2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230559

How Sensitive Are Deep Learning Based Radiotherapy Dose Prediction Models To Variability In Organs At Risk Segmentation?

Amith Kamath[1] *Robert Poel*[1,2] *Jonas Willmann*[3,4] *Nicolaus Andratschke*[3] *Mauricio Reyes*[1]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern
[2] Department of Radiation Oncology, Inselspital, Bern University Hospital
[3] Department of Radiation Oncology, University Hospital Zurich, University of Zurich
[4] Center for Proton Therapy, Paul Scherrer Institute

ABSTRACT

Radiotherapy is a critical component of treatment for brain tumors. Inter-expert variability, differences in protocols, and human errors in segmentation of organ-at-risk (OAR) and target volume contours may necessitate re-planning treatment. This is time-consuming, significantly reduces the efficiency of radiation oncology teams, and hampers timely intervention to curb tumor growth. Hence, automated quality assurance of segmentation results hold much potential. However, such a quality assurance method must be fast and have good levels of sensitivity to radiation dose changes due to contour variations. In this paper, we evaluated a Cascaded 3D UNet deep neural network for dose prediction in brain tumors. Using metrics defined in the openKBP challenge, we report a promising mean dose score or mean absolute error (MAE) of 0.906 and a mean Dose Volume Histogram (DVH) score of 1.942, between predicted versus reference 3D dose volumes on 20 clinical test cases. We further tested the sensitivity of these dose predictions to realistic inter-expert variability in segmentation of the left optic nerve, chosen due to its clinical relevance. We found that the predicted DVH curves for such variations match well with the reference, average prediction dose MAE of 2.039 was close to average inter-expert dose MAE of 2.115, and the correlation coefficient between the predicted and reference dose differences was 0.926, indicating strong sensitivity to contour variations. These encouraging results show the potential of employing such models within a broader automated quality assurance system in the radiotherapy planning workflow. Code to reproduce this is available at `https://github.com/amithjkamath/deepdosesens`

***Index Terms*—** Radiotherapy, Treatment Planning, Deep Learning, U-Net, Automated Dose Prediction.

1. INTRODUCTION

Aggressive tumors like glioblastoma account for 45% of all malignant primary brain tumors [1]. Current treatment is a combination of surgery, adjuvant radiotherapy (RT), and concomitant and adjuvant chemotherapy [2]. The aim of RT planning is to conform dose to the target volume (i.e., tumor or resection cavity, with adjacent areas of potential microscopic spread) while sparing organs-at-risk (OAR). This limits normal tissue toxicity while ensuring optimal tumor control [3].

It is hence critical to have an accurate segmentation of the anatomy to achieve this objective. Radiation oncologists draw contours around OAR and target volumes, either manually or semi-automatically. This process however can take up to seven hours per patient [4]. In a multi-institutional delineation study among radiation oncologists, incorrect target volume segmentation has been reported to have caused 25% of non-compliant treatment plans [5]. Target volume and OAR segmentation are hence amongst the most time-consuming yet error-prone steps in the RT process. Efforts have hence been made to create segmentation standards and develop RT Quality Assurance (RTQA) systems [6].

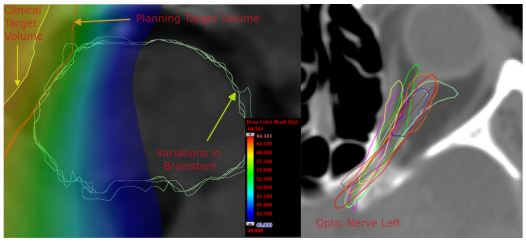

Fig. 1. Visualizing inter-expert variability in OAR contours of brainstem in cyan (left). Orange and yellow contours are around the tumor target volume. Overlaid heat map indicates dose. Various plausible left optic nerve segmentation (right) lead to changes in dose delivered.

Impact of segmentation variability: Fig. 1 (left) shows overlapping cyan lines representing potential choices of brainstem contours due to inter-expert variability. The planning target volume is represented in orange, and the clinical target volume in yellow. A heat map indicating RT dose distribution (color wash) for a treatment plan in Gray (red: high, blue: low dose) is overlaid for the dose context. Variations in these contours are most critical in the border of the

higher dose area, where the dose gradient is most steep, while less critical elsewhere. Fig. 1 (right) shows 10 examples of plausible optic nerve (left) contours. Over-contouring, where volumes are larger than the 'true' extent result in (i) overestimating the OAR dose since less area of the OAR lies within high-dose region, and (ii) potentially under-dosing the tumor target volume, to spare the OAR from excess dose. This negatively impacting tumor control. Conversely, under-contouring i.e. missing 'true' areas of the OAR, would result in a under-estimation of the actual dose. This leads to excess toxicity.

