A Multi-criteria Quality Assessment of Automated Alternative Segmentations for Radiation Therapy of Brain Tumor Patients

Amith Kamath [1] AMITH.KAMATH@UNIBE.CH
[1] *ARTORG Center for Biomedical Engineering Research, University of Bern*
Robert Münger [1] ROBERT.MUENGER@UNIBE.CH
Robert Poel [2] ROBERT.POEL@INSEL.CH
[2] *Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern*
Elias Rüfenacht [1] ELIAS.RUEFENACHT@UNIBE.CH
Alain Jungo [1] ALAIN.JUNGO@UNIBE.CH
Jonas Willmann [3] JONAS.WILLMANN@USZ.CH
[3] *Department of Radiation Oncology, University Hospital Zurich, University of Zurich*
Nicolaus Andratschke [3] NICOLAUS.ANDRATSCHKE@USZ.CH
Mauricio Reyes [1] MAURICIO.REYES@UNIBE.CH

Editors: Under Review for MIDL 2022

Abstract

Medical image segmentation is a crucial part of the radiotherapy (RT) planning workflow for treating brain tumor patients. Consequently, radiotherapy quality assurance (RTQA) of expert-segmentations is performed in clinical routine, especially in clinical trials. However, RTQA is time-consuming and error-prone. Towards automating this, we hypothesize that models that can generate a distribution of segmentations to simulate the variability in expert-segmentation, can be used as a proxy to evaluate the compliance and quality of new segmentations. In this paper, we evaluate a deep learning (Stochastic Segmentation Networks), and a non-deep learning approach (Gaussian Process Sampling Segmentation of Images) to generate this distribution of 'alternative segmentations'. We assess the quality of these alternative segmentations using three complementary criteria: (i) a blinded qualitative assessment of expert-segmentations with computer-generated alternatives, (ii) geometric compliance using Dice similarity coefficient, and a (iii) dosimetric compliance evaluation using dose-volume histogram curves. On an evaluation data set consisting of 40 RT plans (2 subjects, 2 Organs at Risk, and 10 plans each), our results indicate that these methods yield plausible alternative segmentations which could be used to build a deep learning-based RTQA platform. Our results further indicate that geometric compliance needs to be complemented with dosimetry to fully characterize the impact of segmentation deviations for target coverage and organs at risk toxicity.

Keywords: Radiotherapy Quality Assurance, Alternative Segmentations, Clinical Evaluation.

1. Introduction

Glioblastoma is the most common and aggressive brain tumor, accounting for 45% of all malignant primary brain tumors (McFaline-Figueroa and Lee, 2018). Current treatment consists of a combination of surgery, adjuvant radiotherapy (RT), and concomitant and adjuvant chemotherapy (Stupp et al., 2005). An accurate segmentation of the anatomy is critical for conformal RT planning, as it aims to improve tumor control by boosting the dose to the target volume (i.e., tumor or resection cavity, with adjacent areas of potential microscopic spread) while sparing organs at risk (OAR), thus limiting normal tissue toxicity (Scoccianti et al., 2015). Radiation oncologists can choose from more than a dozen commercial software tools to draw boundaries around the target volumes and OAR either manually or semi-automatically (Meyer et al., 2018).

Accurate segmentation of target volumes and OAR is time consuming and prone to variability (Wee et al., 2016; Vogin et al., 2021). In clinical trials, failure to adhere to RT standards may impact local control and overall survival, and jeopardize the therapeutic effectiveness of the studied regime (Ohri et al., 2013). In a multi-center phase III head and neck cancer trial, protocol non-compliance was associated with a 20% poorer survival at two years (Peters et al., 2010). Incorrect target volume segmentation caused 25% (24/97) of the non-compliant treatment plans. This indicates that target volume and OAR segmentation are amongst the most critical steps in the RT process. Therefore, efforts have been made to standardize segmentation and develop RTQA systems (Weber et al., 2012; Niyazi et al., 2016).

Organizations such as the European Organisation for Research and Treatment of Cancer (EORTC) and Radiotherapy Trials Quality Assurance (RTTQA) have introduced QA procedures to eliminate bias due to systematic variations in radiotherapy delivery (Fairchild et al., 2012; Nixon et al., 2013). In current EORTC brain tumor trials, segmentations of target volumes and OAR are manually reviewed by specialized radiation oncologists. This process is referred to as the individual case review. In a trial of several hundreds of patients, this becomes time-consuming for investigators and reviewers, and cost-intensive for the sponsor. Moreover, the extra workload for investigators might discourage trial participation.

This motivates us to explore automated QA methods (referred hereafter as auto-QA) to evaluate OAR segmentations for radiotherapy. Prior literature on automated evaluations of segmentations have either focused on geometric plausibility (Robinson et al., 2017, 2018; Audelan and Delingette, 2019; Xia et al., 2020; Fournel et al., 2021), or on time-saving aspects (Van der Veen et al., 2019; Vaassen et al., 2020). Beyond these, we believe that a multi-criteria, radiotherapy focused auto-QA system must incorporate dosimetric compliance in addition to radiological (i.e., reviewer visual evaluation) and geometric metrics. Dosimetric parameters are indicative of tumor control or normal tissue complication probability and commonly evaluated using dose-volume histograms (DVH) for radiotherapy treatment plans (Drzymala et al., 1991). The DVH curves describe the uniformity of radiation dose each target volume and adjacent OAR receive, summarizing the three-dimensional dose distribution in a two-dimensional format.

