surgery or pre-RT time points; 4) an effective pipeline that performs consistently better than SOC that can be directly implemented in a future clinical trial to determine impact on extending survival outcomes.

2 Related Works

Recent advances in diffusion-weighted and metabolic MRI have enabled voxel-level visualization and characterization of cellular-level measures of tumor involvement [14, 15], yet are largely unused in RT planning outside of a few recent single-arm phase II clinical trials [16, 17, 18, 19, 20, 21].

Increases in apparent diffusion coefficient (ADC) and decreases in fractional anisotropy (FA) using diffusion tensor imaging (DTI) [22, 23, 24] can reflect subclinical tumor invasion, which causes an increase in edema and a decrease in directionality along white matter tracts [25]. Metabolite levels estimated using proton Magnetic Resonance Spectroscopic Imaging (^{1}H-MRSI) and the derived Choline-to-NAA index (CNI) can probe underlying cellular metabolism associated with infiltrative tumor [25, 26], hypoxia [27], tumor progression, and survival [14, 28, 29]. Although these MRSI [16, 17, 18, 21, 30, 31, 32] and other imaging methods such as ^{18}F-FET-PET and ^{11}C-MET-PET [33, 34, 35, 36] have shown great promise in more accurately predicting tumor infiltration for incorporation into RT planning, these studies have been limited to simulations or retrospective analyses, lacking prospective evaluation in a clinical trial.

More recent prospective single-arm phase II studies have used imaging to either guide dose escalation based on ^{18}F-DOPA-PET [37] or choline/NAA > 2 from MRSI [17], or boost regions based on elevated relative cerebral blood volume or hypercellularity volume defined on high b-value diffusion images [38, 39]. Although these studies demonstrated significant improvements in outcome (92% 12-month OS rate) compared to historical controls [38], they relied on images describing the current characteristics of the tumor prior to radiation, without modeling where the microscopic infiltrative tumor would ultimately progress.

Recent works in using deep learning for brain lesion segmentation use Swin-like transformer based approaches [40, 41, 42] on multimodal MRI input. These approaches provide strong performance on BraTS 2021 challenge [43]. However, they generate segmentation of the tumor voxels in the input image and hence do not solve for predicting progressed lesion mask post RT.

3 Methodology

We model CTV generation as a predictive segmentation task where we generate the mask of voxels at the time of recurrence, given multi-modal MRI scans (metabolic, diffusion-weighted, and anatomical) acquired immediately before surgery or RT.

Data As input to the model, we include anatomical T2 FLAIR (FLA) and T1 post-contrast (T1C) images, Apparent Diffusion Coefficient (ADC) and Fractional Anisotropy (FA) maps quantified from diffusion-weighted imaging, and Choline-to-Creatine Index (CCrI) and Choline-to-NAA (CNI) from MRSI (metabolic imaging). We use this specific combination of 6 images as it performs the

best in our ablation study (Appendix Figure 2). At recurrence, we use semi-automated methods to generate anatomical regions of interest (ROIs) – the T1 contrast-enhancing lesion (CEL) and T2 FLAIR hyperintensity lesion (T2L) – to construct the ground truth labels. Appendix B.2 provides the data preparation and processing details.

Progression Model We stack 6 pre-surgery or pre-RT 3D volumes of intensity as 6 channels of a 3D image to form the input x, while we consider the union of both the 3D progression masks as the ground-truth mask y. Since x is a 3D image and y is a 3D image mask, we model this as a segmentation task even though y is the segmentation mask of x in the future (at recurrence) to arrive at an approximation ($\hat{y} = f'(x)$) of the true mapping ($y = f(x)$) to determine the progression CTV. We use 2 SOTA encoder-decoder medical segmentation models from different families: EquiUNet [44, 45] from the UNet family, and SegFormer3D [46] from the Transformer (ViT) family.

Objective Patients with smaller lesions pose the problem of a more imbalanced dataset with larger negative class samples consisting of non-lesion brain tissue. To counteract this imbalance and enhance sensitivity, false negatives require a higher weight, which can be obtained through implementing the flexible Tversky Loss. We propose curating the False Positive weight (α) and False Negative weight (β) based on lesion size to help solve the imbalance problem and determine more optimal decision thresholds for the model, potentially improving its performance. We coin this variation of Tversky Index as the Progression Coverage Coefficient (PCC) in Eq. 1. Our PCC formulation allows balancing the tradeoff between sensitivity and specificity per input sample, whereby patients with large tumors benefit from weighting towards higher specificity (to prevent overtreating normal brain), while patients with small tumors benefit from enforcing a higher sensitivity (to ensure complete coverage of the progressed lesion).

$$PCC = \frac{TP}{TP + \alpha FP + \beta FN} \quad \text{where } \beta = \frac{1}{f+1},\ \alpha = 1 - \beta,\ f = \frac{\#\text{ lesion voxels}}{\#\text{ brain voxels}} \tag{1}$$

Given the ground-truth Y, and the prediction $\hat{Y} = f'(X)$, the PCC Loss can be evaluated as:

$$L_{pcc}(Y, \hat{Y}) = 1 - \left(\frac{\|\hat{Y} \cdot Y\|_1}{\|\hat{Y} \cdot Y\|_1 + \alpha \|\hat{Y} \cdot (1 - Y)\|_1 + \beta \|(1 - \hat{Y}) \cdot Y\|_1} \right) \tag{2}$$

We define the objective function of our task as the combination of PCC Loss (Eq. 2) with the Binary Cross Entropy Loss (L_{bce}) weighted by scalar λ (Eq. 3).

$$L(Y, \hat{Y}; \lambda) = L_{pcc}(Y, \hat{Y}) + \lambda L_{bce}(Y, \hat{Y}) \tag{3}$$

Evaluation We evaluate model performance using 4 metrics: 1) sensitivity, measured to account for the coverage of the tumor, with high sensitivity being a necessary requirement; 2) specificity, measured to evaluate the amount of normal brain spared; 3) Dice coefficient to get a combined sense of precision and recall, and 4) the newly-derived and individualized PCC (Eq. 1), quantified to take into account the tumor size when weighting FPs and FNs. The importance of using PCC, as an evaluation metric as well, is underscored by the fact that the CTV, which was modeled using only the input anatomical lesions, had the highest value for both mIOU and Dice score (Table 1 – "No Prog") but should be the poorest model as it neglects to cover any infiltrating tumor cells that later progress; This is correctly reflected by the model's extremely low sensitivity. The PCC scores for this CTV were low, capturing its poor performance. In contrast, both SOC CTVs had low mIOU and Dice scores but comparatively higher PCC scores. We compare our approach against the two commonly utilized SOC CTV definitions. 1) The Radiation Therapy Oncology Group (RTOG) recommends a more aggressive treatment CTV, which combines CEL and T2L with a 2 cm uniform expansion [47, 48]; 2) The European Organization for Research and Treatment of Cancer (EORTC) recommends a more conservative CTV, which expands the CEL along with the resection cavity uniformly by 1.5 cm [49] [50], excluding any vasogenic edema observed on the FLA image [51].