Treatment dose plan computation is currently done independently after the contouring step, by medical physics experts. If this dose is not protocol compliant, reviewing any potentially incorrect contours and re-computing adds on to delays. This time between image acquisition and RT planning completion is reportedly 9.63 days on average [7]. This has motivated the use of deep learning for online and accurate RT dose predictions [8]. When such models are reliable, it could prevent re-planning by evaluating contour quality prior to or simultaneously with planning. However, to the best of our knowledge, the sensitivity of these deep learning models to local contour variations has not previously been analyzed.

Hypothesis and Contributions: Our main hypothesis is that a deep learning dose prediction model that provides near-instant dosimetry is also sensitive to local contour changes, thereby being an efficient means of segmentation quality assurance. The main benefit of using a deep learning dose prediction model is that it is near-instant (inference time of 15 seconds on a GPU). This enables an interactive segmentation process guided by dose estimates where contours are edited immediately based on dose compliance, as opposed to relying on post-facto dose evaluations leading to delays. To make this feasible in clinical practice, reliable sensitivity of such models to local changes in contours is essential.

We test this sensitivity by constructing simulated expert variations in contours and evaluating the similarity of dose predictions from such models to a reference plan. Our contributions in this paper are therefore threefold:

- Based on a data set of 100 clinical cases, we show that a Cascaded 3D (C3D) UNet dose prediction model [9] achieves a mean dose score of 0.906 and mean DVH score of 1.942, indicating strong potential usage in treatment planning. To the best of our knowledge, this is the first such analysis on dose predictions for glioblastoma.
- Based on a per-OAR (a total of 13) analysis of dose and DVH scores, we find that model performance depends on both size (larger is worse) and proximity to tumor target volumes (closer is worse).
- We further analyze the sensitivity of dose predictions to small yet realistic contour changes of the left optic nerve, selected due to its clinical relevance and sensitivity to radiation. We show a strong correlation of 0.921 between predicted dose versus reference dose differences.

2. MATERIALS AND METHODS

Data: Our data set included imaging and contour data from 100 subjects who were diagnosed with glioblastoma. This included CT imaging data, along with associated binary segmentation masks of 13 OARs (see full list in Table. 1) as well as the Planning Target Volume. Each of these subjects also had a reference dose plan, calculated using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). This reference was a double arc coplanar volumetric modulated arc therapy (VMAT) plan with 6 mega volt flattening filter free beams, optimized (Varian photon optimizer version 15.6.05) to deliver 30 times 2 Gray while maximally sparing the OARs. The dose was calculated with the AAA algorithm [10], normalized so that 100% of the prescribed dose covers 50% of the target volume. Sixty randomly chosen subjects formed the training set, 15 were used as validation (five samples excluded due to missing contours) and the rest of the 20 were used as the test set.

Model: We used a two-level C3D U-Net [9] as the dose prediction network (i.e, the input to the second U-Net is the output of the first concatenated with the input to the first U-Net). The model input was a normalized CT volume and binary segmentation masks for each of the 13 OARs and target volume, and predicted a continuous valued dose volume (upscaled from [0, 1] to 0 to 70 Gray) of the same dimension as the input. The loss was computed as

$$Loss = 0.5 * L1(reference, A) + L1(reference, B) \quad (1)$$

where A and B were the outputs of the first and second U-Nets respectively, $reference$ was the reference dose and $L1$ refers to the L1 loss. All volumes were resampled to 128^3 voxels, due to GPU memory constraints. The hyperparameters for training the C3D model were unchanged from the original implementation [9], except the number of input binary masks was updated to 14, to match the number of OARs in our data set. The weights were randomly initialized using the 'He' method. Training ran for 80000 iterations and the model with the best validation dose score was saved. All experiments were run with PyTorch 1.12 on an NVIDIA RTX A5000 GPU, and each training run took 24 hours. We trained the model five times with the same hyperparameter set but different random seed initialization to ensure reliable convergence.

Metrics: We adopted the dose and DVH score as evaluation metrics, from openKBP [8], an international challenge designed for head and neck tumors. Dose scores indicate the mean absolute error (MAE) of predicted versus reference dose within a mask (either body, brain, or an OAR). DVH scores are the average of the MAE between prediction and reference for mean and 0.1CC dose of OARs, and the average of MAE for $1st$, $95th$ and $99th$ percentile of the dose for tumor target volume, also computed within their masks.

