Translating automated brain tumour phenotyping to clinical neuroimaging

James K. Ruffle FRCR MSc¹, Samia Mohinta MSc¹, Robert Gray PhD¹, Harpreet Hyare FRCR PhD¹ and Parashkev Nachev FRCP PhD¹

Author Affiliations:

1. Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK

Running title:

Clinical translation of automated brain tumour phenotyping

Correspondence to:

Dr James K Ruffle

Email: j.ruffle@ucl.ac.uk

Address: Institute of Neurology, UCL, London WC1N 3BG, UK
Funding:

JKR was supported by the Guarantors of Brain, the NHS Topol Digital Fellowship and the UCL CDT i4health. PN is supported by the Wellcome Trust (213038/Z/18/Z) and the UCLH NIHR Biomedical Research Centre. HH is supported by the UCLH NIHR Biomedical Research Centre.

Conflict of interest:

None to declare.

Authorship: Conceptualization: JR, HH, PN; Methodology: JR, SM, RG; Software: JR, SM; Validation: JR, HH; Formal analysis: JR, SM; Manuscript writing, reviewing, and editing: JR, SM, RG, HH, PN. All authors have been involved in the writing of the manuscript and have read and approved the final version.

Manuscript Type: Original article

Total word count: 3500

Figures: 6

References: 30

Data sharing:

All BraTS 2021 challenge data is readily available from the challenge website: http://braintumorsegmentation.org. Modelling code is readily available from the nnU-Net authors here: https://github.com/MIC-DKFZ/nnUNet. All trained model weights are available upon request. Patient imaging data from our external validation site is not available for dissemination under the ethical framework that governs its use.

Abbreviations: T1CE, contrast-enhanced T1 image; TA, acquisition time.
Abstract

**Background:** The complex heterogeneity of brain tumours is increasingly recognized to demand data of magnitudes and richness only fully-inclusive, large-scale collections drawn from routine clinical care could plausibly offer. This is a task contemporary machine learning could facilitate, especially in neuroimaging, but its ability to deal with incomplete data common in real world clinical practice remains unknown. Here we apply state-of-the-art methods to large scale, multi-site MRI data to quantify the comparative fidelity of automated tumour segmentation models replicating the various levels of completeness observed in clinical reality.

**Methods:** We compare deep learning (nnU-Net-derived) tumour segmentation models with all possible combinations of T1, contrast-enhanced T1, T2, and FLAIR imaging sequences, trained and validated with five-fold cross-validation on the 2021 BraTS-RSNA glioma population of 1251 patients, and tested on a diverse, real-world 50 patient sample.

**Results:** Models trained on incomplete data segmented lesions well, often equivalently to those trained on complete data, exhibiting Dice coefficients of 0.907 (single sequence) to 0.945 (full datasets) for whole tumours, and 0.701 (single sequence) to 0.891 (full datasets) for component tissue types. Incomplete data segmentation models could accurately detect enhancing tumour in the absence of contrast imaging, quantifying its volume with an R² between 0.95-0.97.

**Conclusions:** Deep learning segmentation models characterize tumours well when missing data and can even detect enhancing tissue without the use of contrast. This suggests translation to clinical practice, where incomplete data is common, may be easier than hitherto believed, and may be of value in reducing dependence on contrast use.

**Key words:** Tumour segmentation; Medical imaging; Magnetic Resonance Imaging; Deep Learning; Artificial Intelligence.
**What is already known on this topic**

The management of brain tumours can only become personalised once their characteristics are captured in detail fine enough to be individuating: a task, in the context of brain imaging, only automated systems could conceivably accomplish. While deep learning-based models have been shown to have the necessary power, their applicability to the partial, low-quality acquisition sets usual in real-world clinical practice remains untested. Consider, for example, patients with renal failure who cannot receive contrast, artefact-spoiled sequences, or patients undergoing single-sequence intraoperative imaging. To judge the translatability of deep learning into clinical practice, we must therefore ascertain how well tumour segmentation models perform with incomplete imaging data, and what features of the lesion are detectable under such circumstances.

**What this study adds**

Brain tumour segmentation models with missing sequences identify lesions well and even identify enhancing tumour without post-contrast imaging, providing a translational opportunity in clinical situations where incomplete data is frequent.

**How this study might affect research, practice, or policy**

These findings open the door to the application of such models to large-scale, incomplete, real-world clinical data essential for illuminating complex patterns of individual tumour heterogeneity, and potentially enables the characterization of enhancing regions without the use of contrast.
Introduction

Progress in neuro-oncology is increasingly recognized to be obstructed by the marked heterogeneity—genetic, pathological, and clinical—of brain tumours. If the treatment susceptibilities and outcomes of individual patients differ widely, determined by the interactions of many multimodal characteristics, then large-scale, fully-inclusive, richly phenotyped data—including imaging—will be needed to predict them at the individual level. Such data can realistically be acquired only in the routine clinical stream, where its quality is inevitably degraded by the constraints of real-world clinical care. Although contemporary machine learning could theoretically provide a solution to this task, especially in the domain of imaging, its ability to cope with realistic, incomplete, low-quality data is yet to be determined.

Over the last few decades, lesion segmentation has formed a cornerstone of innovation across the domains of neuro-oncology, medical imaging, biomedical engineering, and machine and deep learning. The ability to segment an anatomical or pathological lesion in three-dimensions confers the ability to evaluate it quantitatively—moving beyond visual qualitative assessment—with greater richness and fidelity than conventional two-dimensional measurements repeatedly shown to be often spurious and inconsistent between radiologists, and with greater sensitivity to the heterogeneity of the underlying pathological patterns. Enabling radiological image segmentation opens a wide array of possibilities for downstream innovation in neuro-oncological healthcare and research, ranging from clinical stratification, outcome prediction, response assessment, treatment allocation and risk quantification, many of which have already shown great promise. The underlying goal is to enhance the individual fidelity of data-driven decision making, facilitating better patient-centred care, a remit especially warranted in neuro-oncology.

