Physics and Imaging in Radiation Oncology 31 (2024) 100620

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

The effect of editing clinical contours on deep-learning segmentation accuracy of the gross tumor volume in glioblastoma

Kim M. Hochreuter [a,b], Jintao Ren [a,b,c], Jasper Nijkamp [a,b], Stine S. Korreman [a,b,c], Slávka Lukacova [b,c], Jesper F. Kallehauge [a,b,*,1], Anouk K. Trip [a,1]

[a] Danish Centre for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[b] Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[c] Department of Oncology, Aarhus University Hospital, Aarhus, Denmark

ARTICLE INFO

Keywords:
Radiotherapy
GTV
Deep learning
Editing labels
Glioblastoma

ABSTRACT

Background and purpose: Deep-learning (DL) models for segmentation of the gross tumor volume (GTV) in radiotherapy are generally based on clinical delineations which suffer from inter-observer variability. The aim of this study was to compare performance of a DL-model based on clinical glioblastoma GTVs to a model based on a single-observer edited version of the same GTVs.

Materials and methods: The dataset included imaging data (Computed Tomography (CT), T1, contrast-T1 (T1C), and fluid-attenuated-inversion-recovery (FLAIR)) of 259 glioblastoma patients treated with post-operative radiotherapy between 2012 and 2019 at a single institute. The clinical GTVs were edited using all imaging data. The dataset was split into 207 cases for training/validation and 52 for testing.

GTV segmentation models (nnUNet) were trained on clinical and edited GTVs separately and compared using Surface Dice with 1 mm tolerance ($sDSC_{1mm}$). We also evaluated model performance with respect to extent of resection (EOR), and different imaging combinations (T1C/T1/FLAIR/CT, T1C/FLAIR/CT, T1C/FLAIR, T1C/CT, T1C/T1, T1C). A Wilcoxon test was used for significance testing.

Results: The median (range) $sDSC_{1mm}$ of the clinical-GTV-model and edited-GTV-model both evaluated with the edited contours, was 0.76 (0.43–0.94) vs. 0.92 (0.60–0.98) respectively ($p < 0.001$). $sDSC_{1mm}$ was not significantly different between patients with a biopsy, partial, and complete resection. T1C as single input performed as good as use of imaging combinations.

Conclusions: High segmentation accuracy was obtained by the DL-models. Editing of the clinical GTVs significantly increased DL performance with a relevant effect size. DL performance was robust for EOR and highly accurate using only T1C.

1. Introduction

Glioblastoma (GBM) is the most common primary brain cancer in adults [1]. The current standard-of-care consists of a maximal safe resection, followed by chemoradiotherapy and adjuvant chemotherapy [2]. The ESTRO-EANO guideline for radiotherapy (RT), defines the gross tumor volume (GTV) as follows "GTV is defined as T1 contrast-enhancing tumor (for biopsy only patients) and/or resection cavity plus residual contrast enhancing tumor, if present." Rarely, a T2 or T2-weighted fluid-attenuated-inversion-recovery (FLAIR)-based definition is applied [3,4]. The clinical target volume (CTV) is defined as the GTV plus a 2 cm isotropic margin along the white matter tracts and reduced at anatomical barriers. In the guideline of 2023, the CTV margin was reduced to 1.5 cm. A planning target volume margin of 2–3 mm is typically added and the prescribed dose is typically 60 Gy in 30 fractions.

Delineation of the GTV can be a time-consuming task and is associated with inter-observer variation (IOV) [5–7], which affects the treatment accuracy. One potential approach to achieve consistent delineations while saving time is to use deep-learning (DL) segmentation models. These models have been extensively explored in many contexts within RT [8–14]. However, for GTV segmentation in the GBM setting

* Corresponding author at: Danish Centre for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark.
E-mail address: jespkall@rm.dk (J.F. Kallehauge).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.phro.2024.100620
Received 30 April 2024; Received in revised form 29 July 2024; Accepted 1 August 2024
Available online 5 August 2024

there is only one report [8]. In that report by Ramesh et al., five different U-Net approaches were compared using contrast-T1 (T1C) and FLAIR as image modalities. Their best-performing model achieved a mean Dice Similarity Coefficient (DSC) of 0.73 and a maximum Hausdorff Distance (HD) of 10.75 mm.

To improve DL-segmentation models, many papers focus on optimizing model parameters and architecture [8,12,15–18]. However, in the setting of supervised learning, a determining factor for segmentation accuracy lies in the quantity and quality of the imaging data, and the provided ground truth delineations. Delineations and scans are often sourced from historical patient data due to its accessibility, but can contain noisy labels impacting the performance of DL-segmentation models [19–21]. Noisy labels can arise due to IOV [5–7] and labels that do not precisely fit the corresponding imaging data. In the setting of GBM in which large CTV margins are advised, the position of the GTV contour does not necessarily impact the position of the CTV contour. Curating datasets and editing contours has been shown to improve DL-segmentation performance of organs at risk in brain and head & neck [13,22]. While curation is a well-known strategy to reduce bias and increase accuracy, there are no established guidelines on best practices, and the extent of editing practices in current literature is unclear. Additionally, a head-to-head comparison of the impact of editing ground truth delineations has not been done yet.

GBM patients first undergo a maximal safe resection resulting in a great range of surgical extent across patients, i.e. from biopsy to complete resection [23]. This in turn results in a varying appearance of the GTV for *post*-operative RT [3,4], of which the possible impact on DL-model performance is currently unknown.