4 Experiments

We conduct experiments on two different datasets: 1) **pre-surgery data** — 92 patients newly-diagnosed with GBM whose MRI scans were acquired 1-3 days before surgery; 2) **pre-RT data**

Table 1: **Model performance comparison across datasets.** This compares both our deep learning approaches, EquiUNet (here UNet) and SegFormer3D (here ViT), against the SOC baselines and the case where we generate a CTV with no progression (here No Prog). The numbers show the mean across the test patients along with the standard deviation. PCC more accurately captures lesion segmentation performance than traditional metrics like Dice and mIOU as the latter find No Prog the best. Our approaches outperform SOC CTVs on PCC.

Data	Model	Sensitivity	Specificity	mIOU	Dice	PCC
Pre-Surgery	No Prog	0.60 ± 0.22	1.00 ± 0.00	0.60 ± 0.22	0.73 ± 0.19	0.62 ± 0.22
	EORTC	0.77 ± 0.17	0.92 ± 0.04	0.43 ± 0.12	$\mathbf{0.59 \pm 0.12}$	0.74 ± 0.16
	RTOG	0.90 ± 0.12	0.80 ± 0.06	0.27 ± 0.10	0.42 ± 0.11	0.81 ± 0.11
	UNet	0.82 ± 0.14	$\mathbf{0.93 \pm 0.03}$	$\mathbf{0.44 \pm 0.17}$	0.59 ± 0.17	0.80 ± 0.13
	Vit	$\mathbf{0.92 \pm 0.10}$	0.82 ± 0.05	0.29 ± 0.11	0.44 ± 0.14	$\mathbf{0.82 \pm 0.09}$
Pre-RT	No Prog	0.46 ± 0.18	1.00 ± 0.00	0.46 ± 0.18	0.61 ± 0.17	0.47 ± 0.18
	EORTC	0.88 ± 0.11	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.28 \pm 0.13}$	$\mathbf{0.43 \pm 0.16}$	0.82 ± 0.10
	RTOG	$\mathbf{0.96 \pm 0.09}$	0.78 ± 0.10	0.18 ± 0.06	0.30 ± 0.09	0.83 ± 0.08
	UNet	0.93 ± 0.08	0.88 ± 0.04	0.27 ± 0.13	0.41 ± 0.16	$\mathbf{0.85 \pm 0.07}$
	ViT	0.90 ± 0.15	0.82 ± 0.10	0.20 ± 0.06	0.34 ± 0.08	0.80 ± 0.12

— 101 patients with GBM post-surgical resection and scanned within 1 week of beginning RT. All patients were diagnosed with a pathologically confirmed primary GBM according to WHO 2016 criteria and followed up with clinical MRI scans until progression was confirmed. Appendix B.1 contains further details about data acquisition. Pre-surgery data was split into 54/27/11 (train/val/test) sets, while the pre-RT dataset was split 67/16/18. We carried out experiments for both these datasets with two models, EquiUNet and SegFormer3D, comparing them against the 2 SOC definitions: RTOG CTV and EORTC CTV. Appendix C.1 describes model training hyperparameters.

5 Results

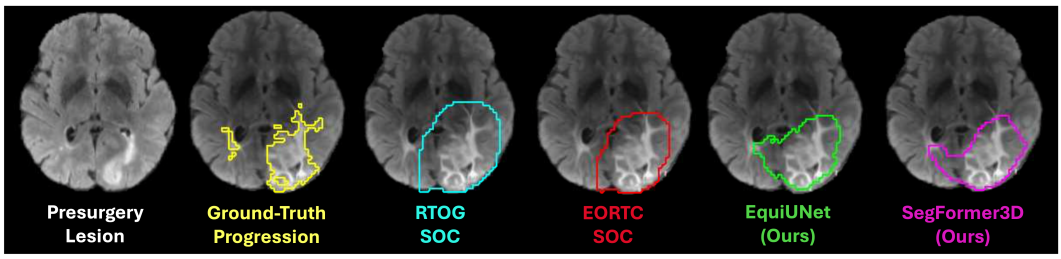
Table 1 compares the results of our models to the SOC CTVs based on the necessary and sufficient metrics for both the pre-surgery and pre-RT data versions. As the RTOG CTV is the most aggressive treatment, it has the highest sensitivity for pre-RT; while EORTC is more conservative in nature and therefore has the highest specificity when generated on the pre-RT dataset. In the case of pre-surgery scans, Segformer3D outperforms both the SOC CTVs in terms of sensitivity and PCC; while EquiUNet has the highest specificity but misses parts of the tumor, as evident by the lower sensitivity and PCC. In the case of pre-RT scans, EquiUNet performs the best overall in terms of having comparably high sensitivity to the RTOG CTV, but with similar specificity to the EORTC CTV, resulting in the highest PCC.

Figure 1 visually demonstrates the improvements that the deep learning approaches bring. For the example from the pre-surgery data (Fig.1a), both the SOC CTVs miss part of the contralateral progressed lesion (yellow) that the deep learning approaches were able to capture, while also sparing more normal brain. In the pre-RT example (Fig.1b), the RTOG CTV has the highest sensitivity, but unnecessarily treats a large portion of unaffected occipital lobe tissue. The more conservative EORTC CTV exhibits higher specificity, but misses a part of the progressed lesion in the anterior portion of the temporal lobe that is covered by the deep learning approaches. EquiUNet has better coverage while still being specific and sparing the normal-appearing brain.