The segmentation of brain tumours remains a particularly challenging task owing to the marked heterogeneity of their imaging appearances: spatial distribution, morphology, signal characteristics, and impact on adjacent healthy anatomical structures. Its difficulty has even inspired an international competition for cutting-edge deep learning groups to create the best segmentation model. Known as the Brain Tumour Segmentation challenge (BraTS), it is attracting increasing attention as well as support from both the Radiological Society of North America and the American Society of Neuroradiology, providing large scale data with
multimodal MRI – FLAIR, T1, T2 and contrast-enhanced T1 (T1CE) sequences – as well as the labelled ground-truths of oedema, non-enhancing and enhancing tumor.

But while benchmark tasks have unquestionably aided the advancement of lesion segmentation – indeed of machine vision generally – they have compelled a research focus on developing uniformly multimodal models trained on complete acquisition sets, often rare in real-world clinical practice. The causes of incomplete data are legion, but common examples include patient contraindications to contrast, corruption by image artefacts, and image acquisition constraints such as those imposed in pre-operative stealth studies. This raises the questions of how well contemporary segmentation modelling architectures perform when trained on incomplete data, and what features of the lesion are correctly identifiable under such circumstances.

To that end, we aimed to systematically quantify and answer these questions with the largest and most comprehensive study of its kind based on the application of state-of-the-art deep learning tumour segmentation models to large-scale MR imaging of brain tumours. We hypothesized that incremental gains in machine model performances would be evident with additional MRI sequence data provided, but nevertheless that reasonable performances would still occur in such situations of missing data.

**Materials and methods**

**Data**

The study was approved by the University College London Hospital / National Hospital for Neurology and Neurosurgery ethics committee. We received ethical permission for the consentless analysis of irrevocably anonymized data collected in the course of routine clinical care.

We used the BraTS 2021 challenge data for all model training. This dataset is described in detail elsewhere. In brief, it includes a large retrospective sample of brain tumour MRI scans, acquired from numerous institutions, with heterogeneous equipment, protocols, and image quality. The following sequences are included: T1-weighted, T2-weighted, FLAIR and post-contrast T1 (T1CE), with a pre-processing pipeline consisting of image co-registration, sampling to a 1mm$^3$ isotropic space and skull-stripping. Lesions are segmented with an
ensemble of previous top-ranking BraTS algorithms with subsequent manual refinement and checking by a panel of neuroradiologists with more than 15 years of experience. We used the training set of 1251 individuals of the BraTS 2021 challenge data—comprising 5004 separate images—as this group included all ground-truth labels for model validation.

Having trained and evaluated a set of models on the BraTS 2021 challenge data, we sought to separately evaluate their performance on an additional held-out population from our own centre. The aim of this was to provide an additional robust safeguard of model performance with international and external validation. Specifically, we acquired retrospective imaging for a random sample of 50 individuals who underwent gadolinium-enhanced MRI head studies between 2006 and 2021 for a known glioblastoma. The random allocation of year selected was to further instil heterogeneity to our sample, as data would be acquired over one of 11 possible MRI scanners of various field strengths and manufacturers. Moreover, of our 50 participants—we also chose to include 10 of those with post-operative imaging and evidential tumour recurrence. The decision to do this was to further instil difficulty in the task, as a model would need to recognize post-operative resection/surgical bed as separate from the subsequent disease recurrence.

Most of our sample, by virtue of the acquisitions available at the year the study was undertaken, did not include volumetric imaging. Therefore, we utilized a multimodal imaging pipeline that utilizes a generative model for super-resolution of clinical MRI data. This pipeline yielded data in similar formulation to that of the BraTS challenge data with 1mm³ isotropic and skull-stripped multi-sequence data. Lesions were hand-labelled with ITK-SNAP by a neuroradiology fellow with 3 years of experience working with brain tumour imaging, with additional aids of the ITK-SNAP semi-automated segmentation tools, namely Random Forest based classifiers with subsequent manual refinement.

Tumour annotations conform to established tissue class labels comprising gadolinium-enhancing tumour, peritumoural oedema/invaded tissue and non-enhancing tumour/necrotic tumour core. The detailed description of these components is beyond the scope of this article and is discussed elsewhere. In brief, enhancing tumour refers to regions with visible enhancement on a T1CE sequence after gadolinium administration. Non-enhancing tumour/necrotic tumour core refers to the part of the tumour that does not enhance after gadolinium, typically deep to the enhancement, while oedema/invaded tissue refers to the
peritumoural oedematous and/or infiltrated brain parenchyma, typified by hyperintensity on T2/FLAIR images.

Algorithm
Our task here was not to propose a new architecture superior to those already evidenced by the BraTS 2021 challenge. Rather, we sought to characterize, evaluate, and quantify the variation in model performance with increasingly incomplete data, as a proxy index of translational potential across the variety of clinical situations where full complete datasets rarely occur. We chose the nnU-Net self-configuring deep learning biomedical image segmentation modelling architecture\textsuperscript{24}, which notably won the 2020 BraTS challenge\textsuperscript{25}. In brief, this segmentation method is able to automatically configure itself, including in pre-processing, architecture, training and post-processing across any task, and has been shown to be a superior methodology across a range of public datasets and tasks, including brain tumour segmentation\textsuperscript{24}. Our choice was guided by its excellent performance, and the simple, largely automated processing and training cycle, which made development across many models at scale feasible.