Besides data curation, multi-modal imaging data can also be leveraged to improve segmentation accuracy. In the BraTS challenge [24], which encompasses GBM tumor segmentation in the *pre*-operative setting, T1, T1C, T2 and T2-FLAIR data are combined to improve segmentation accuracy. However, in the *post*-operative setting there is a limited knowledge concerning the influence of multi-modal imaging on DL-model performance [25].

The primary aim of this study was to investigate how editing clinical delineations affects DL segmentation performance of the GTV in post-operative RT for GBM. This was done by comparing a DL-model trained on clinical GTVs to a DL-model trained on single-observer edited GTVs. The secondary aims were to evaluate the robustness of the DL-model performance to varying extent of resection (EOR) and to assess the need for different multi-modal imaging combinations of Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) sequences.

2. Materials and methods

2.1. Study design and approvals

This single-center study used retrospective data for DL-segmentation of the GTV in GBM patients. It was exempt from review by the Central Denmark Region Committees on Health Research Ethics, approved by the Danish Patient Safety Authority (Reg.no. 31-1521-174) and the Danish Neuro-Oncology Registry (Reg.no. DNOR-2020-03-02), and registered with the Central Denmark Region (Reg.no. 1-16-02-74-20).

2.2. Patient data

Using the Danish Neuro-Oncology Registry, we identified 406 eligible patients treated with post-operative RT at the Department of Oncology, Aarhus University Hospital, between 2012 and 2019. The registry was furthermore used to extract all patient-related variables (including EOR, e.g. complete resection (CR), partial resection (PR) and biopsy). In clinical routine for this patient group, at least the T1C from the planning-MRI was rigidly registered to the planning-CT, with all other imaging available in side-by-side viewing. Each GTV used for clinical treatment has been reviewed before start of RT by a clinical oncologist together with a neuro-radiologist. Over the studied period at least five different clinical oncologists have delineated clinical GTVs.

Of the 406 eligible patients, 114 were excluded from this study due to incomplete data (no dedicated planning MRI: n = 107, no complete clinical, RT, and/or imaging data (transfer): n = 7). For this study the T1, T1C, and FLAIR images of the remaining 292 patients, were rigidly registered to the planning-CT using mutual information as cost function and a region of interest conforming to the skull. The clinically used GTV was subsequently identified and reviewed by a single radiation oncologist (A.K.T. 12 years of experience in radiation oncology, whereof 4.5 years with CNS focus and 3 years as board certified radiation oncologist). With review, another 33 patients were excluded (GTV not delineated: n = 7, GTV based on FLAIR abnormalities: n = 26, Supplementary Fig. 1A and B), resulting in a final dataset of 259 patients. Of those, 95 patients had undergone a CR, 85 a PR, and 80 a biopsy. For more details see Supplementary material – Patient Population.

2.3. Edited GTVs

The clinically used GTVs were edited by a single radiation oncologist (A.K.T.), who had not been involved in delineating the clinically used GTVs. The editing goal was to produce a contour that optimally conformed to the GTV definition (strive to fit the contour with the information in the imaging) and that could be used clinically (Fig. 1A–D). Editing was done using all the co-registered planning-MRI sequences and the planning CT. If in doubt, the pre- and post-operative MRI scans were consulted in a separate viewing window.

2.4. Deep learning

Each image was resampled to a uniform spacing of 0.5 × 0.5 × 1 mm, and skull stripped using a mask of the clinical brain delineation combined with the GTV in a binary union. To account for minor delineation

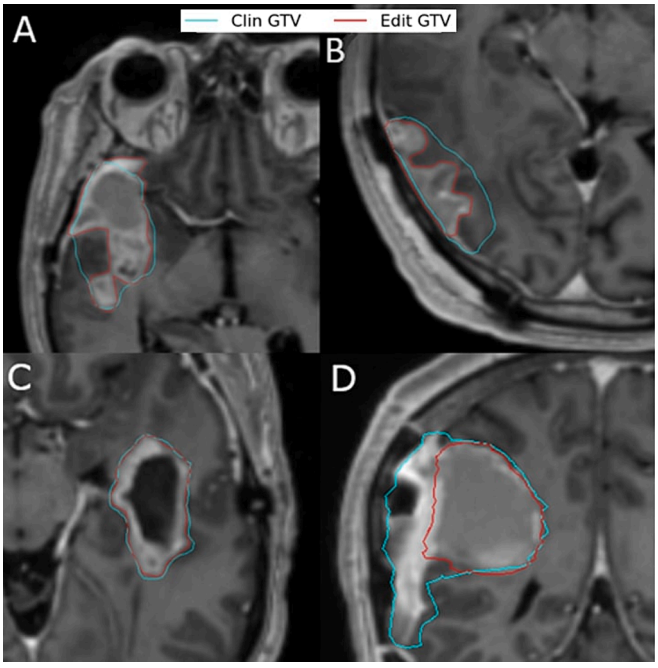

Fig. 1. Four examples of clinical and edited GTVs. Clinical GTVs are depicted in red, edited GTVs are depicted in turquoise. A: Edited GTV follows surgical cavity and residual contrast enhancement. B: Edited GTV follows contrast enhancement. C: Minimal tightening of the clinical GTV. D: Edited GTV excludes postoperative hematoma. GTV: Gross Tumor Volume, Clin GTV: Clinical GTV, Edit GTV: Edited GTV. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

errors, the mask was dilated by 2 mm in all directions. The data was randomly split into a train and test set (80:20), with stratification for extent of resection (biopsy/PR/CR) (Supplementary Table 2). The train and test sets were identical for all explored models. The training and test procedures were also similar for all models.