We hypothesize that a distribution of multiple automated segmentations can be used to simulate the variability in expert-segmentation. Through this, we hope to analyze the variability of RT plans this family of segmentations generates, for a broader evaluation base to flag non-compliance in the dosimetric sense. Towards this broad long-term target of auto-QA for RT planning, our first step involves testing this hypothesis with two methods - Stochastic Segmentation Networks (SSN) (Monteiro et al., 2020) and Gaussian Process Sampling Segmentation of Images (GPSSI) (Lê et al., 2015) which generate what we henceforth call ‘alternative segmentations’. The major contributions of this paper are threefold - we show:

(i) In Section 3.1, that in a blinded qualitative assessment of alternatives vs. reference, expert-clinician reviewers do not find significant differences between Deep Learning generated segmentations versus expert-annotated reference.

(ii) In Section 3.2 that the variations in Dice similarity coefficients (DSC) are not always perceptible from an expert-clinician reviewer perspective. Furthermore, variations in DSC may not reflect linearly as variations in RT plans depending on the OAR in question.

(iii) In Section 3.3 that the variations in DVH for various OAR and segmentations have a complex relationship to their geometric evaluations, which needs further exploration to build clinically relevant models.

2. Methods

This section describes our study design, followed by a description of the data set used. SSN and GPSSI are briefly described next, followed by the radiotherapy planning process.

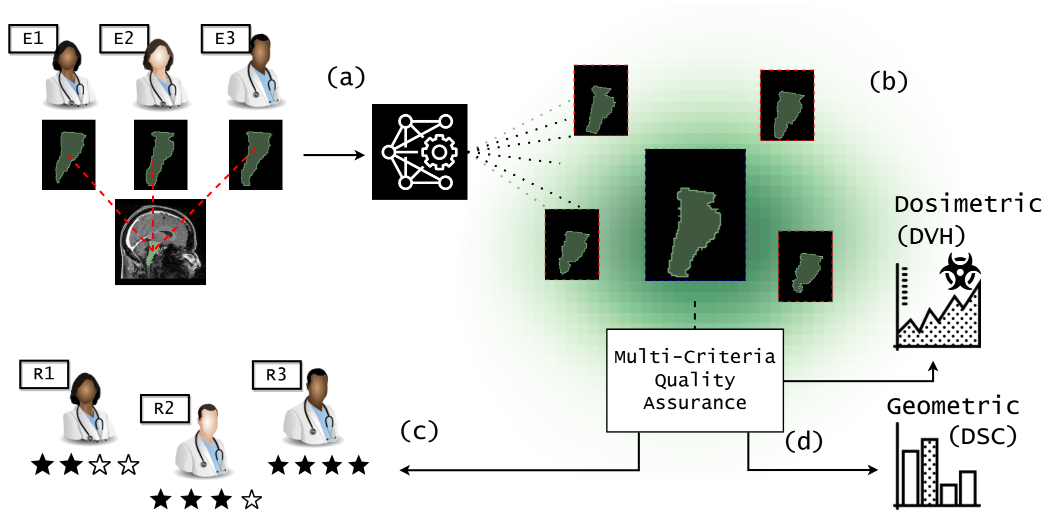

Figure 1: How good are automated alternative segmentations for RTQA? Training data annotated by three experts (E1, E2, E3) are fused via majority voting, to (a) train an alternative segmentation model (SSN/GPSSI). From a new segmentation under QA, the model generates a distribution of alternatives (b), which are evaluated based on: a blinded qualitative assessment of generated segmentations versus reference (c) by three reviewers R1, R2, R3, followed by computation (d) of geometric (DSC) and dosimetric (DVH) metrics.

Fig. 1 describes our study design. 30 glioblastoma cases are independently labeled by three expert radiation oncologists - E1, E2 and E3. For each case, OARs are segmented following standard clinical guidelines (Scoccianti et al., 2015). In this study, we focus specifically on the brainstem and the hippocampus (left and right separately), due to their delineation complexity, their clinical relevance, and adequacy (Bartel et al., 2019) for a first analysis of the proposed auto-QA approach. Majority voting is used to fuse expert-segmentations and yield the reference to train a deep neural network using SSN (described in Section 2.2). We also use GPSSI (described in Section 2.3) to generate alternative segmentations on the same cases. These auto-generated segmentations are scored by three reviewers - R1, R2 and R3 - who also specialize in radiation oncology. A ten point Likert-scale of 1 (not usable) to 10 (experts' choice) is used for the scoring.

In order to compare the reviewers' scores to a reference segmentation of known quality, we replace one of the alternative segmentation with the fused expert segmentation. We do not inform the reviewers (R1, R2, R3) that this reference segmentation was included. We refer to this as a blinded qualitative assessment to measure how such computer generated alternatives would compare to reference segmentation based on the reviewers' qualitative assessment. Fig. 6 shows an example of how alternatives for a single case were presented to reviewers for recording their scores.

2.1. Description of Data set and Reference

The data set we use includes 30 newly diagnosed glioblastoma cases, each containing volumes of magnetic resonance images (MRI) with T1, T2, T1c and FLAIR sequences, as well as computed tomography (CT) scans. The MRI volumes contain 256-by-256-by-256 voxels, in 1mm isotropic resolution. Each of these have segmentations for the brainstem and hippocampus (both left and right), manually drawn independently by E1, E2 and E3. A majority voting of these segmentations was used as the reference for training SSNs, and as initial contours for GPSSI. We register these structures with the corresponding CT scans for RT planning.

2.2. Stochastic Segmentation Networks

SSN (Monteiro et al., 2020) is a probabilistic method for modeling aleatoric uncertainty using generic deep learning-based image segmentation architectures. It does so by efficiently representing correlations between pixels by modeling the logit map as a low-rank multivariate normal distribution.

We use Deep medic (Kamnitsas et al., 2017) as the base segmentation network. Four 3D input channels (FLAIR, T1, T1c, T2) and four labels (Background, brainstem, hippocampus left, hippocampus right) were used as inputs and outputs. For training, a random selection of 20 out of the 30 cases were used. Five cases were used for validation. For evaluating alternative segmentations, 20 random cases are chosen, which could include cases from the training and validation sets. The batch size for training was set to 10, with 20 patches per volume of size 110-by-110-by-110. For data augmentation, we used random elastic deformations on the entire input volume, followed by a random histogram deformation and a random patch rotation on each patch. These settings are inspired from the ones used in the Brain Tumor experiments in (Monteiro et al., 2020) and is implemented using: https://github.com/biomedia-mira/stochastic_segmentation_networks.