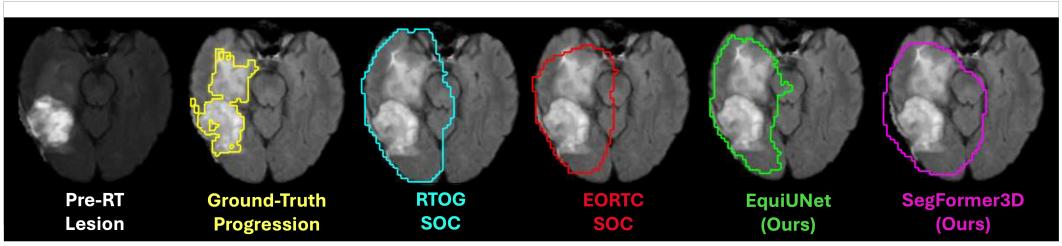
Results from our comparative study between different objective functions (Appendix Figure 3) demonstrate that models trained using our combined objective L_{pcc_bce} achieve higher sensitivity.

6 Discussion

Our results demonstrate the performance gain achieved by our multimodal segmentation-based approach for CTV generation over the SOC CTV definitions. Comparable coverage of the progressed region was achieved while sparing more normal brain. Depending on which time point MRI scan is



(a) Pre-surgery.



(b) Pre-RT.

Figure 1: T2-FLAIR scans from (a) pre-surgery data and (b) pre-RT data comparing performance of our proposed deep learning approach against the baselines and the ground-truth. The left-most scan in each depicts the input scan. Our approaches capture progressed voxels that are missed in the SOC CTVs, while remaining specific and sparing the normal-appearing brain.

used for treatment planning, a different model is recommended. For the pre-surgery case, we see that SegFormer3D, in general, performs better than EquiUNet, while EquiUNet performs better using pre-RT MRI exams. Given that ViTs tend to perform better when more information is available, it is not surprising that they performed better on pre-surgery data, where there is more tumor available to learn from before it is resected. Achieving at-par performance with Vision Transformers even in our low-resource setting is a strong signal for its potential as available data increases over time. Our experiments demonstrate the effectiveness of incorporating multiple, biologically relevant MRI sequences using deep learning with a lesion-size-aware objective for better RT planning, potentially leading to better outcomes for these patients in the future.

7 Conclusion

Our study demonstrates the ability to predict regions of tumor progression and automatically generate clinical target volumes for RT planning with improved sparing of the normal-appearing brain as compared to two different SOC recommendations without compromising on the coverage. Our work also demonstrates that using PCC as an objective improves performance. While as a metric, it more accurately captures segmentation performance where traditional metrics such as Dice score and mIOU fail for this spatially imbalanced task. Future studies will prospectively validate these findings in additional cohorts that include patients who were originally treated according to the most recent EORTC and RTOG guidelines before incorporating this approach in a clinical trial.

Limitations The primary limitation of our work is the lack of data. Even though we verify our approach on two datasets, they are both low-resource and are inherently different, so while our dataset is much larger than most studies, it is still considered small for deep learning tasks. Although the inclusion of multiple modalities gives us a more informed model, the metabolic MRSI is low-resolution and is not routinely performed in clinical practice, and this could limit widespread adoption. However, since the acquisition of this data, higher resolution MRSI has become more widely available within clinically reasonable scan times, and there has been more interest in developing open-source packages for post-processing [27, 52, 53, 54].

References

[1] Roger Stupp, Warren P. Mason, Martin J. van den Bent, Michael Weller, Barbara Fisher, Martin J.B. Taphoorn, Karl Belanger, Alba A. Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C. Janzer, Samuel K. Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J. Gregory Cairncross, Elizabeth Eisenhauer, and René O. Mirimanoff. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[2] Roger Stupp, Sophie Taillibert, Andrew Kanner, William Read, David M. Steinberg, Benoit Lhermitte, Steven Toms, Ahmed Idbaih, Manmeet S. Ahluwalia, Karen Fink, Francesco Di Meco, Frank Lieberman, Jay-Jiguang Zhu, Giuseppe Stragliotto, David D. Tran, Steven Brem, Andreas F. Hottinger, Eilon D. Kirson, Gitit Lavy-Shahaf, Uri Weinberg, Chae-Yong Kim, Sun-Ha Paek, Garth Nicholas, Jordi Bruna, Hal Hirte, Michael Weller, Yoram Palti, Monika E. Hegi, and Zvi Ram. Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: A randomized clinical trial. *JAMA*, 318(23):2306–2316, 12 2017.

[3] Olivier L Chinot, Wolfgang Wick, Warren Mason, Roger Henriksson, Frank Saran, Ryo Nishikawa, Antoine F Carpentier, Khe Hoang-Xuan, Petr Kavan, Dana Cernea, Alba A Brandes, Magalie Hilton, Lauren Abrey, and Timothy Cloughesy. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N. Engl. J. Med.*, 370(8):709–722, February 2014.

[4] A Giese, R Bjerkvig, M E Berens, and M Westphal. Cost of migration: invasion of malignant gliomas and implications for treatment. *J. Clin. Oncol.*, 21(8):1624–1636, April 2003.

[5] Marion Rapp, Jessica Baernreuther, Bernd Turowski, Hans-Jakob Steiger, Michael Sabel, and Marcel A Kamp. Recurrence pattern analysis of primary glioblastoma. *World Neurosurg.*, 103:733–740, July 2017.

[6] Brian J Gebhardt, Michael C Dobelbower, William H Ennis, Asim K Bag, James M Markert, and John B Fiveash. Patterns of failure for glioblastoma multiforme following limited-margin radiation and concurrent temozolomide. *Radiat. Oncol.*, 9(1):130, June 2014.

[7] Giuseppe Minniti, Dante Amelio, Maurizio Amichetti, Maurizio Salvati, Roberta Muni, Alessandro Bozzao, Gaetano Lanzetta, Stefania Scarpino, Antonella Arcella, and Riccardo Maurizi Enrici. Patterns of failure and comparison of different target volume delineations in patients with glioblastoma treated with conformal radiotherapy plus concomitant and adjuvant temozolomide. *Radiother. Oncol.*, 97(3):377–381, December 2010.