We made use of the no-new-Unet (nnUNet) framework [26], where network structure and hyper-parameters are automatically configured (see Supplementary material – Deep Learning). We trained one model using the clinical GTVs (clinical-GTV-model) and one model using the edited GTVs (edited-GTV-model). We repeated this process six times using different combinations of image modalities: T1C/T1/FLAIR/CT, T1C/FLAIR/CT, T1C/FLAIR, T1C/CT, T1C/T1 and T1C-only.

2.5. Statistical analysis

To evaluate model segmentation performance, we used Surface DSC with a 1 mm tolerance ($sDSC_{1mm}$), DSC, Mean Surface Distance (MSD) and 95th percentile HD (HD95), calculated with Medpy and Surface Distance Python libraries. Significance was set at 0.01. All statistical analyses were conducted using R.

We first evaluated model performance in a head-to-head comparison, between the clinical- and edited-GTV-model using the Wilcoxon signed rank test. We compared the predictions made by the clinical-GTV-model evaluated on the clinical test set, to the predictions made by the edited-GTV-model evaluated on the edited test set (Supplementary Fig. 2A). We furthermore compared the predictions made by the clinical-GTV-model evaluated on the edited test set, to the predictions made by the edited-GTV-model evaluated on the edited test set (Supplementary Fig. 2B). This step was also done using the clinical test set (Supplementary Fig. 2C).

We investigated if editing influenced the evaluation of the clinical- and edited-GTV-model using the Wilcoxon signed rank test. We compared the predictions made by the clinical-GTV-model evaluated on the clinical test set, to the predictions made by the clinical-GTV-model evaluated on the edited test set (Supplementary Fig. 2D). For the same purpose, we compared the predictions made by the edited-GTV-model evaluated on the clinical test set, to the predictions made by the edited-GTV-model evaluated on the edited test set (Supplementary Fig. 2E).

To investigate whether the EOR influenced the DL-model performance, we compared the type of surgery for both the clinical and edited-GTV-model using the Kruskal-Wallis test. Furthermore, to evaluate the impact of imaging modalities on the DL-model performance, we compared all models utilizing multiple image modalities to the T1C model using the Wilcoxon signed rank test with Holm-Bonferroni correction. In each group (clinical and edited-GTV-model) six tests were corrected.

3. Results

3.1. Clinical and edited GTVs

All GTVs were edited. Edits included for example tightening of the delineation, excluding subdural hematoma, and expanding to encompass contrast enhancement and cavity (Fig. 1). The majority of the GTVs underwent small volume changes; the median (range) signed and relative volume change was 0.85 (-37.0–32.5) cc and 0.03 (-74.3–0.7) %, respectively (Supplementary Fig. 2). In total, 65 % of all cases had a decrease in volume in the edited version of the GTV, with the remaining 35 % increasing in volume.

3.2. Effect of editing on DL-model performance

In this part of the study, we compared the clinical- and edited-GTV-model trained using all available modalities (T1C/T1/FLAIR/CT).

A. When both models were evaluated on their own test sets, the $sDSC_{1mm}$ of the edited-GTV-model was significantly higher than that of the clinical-GTV-model: median (range) 0.92 (0.60–0.98) vs 0.69 (0.26–0.89), p-value < 0.001 (Table 1, Fig. 2A, Fig. 3).
B. When both models were evaluated on the same *edited* test set, the $sDSC_{1mm}$ of the edited-GTV-model remained significantly higher than that of the clinical-GTV-model; 0.92 (0.60–0.98) vs 0.76 (0.43–0.94), p-value < 0.001 (Fig. 2B).
C. When the two models were evaluated on the same *clinical* test set, the $sDSC_{1mm}$ of the models was not significantly different: 0.69 (0.26–0.89) for the clinical-GTV-model vs 0.70 (0.18–0.88) for the edited-GTV-model, p-value = 0.34 (Fig. 2C).

Results were consistent across all metrics (Supplementary Tables 3 and 4).

3.3. Effect of editing on DL-model evaluation

In this part of the study, we evaluate the performance of the clinical- and edited-GTV-models trained using all available modalities (T1C/T1/FLAIR/CT).

D. The $sDSC_{1mm}$ of the *clinical*-GTV-model evaluated on the *edited* test set was significantly higher than when evaluated on the *clinical* test set; 0.76 (0.43–0.94) vs 0.69 (0.26–0.89), p-value < 0.001 (Fig. 2D).
E. The $sDSC_{1mm}$ of the *edited*-GTV-model evaluated on the *clinical* test set was significantly lower than when evaluated on the *edited* test set; 0.70 (0.18–0.88) vs 0.92 (0.60–0.98), p-value < 0.001 (Fig. 2E).

Results were consistent across all metrics (Supplementary Tables 3 and 4).

3.4. Model performance stratified for extent of resection

In this part of the study, we evaluated the performance of the clinical- and edited-GTV-models trained using all available modalities (T1C/T1/FLAIR/CT). In the test set, 19 patients had undergone a CR, 17 a PR, and 16 a biopsy. The median (range) $sDSC_{1mm}$ of the clinical-GTV-model predictions tested on the clinical test set, was 0.75 (0.37–0.89), 0.69 (0.26–0.87), and 0.60 (0.43–0.80) for CR, PR, and biopsy, respectively (p-value = 0.074, Fig. 4). The median (range) $sDSC_{1mm}$ of the edited-GTV-model predictions tested on the edited test set, was 0.93 (0.77–0.96), 0.89 (0.60–0.96), and 0.92 (0.72–0.96) for CR, PR, and biopsy, respectively (p-value = 0.223). Results were consistent across all metrics (Supplementary Table 5).