Each nnU-Net\textsuperscript{24} is in particular a self-configuring U-Net\textsuperscript{26}, incorporating the standard encoder-decoder architecture and skip connections, instance normalization and leaky rectified linear units. The nnU-Net approach employs a polynomially decaying learning rate, initially set to 0.01, with stochastic gradient descent optimization. The loss function is a weighted sum of the Sørenson-Dice coefficient and cross-entropy. Training data is augmented on the fly, including with rotations, scaling, Gaussian noise and blur, brightness and contrast shifting and gamma correction. Patch and batch size are also self-configured. Model training utilizes 1000 epochs, with foreground oversampling to mitigate the impact of class imbalances. We use 5-fold cross-validation for each experiment with the BraTS 2021 challenge data, as well as evaluated external/international out-of-sample performance of models with the additional data from our own centre.

Performance evaluation
We trained all possible combinations of the MRI sequences T1, T2, FLAIR and T1CE as separate models. This included all models using only a single sequence, two sequences, three
sequences and finally a complete four-sequence model with no missing data. We also trained separate models for abnormality detection (i.e., a binary lesion mask to detect and segment the whole tumour) as well as tumour segmentation with the tissue classes of oedema, enhancing and non-enhancing tumour. This approach comprised 30 different models in total.

Performance was principally quantified by the Sørenson-Dice coefficient between ground truth and inferred labels\cite{27, 28}, in accordance with typical research practices\cite{2, 7}. This metric derives the area of overlap between the model prediction and the labelled ground-truth. We also quantified overall model accuracy, false discovery rate, false negative rate, false omission rate, false positive rate, negative predictive value, precision, and recall. All listed metrics were derived for whole tumour and the separate tissue constituents of oedema, enhancing and non-enhancing tumour. We additionally constructed regression models between ground truth tumour volumes and model predictions, reporting the $R^2$. Lastly, we acquired the acquisition times of contemporaneous imaging protocols at our centre for a given imaging sequence, to allow comparison between a gain in model performance aligned to the time it would take to be acquired.

**Data and code availability**

All BraTS 2021 challenge data is readily available from the challenge website here: http://braintumorsegmentation.org\cite{7, 19}. Modelling code is readily available from the nnU-Net authors here: https://github.com/MIC-DKFZ/nnUNet\cite{24}. All trained model weights are available upon request. Patient imaging data from our external validation site is not available for dissemination under the ethical framework that governs its use.

**Compute**

All models were trained on an NVIDIA DGX-1 with 8 16GB Tesla P100 GPUs. With approximately 3.5 days to train a single model, the task required just over 13 days utilization of all cards.
Results

Incremental performance with sequence addition

All models performed subjectively well for segmentation of the whole tumour when reviewed visually, despite varying degrees of missing data, with objective performance ranging from a Dice coefficient of 0.907 (single sequence) to 0.945 (complete sequence set) (Figure 1). Results for segmentation of the oedema, enhancing, and non-enhancing components were more variable, with Dice coefficients ranging from 0.701 (single sequence) to 0.891 (full datasets). Of note, the models that performed the poorest typically struggled in the segmentation on the non-enhancing tumour component, particularly affecting single sequence models of T1, T2 and FLAIR, two and three-sequence models employing combinations of the former (i.e., with the omission of contrast). There was no evidence of model over-fitting. We provide the full breakdown of Dice coefficients for all models in Figure 1. Example image segmentations across the range of all models are provided in Figure 2 and 3, which visually illustrate excellent coverage of the lesion by the models, with relatively little error.

Trade-off between acquisition time and segmentation fidelity

We aligned the acquisition times of all possible combinations of sequences using contemporaneous scanner protocol data at our centre, and from which determined the gain in model fidelity in Dice per scanning minute (Figure 1). As to be expected, it illustrated an increment in acquisition time with additional imaging sequences. However, it also demonstrated that certain combinations of sequences appeared to offer greater gains in Dice score compared to others, offering an insight into the efficiency of data acquisition in this clinical context. For instance, it was noted that whilst a single volumetric T1CE acquisition (proxy for a contrast-enhanced MRI stealth study for neurosurgical planning) took 3.1 minutes, achieving a whole tumour Dice coefficient of 0.908 and reasonable performance on individual components (see Figure 1), the addition of FLAIR raised total scanning time to only 4.9 minutes while improving whole tumour Dice to 0.943, just shy of the best performing model with all four sequences (Dice coefficient 0.945). Similarly, the three-sequence acquisition of FLAIR + T1CE + T2 (i.e., neglecting the pre-contrast T1) achieved Dice coefficients for whole tumour segmentation essentially equivalent to that of the complete four sequences, and reduced
scanning time by 33%, from 9.48 to 6.38 minutes. We do, of course, note the omission of a pre-contrast T1 brings its own issues in delineating contrast from, for example, haemorrhage, but is nonetheless a striking illustration of how models with missing data still achieved comparable performance.

**Segmenting enhancing tumour without contrast-enhanced imaging**

Interestingly, we discovered that models missing contrast-enhanced imaging could still delineate tumours relatively well (Figure 4-5). Models without contrast imaging segmented whole tumour lesions with Dice coefficients ranging from 0.907 (single sequence – T1) to 0.942 (three sequences: FLAIR + T1 + T2). Of note, this latter performance was only just shy of the best performing full four sequence model with Dice of 0.945. Furthermore, models missing the contrast-enhanced T1 sequence could still identify the enhancing tumour component well, with Dice coefficients ranging from 0.756 (single sequence – T1) to 0.790 (three sequences: FLAIR + T1 + T2) (Figure 4-5). This included the model’s ability to identify and segment lesions where the focus of enhancing tumour was less than 7mm in diameter (Figure 5). The volume of enhancing tumour was highly significantly correlated to that of all model predictions, even despite contrast-enhanced imaging not being provided. The correlation between actual enhancing tumour volume to that of the model predictions with the following inputs were as follows: FLAIR alone (R² 0.964); T1 alone (R² 0.953); T2 alone (R² 0.966); FLAIR + T1 (R² 0.973); FLAIR + T2 (R² 0.976); T1 + T2 (R² 0.962); FLAIR + T1 + T2 (R² 0.972).