[8] Seo Hee Choi, Jun Won Kim, Jee Suk Chang, Jae Ho Cho, Se Hoon Kim, Jong Hee Chang, and Chang-Ok Suh. Impact of including peritumoral edema in radiotherapy target volume on patterns of failure in glioblastoma following temozolomide-based chemoradiotherapy. *Sci. Rep.*, 7:42148, February 2017.

[9] Michael C Dobelbower, Omer L Burnett, Iii, Robert A Nordal, Louis B Nabors, James M Markert, Mark D Hyatt, and John B Fiveash. Patterns of failure for glioblastoma multiforme following concurrent radiation and temozolomide. *J. Med. Imaging Radiat. Oncol.*, 55(1):77–81, February 2011.

[10] Christopher H Chapman, Jared H Hara, Annette M Molinaro, Jennifer L Clarke, Nancy Ann Oberheim Bush, Jennie W Taylor, Nicholas A Butowski, Susan M Chang, Shannon E Fogh, Penny K Sneed, Jean L Nakamura, David R Raleigh, and Steve E Braunstein. Reirradiation of recurrent high-grade glioma and development of prognostic scores for progression and survival. *Neurooncol. Pract.*, 6(5):364–374, September 2019.

[11] Ilwoo Park, Gregory Tamai, Michael C Lee, Cynthia F Chuang, Susan M Chang, Mitchel S Berger, Sarah J Nelson, and Andrea Pirzkall. Patterns of recurrence analysis in newly diagnosed glioblastoma multiforme after three-dimensional conformal radiation therapy with respect to pre-radiation therapy magnetic resonance spectroscopic findings. *Int. J. Radiat. Oncol. Biol. Phys.*, 69(2):381–389, October 2007.

[12] Y R Lawrence, M Wang, A P Dicker, D Andrews, W J Curran, Jr, J M Michalski, L Souhami, W-Ka Yung, and M Mehta. Early toxicity predicts long-term survival in high-grade glioma. *Br. J. Cancer*, 104(9):1365–1371, April 2011.

[13] Walter Stummer, Hanns-Jürgen Reulen, Thomas Meinel, Uwe Pichlmeier, Wiebke Schumacher, Jörg-Christian Tonn, Veit Rohde, Falk Oppel, Bernd Turowski, Christian Woiciechowsky, Kea Franz, Torsten Pietsch, and ALA-Glioma Study Group. Extent of resection and survival in glioblastoma multiforme: identification of and adjustment for bias. *Neurosurgery*, 62(3):564–76; discussion 564–76, March 2008.

[14] Sarah J Nelson. Assessment of therapeutic response and treatment planning for brain tumors using metabolic and physiological MRI. *NMR Biomed.*, 24(6):734–749, July 2011.

[15] Sarah J Nelson and Soonmee Cha. Imaging glioblastoma multiforme. *Cancer J.*, 9(2):134–145, March 2003.

[16] J Scott Cordova, Shravan Kandula, Saumya Gurbani, Jim Zhong, Mital Tejani, Oluwatosin Kayode, Kirtesh Patel, Roshan Prabhu, Eduard Schreibmann, Ian Crocker, Chad A Holder, Hyunsuk Shim, and Hui-Kuo Shu. Simulating the effect of spectroscopic MRI as a metric for radiation therapy planning in patients with glioblastoma. *Tomography*, 2(4):366–373, December 2016.

[17] Douglas B Einstein, Barry Wessels, Barbara Bangert, Pingfu Fu, A Dennis Nelson, Mark Cohen, Stephen Sagar, Jonathan Lewin, Andrew Sloan, Yiran Zheng, Jordonna Williams, Valdir Colussi, Robert Vinkler, and Robert Maciunas. Phase II trial of radiosurgery to magnetic resonance spectroscopy-defined high-risk tumor volumes in patients with glioblastoma multiforme. *Int. J. Radiat. Oncol. Biol. Phys.*, 84(3):668–674, November 2012.

[18] N Andres Parra, Andrew A Maudsley, Rakesh K Gupta, Fazilat Ishkanian, Kris Huang, Gail R Walker, Kyle Padgett, Bhaswati Roy, Joseph Panoff, Arnold Markoe, and Radka Stoyanova. Volumetric spectroscopic imaging of glioblastoma multiforme radiation treatment volumes. *Int. J. Radiat. Oncol. Biol. Phys.*, 90(2):376–384, October 2014.

[19] Donggeon Heo, Jisoo Lee, Roh-Eul Yoo, Seung Hong Choi, Tae Min Kim, Chul-Kee Park, Sung-Hye Park, Jae-Kyung Won, Joo Ho Lee, Soon Tae Lee, Kyu Sung Choi, Ji Ye Lee, Inpyeong Hwang, Koung Mi Kang, and Tae Jin Yun. Deep learning based on dynamic susceptibility contrast MR imaging for prediction of local progression in adult-type diffuse glioma (grade 4). *Sci. Rep.*, 13(1):13864, August 2023.

[20] Hatef Mehrabian, Kimberly L Desmond, Hany Soliman, Arjun Sahgal, and Greg J Stanisz. Differentiation between radiation necrosis and tumor progression using chemical exchange saturation transfer. *Clin. Cancer Res.*, 23(14):3667–3675, July 2017.

[21] Alexandra Deviers, Soléakhéna Ken, Thomas Filleron, Benjamin Rowland, Andrea Laruelo, Isabelle Catalaa, Vincent Lubrano, Pierre Celsis, Isabelle Berry, Giovanni Mogicato, Elizabeth Cohen-Jonathan Moyal, and Anne Laprie. Evaluation of the lactate-to-n-acetyl-aspartate ratio defined with magnetic resonance spectroscopic imaging before radiation therapy as a new predictive marker of the site of relapse in patients with glioblastoma multiforme. *Int. J. Radiat. Oncol. Biol. Phys.*, 90(2):385–393, October 2014.

[22] Stelios Angeli, Kyrre E Emblem, Paulina Due-Tonnessen, and Triantafyllos Stylianopoulos. Towards patient-specific modeling of brain tumor growth and formation of secondary nodes guided by DTI-MRI. *NeuroImage Clin.*, 20:664–673, August 2018.

[23] J C L Alfonso, K Talkenberger, M Seifert, B Klink, A Hawkins-Daarud, K R Swanson, H Hatzikirou, and A Deutsch. The biology and mathematical modelling of glioma invasion: a review. *J. R. Soc. Interface*, 14(136):20170490, November 2017.