2.3. Gaussian Process Sampling for Segmenting Images

For comparison, we investigate GPSSI, a non-deep learning approach (Lê et al., 2015). The reference segmentation is used as input to define a probability distribution of segmentations as a Gaussian process with a squared exponential covariance. This produces spatially coherent alternative segmentations, while using a single parameter to vary the samples, corresponding to the DSC between the initial segmentation and the generated alternatives. Sampled variations account for the image intensity by combining signed distance functions with signed geodesic distances. This ensures that the variation is less likely to perturb those regions with higher contrast, as compared to those with a lower geodesic distance, i.e., lower contrast variations. We use the implementation from https://github.com/alainjungo/gpssi.

2.4. Radiotherapy Treatment Planning

We use the Eclipse treatment planning system (Varian Medical Systems, Palo Alto, CA) to generate treatment plans for two glioblastoma cases. Each include reference target volumes, OARs, and 10 alternative segmentations for brainstem and hippocampus. First, a clinically acceptable treatment 'reference plan' is created, and optimized on the reference target volumes and OAR. Next, the reference plan is re-optimized for each of the alternative segmentations of brainstem and hippocampus, creating one 'alternative plan' per alternative segmentation. To avoid plan optimization parameters from being sources of confounding, the same parameters that yield a clinically acceptable reference plan are used for each alternative plan. To assess the impact of segmentation variability on dosimetric parameters, the DVH curves for brainstem, hippocampus and planning target volume (PTV) from each of the alternative plans and the reference plan are compared in terms of applied dose at specific ratios of total structure volume. The parameters for plan optimization used for each case is available on request.

3. Results

Qualitative results are presented in Section 3.1, followed by a geometric assessment in Section 3.2, and finally a dosimetric assessment in Section 3.3.

3.1. Qualitative Results

Qualitative scores are recorded on an 10-point Likert-scale of 1 (not usable) to 10 (experts' choice). Intermediate levels are marked as 3 ("OK but needs lots of corrections"), 5 ("beginner") and 7 ("intermediate").

Fig. 2 (left) summarizes scores by each reviewer R1, R2 and R3, per alternative generated using GPSSI (as shown in the questionnaire). We observed that the mean rating of all the reviewers for the reference is higher than that for any of the five GPSSI generated alternatives (see Table 1 for more details). Fig. 2 (right) shows the average scores for each of the 20 cases. Overall, we observe a good level of agreement among raters, indicating a robust assessment.

Fig. 3 shows the corresponding qualitative scores for alternatives generated using SSN. In contrast to GPSSI, two of the three reviewers gave a higher mean rating to segmentations

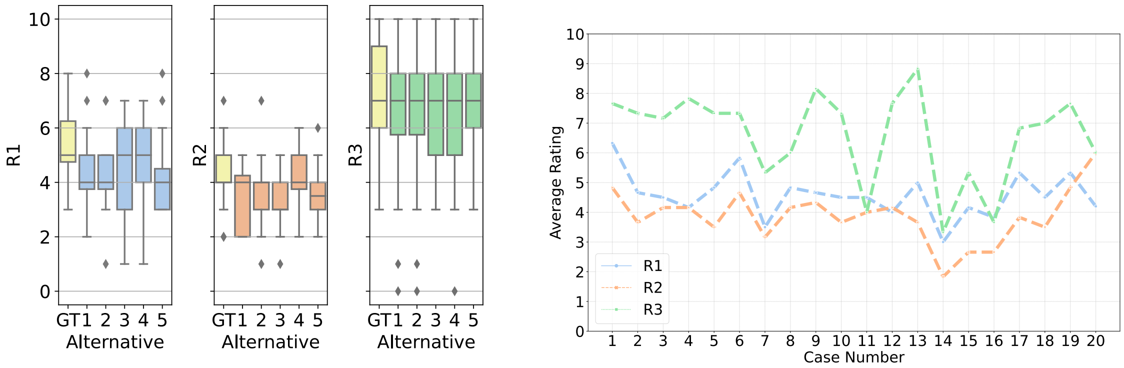

Figure 2: Qualitative evaluation of GPSSI-generated alternatives by reviewers (blue - R1, orange - R2, green - R3). Left summarizes score per alternative (yellow reference + 5 alternatives). Right shows the average score trend for 20 cases.

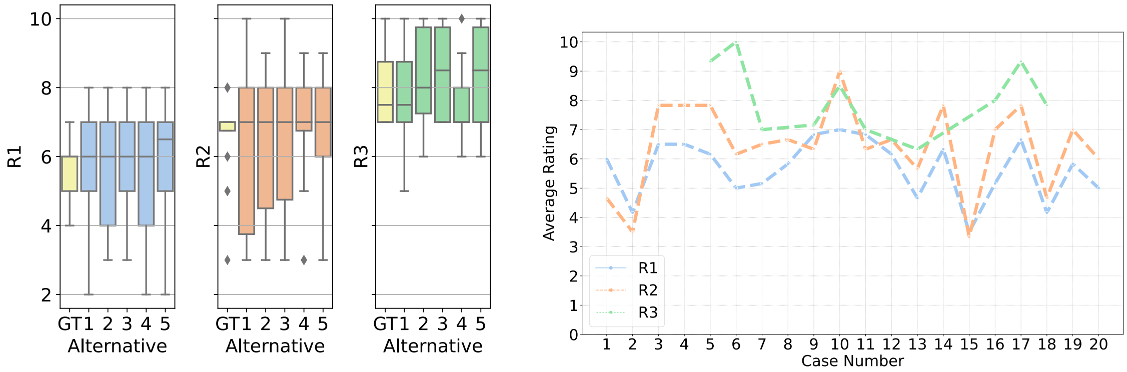

Figure 3: Qualitative evaluation of SSN-generated alternatives (in the same format as Fig. 2) Note: Reviewer R3 scored only 10 out of 20 cases - some green points are hence missing.

generated by SSN than the reference - indicating the plausibility of SSN as generator of alternative segmentations. However, we also observe a larger standard deviation of SSN-based segmentations (see Table 2 for more details).