3.5. Model performance stratified for image modalities

The median (range) $sDSC_{1mm}$ of the clinical- and edited-GTV-models based on multiple image modalities as input, evaluated on their own test sets, did not differ significantly compared to the clinical- and edited-

Table 1
Deep learning model results for the clinical-GTV-model trained on full imaging data and evaluated on the clinical test set and the edited-GTV-model trained on full imaging data and evaluated on the edited test. GTV: Gross Tumor Volume, $sDSC_{1mm}$: Surface Dice Similarity Coefficient at 1 mm tolerance, DSC: Dice Similarity Coefficient, HD95: 95th percentile Hausdorff Distance, MSD: Mean Surface Distance, T1C: Contrast enhanced T1, FLAIR: T2-weighted fluid-attenuated-inversion-recovery, CT: Computed Tomography.

Model: T1C+T1 + FLAIR+CT	Clinical-GTV-model Median (range)	Edited-GTV-model Median (range)
$sDSC_{1mm}$	0.69 (0.26–0.89)	0.92 (0.60–0.98)
DSC	0.90 (0.60–0.96)	0.95 (0.85–0.97)
HD95 (mm)	3.0 (1.4–22.1)	1.4 (1.0–11.1)
MSD (mm)	0.9 (0.5–3.6)	0.4 (0.2–2.8)

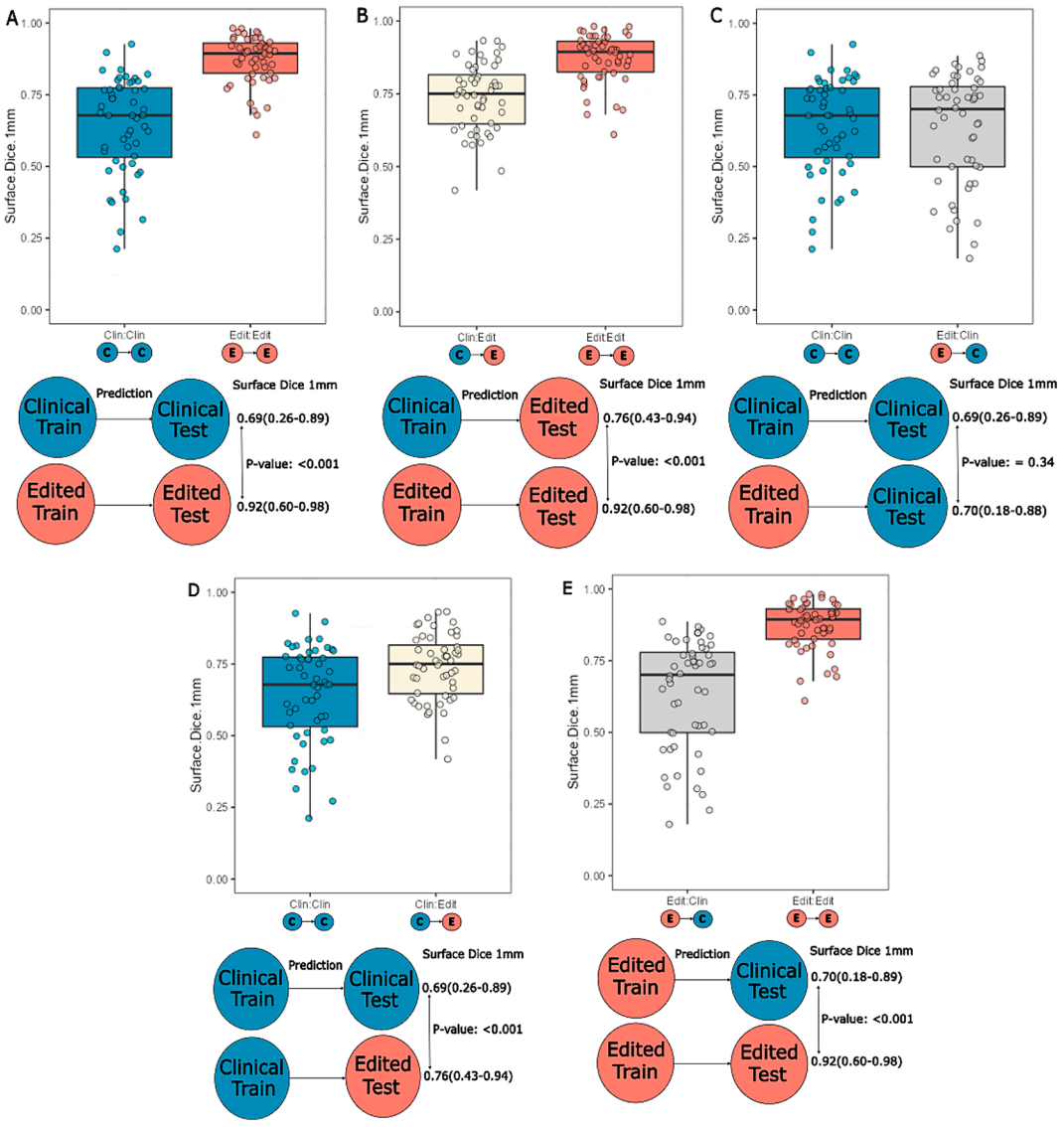

Fig. 2. Effect of editing on DL model performance and evaluation. (A, B and C): Comparisons of $sDSC_{1mm}$ where the models are fixed and the ground truths are changed. (D) and (E): Comparisons of $sDSC_{1mm}$ where the ground truth is fixed and the models are changed. Beneath each histogram figure is a visualization of the comparison made, along with the median $sDSC_{1mm}$, range and p-value for the hypothesis of no difference between groups. $sDSC_{1mm}$: Surface Dice Similarity Coefficient at 1 mm tolerance. Clin:Clin, Model trained on clinical data and evaluated on clinical test set. Clin:Edit, Model trained on clinical data and evaluated on edited test set. Edit:Edit, Model trained on edited data and evaluated on edited test set. Edit:Clin, Model trained on edited data and evaluated on clinical test set.