[24] Vishnu Anand Cuddapah, Stefanie Robel, Stacey Watkins, and Harald Sontheimer. A neurocentric perspective on glioma invasion. *Nat. Rev. Neurosci.*, 15(7):455–465, July 2014.

[25] T R McKnight, S M Noworolski, D B Vigneron, and S J Nelson. An automated technique for the quantitative assessment of 3D-MRSI data from patients with glioma. *J. Magn. Reson. Imaging*, 13(2):167–177, February 2001.

[26] Tracy R McKnight, Mary H von dem Bussche, Daniel B Vigneron, Ying Lu, Mitchel S Berger, Michael W McDermott, William P Dillon, Edward E Graves, Andrea Pirzkall, and Sarah J Nelson. Histopathological validation of a three-dimensional magnetic resonance spectroscopy index as a predictor of tumor presence. *J. Neurosurg.*, 97(4):794–802, October 2002.

[27] Ilwoo Park, Albert P Chen, Matthew L Zierhut, Esin Ozturk-Isik, Daniel B Vigneron, and Sarah J Nelson. Implementation of 3 T lactate-edited 3D 1H MR spectroscopic imaging with flyback echo-planar readout for gliomas patients. *Ann. Biomed. Eng.*, 39(1):193–204, January 2011.

[28] Sarah J Nelson, Edward Graves, Andrea Pirzkall, Xiaojuan Li, Antionette Antiniw Chan, Daniel B Vigneron, and Tracy R McKnight. In vivo molecular imaging for planning radiation therapy of gliomas: an application of 1H MRSI. *J. Magn. Reson. Imaging*, 16(4):464–476, October 2002.

[29] Mekhail Anwar, Annette M Molinaro, Olivier Morin, Susan M Chang, Daphne A Haas-Kogan, Sarah J Nelson, and Janine M Lupo. Identifying voxels at risk for progression in glioblastoma based on dosimetry, physiologic and metabolic MRI. *Radiat. Res.*, 188(3):303–313, September 2017.

[30] Daniel R Wahl, Michelle M Kim, Madhava P Aryal, Holly Hartman, Theodore S Lawrence, Matthew J Schipper, Hemant A Parmar, and Yue Cao. Combining perfusion and high b-value diffusion MRI to inform prognosis and predict failure patterns in glioblastoma. *Int. J. Radiat. Oncol. Biol. Phys.*, 102(4):757–764, November 2018.

[31] Priyanka P Pramanik, Hemant A Parmar, Aaron G Mammoser, Larry R Junck, Michelle M Kim, Christina I Tsien, Theodore S Lawrence, and Yue Cao. Hypercellularity components of glioblastoma identified by high b-value diffusion-weighted imaging. *Int. J. Radiat. Oncol. Biol. Phys.*, 92(4):811–819, July 2015.

[32] Jatta Berberat, Jane McNamara, Luca Remonda, Stephan Bodis, and Susanne Rogers. Diffusion tensor imaging for target volume definition in glioblastoma multiforme. *Strahlenther. Onkol.*, 190(10):939–943, October 2014.

[33] Stefan Rieken, Daniel Habermehl, Frederik L Giesel, Christoph Hoffmann, Ute Burger, Harald Rief, Thomas Welzel, Uwe Haberkorn, Jürgen Debus, and Stephanie E Combs. Analysis of FET-PET imaging for target volume definition in patients with gliomas treated with conformal radiotherapy. *Radiother. Oncol.*, 109(3):487–492, December 2013.

[34] Masayuki Matsuo, Kazuhiro Miwa, Osamu Tanaka, Jun Shinoda, Hironori Nishibori, Yusuke Tsuge, Hirohito Yano, Toru Iwama, Shinya Hayashi, Hiroaki Hoshi, Jitsuhiro Yamada, Masayuki Kanematsu, and Hidefumi Aoyama. Impact of [11c]methionine positron emission tomography for target definition of glioblastoma multiforme in radiation therapy planning. *Int. J. Radiat. Oncol. Biol. Phys.*, 82(1):83–89, January 2012.

[35] Sean Miller, Pin Li, Matthew Schipper, Larry Junck, Morand Piert, Theodore S Lawrence, Christina Tsien, Yue Cao, and Michelle M Kim. Metabolic tumor volume response assessment using (11)c-methionine positron emission tomography identifies glioblastoma tumor subregions that predict progression better than baseline or anatomic magnetic resonance imaging alone. *Adv. Radiat. Oncol.*, 5(1):53–61, January 2020.

[36] Kazuhiro Miwa, Masayuki Matsuo, Shin-Ichi Ogawa, Jun Shinoda, Yoshitaka Asano, Takeshi Ito, Kazutoshi Yokoyama, Jitsuhiro Yamada, Hirohito Yano, and Toru Iwama. Hypofractionated high-dose irradiation with positron emission tomography data for the treatment of glioblastoma multiforme. *Biomed Res. Int.*, 2014:407026, May 2014.

[37] Nadia Nicole Laack, Deanna Pafundi, S Keith Anderson, Timothy Kaufmann, Val Lowe, Christopher Hunt, Diane Vogen, Elizabeth Yan, Jann Sarkaria, Paul Brown, Sani Kizilbash, Joon Uhm, Michael Ruff, Mark Zakhary, Yan Zhang, Maasa Seaberg, Hok Seum Wan Chan Tseung, Brian Kabat, Bradley Kemp, and Debra Brinkmann. Initial results of a phase 2 trial of 18F-DOPA PET-guided dose-escalated radiation therapy for glioblastoma. *Int. J. Radiat. Oncol. Biol. Phys.*, 110(5):1383–1395, August 2021.

[38] Michelle M Kim, Yilun Sun, Madhava P Aryal, Hemant A Parmar, Morand Piert, Benjamin Rosen, Charles S Mayo, James M Balter, Matthew Schipper, Nicolette Gabel, Emily M Briceño, Daekeun You, Jason Heth, Wajd Al-Holou, Yoshie Umemura, Denise Leung, Larry Junck, Daniel R Wahl, Theodore S Lawrence, and Yue Cao. A phase 2 study of dose-intensified chemoradiation using biologically based target volume definition in patients with newly diagnosed glioblastoma. *Int. J. Radiat. Oncol. Biol. Phys.*, 110(3):792–803, July 2021.