3.2. Dice Similarity Coefficient Analysis

To evaluate geometric compliance, we generate 20 alternatives using SSN/GPSSI models and calculate DSC versus expert-segmentations from E1, E2 and E3 for each of 20 cases (see individual case results for SSN in Fig. 8). Fig. 4 shows a box plot of DSC for all 400 of these alternatives generated using SSN (left) and GPSSI (right). The DSC ranges for SSN are higher than those of GPSSI - following the same trend as the qualitative assessment in Section. 3.1. We observe that the mean DSC for brainstem (blue) is consistently higher than the mean DSC of hippocampus (orange) for both, while the variance of DSC for the brainstem is smaller than that for the hippocampus. Contrasting these numbers to the qualitative analysis indicates that an auto-generated sample with a low DSC could still be rated in the same range as one with a high DSC. This indicates a need to investigate its broader utility in RT planning.

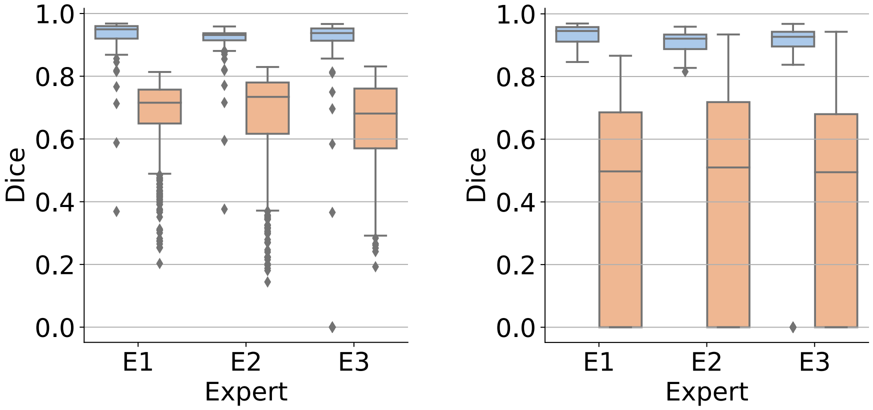

Figure 4: DSC variations using SSN (left) and GPSSI (right) for both brainstem (blue) and hippocampus (orange) for 400 alternatives (20 each for 20 cases) against each of three expert-references by E1, E2 and E3.

3.3. Dose Volume Histogram Analysis

We analyze DVH curves for two cases using 10 SSN-generated alternative segmentations. Fig. 5 shows the DVH curves for the left hippocampus in these two cases. Two of these alternatives included spurious false positive voxels, marked alternative 4 and 8 in Fig. 5 (right) (see Fig. 7 for more details).

A hippocampal D40%, i.e., dose to 40% of the hippocampus, of more than 7.3 Gray (Gy) has been established as a dose constraint for neurocognitive impairment in prospective trials on RT for brain tumors (Gondi et al., 2012). We observe that in case # 10 (left), 40% of the hippocampus gets a dose higher than the recommended 7.3 Gy. At 40%, the dose varies from 12 Gy for the reference segmentation to between 15 to 16 Gy for the alternatives. Similarly, in case # 7 (right), alternative segmentations generally lead to dose variations of up to 10 Gy. Upon comparing the DSC for these two cases (see highlighted cases in Fig. 8), these DVH variations show the need for using broader complementary QA metrics. Even though we observe no major dosimetric variations from outlier (false positive voxel regions) detected (i.e., below 2 Gy difference), more research is needed to investigate the impact of outlier size and location on RTQA.

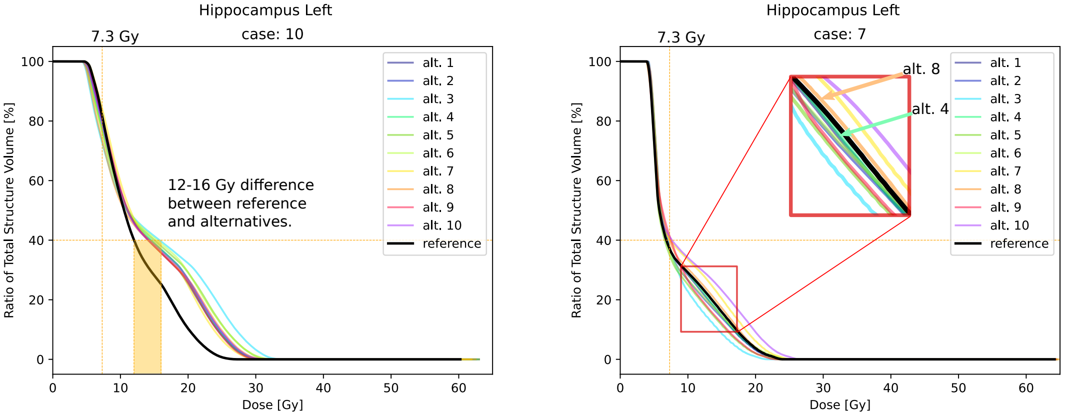

Figure 5: Example of DVH variations for cases # 10 and # 7. While these cases show high DSC, the corresponding DVH curves present large dosimetric variations.