GTV-model using T1C-only (p-values > 0.01, Fig. 5). Results were consistent across all metrics (Supplementary Table 6).

4. Discussion

We performed the first head-to-head comparison of a DL-model with clinically used GTVs vs a DL-model with edited GTVs as labelled input to quantify the effect of editing contours. Our results show that editing the ground truth contours had a significantly positive impact on DL-segmentation performance. Furthermore, model performance did not depend on the extent of surgical resection. In addition, using multiple MR sequences and CT as input did not improve the performance. Lastly, our results suggest that edited contours in the test set, will increase sensitivity for evaluation of DL-model performance.

In comparison to the one other GBM GTV-model reported in the literature [8], our DL-model trained on clinical GTVs performed on a similar level, while ours trained on edited GTVs demonstrated a higher performance. Ramesh et al. had a similar size cohort of 225 patients for training and 30 for testing. Their training data underwent a similar editing procedure as in our study. They used a partially independent test set for which it is unclear if the delineations were edited. They reported a *mean* DSC of 0.73 and maximum HD of 10.75 mm where we observed a *mean* DSC/maximum HD of 0.94/8.7 mm in our edited-GTV model (Supplementary Table 7). Furthermore, their performance metrics from training were also lower than our test performance, suggesting that their model was not able to fully learn the task. Lastly, our DL-models performed well even in comparison with DL-models on scans made *before* surgery (top 10 contenders of the 2021 BraTS challenge, median DSC scores ranged from 0.94 to 0.945).

Our clinical- and edited-GTV-model had a median DSC of 0.90 and 0.95 respectively on the *edited* test set, obtaining a 0.05 increase in DL-model performance by editing the contours to fit with the imaging information. This improvement is substantial compared to DSC improvements achieved by network modifications within the BraTS challenge. For instance, Isensee et al. investigated modifications to the network structure and obtained an 0.006 increase in median DSC from 0.906 to 0.912 [27]. Luu and Park obtained an increase of 0.003 in median DSC from 0.925 to 0.928 [28]. Outside the field of GBM, other initiatives to