[39] Michelle M Kim, Hemant A Parmar, Madhava P Aryal, Charles S Mayo, James M Balter, Theodore S Lawrence, and Yue Cao. Developing a pipeline for multiparametric MRI-guided radiation therapy: Initial results from a phase II clinical trial in newly diagnosed glioblastoma. *Tomography*, 5(1):118–126, March 2019.

[40] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 272–284. Springer International Publishing, Cham, 2022.

[41] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. SwinUNETR-V2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 416–426. Springer Nature Switzerland, Cham, 2023.

[42] Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. SwinBTS: A method for 3D multimodal brain tumor segmentation using swin transformer. *Brain Sci.*, 12(6):797, June 2022.

[43] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe Campos Kitamura, Sarthak Pati, Luciano Prevedello, Jeffrey Rudie, Chiharu Sako, Russell Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B Freymann, Justin S Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-André Weber, Abhishek Mahajan, Bjoern Menze, Adam E Flanders, and Spyridon Bakas. RSNA-ASNR-MICCAI-BraTS-2021, 2023.

[44] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D u-net: Learning dense volumetric segmentation from sparse annotation. 2016.

[45] Theophraste Henry, Alexandre Carre, Marvin Lerousseau, Theo Estienne, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Brain tumor segmentation with self-ensembled, deeply-supervised 3D u-net neural networks: a BraTS 2020 challenge solution. 2020.

[46] Shehan Perera, Pouyan Navard, and Alper Yilmaz. SegFormer3D: an efficient transformer for 3D medical image segmentation. 2024.

[47] Alvin R Cabrera, John P Kirkpatrick, John B Fiveash, Helen A Shih, Eugene J Koay, Stephen Lutz, Joshua Petit, Samuel T Chao, Paul D Brown, Michael Vogelbaum, David A Reardon, Arnab Chakravarti, Patrick Y Wen, and Eric Chang. Radiation therapy for glioblastoma: Executive summary of an american society for radiation oncology Evidence-Based clinical practice guideline. *Pract. Radiat. Oncol.*, 6(4):217–225, July 2016.

[48] Arnab Chakravarti, Meihua Wang, H Ian Robins, Tim Lautenschlaeger, Walter J Curran, David G Brachman, Christopher J Schultz, Ali Choucair, Marisa Dolled-Filhart, Jason Christiansen, Mark Gustavson, Annette Molinaro, Paul Mischel, Adam P Dicker, Markus Bredel, and Minesh Mehta. RTOG 0211: a phase 1/2 study of radiation therapy with concurrent gefitinib for newly diagnosed glioblastoma patients. *Int. J. Radiat. Oncol. Biol. Phys.*, 85(5):1206–1211, April 2013.

[49] Maximilian Niyazi, Michael Brada, Anthony J Chalmers, Stephanie E Combs, Sara C Erridge, Alba Fiorentino, Anca L Grosu, Frank J Lagerwaard, Giuseppe Minniti, René-Olivier Mirimanoff, Umberto Ricardi, Susan C Short, Damien C Weber, and Claus Belka. ESTRO-ACROP guideline "target delineation of glioblastomas". *Radiother. Oncol.*, 118(1):35–42, January 2016.

[50] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V Miller, Steve Pieper, and Ron Kikinis. 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging*, 30(9):1323–1341, November 2012.

[51] Maximilian Niyazi, Nicolaus Andratschke, Martin Bendszus, Anthony J Chalmers, Sara C Erridge, Norbert Galldiks, Frank J Lagerwaard, Pierina Navarria, Per Munck Af Rosenschöld, Umberto Ricardi, Martin J van den Bent, Michael Weller, Claus Belka, and Giuseppe Minniti. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. *Radiother. Oncol.*, 184(109663):109663, July 2023.

[52] Jason C Crane, Marram P Olson, and Sarah J Nelson. SIVIC: Open-source, standards-based software for DICOM MR spectroscopy workflows. *Int. J. Biomed. Imaging*, 2013:169526, July 2013.

[53] Wolfgang Bogner, Ricardo Otazo, and Anke Henning. Accelerated MR spectroscopic imaging-a review of current and emerging techniques. *NMR Biomed.*, 34(5):e4314, May 2021.

[54] Mohammad Sabati, Jiping Zhan, Varan Govind, Kristopher L Arheart, and Andrew A Maudsley. Impact of reduced k-space acquisition on pathologic detectability for volumetric MR spectroscopic imaging. *J. Magn. Reson. Imaging*, 39(1):224–234, January 2014.

[55] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum. Brain Mapp.*, 40(17):4952–4964, December 2019.

[56] Julio M Duarte-Carvajalino, Guillermo Sapiro, Noam Harel, and Christophe Lenglet. A framework for linear and non-linear registration of diffusion-weighted MRIs using angular interpolation. *Front. Neurosci.*, 7:41, April 2013.

[57] M Jenkinson and S Smith. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.*, 5(2):143–156, June 2001.

[58] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, October 2002.

[59] Tony C. W. Mok and Albert C. S. Chung. Robust image registration with absent correspondences in pre-operative and follow-up brain mri scans of diffuse glioma patients. In Spyridon Bakas, Alessandro Crimi, Ujjwal Baid, Sylwia Malec, Monika Pytlarz, Bhakti Baheti, Maximilian Zenk, and Reuben Dorent, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 231–240, Cham, 2023. Springer Nature Switzerland.

[60] Michelle M Kim, Corey Speers, Pin Li, Matthew Schipper, Larry Junck, Denise Leung, Daniel Orringer, Jason Heth, Yoshie Umemura, Daniel E Spratt, Daniel R Wahl, Yue Cao, Theodore S Lawrence, and Christina I Tsien. Dose-intensified chemoradiation is associated with altered patterns of failure and favorable survival in patients with newly diagnosed glioblastoma. *J. Neurooncol.*, 143(2):313–319, June 2019.

A Additional Experiments

Additional experiments were performed to find the best combination of MRI input modalities and the effectiveness of using the PCC + BCE combination loss function.