4. Discussions and Conclusion

We postulate that alternative segmentations can be useful to estimate ranges of plausible segmentations for RTQA as they provide a richer set of reference information than single-point comparisons (e.g., comparison to a pseudo-reference) or direct estimations of specific QA metrics (e.g., DSC). Towards such a multi-criteria RTQA platform using alternative segmentations, we evaluate the performance and plausibility of alternative segmentations generated by two approaches - SSN (deep learning-based) and GPSSI (non-deep learning). Based on radiological (i.e., reviewer visual evaluation), geometric (i.e., DSC), and dosimetric (i.e., DVH) metrics, we find that deep learning models like SSN have the potential to generate distributions of alternative segmentations.

However, we also observe some false positive outlier regions, which need further investigation for validating such approaches for RTQA. Following recent findings (Poel et al., 2021; Kofler et al., 2021), and community efforts (Reinke et al., 2021), we demonstrate that conventional loss metrics used to train and assess models may not entirely address the needs for RTQA (see Fig. 9 for an example). We hope this inspires more work to design clinically-oriented multi-criteria metrics for RT segmentation review.

In this initial study, we focused on alternative segmentations of the brainstem and hippocampus as two representative structures for RTQA. Future work will incorporate other OARs within a larger multi-center clinical trial data set of cases from the EORTC. We further aim at building on this work to create an automatic RTQA workflow, where input segmentations can be evaluated with radiological, geometric and dosimetric measures. In this sense, we would like to explore how combining metrics towards a single score could help improve the clinical workflow. Beyond the area of RT, we hope this work stimulates further research in other domains where deep learning based quality assurance can play an important role.

Acknowledgments

We thank the Swiss Cancer League (grant number KFS-5127-08-2020) for funding this work.

References

Benoît Audelan and Hervé Delingette. Unsupervised quality control of image segmentation based on bayesian learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 21–29. Springer, 2019.

F Bartel, Marcel van Herk, H Vrenken, F Vandaele, Stefan Sunaert, K De Jaeger, NJ Dollekamp, C Carbaat, E Lamers, EMT Dieleman, et al. Inter-observer variation of hippocampus delineation in hippocampal avoidance prophylactic cranial irradiation. *Clinical and Translational Oncology*, 21(2):178–186, 2019.

RE Drzymala, Radhe Mohan, L Brewster, J Chu, Michael Goitein, W Harms, and M Urie. Dose-volume histograms. *International Journal of Radiation Oncology* Biology* Physics*, 21(1):71–78, 1991.

Alysa Fairchild, Edwin Aird, Paul A Fenton, Vincent Gregoire, Akos Gulyban, Denis Lacombe, Oscar Matzinger, Philip Poortmans, Pascal Ruyskart, Damien C Weber, et al. EORTC radiation oncology group quality assurance platform: establishment of a digital central review facility. *Radiotherapy and Oncology*, 103(3):279–286, 2012.

Joris Fournel, Axel Bartoli, David Bendahan, Maxime Guye, Monique Bernard, Elisa Rauseo, Mohammed Y Khanji, Steffen E Petersen, Alexis Jacquier, and Badih Ghattas. Medical image segmentation automatic quality control: A multi-dimensional approach. *Medical Image Analysis*, 74:102213, 2021.

Vinai Gondi, Bruce P Hermann, Minesh P Mehta, and Wolfgang A Tomé. Hippocampal dosimetry predicts neurocognitive function impairment after fractionated stereotactic radiotherapy for benign or low-grade adult brain tumors. *International Journal of Radiation Oncology* Biology* Physics*, 83(4):e487–e493, 2012.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

Florian Kofler, Ivan Ezhov, Fabian Isensee, Fabian Balsiger, Christoph Berger, Maximilian Koerner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *arXiv preprint arXiv:2103.06205*, 2021.

Matthieu Lê, Jan Unkelbach, Nicholas Ayache, and Hervé Delingette. GPSSI: Gaussian process for sampling segmentations of images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 38–46. Springer, 2015.

J Ricardo McFaline-Figueroa and Eudocia Q Lee. Brain tumors. *The American Journal of Medicine*, 131(8):874–882, 2018.

Philippe Meyer, Vincent Noblet, Christophe Mazzara, and Alex Lallement. Survey on deep learning for radiotherapy. *Computers in Biology and Medicine*, 98:126–146, 2018.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *arXiv preprint arXiv:2006.06015*, 2020.

Lisette Sheena Nixon, S Mukherjee, L Wills, T Millin, SE Bridges, RA Abrams, G Joseph, G Griffiths, C Hurt, and J Staffurth. The SCALOP trial plan assessment form (PAF) as a tool for radiation therapy trials quality assurance (RTTQA). *International Journal of Radiation Oncology, Biology, Physics*, 87(2):S306, 2013.

Maximilian Niyazi, Michael Brada, Anthony J Chalmers, Stephanie E Combs, Sara C Erridge, Alba Fiorentino, Anca L Grosu, Frank J Lagerwaard, Giuseppe Minniti, René-Olivier Mirimanoff, et al. ESTRO-ACROP guideline "target delineation of glioblastomas". *Radiotherapy and Oncology*, 118(1):35–42, 2016.

Nitin Ohri, Xinglei Shen, Adam P Dicker, Laura A Doyle, Amy S Harrison, and Timothy N Showalter. Radiotherapy protocol deviations and clinical outcomes: A meta-analysis of cooperative group clinical trials. *Journal of the National Cancer Institute*, 105(6): 387–393, 2013.

Lester J Peters, Brian O'Sullivan, Jordi Giralt, Thomas J Fitzgerald, Andy Trotti, Jacques Bernier, Jean Bourhis, Kally Yuen, Richard Fisher, and Danny Rischin. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *Journal of Clinical Oncology*, 28(18):2996–3001, 2010.

Robert Poel, Elias Rüfenacht, Evelyn Hermann, Stefan Scheib, Peter Manser, Daniel M Aebersold, and Mauricio Reyes. The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. *Medical Image Analysis*, 73:102161, 2021.