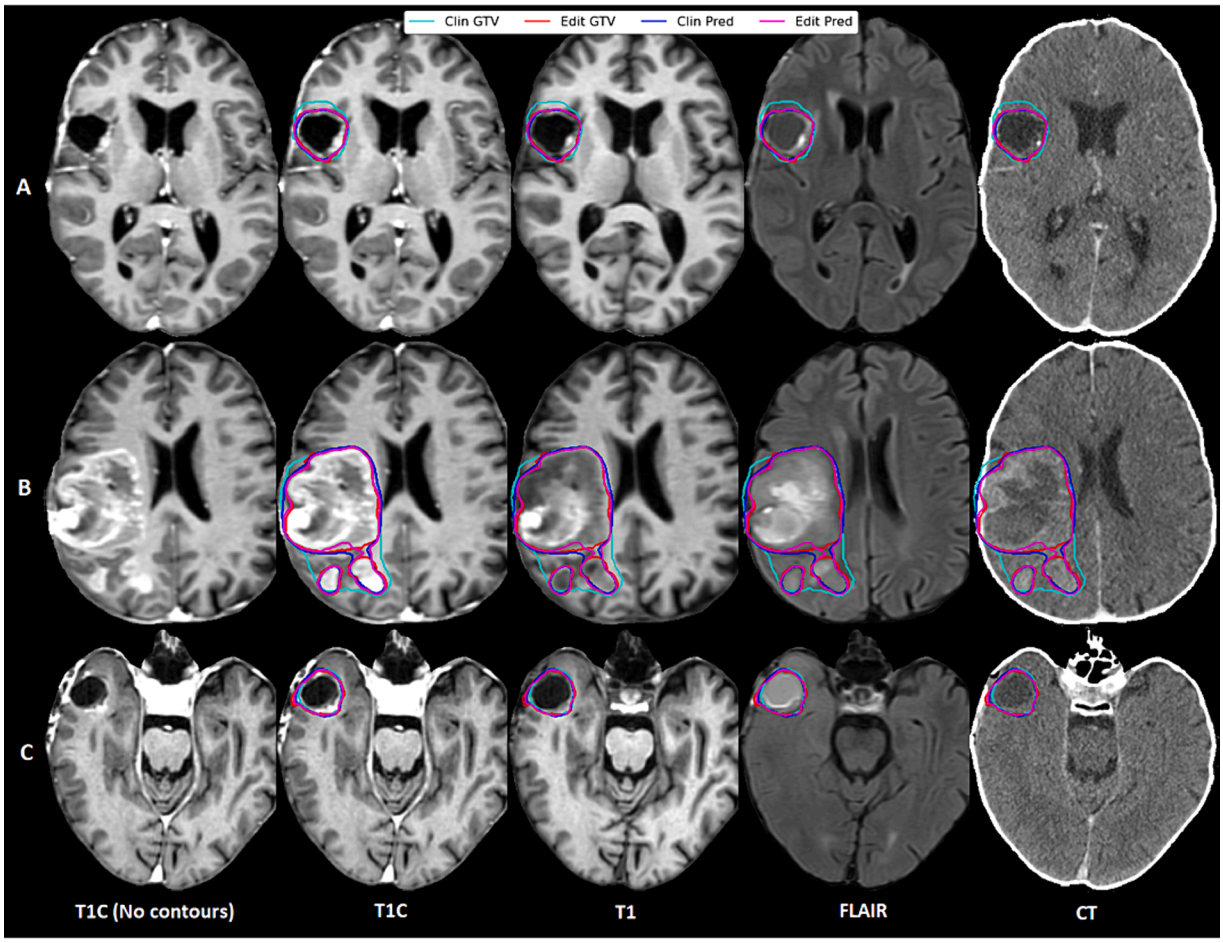

Fig. 3. Three examples of clinical and edited GTV segmentations together with clinical- and edited-GTV-model predictions, provided by the model using full imaging data. Each row depicts four image modalities from a patient; from left to right contrast enhanced T1 (T1C) without and with contours, T1, Fluid attenuated inversion recovery (FLAIR) and Computed Tomography (CT). Turquoise: Clinical ground truth (Clin GT), Red: Edited ground truth (Edit GT), Blue: Clinical deep learning prediction (Clin Pred), Magenta: Edit deep learning prediction (Edit Pred). Clin GT: Clinical Ground Truth, Edit GT: Edited Ground Truth, Clin Pred: Prediction using model trained on clinical ground truth, Edit Pred: Prediction using model trained on edited ground truth. T1C: Contrast enhanced T1, FLAIR: T2-weighted fluid-attenuated-inversion-recovery, CT: Computed Tomography. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

improve the performance of DL-models, such as multimodality imaging in Head & Neck GTV have achieved an increase from 0.58 to 0.74 mean DSC [14]. These findings support the hypothesis that the input imaging and labels are an important driving factor in DL-model performance using nnUNet.

Editing also had an influence on the evaluation of the DL-models, as both models had better performance metrics on the edited test set (Fig. 2D and E). Simultaneously, the *edited*-GTV-model performed on a similar level compared to the clinical-GTV-model when evaluated on the *clinical* test set (Fig. 2C). This leads us to conclude that using non-curated clinical data in a test set, can render the test set less sensitive to measure improvements in DL performance. Interestingly, these findings may also indicate that a DL-model can only learn from information present in the imaging, but not from random noise. This is supported by our finding that although the clinical-GTV-model was trained on multi-observer input delineations, its predictions were closer to the single-observer edited contours.

The EOR is a marker for heterogeneity in the imaging data and GTV definition, that could potentially be disruptive in the training procedure. However, we saw no significant difference in performance of the DL-models between patients with a biopsy, PR or CR. This indicates that we had an adequate number of patients in the DL-training process to effectively grasp and learn the GTV definition across varying EOR.

We considered six different combinations of image modalities, with T1C being present in all of them. We have not tested combinations without T1C due to its importance in the GTV definition. There was no statistical difference in performance between the model using only T1C and the more extensive combinations, similar to literature regarding brain metastases [10]. This suggests that using sequences on top of T1C has a very limited impact on DL-model performance. From our findings, we may also conclude that the clinical-GTV-model performance is not driven by other sequences than the edited-GTV model. A simple model based on a single image sequence is easier to implement and maintain in clinical practice.

We identified the following limitations for this study. With the chosen methodology, we cannot disentangle the effects of working with a single-observer and editing on DL-model performance. Nevertheless, we hypothesize that the fitting of the contours to the imaging information is responsible for a large part of the observed effect, as the multi-observer clinical-model generalized better to the *edited* contours. Our findings are furthermore limited to the specific GTV definition used, i.e. the GTV was based on the surgical cavity and any residual contrast enhancement. Although this definition is suited for the vast majority of GBM patients, it excludes patients in whom the GTV should be defined by FLAIR abnormalities as well [4]. Quantifying IOV would have made it possible to measure if the automatic segmentations meet the clinical threshold. In addition to this, automatic segmentations that exceed IOV could be a sign that the model is too observer specific. Training a model