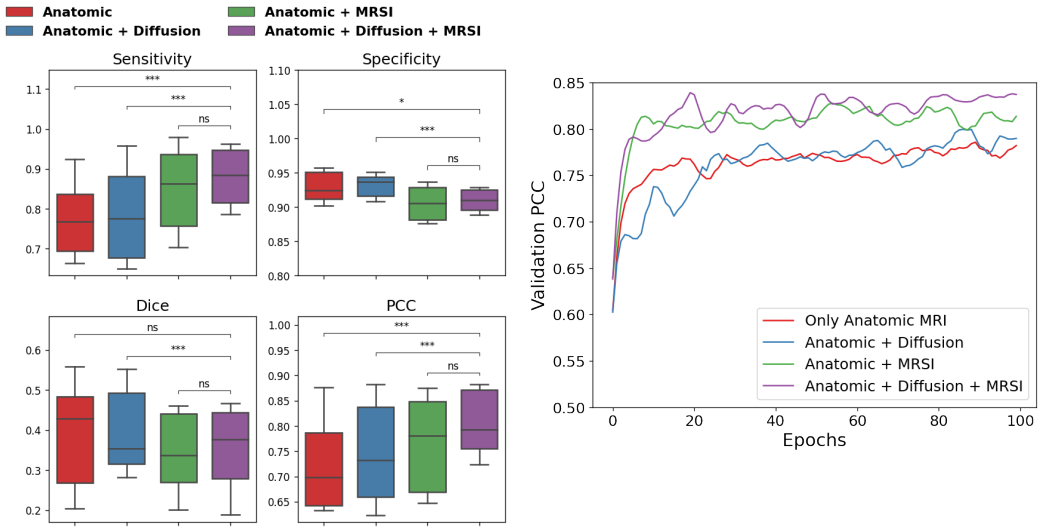


Figure 2: **Model performance comparison among different MRI input modalities.** (Left) Sensitivity, Specificity, Dice, and PCC metric comparison for pre-RT data. All models were trained using the PCC + BCE loss function. Wilcoxon signed rank tests were used with significant levels defined as *, **, *** for p-values < 0.05, 0.01, and 0.001, respectively. (Right) The PCC score on the validation set. The model trained using all MRI modalities achieved a significantly higher PCC by improving the sensitivity of the model.

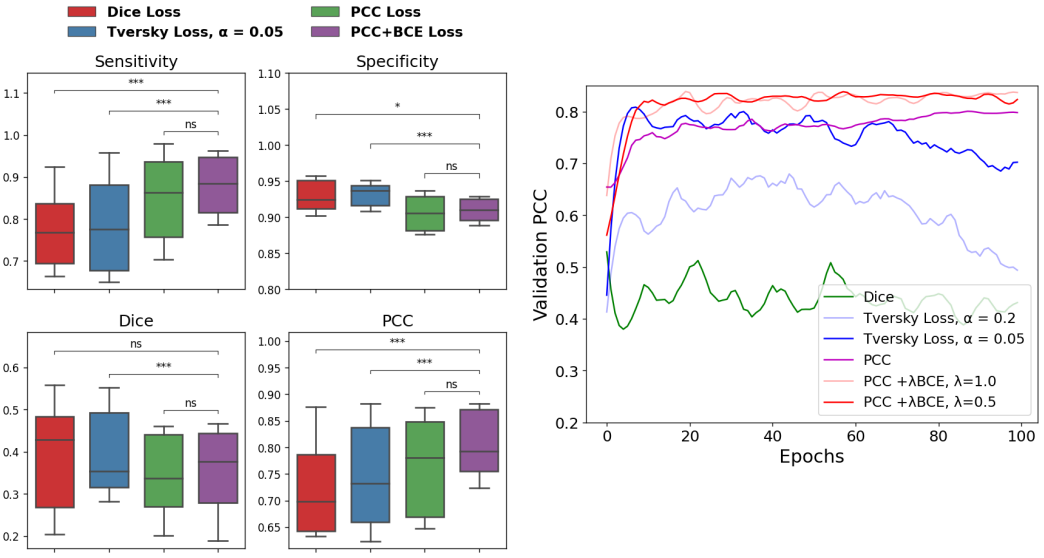


Figure 3: **Comparison of models with different loss functions.** (Left) Sensitivity, Specificity, Dice, and PCC of EquiUNet models trained with different loss functions for pre-RT data. Significant levels from a Wilcoxon rank sum test were defined as *, **, *** for p-values < 0.05, 0.01, and 0.001, respectively. All models were trained and optimized separately using anatomic + diffusion + metabolic as inputs. (Right) The PCC score on the validation set. Models trained using the PCC + BCE combination loss function converged faster and significantly outperformed other variations in terms of sensitivity.

B Data

B.1 Acquisition

A total of 193 patients who were newly diagnosed with a pathologically confirmed primary GBM according to WHO 2016 criteria were included in this retrospective analysis. All patients received SOC treatment, including surgical resection followed by external beam RT (total dose of 60 Gy in 2 Gy fractions over a course of 6 weeks), concomitant daily temozolomide (75 mg/m2), and six cycles of maintenance adjuvant temozolomide chemotherapy (total 150–200 mg/m2). All patients gave informed consent to participate in the research according to guidelines established by the Institutional Review Board (IRB) of the organization.

Out of these, 92 patients received a baseline MRI scan pre-surgical resection, and the remaining 101 patients received the scan post-surgical resection but within 1 week prior to initiating radiotherapy and chemotherapy. These included T2-FLAIR imaging, pre- and post-contrast T1-weighted imaging, DWI, and MRSI. After the course of radiotherapy and chemotherapy, patients were followed serially with clinical MRI scans every two months (including at least pre- and post-contrast T1-weighted and T2-FLAIR imaging) until progression.