**International clinical validation**

**Whole tumour segmentation**

We evaluated the performance of all trained models on an out-of-sample cohort of 50 patients from our own centre in which lesions were hand-labelled. The cross-validation performances of all models from the BraTS data were well reproducible on our own data, with Dice coefficients for all models significantly correlated (r = 0.97, p < 0.0001) (Figure 6). This was despite multiple steps taken to deliberately make data more heterogenous and liable to error
(see the above methods). As expected, models with single imaging modalities, such as T1 and T2 sequences, performed worst, with incremental gains in performance with alternative and supplementary modalities.

**Tissue class segmentation**

We manually reviewed the tissue segmentations of our own data predicted by the complete four-sequence model and determined the model’s performances classifying tumour by subclasses of non-enhancing tumour, enhancing tumour and oedema were qualitatively more accurate than our semi-automated hand-segmentation. Akin to the method employed by the BraTS 2021 challenge\(^1\), we then utilized these complete model predictions as our new ground-truth with subsequent manual checking and refinement where required. We then compared the performance of all other models, i.e., those without four sequences, to this revised ground-truth. Model performances were again highly reproducible between the BraTS 2021 challenge data and that of our own external sample, with Dice coefficients significantly correlated (r = 0.95, \( p < 0.0001 \)) (Figure 6). As is usually the case in brain tumour segmentation models, segmentations for the non-enhancing tumour component fared worst – especially those with single imaging modalities, whilst prediction of enhancing tumour or oedema fared much better. Supplementary Tables 1-4 illustrates an extensive performance breakdown of all models, for all lesion formulations, including accuracy, Dice, false discovery rate, false negative rate, false omission rate, false positive rate, negative predictive value, precision, and recall.

**Discussion**

We have systematically surveyed the ability of state-of-the-art tumour segmentation models in delineating and quantifying brain tumour components in real-world clinical situations of incomplete and/or low-quality data. We reveal there is surprisingly little variation in the performance of segmenting a whole tumour with the number of modelled imaging modalities. Greater variation is observed when segmenting tumour components: a clear pattern of incremental improvement with the addition of further sequences emerges. These findings open the door both to the application of segmentation models to large-scale historical data, for the purpose of building treatment and outcome predictive models, and their deployment to real-world clinical care.
Strikingly, we find that segmentation models trained with contrast-enhancing imaging missing still characterize the anatomy of enhancing tumour components remarkably well. This includes quantification of the volumetric burden of enhancing tumour with an accuracy approaching a perfect result. Out-of-sample validation illustrates strong generalizability of these findings, including across super-resolved non-isotropic acquisitions and tumour recurrences on complex postoperative imaging of limited quality. Our analyses show that current segmentation models generalize surprisingly well to real-world clinical imaging varying in quality and sequence completeness.

**Additive value of multiple sequences**

Model fidelity unsurprisingly rose with the number of modelled sequences. What is, however, surprising is the ability of models based on limited data to delineate lesions very well. This is particularly striking in the segmentation of brain tumours as a whole (the ‘Whole Tumour’ label in our models), where only marginal differences in Dice coefficient were seen across the range of sequence combinations. We can conclude that even single sequences may be sufficient for segmenting brain tumours with fidelity adequate for many downstream tasks.

The segmentation of tumour compartments—oedema, enhancing and non-enhancing tumour—however presents a more complex picture. Single sequence models of oedema and enhancing tumour perform best with FLAIR and T1CE sequences, respectively. But models of two or three sequences exhibit less intuitive behaviour. Adding FLAIR to T1CE achieves whole tumour performance very close to that of the full, four sequence model, despite receiving only half the data. To that end, single T1CE MRI studies (such as in stealth imaging) may therefore benefit from the addition of a FLAIR sequence to enable more optimum visualization of the entire lesion to aid pre-operative planning. A two-sequence model of T1 & T1CE can delineate oedema well without the T2 or FLAIR typically used to identify it. Overall, these findings illustrate the ability of contemporary machine vision models to extract information from multiple sequences with greater efficiency than intuitive perception may suggest.29, 30
Segmenting enhancing tumour without contrast-enhanced imaging

Strikingly, we found that models missing the contrast enhancing sequence (T1CE) can still segment what has been hand-labelled by experienced neuroradiologists with full imaging datasets as the enhancing component of the tumour well, invariantly to the size of the enhancing component. This introduces the possibility—across both research and clinical practice—to make approximate inferences about the anatomy of enhancing components without the use of contrast. Moreover, that a model can identify what has been termed the ‘enhancing’ tumor\textsuperscript{18, 30}, without any information about its enhancing properties, reveals the presence of non-intuitive imaging features that could render the enhancing component quantifiable without the use of contrast. This challenges the current dogma of ‘enhancing tumour’, given a machine can identify it without the administration of intravenous agents ordinarily required to reveal it. Further investigation of this possibility is warranted, including the detectability of the presence of any degree of enhancement. These findings also illustrate a clinically important opportunity in oncological imaging when contrast enhanced-imaging cannot be acquired, not least in situations of repeated follow-up where the over-use of contrast should ideally be limited, for example to minimize Gadolinium retention in paediatric patients.