Annika Reinke, Matthias Eisenmann, Minu Dietlinde Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. Common limitations of performance metrics in biomedical image analysis. In *Medical Imaging with Deep Learning*, 2021.

Robert Robinson, Vanya V Valindria, Wenjia Bai, Hideaki Suzuki, Paul M Matthews, Chris Page, Daniel Rueckert, and Ben Glocker. Automatic quality control of cardiac MRI segmentation in large-scale population imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 720–727. Springer, 2017.

Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V Valindria, Mihir M Sanghvi, Nay Aung, José M Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, et al. Real-time prediction of segmentation quality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–585. Springer, 2018.

Silvia Scoccianti, Beatrice Detti, Davide Gadda, Daniela Greto, Ilaria Furfaro, Fiammetta Meacci, Gabriele Simontacchi, Lucia Di Brina, Pierluigi Bonomo, Irene Giacomelli, et al. Organs at risk in the brain and their dose-constraints in adults and in children: A radiation oncologist's guide for delineation in everyday practice. *Radiotherapy and Oncology*, 114(2):230–238, 2015.

Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020.

J Van der Veen, S Willems, S Deschuymer, D Robben, W Crijns, F Maes, and S Nuyts. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology*, 138:68–74, 2019.

Guillaume Vogin, Liza Hettal, Clarisse Bartau, Juliette Thariat, Marie-Virginie Claeys, Guillaume Peyraga, Paul Retif, Ulrike Schick, Delphine Antoni, Zsuzsa Bodgal, et al. Cranial organs at risk delineation: heterogenous practices in radiotherapy planning. *Radiation Oncology*, 16(1):1–11, 2021.

Damien C Weber, Milan Tomsej, Christos Melidis, and Coen W Hurkmans. QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. *Radiotherapy and Oncology*, 105(1):4–8, 2012.

Chan Woo Wee, Wonmo Sung, Hyun-Cheol Kang, Kwan Ho Cho, Tae Jin Han, Bae-Kwon Jeong, Jae-Uk Jeong, Haeyoung Kim, In Ah Kim, Jin Hee Kim, et al. Evaluation of variability in target volume delineation for newly diagnosed glioblastoma: A multi-institutional study from the Korean Radiation Oncology Group. *Radiation Oncology*, 10 (1):1–9, 2016.

Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020.

Appendix A. Qualitative Analysis

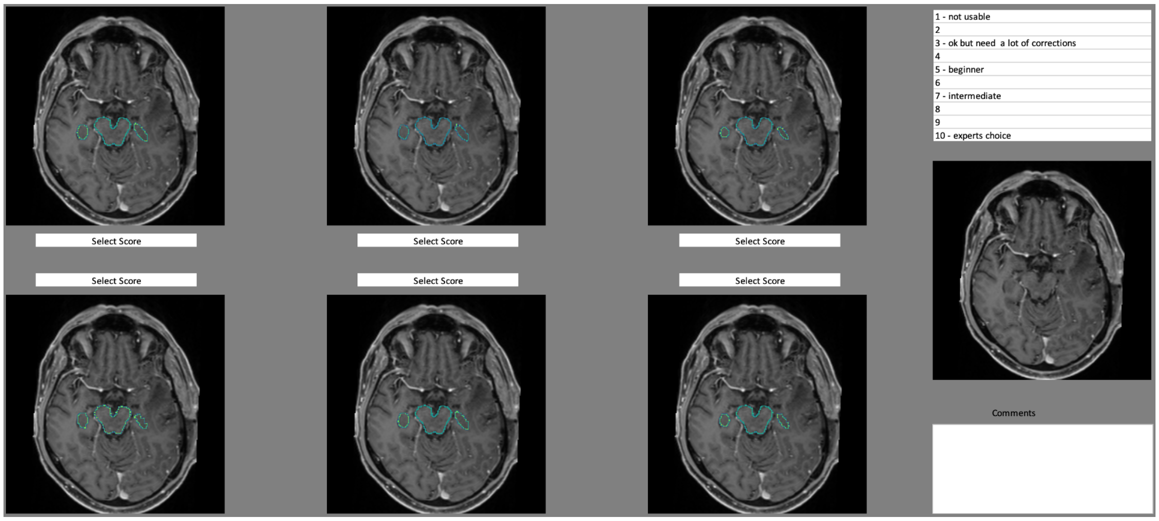

Figure 6: Questionnaire for reviewers to evaluate segmentations. For each case, six different alternatives were displayed and reviewers were asked to score them using Likert scale (1 - not usable; 10 - experts' choice). Human-expert reference was included in a random location for a blinded qualitative assessment.

As shown in Fig. 6, we show a grid of six slices of segmentations of both the brainstem and hippocampus for the reviewers to enter their score. The scores are between 1 and 10, and each of the six choices gets a score. This is repeated for 20 cases - once for GPSSI generated alternatives, and again for SSN. Statistics of the results are shown in Table. 1 and Table. 2.

Reviewer	Reference	Alt. 1	Alt. 2	Alt. 3	Alt. 4	Alt. 5
R1	**5.45 (0.61)**	4.50 (0.64)	4.30 (0.65)	4.45 (0.82)	4.70 (0.65)	4.25 (0.71)
R2	**4.25 (0.55)**	3.55 (0.52)	3.70 (0.59)	3.40 (0.41)	4.00 (0.45)	3.70 (0.43)
R3	**7.15 (0.89)**	6.25 (1.09)	6.50 (1.18)	6.50 (0.80)	6.50 (1.15)	6.80 (0.82)

Table 1: Mean (Standard deviation) of reviewer scores for reference and GPSSI generated alternatives.

For GPSSI, the mean scores are the highest for the reference (majority voting of expert-segmentations) as compared to any of the other five alternatives shown. This is shown in Table. 1. Additionally, the difference between the mean score for the reference segmentation and the next best is 0.75 (R1), 0.25 (R2) and 0.35 (R3) points.