The MR examinations were performed on a 3 T GE Signa scanner using an eight-channel phased-array head coil. Standard anatomical imaging included T2-weighted FLAIR and 3D T1-weighted IR-SPGR imaging pre- and post-injection of a gadolinium-based contrast agent. For pre-RT MRI scan, diffusion-tensor images were obtained with b = 1000 s/mm^2, 6-directional diffusion-weighted echo-planar imaging sequence, and 4 b_0 images (repetition-time[TR]/echo-time[TE] = 1000/108 ms, voxel size=1.7–2.0 $\times$ 1.7–2.0 $\times$ 2.0–3.0 mm). Lactate-edited 3D ^{1}H-MRSI was acquired using point-resolved spectroscopy volume localization and very selective saturation bands to avoid chemical shift artifacts as well as to suppress residual lipid signals (excited volume = 80 $\times$ 80 $\times$ 40 mm, TE/TR = 144/1100-1250 ms, over- PRESS-factor = 1.5, nominal voxel size = 1 $\times$ 1 $\times$ 1 cm, flyback echo-planar readout in SI, total acquisition time = 9.5 min, sweep-width=988 Hz, and 712 dwell-points).

B.2 Processing

Image Construction Anatomical ROIs included the T1 contrast-enhancing lesion (CEL), T2 FLAIR hyperintensity lesion (T2L), non-enhancing lesion (NEL; defined as CEL subtracted from the T2L) and normal-appearing voxels (NAV; defined as normal brain tissue from a skull-stripped brain mask obtained using the HD-BET brain extraction tool [55] after subtraction of cavity, ventricles, and lesion ROIs). CEL, T2L, and NEL ROIs were semi-automatically delineated on the pre- and post-contrast T1-weighted images (CEL) and T2-weighted FLAIR images (T2L), using in-house software, before manual inspection and editing by a trained senior research specialist in radiology. All exams in the test set, as well as whenever there was a question on the boundary, were also verified by a study neuroradiologist. From the DTI data, maps of ADC and FA were calculated using FMRIB's Diffusion Toolkit [56] and normalized to the mode of intensities in normal-appearing brain tissue. Spectroscopic data were reconstructed and post processed using in-house software to generate metabolite peak height maps, choline-to-NAA index (CNI), choline-to-creatine index (CCrI), and creatine-to-NAA index (CrNI) from baseline-subtracted, frequency- and phase-corrected spectra on a voxel-by-voxel basis [27, 52].

Image Alignment All images from the pre-surgery and pre-RT timepoints were rigidly aligned to their respective post-contrast T1-weighted image using Slicer's BRAINSFit tool with B-spline warping [50], or FMRIB's FLIRT [57, 58], before being resampled to 3$\times$3$\times$3 mm resolution to mitigate any potential errors due to any residual misalignment. In order to allow for accurate matching of voxels between the input (pre-RT or pre-surgery) and the corresponding progression scans, a deep learning method specifically trained on serial post-resection glioma data with tissue shift [59] as part of the BraTS-Reg 2022 challenge was utilized to align anatomical images at progression to the corresponding input scan, and the resulting transformation matrix was applied to all images and ROI files from the progression scan. This software, which ranked 1st place in the 2022 MICCAI BraTS-Reg challenge, utilized a 3-step deep-learning-based approach to match voxels between pre-treatment and progression scans that consists of: (1) multi-level affine pre-alignment, (2) a conditional deep Laplacian pyramid image registration network (cLapIRN) with forward-backward

Table 2: **Model hyperparameters.** Optimal values that were used to produce the results in Section 5 for EquiUNet, and SegFormer3D for both the datasets.

Data	Model	Epochs	LR	λ	Optim	Warmup
Pre-Surgery	EquiUNet	150	5×10^{-5}	0.6	ranger	-
	SegFormer3D	300	1×10^{-4}	0.9	adamw	20
Pre-RT	EquiUNet	150	5×10^{-5}	0.5	ranger	-
	SegFormer3D	300	1×10^{-4}	1.2	adamw	20

consistency constraints, and (3) a non-linear instance optimization with inverse consistency. The resulting transformation matrix was then applied to all images and ROI files from the progression scan. All outputs were visually inspected by a senior research scientist with over 20 years' experience in verifying serial alignments. In the few cases where the quality of alignment was deemed not sufficient (<5%), non-rigid registration with B-spline warping using Slicer's BrainsFit [60] and intermediate scans was first applied, followed by the deep learning model. We found that this process was able to adequately handle alignment and any tissue shift from the shrinkage of the cavity, visually.

B.3 Release

The imaging dataset used in this study cannot be shared publicly due to patient privacy and ethical restrictions. However, the results generated from this study can be made available from the corresponding author upon reasonable request.

C Training

C.1 Hyperparameters

Hyperparameter optimization was performed separately for each model. All images were first normalized using min-max normalization, which resulted in better performance compared to z-score normalization. Data augmentation included random flipping and rotation per batch, adding Gaussian noise, as well as channel shuffling and channel dropping. Hyperparameter searching was performed to identify the optimal values of the number of training epochs, learning rate (LR), BCE multiplier λ from 3, optimizer (optim), and number of epochs used for warmup before training for training epochs as listed in Table 2.

C.2 Compute Resources

All models were trained using a single Nvidia TITAN Xp GPU with 12GB VRAM for 12 hours.

C.3 Release

Code and trained model weights will be made available upon reasonable request according to the guidelines established by the organization and agencies that provided funding for this study, once all studies using the data have been published.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are backed by the results (Section 5) of the experiments in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Limitations paragraph of Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiments are provided in Section 4 and detailed information regarding data acquisition and training hyperparameters is in Appendix B and C respectively.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The imaging dataset used in this study cannot be shared publicly due to patient privacy and ethical restrictions. However, the results generated from this study can be made available from the corresponding author upon reasonable request. Code will be made available upon reasonable request according to the guidelines established by the organization and agencies that provided funding for this study once all studies using the data have been published. The detailed instructions to faithfully reproduce the experiments are provided in Section 4 and Appendix B.3 and C.3.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The training and test details are included in Section 4 and Appendix C.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All the results in Section 5 show the mean of metrics measured over the test set along with the standard deviation. Table 1 reflects this information.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are discussed in Appendix C.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics. Data collection is under an IRB, and the model promotes further research in societal good for improving RT planning outcomes. Data is not being released publicly to preserve privacy.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impact of the work has been discussed in Section 6.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The data and the trained model weights are not being released publicly due to patient privacy and ethical restrictions as discussed in Appendix B and C.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets. The dataset used is an in-house dataset covered under an IRB.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The work uses new assets under an IRB but does not intend to release those assets due to patient privacy and ethical concerns.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper only involves the use of retrospective patient scans that are covered under an IRB.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Appendix B reflects that the dataset was obtained under an IRB.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.