Study limitations

In our systematic evaluation of the ability of deep learning models to identify brain tumours with varying degrees of missing data, we opted to use one single self-configuring architecture – nnU-Net\textsuperscript{24, 25}. The use of this software is well justified given its validated performance across many domains of medical imaging. But segmentation models are a rapidly evolving field, and so it is possible that other architectures might perform differently, perhaps even superiorly, to that used here. It is however important to note that our aim was not to identify the definitive ‘best’ tumour segmentation model, instead to quantify the similarity and discrepancies between when varying degrees of data missingness were provided to it. Rather, it was our aim to derive how such models could perform in real-world clinical situations where ‘perfect’ data rarely exists, quantifying its appropriateness for translation to the clinical frontline.
Conclusion

Automated segmentation models can characterize tumours in real-world clinical situations of incomplete imaging data remarkably well. Such models are also able to identify enhancing tumour *without* the use of contrast-enhanced imaging, potentially providing clinical guidance in circumstances where contrast administration is contra-indicated or where its repeated use should be minimized. This opens the way to quantifying enhancing components without the administration of intravenous agents, not least invites a revision of the notion of tumour enhancement if the same information can be extracted without contrast. Its applicability includes not just prospective scenarios wherein a full scan may not be possible such as patients unable to receive intravenous contrast, but also applies to historical datasets where certain sequences might not have been acquired. Out-of-sample validation illustrates strong generalizability, across non-isotropic acquisitions and even on complex postoperative imaging where tumours have recurred. Translation of such models to the clinical frontline for response assessment - where complete data is a rarity - may be easier than hitherto believed.
References

1. Peng J, Kim DD, Patel JB, et al. Corrigendum to: Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-Oncology* 2021;23(12):2124-24. doi: 10.1093/neuonc/noab226

2. Xue J, Wang B, Ming Y, et al. Deep learning–based detection and segmentation-assisted management of brain metastases. *Neuro-Oncology* 2020;22(4):505-14. doi: 10.1093/neuonc/noz234

3. Lu S-L, Xiao F-R, Cheng JC-H, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. *Neuro-Oncology* 2021;23(9):1560-68. doi: 10.1093/neuonc/noab071

4. Lenchik L, Heacock L, Weaver AA, et al. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. *Acad Radiol* 2019;26(12):1695-706. doi: 10.1016/j.acra.2019.07.006 [published Online First: 20190810]

5. Suetens P, Bellon E, Vandermeulen D, et al. Image segmentation: methods and applications in diagnostic radiology and nuclear medicine. *Eur J Radiol* 1993;17(1):14-21. doi: 10.1016/0720-048x(93)90023-g

6. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26(3):839-51. doi: 10.1016/j.neuroimage.2005.02.018 [published Online First: 2005/06/16]

7. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* 2015;34(10):1993-2024. doi: 10.1109/TMI.2014.2377694 [published Online First: 2014/12/11]
8. Zhao B, James LP, Moskowitz CS, et al. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* 2009;252(1):263-72. doi: 10.1148/radiol.2522081593

9. McNitt-Gray MF, Kim GH, Zhao B, et al. Determining the Variability of Lesion Size Measurements from CT Patient Data Sets Acquired under "No Change" Conditions. *Transl Oncol* 2015;8(1):55-64. doi: 10.1016/j.tranon.2015.01.001

10. Dempsey MF, Condon BR, Hadley DM. Measurement of tumor "size" in recurrent malignant glioma: 1D, 2D, or 3D? *AJNR Am J Neuroradiol* 2005;26(4):770-6.

11. Mandal AS, Romero-Garcia R, Hart MG, et al. Genetic, cellular, and connectomic characterization of the brain regions commonly plagued by glioma. *Brain* 2020;143(11):3294-307. doi: 10.1093/brain/awaa277 [published Online First: 2020/12/06]

12. Topol E. The Topol Review: Preparing the healthcare workforce to deliver the digital future. In: NHS, ed., 2019.

13. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nature Medicine* 2022 doi: 10.1038/s41591-021-01614-0

14. Ruffle JK, Farmer AD, Aziz Q. Artificial Intelligence Assisted Gastroenterology - Promises and Pitfalls. *Am J Gastroenterol* 2019;114(3):422-28. doi: 10.1038/s41395-018-0268-4.

15. Chow D, Chang P, Weinberg BD, et al. Imaging Genetic Heterogeneity in Glioblastoma and Other Glial Tumors: Review of Current Methods and Future Directions. *American Journal of Roentgenology* 2017;210(1):30-38. doi: 10.2214/AJR.17.18754

16. Molina D, Perez-Beteta J, Luque B, et al. Tumour heterogeneity in glioblastoma assessed by MRI texture analysis: a potential marker of survival. *Br J Radiol* 2016;89(1064):20160242. doi: 10.1259/bjr.20160242 [published Online First: 20160620]
17. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol 2021;23(8):1231-51. doi: 10.1093/neuonc/noab106

18. TCGA. VASARI Research Project 2020 [Available from: https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project accessed 25/01/2022 2022.

19. Baid U, Ghodasara S, Bilello M, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification2021.

20. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4:170117. doi: 10.1038/sdata.2017.117 [published Online First: 20170905]

21. Bakas S, Reyes M, Jakab As, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. ArXiv 2018;abs/1811.02629

22. Nixon M, Mahmoodi S, Zwiggelaar R, eds. MRI Super-Resolution Using Multi-channel Total Variation. Medical Image Understanding and Analysis; 2018 2018//; Cham. Springer International Publishing.

23. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 16-20 Aug. 2016.

24. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203-11. doi: 10.1038/s41592-020-01008-z [published Online First: 2020/12/09]

25. nnU-Net for Brain Tumor Segmentation. BrainLes@MICCAI; 2020.
26. Navab N, Hornegger J, Wells WM, et al., eds. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; 2015 2015//; Cham. Springer International Publishing.