In contrast, all three reviewers gave the highest mean scores to one of the SSN generated segmentations, indicating its potential to be used to simulate alternative segmentations.

Reviewer	Reference	Alt. 1	Alt. 2	Alt. 3	Alt. 4	Alt. 5
R1	5.70 (0.40)	5.50 (0.76)	5.55 (0.73)	5.85 (0.67)	5.55 (0.67)	**5.90 (0.78)**
R2	**6.65 (0.52)**	6.20 (0.96)	6.30 (0.93)	6.20 (0.92)	6.60 (0.77)	**6.65 (0.77)**
R3	7.90 (0.68)	7.80 (0.96)	8.30 (0.88)	**8.40 (0.84)**	7.60 (0.73)	8.30 (0.93)

Table 2: Mean (Standard deviation) of reviewer scores for reference and SSN generated alternatives.

This is shown in Table. 2. It is interesting also that reviewer R3, the most senior reviewer, had the highest mean rating, as well as the highest corresponding standard deviation for both of these methods.

Appendix B. Complexity of relationship between DSC and DVH

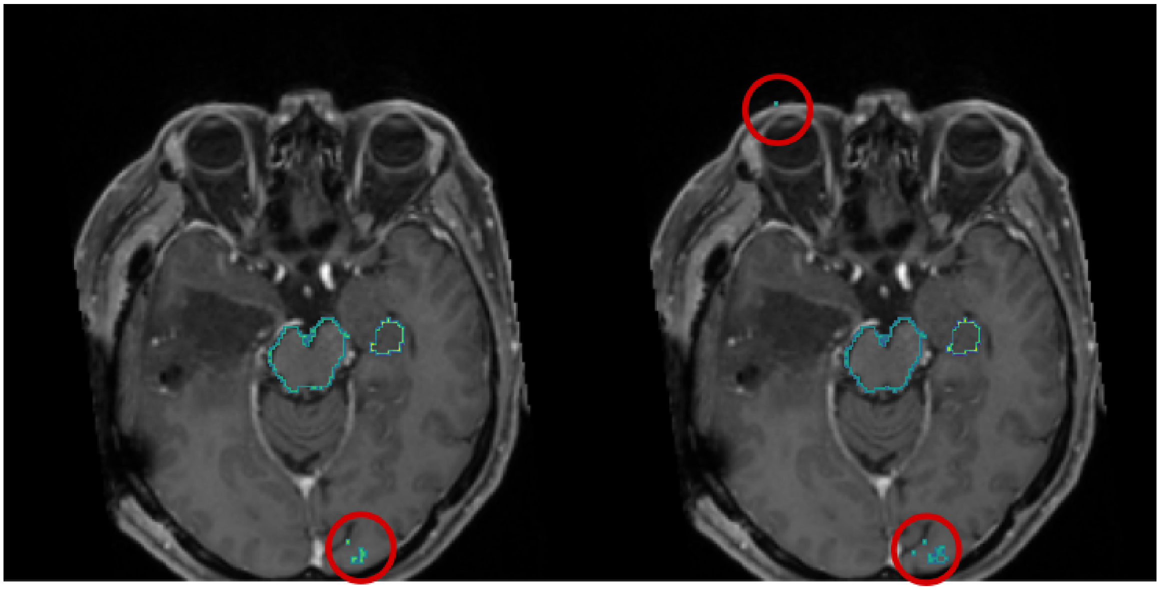

Figure 7: False positive outlier regions in the segmentation for case # 7 (alternatives 4 and 8) do not show up as large variations either in DSC or DVH.

Fig. 7 shows the two alternatives that erroneously generated segmentations near the eye and in the posterior area (highlighted inside the red circles). DSC is known to be sensitive to the relative size of the false positives with respect to the actual region of interest, whereas DVH appears to be sensitive to the relative location of these regions to the target volume. Novel metrics are needed to reliably identify such scenarios and other potentially unknown relationships, where it is unclear what the impact of these regions are.