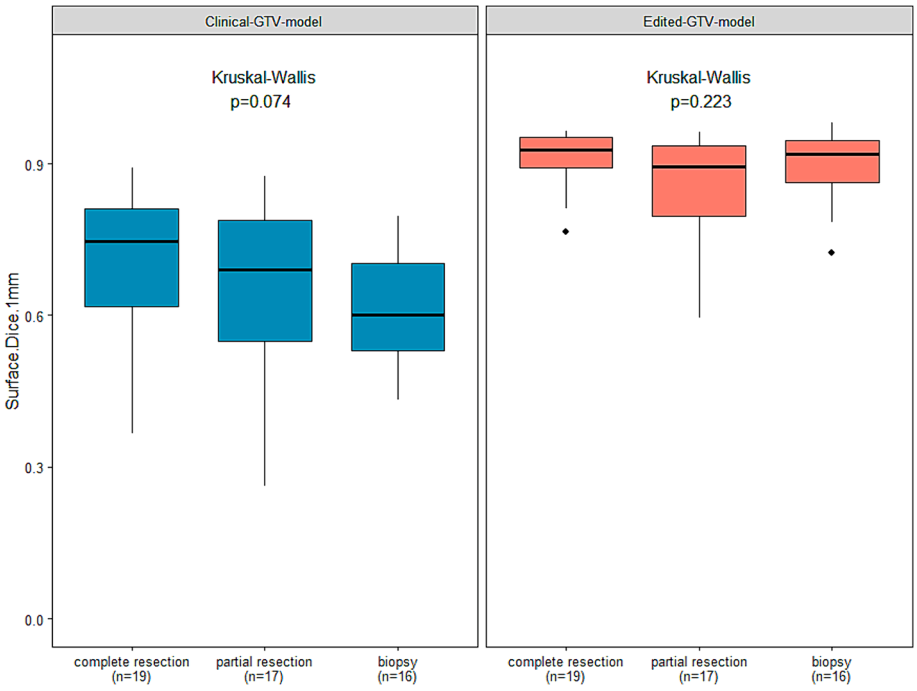

Fig. 4. Clinical- and edited-GTV-model performance differentiated for extent of resection. Boxplot depicting $sDSC_{1mm}$ of the clinical-GTV-model on the left and the edited-GTV-model on the right, both using all modalities (T1C, T1, FLAIR, CT), for patients who had undergone a complete resection (left), partial resection (middle) and biopsy (right). The p-values for the difference comparison are shown above. $sDSC_{1mm}$: Surface Dice Similarity Coefficient at 1 mm tolerance, GTV: Gross Tumor Volume, T1C: Contrast enhanced T1, FLAIR: T2-weighted fluid-attenuated-inversion-recovery, CT: Computed Tomography.

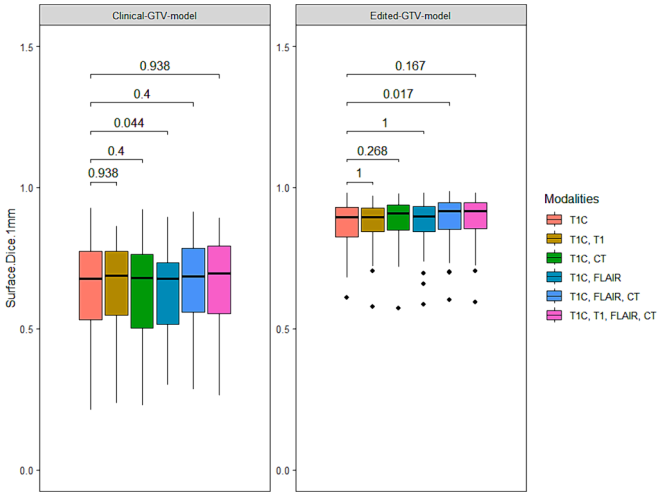

Fig. 5. Clinical- and edited-GTV-model performance differentiated for selected combinations of image modalities used in the training procedure. Boxplot comparing $sDSC_{1mm}$ of the clinical-GTV-model on the left and the edited-GTV-model on the right, trained using the following combinations of image modalities: T1C (red), T1C, T1 (beige), T1C, FLAIR (teal), T1C, FLAIR, CT (blue) and T1C, T1, FLAIR, CT (pink). The p-values for the difference comparison are shown above. $sDSC_{1mm}$: Surface Dice Similarity Coefficient at 1 mm tolerance, GTV: Gross Tumor Volume, T1C: Contrast enhanced T1, FLAIR: T2-weighted fluid-attenuated-inversion-recovery, CT: Computed Tomography. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

on delineations edited by a single observer will skew the predictions towards the observer's delineation style (bias). This could limit the clinical applicability of the model. Nevertheless, an argument supporting the clinical applicability of our single-observer edited-GTV-model would be that the predictions from the clinical-GTV-model were closer to the edited than to the clinical contours. Finally, the increase in DL-performance with editing was not evaluated for quality. Therefore, it remains to be investigated prospectively whether this increase reduces the need of manual corrections for an automatic segmentation to be clinically acceptable.

To conclude, we achieved a high GTV segmentation accuracy for post-operative GBM RT, where editing had a significant positive effect on the DL-model performance with a relevant effect size. Our model was not dependent on varying EOR. Finally, to achieve this result, only T1C was needed as an input image.

CRediT authorship contribution statement

Kim M. Hochreuter: Methodology, Investigation, Formal analysis, Data curation, Writing – original draft, Visualization. **Jintao Ren:** Methodology, Validation, Data curation, Writing – review & editing. **Jasper Nijkamp:** Methodology, Validation, Data curation, Writing – review & editing. **Stine S. Korreman:** Methodology, Writing – review & editing. **Slávka Lukacova:** Resources, Writing – review & editing. **Jesper F. Kallehauge:** Methodology, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Anouk K. Trip:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jasper Nijkamp declares that he is an associate editor on the editorial board of PhiRO.

Acknowledgements

The authors would like to acknowledge the Danish Clinical Quality

Program – National Clinical Registries.

Funding

The work in this manuscript was supported by the Danish Cancer Society (grant no. R302-A17263-B395), the Danish Comprehensive Cancer Center – Brain Tumor Center which is funded by the Danish Cancer Society (grant no. R295-A16770) and the Danish Comprehensive Cancer Center (grant no. 2021-08).

Data statement

The T1C and full model trained using edited delineations can be accessed at: https://gitlab.com/dcpt-research/glioblastoma-gtv-segmentation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2024.100620.