27. A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons; 1948.

28. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology* 1945;26(3):297-302. doi: 10.2307/1932409

29. Bink A, Benner J, Reinhardt J, et al. Structured Reporting in Neuroradiology: Intracranial Tumors. *Front Neurol* 2018;9:32. doi: 10.3389/fneur.2018.00032 [published Online First: 20180206]

30. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 2010;28(11):1963-72. doi: 10.1200/JCO.2009.26.3541 [published Online First: 20100315]
Tables & Figures

Figure 1: Performance of all model combinations. A) Heatmap illustrates the validation Dice coefficient across all models, for both whole tumour and the individual components. Models are partitioned into those which utilized just one sequence, two, three and finally the complete four-sequence model. A brighter orange/white box depicts a better performing model as per the Dice coefficient. B) Second heatmap depicts the relative acquisition time (in minutes) for the sequences used for a given model, with a more green/yellow box illustrating a longer acquisition time. C) Third heatmap illustrates the performance gain in Dice coefficient per minute of acquisition time. Colour keys are given at the right of the plot. Abbreviations: T1CE, contrast-enhanced T1 image; TA, acquisition time.
Figure 2: Example segmentation results. A) Left upper panel illustrates stacked axial slices of a given lesion for all imaging sequences, with B) corresponding radiologist-labelled ground-truth in the left lower panel. C) Right panel illustrates the tumour segmentation predictions across all model formulations, aligned to the number of sequences supplied. Abbreviations: T1CE, contrast-enhanced T1 image.
Figure 3: Minimal error in example segmentation results. A) Left upper panel illustrates stacked axial slices of a given lesion for all imaging sequences, with B) corresponding radiologist-labelled ground-truth in the left lower panel. C) Right panel illustrates the tissue-specific error in tumour segmentation predictions across all model formulations, aligned to the number of sequences supplied. Abbreviations: T1CE, contrast-enhanced T1 image.
Figure 4: Segmenting enhancing tumour without contrast. A) Top panel illustrates axial slices of the lesion across the four sequences. B) Second panel illustrates the radiologist hand-labelled ground truth for the three tissue classes – of note red depicts enhancing tumour. C) Third panel illustrates predictions of enhancing tumour segmentation for four models with the following input data: i) FLAIR alone; ii) FLAIR and T1; iii) FLAIR, T1 and T2; and iv) FLAIR, T1, T2 and T1CE. Of note, only the final model is exposed to contrast-enhanced imaging, although the other three models still reasonably identify the location of the enhancing component. D) Fourth panel illustrates the component of enhancing tumour that is missed by the model. Abbreviations: T1CE, contrast-enhanced T1 image.
**Figure 5:** Further examples of segmenting enhancing tumour without contrast. A-C) Left two columns and rows of each panel illustrate the anatomical imaging for three randomly selected cases, whilst the third column of each panel illustrates the hand-labelled ground truth shown with the overlayed T1CE image, and finally the model prediction where contrast imaging was not provided. Of note, the case in panel B comprised a tumour with only a 7mm diameter enhancing component. D) The volume of enhancing tumour is highly significantly correlated to that of all model predictions, even when contrast-enhanced imaging is not provided. Abbreviations: T1CE, contrast-enhanced T1 image.
Figure 6: International validation. A) Scatterplot illustrates the strong relationship between radiologist-labelled lesions from a disparate international centre. Only relationship between whole tumour hand-segmentations and model predictions are shown here, as it transpired that the complete four sequence model more accurately delineated tissue classes than when hand-labelled. B) Scatterplot illustrates the strong relationship between model performances from the validation set, and when re-evaluated on our own data. For this plot, the complete four sequence model was utilized as the ground truth for the tissue subclasses of the international validation data. Abbreviations: T1CE, contrast-enhanced T1 image.
### Supplementary Table 1: Detailed performance metrics of all sequence combinations across the models in segmenting whole tumour

| Sequences          | Whole Tumour: Accuracy | Whole Tumour: Dice | Whole Tumour: False Discovery Rate | Whole Tumour: False Negative Rate | Whole Tumour: False Omission Rate | Whole Tumour: False Positive Rate | Whole Tumour: Negative Predictive Value | Whole Tumour: Precision | Whole Tumour: Recall |
|--------------------|------------------------|-------------------|-----------------------------------|----------------------------------|----------------------------------|-----------------------------------|----------------------------------------|-----------------------|-------------------|
| FLAIR              | 0.998                  | 0.93              | 0.052                             | 0.083                            | 0.001                            | 0.001                            | 0.999                   | 0.948                 | 0.917            |
| T1                 | 0.994                  | 0.803             | 0.181                             | 0.19                             | 0.003                            | 0.003                            | 0.997                   | 0.821                 | 0.81             |
| T2                 | 0.996                  | 0.89              | 0.105                             | 0.104                            | 0.002                            | 0.002                            | 0.998                   | 0.895                 | 0.896            |
| T1CE               | 0.994                  | 0.802             | 0.179                             | 0.199                            | 0.003                            | 0.003                            | 0.997                   | 0.821                 | 0.801            |
| FLAIR + T1         | 0.998                  | 0.942             | 0.055                             | 0.057                            | 0.001                            | 0.001                            | 0.999                   | 0.945                 | 0.943            |
| FLAIR + T2         | 0.998                  | 0.955             | 0.034                             | 0.053                            | 0.001                            | 0.001                            | 0.999                   | 0.966                 | 0.947            |
| FLAIR + T1CE       | 0.998                  | 0.94              | 0.044                             | 0.067                            | 0.001                            | 0.001                            | 0.999                   | 0.956                 | 0.933            |
| T1 + T2            | 0.997                  | 0.906             | 0.086                             | 0.093                            | 0.002                            | 0.001                            | 0.998                   | 0.914                 | 0.907            |
| T1 + T1CE          | 0.995                  | 0.839             | 0.142                             | 0.166                            | 0.003                            | 0.002                            | 0.997                   | 0.858                 | 0.834            |
| T1CE + T2          | 0.997                  | 0.903             | 0.097                             | 0.086                            | 0.002                            | 0.002                            | 0.998                   | 0.903                 | 0.914            |
| FLAIR + T1 + T2    | 0.999                  | 0.961             | 0.041                             | 0.036                            | 0.001                            | 0.001                            | 0.999                   | 0.959                 | 0.964            |
| FLAIR + T1 + T1CE  | 0.998                  | 0.951             | 0.053                             | 0.04                             | 0.001                            | 0.001                            | 0.999                   | 0.947                 | 0.96             |
| FLAIR + T1CE + T2  | 0.999                  | 0.964             | 0.046                             | 0.024                            | 0                             | 0.001                            | 1                       | 0.954                 | 0.976            |
| T1 + T1CE + T2     | 0.997                  | 0.908             | 0.091                             | 0.084                            | 0.001                            | 0.001                            | 0.999                   | 0.909                 | 0.916            |
## Supplementary Table 2: Detailed performance metrics of all sequence combinations across the models in segmenting oedema