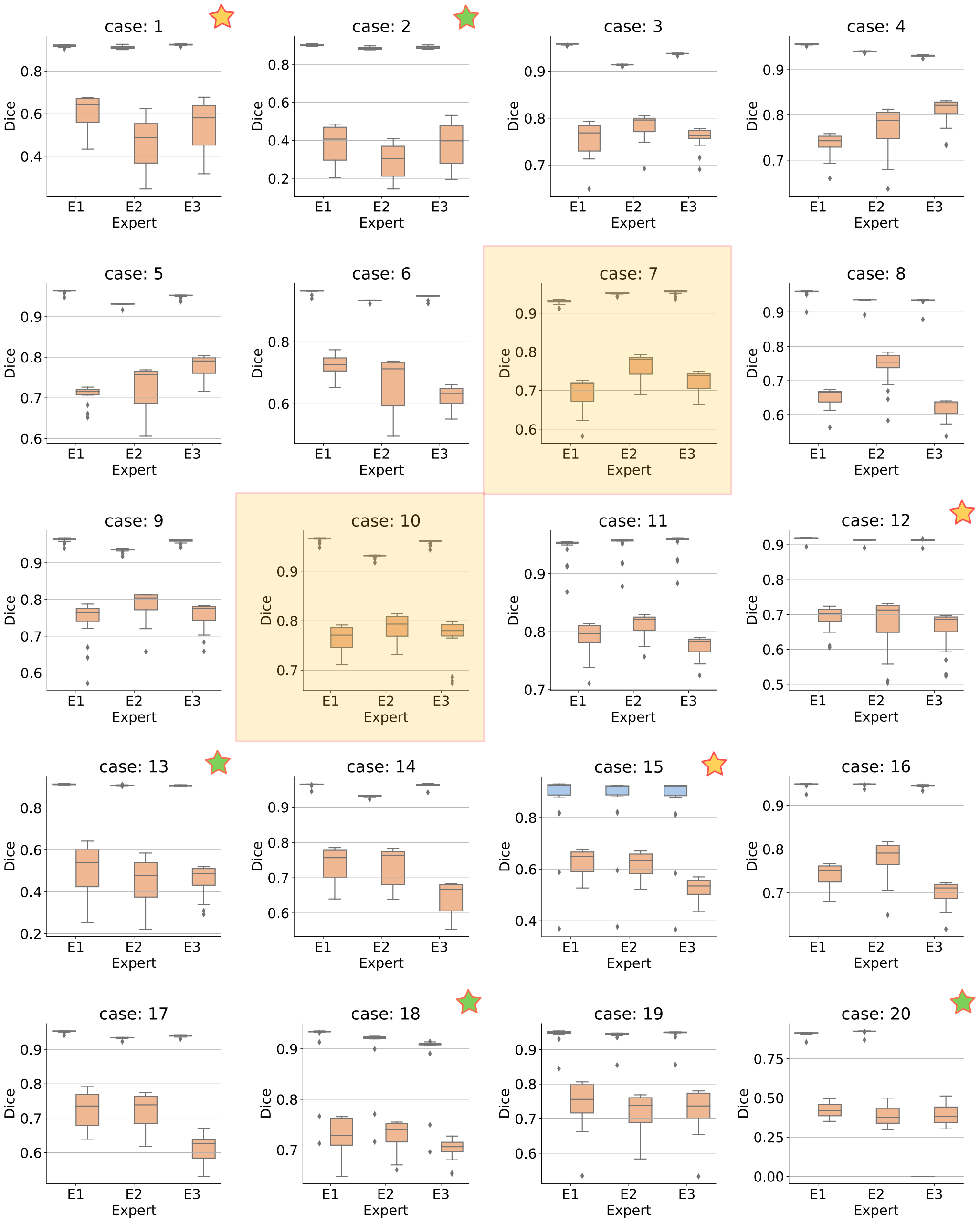

Figure 8: DSC for all 20 cases using SSN - measured against reference by E1, E2 and E3 (20 alternatives generated for each case). Brainstem shown in blue, hippocampus in orange. Case # 7 and # 10 are highlighted to compare with the corresponding DVH curves, presented in Fig. 5. Orange stars indicate cases from the test set, green indicates cases from the validation set, all others are from the training set.

Fig. 8 contains plots of the DSC for each of the 20 cases - generated using SSN. These include both in-training and out-of-training cases - for comparison. The brainstem is represented in blue and hippocampus in orange. The brainstem DSC is consistently higher than that of the hippocampus. Cases # 7 and # 10 are highlighted, as they correspond to the DVH curves presented in Fig. 5 - specifically for the hippocampus.

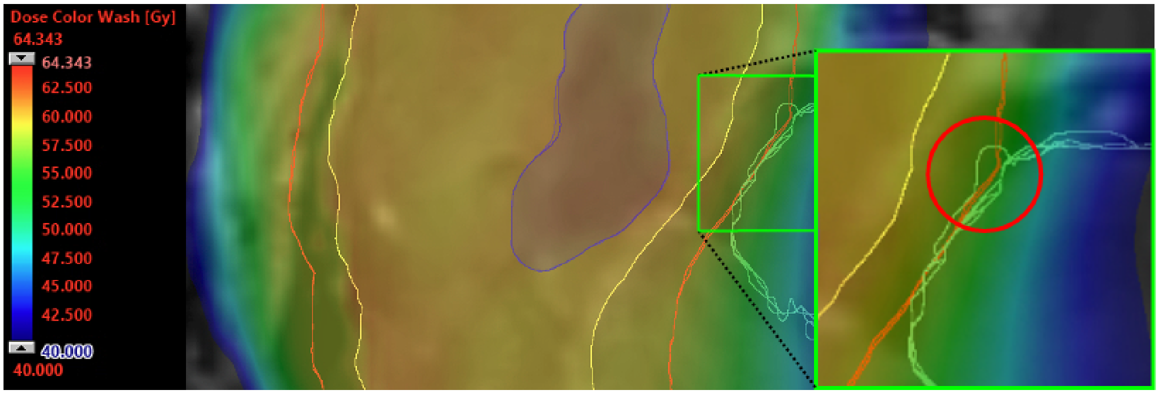

Figure 9: Visualization (using Eclipse) of dose impacts on OARs - target to the left of brainstem; color wash heatmap scale in Gray

Fig. 9 is a visualization of the dose distribution in a treatment plan (color wash). The heatmap indicates the complex relationship between the segmentations of target volume in the temporal lobe and the brainstem. Variations in the segmentation of OARs, such as the brainstem (multiple overlapping but different cyan lines) and the target volumes (PTV: red line, clinical target volume/CTV: yellow line, gross tumor volume/GTV: blue line) in the border of the high dose area, where the dose gradient is most steep (see scale on the left; red: high dose, blue: low dose), are most critical.

In the zoomed-in section, one of the alternative brainstem segmentations is depicted overlapping with the high-dose PTV - a potential case of over-contouring or segmentation error. Such a variation would result in (i) overestimation of the dose to the brainstem, since less area of the true brainstem lies within the high-dose area, and therefore (ii) potentially an under-dosing of the target volume, in order to spare the brainstem from excess dose during plan optimization. This could negatively impact tumor control.

On the other hand, under-contouring of the brainstem, i.e. missing certain areas, would result in a under-estimation of the dose to the brainstem. When the actual OAR dose is exceeding tolerance limits, excess toxicities might occur. Traditional metrics for volume comparison such as DSC are insensitive to the spatial location of segmentation variation in relation of the dose distribution, and might thus be insufficient to detect these clinically meaningful variations.