References

[1] Delgado-López PD, Corrales-García EM. Survival in glioblastoma: a review on the impact of treatment modalities. Clin Transl Oncol 2016;18:1062–71. https://doi.org/10.1007/s12094-016-1497-x.
[2] Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med 2005;352:987–96. https://doi.org/10.1056/NEJMoa043330.
[3] Niyazi M, Brada M, Chalmers AJ, Combs SE, Erridge SC, Fiorentino A, et al. ESTRO-ACROP guideline "target delineation of glioblastomas". Radiother Oncol 2016;118:35–42. https://doi.org/10.1016/j.radonc.2015.12.003.
[4] Niyazi M, Andratschke N, Bendszus M, Chalmers AJ, Erridge SC, Galldiks N, et al. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. Radiother Oncol 2023;184:109663. https://doi.org/10.1016/j.radonc.2023.109663.
[5] Wee CW, Sung W, Kang H-C, Cho KH, Han TJ, Jeong B-K, et al. Evaluation of variability in target volume delineation for newly diagnosed glioblastoma: a multi-institutional study from the Korean Radiation Oncology Group. Radiat Oncol 2016;10:137. https://doi.org/10.1186/s13014-015-0439-z.
[6] Tseng C-L, Stewart J, Whitfield G, Verhoeff JJC, Bovi J, Soliman H, et al. Glioma consensus contouring recommendations from a MR-Linac International Consortium Research Group and evaluation of a CT-MRI and MRI-only workflow. J Neurooncol 2020;149:305–14. https://doi.org/10.1007/s11060-020-03605-6.
[7] Kruser TJ, Bosch WR, Badiyan SN, Bovi JA, Ghia AJ, Kim MM, et al. NRG brain tumor specialists consensus guidelines for glioblastoma contouring. J Neurooncol 2019;143:157–66. https://doi.org/10.1007/s11060-019-03152-9.
[8] Ramesh KK, Xu KM, Trivedi AG, Huang V, Sharghi VK, Kleinberg LR, et al. A fully automated post-surgical brain tumor segmentation model for radiation treatment planning and longitudinal tracking. Cancers 2023;15:3956. https://doi.org/10.3390/cancers15153956.
[9] Ermiş E, Jungo A, Poel R, Blatti-Moreno M, Meier R, Knecht U, et al. Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. Radiat Oncol 2020;15:100. https://doi.org/10.1186/s13014-020-01553-z.
[10] Buchner JA, Peeken JC, Etzel L, Ezhov I, Mayinger M, Christ SM, et al. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. Radiother Oncol 2023;188:109901. https://doi.org/10.1016/j.radonc.2023.109901.
[11] Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother Oncol 2021;160:175–84. https://doi.org/10.1016/j.radonc.2021.04.019.
[12] Henderson EGA, Osorio EMV, van Herk M, Green AF. Optimising a 3D convolutional neural network for head and neck computed tomography segmentation with limited training data. Phys Imaging Radiat Oncol 2022;22:44–50. https://doi.org/10.1016/j.phro.2022.04.003.
[13] Henderson EGA, Vasquez Osorio EM, Van Herk M, Brouwer CL, Steenbakkers RJHM, Green AF. Accurate segmentation of head and neck radiotherapy CT scans with 3D CNNs: consistency is key. Phys Med Biol 2023;68:085003. https://doi.org/10.1088/1361-6560/acc309.
[14] Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol 2021;60:1399–406. https://doi.org/10.1080/0284186X.2021.1949034.
[15] Holtzman Gazit M, Faran R, Stepovoy K, Peles O, Shamir RR. Post-operative glioblastoma multiforme segmentation with uncertainty estimation. Front Hum Neurosci 2022;16.
[16] Ghaffari M, Samarasinghe G, Jameson M, Aly F, Holloway L, Chlap P, et al. Automated post-operative brain tumour segmentation: a deep learning model based on transfer learning from pre-operative images. Magn Reson Imaging 2022;86:28–36. https://doi.org/10.1016/j.mri.2021.10.012.
[17] Ren J, Huynh B-N, Groendahl AR, Tomic O, Futsaether CM, Korreman SSPET. Normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, editors. Head neck tumor segmentation outcome predict, vol. 13209. Cham: Springer International Publishing; 2022. p. 83–91. https://doi.org/10.1007/978-3-030-98253-9_7.
[18] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge; 2018. https://doi.org/10.48550/ARXIV.1811.02629.
[19] Jungo A, Meier R, Ermis E, Blatti-Moreno M, Herrmann E, Wiest R, et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation; 2018. https://doi.org/10.48550/ARXIV.1806.02562.
[20] Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 2020;65:101759. https://doi.org/10.1016/j.media.2020.101759.
[21] AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med Phys 2018;45:1150–8. https://doi.org/10.1002/mp.12752.
[22] Deeley MA, Chen A, Datteri RD, Noble J, Cmelak A, Donnelly E, et al. Segmentation editing improves efficiency while reducing inter-expert variation and maintaining accuracy for normal brain tissues in the presence of space-occupying lesions. Phys Med Biol 2013;58:4071–97. https://doi.org/10.1088/0031-9155/58/12/4071.
[23] Karschnia P, Young JS, Dono A, Häni L, Sciortino T, Bruno F, et al. Prognostic validation of a new classification system for extent of resection in glioblastoma: a report of the RANO *resect* group. Neuro-Oncol 2023;25:940–54. https://doi.org/10.1093/neuonc/noac193.
[24] Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification 2021.
[25] Helland RH, Ferles A, Pedersen A, Kommers I, Ardon H, Barkhof F, et al. Segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks. In Review 2023. https://doi.org/10.21203/rs.3.rs-2943128/v1.
[26] Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: self-adapting framework for U-net-based medical image segmentation; 2018. https://doi.org/10.48550/arXiv.1809.10486.
[27] Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. nnU-net for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion Glioma Mult. Scler. Stroke Trauma. Brain Inj. Cham: Springer International Publishing; 2021. p. 118–32. https://doi.org/10.1007/978-3-030-72087-2_11.
[28] Luu HM, Park S-H. Extending nn-UNet for brain tumor segmentation. In: Crimi A, Bakas S, editors. Brainlesion Glioma Mult. Scler. Stroke Trauma. Brain Inj. Cham: Springer International Publishing; 2022. p. 173–86. https://doi.org/10.1007/978-3-031-09002-8_16.