| Sequences       | Oedema: Accuracy | Oedema: Dice | Oedema: False Discovery Rate | Oedema: False Negative Rate | Oedema: False Omission Rate | Oedema: False Positive Rate | Oedema: Negative Predictive Value | Oedema: Precision | Oedema: Recall |
|-----------------|------------------|--------------|-----------------------------|---------------------------|---------------------------|----------------------------|-------------------------------|------------------|----------------|
| FLAIR           | 0.995            | 0.807        | 0.196                       | 0.163                     | 0.002                     | 0.003                     | 0.998                        | 0.804            | 0.837          |
| T1              | 0.99             | 0.62         | 0.332                       | 0.356                     | 0.004                     | 0.006                     | 0.996                        | 0.668            | 0.644          |
| T2              | 0.994            | 0.763        | 0.237                       | 0.21                      | 0.003                     | 0.003                     | 0.997                        | 0.763            | 0.79           |
| T1CE            | 0.994            | 0.695        | 0.266                       | 0.304                     | 0.003                     | 0.003                     | 0.997                        | 0.734            | 0.696          |
| FLAIR + T1      | 0.996            | 0.836        | 0.146                       | 0.162                     | 0.002                     | 0.002                     | 0.998                        | 0.854            | 0.838          |
| FLAIR + T2      | 0.996            | 0.846        | 0.13                        | 0.157                     | 0.002                     | 0.002                     | 0.998                        | 0.87             | 0.843          |
| FLAIR + T1CE    | 0.998            | 0.923        | 0.073                       | 0.077                     | 0.001                     | 0.001                     | 0.999                        | 0.927            | 0.923          |
| T1 + T2         | 0.995            | 0.794        | 0.198                       | 0.194                     | 0.002                     | 0.003                     | 0.998                        | 0.802            | 0.806          |
| T1 + T1CE       | 0.994            | 0.751        | 0.254                       | 0.224                     | 0.003                     | 0.003                     | 0.997                        | 0.746            | 0.776          |
| T1CE + T2       | 0.997            | 0.852        | 0.13                        | 0.153                     | 0.002                     | 0.001                     | 0.998                        | 0.87             | 0.847          |
| FLAIR + T1 + T2 | 0.997            | 0.866        | 0.117                       | 0.139                     | 0.002                     | 0.002                     | 0.998                        | 0.883            | 0.861          |
| FLAIR + T1CE +  | 0.998            | 0.921        | 0.054                       | 0.098                     | 0.001                     | 0.001                     | 0.999                        | 0.946            | 0.902          |
| T1CE + T2       | 0.999            | 0.945        | 0.055                       | 0.053                     | 0.001                     | 0.001                     | 0.999                        | 0.945            | 0.947          |
| FLAIR + T1CE +  | 0.997            | 0.866        | 0.123                       | 0.138                     | 0.002                     | 0.001                     | 0.998                        | 0.877            | 0.862          |
| T1 + T1CE + T2  | 0.997            | 0.866        | 0.123                       | 0.138                     | 0.002                     | 0.001                     | 0.998                        | 0.877            | 0.862          |
Supplementary Table 3: Detailed performance metrics of all sequence combinations across the models in segmenting enhancing tumour

| Sequences       | Enhancing Tumour: Accuracy | Enhancing Tumour: Dice | Enhancing Tumour: False Discovery Rate | Enhancing Tumour: False Negative Rate | Enhancing Tumour: False Omission Rate | Enhancing Tumour: False Positive Rate | Enhancing Tumour: Negative Predictive Value | Enhancing Tumour: Precision | Enhancing Tumour: Recall |
|-----------------|---------------------------|------------------------|----------------------------------------|---------------------------------------|---------------------------------------|----------------------------------------|---------------------------------------------|---------------------------|-------------------------|
| FLAIR           | 0.996                     | 0.421                  | 0.441                                  | 0.56                                  | 0.002                                 | 0.002                                  | 0.998                                       | 0.559                     | 0.44                    |
| T1              | 0.996                     | 0.414                  | 0.484                                  | 0.52                                  | 0.002                                 | 0.002                                  | 0.998                                       | 0.516                     | 0.48                    |
| T2              | 0.997                     | 0.478                  | 0.426                                  | 0.495                                 | 0.002                                 | 0.001                                  | 0.998                                       | 0.574                     | 0.505                   |
| T1CE            | 0.999                     | 0.829                  | 0.139                                  | 0.167                                 | 0                                     | 0                                     | 1                                           | 0.861                     | 0.833                   |
| FLAIR + T1      | 0.996                     | 0.462                  | 0.421                                  | 0.478                                 | 0.002                                 | 0.002                                  | 0.998                                       | 0.579                     | 0.522                   |
| FLAIR + T2      | 0.997                     | 0.522                  | 0.41                                  | 0.453                                 | 0.002                                 | 0.001                                  | 0.998                                       | 0.59                      | 0.547                   |
| FLAIR + T1CE    | 0.999                     | 0.857                  | 0.093                                  | 0.138                                 | 0                                     | 0                                     | 1                                           | 0.907                     | 0.862                   |
| T1 + T2         | 0.997                     | 0.49                   | 0.413                                  | 0.479                                 | 0.002                                 | 0.001                                  | 0.998                                       | 0.587                     | 0.521                   |
| T1 + T1CE       | 0.999                     | 0.856                  | 0.08                                  | 0.156                                 | 0                                     | 0                                     | 1                                           | 0.92                      | 0.844                   |
| T1CE + T2       | 0.999                     | 0.848                  | 0.111                                  | 0.154                                 | 0                                     | 0                                     | 1                                           | 0.889                     | 0.846                   |
| FLAIR + T1 + T2 | 0.997                     | 0.521                  | 0.409                                  | 0.443                                 | 0.002                                 | 0.001                                  | 0.998                                       | 0.591                     | 0.557                   |
| FLAIR + T1CE +  |                           |                        |                                        |                                       |                                       |                                       |                                              |                           |                         |
| T1CE            | 1                         | 0.887                  | 0.058                                  | 0.138                                 | 0                                     | 0                                     | 1                                           | 0.942                     | 0.862                   |
| FLAIR + T1CE +  |                           |                        |                                        |                                       |                                       |                                       |                                              |                           |                         |
| T2              | 1                         | 0.894                  | 0.083                                  | 0.099                                 | 0                                     | 0                                     | 1                                           | 0.917                     | 0.901                   |
| T1 + T1CE + T2  | 0.999                     | 0.856                  | 0.103                                  | 0.149                                 | 0                                     | 0                                     | 1                                           | 0.897                     | 0.851                   |
| Sequences | Non-enhancing Tumour: Accuracy | Non-enhancing Tumour: Dice | Non-enhancing Tumour: False Discovery Rate | Non-enhancing Tumour: False Negative Rate | Non-enhancing Tumour: False Omission Rate | Non-enhancing Tumour: False Positive Rate | Non-enhancing Tumour: Negative Predictive Value | Non-enhancing Tumour: Precision | Non-enhancing Tumour: Recall |
|-----------|-------------------------------|--------------------------|------------------------------------------|------------------------------------------|------------------------------------------|------------------------------------------|-------------------------------------------|----------------------------|--------------------------|
| FLAIR     | 0.997                         | 0.358                    | 0.528                                    | 0.59                                     | 0.002                                    | 0.001                                    | 0.998                                     | 0.472                     | 0.41                     |
| T1        | 0.998                         | 0.325                    | 0.434                                    | 0.698                                    | 0.002                                    | 0.001                                    | 0.998                                     | 0.566                     | 0.302                    |
| T2        | 0.998                         | 0.378                    | 0.523                                    | 0.572                                    | 0.001                                    | 0.001                                    | 0.999                                     | 0.477                     | 0.428                    |
| T1CE      | 0.999                         | 0.711                    | 0.159                                    | 0.315                                    | 0.001                                    | 0                                    | 0.999                                     | 0.841                     | 0.685                    |
| FLAIR + T1| 0.998                         | 0.382                    | 0.421                                    | 0.601                                    | 0.001                                    | 0.001                                    | 0.999                                     | 0.579                     | 0.399                    |
| FLAIR + T2| 0.998                         | 0.432                    | 0.5                                      | 0.484                                    | 0.001                                    | 0.001                                    | 0.999                                     | 0.5                       | 0.516                    |
| FLAIR + T1CE| 0.999                         | 0.766                    | 0.094                                    | 0.274                                    | 0                                        | 0                                       | 1.0                                      | 0.906                     | 0.726                    |
| T1 + T2   | 0.998                         | 0.459                    | 0.469                                    | 0.494                                    | 0.001                                    | 0.001                                    | 0.999                                     | 0.531                     | 0.506                    |
| T1 + T1CE | 0.999                         | 0.734                    | 0.144                                    | 0.272                                    | 0                                        | 0                                       | 0.999                                     | 0.856                     | 0.728                    |
| T1CE + T2 | 0.999                         | 0.743                    | 0.183                                    | 0.234                                    | 0                                        | 0.001                                    | 1.0                                      | 0.817                     | 0.766                    |
| FLAIR + T1 + T2 | 0.998                     | 0.454                    | 0.479                                    | 0.489                                    | 0.001                                    | 0.001                                    | 0.999                                     | 0.521                     | 0.511                    |
| FLAIR + T1CE + T1CE | 0.999                   | 0.773                    | 0.084                                    | 0.269                                    | 0                                        | 0                                       | 1.0                                      | 0.916                     | 0.731                    |
| FLAIR + T1CE + T2 | 0.999                   | 0.787                    | 0.117                                    | 0.219                                    | 0                                        | 0                                       | 1.0                                      | 0.883                     | 0.781                    |
| T1 + T1CE + T2 | 0.999                   | 0.79                     | 0.159                                    | 0.191                                    | 0                                        | 0                                       | 1.0                                      | 0.841                     | 0.809                    |