Sensitivity experiments: To analyze the sensitivity of the

trained model to inter-expert variability in segmentation, we manually modified the left optic nerve OAR for a single subject in the test set, to create ten variations (as shown in Fig. 1 (right)), validated by radiation oncologists for plausibility. For each of these, a new reference dose was computed using the same settings as used for the training data. The left optic nerve was chosen because of its proximity to the tumor target, resulting in large dose changes even for small contour changes. We then compared the predicted (P_i) and reference dose (R_i) qualitatively with DVH curves, and quantitatively by analyzing the difference in mean OAR dose between reference and predicted dose, for nine variations against a reference (index 0 without loss of generality).

3. RESULTS

Over five training runs, we report a mean dose score of 0.906 (std. 0.009), and a mean DVH score of 1.942 (std. 0.041) on 20 test subjects, which was in the same range as the winning entry [9] of the openKBP challenge. For subsequent analysis, we used the best performing model (out of five) with a dose score of 0.891 and DVH score of 1.919.

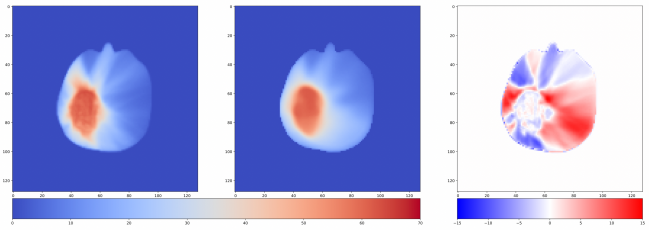

Fig. 2. Comparison of dose predictions: Reference (left), model prediction (middle) (range of these values are from 0 to 70 Gray), and differences (right) (range of these differences is -15 to 15 Gray). For the difference image (right), darker blue regions are underestimates, and darker red are overestimates.

Fig. 2 demonstrates the dose prediction in the axial plane. The model tracked the reference well and avoided higher dose in the eye region (the two blue streaks at the top in Fig. 2), while also effectively enveloping the shape of the tumor target. The difference between the two (right panel) consist mostly of radial streaks which were hardware specific artifacts that were not clinically pertinent to our assessment. Across the 20 test set subjects, the dose score varied between 0.470 and 2.167, where higher scores are typically due to larger tumor target volumes, because of higher overall dose to the entire anatomy. The DVH score varied between 0.451 and 4.203.

Table 1 shows the per-OAR results as mean (standard deviation) of dose and DVH scores. Larger OARs like the brainstem yielded better metrics, while smaller e.g., lacrimal glands are worse, leading to more over/under-estimates. Proximity to tumor target volume was also an important factor in the dose score, where closer OARs had higher scores. Dose differences within tumor target volumes were nonetheless always under 2.5 Gray, which was less than 5% of prescribed dose.

Table 1. Mean (std.) of dose and DVH scores for 13 OARs in 20 test dose predictions. Lower values are better.

OAR	Dose Score	DVH Score
Brainstem	1.399 (1.392)	2.025 (1.746)
Chiasm	2.985 (2.418)	2.798 (2.469)
Cochlea L	1.856 (4.728)	1.036 (2.347)
Cochlea R	2.433 (5.109)	1.406 (2.673)
Eye L	1.487 (2.194)	1.707 (2.517)
Eye R	2.210 (3.939)	2.836 (4.832)
Hippocampus L	2.101 (1.743)	1.976 (1.618)
Hippocampus R	2.601 (2.945)	2.381 (2.166)
Lacrimal Gland L	1.448 (1.320)	1.617 (1.404)
Lacrimal Gland R	1.938 (2.011)	1.912 (2.069)
Optic Nerve L	2.121 (2.464)	2.475 (3.122)
Optic Nerve R	2.266 (2.342)	2.072 (2.135)
Pituitary	1.889 (1.780)	1.932 (1.689)
Overall	0.891 (0.376)	1.919 (1.216)

Table 2. Sensitivity analysis: R_i is the reference mean dose and P_i is the predicted mean dose for index 'i', both for optic nerve left. DSC(i) is the Dice Similarity Coefficient between index 'i' and '0'. Dose difference (ΔD) reported in Gray.

Index (i)	ΔD: $\lvert R_i - R_0 \rvert \downarrow$	ΔD: $\lvert P_i - P_0 \rvert$	DSC(i)
1	0.145	0.418	0.325
2	0.283	0.222	0.627
3	0.357	1.032	0.783
4	0.435	0.519	0.363
5	2.089	3.171	0.590
6	2.402	2.487	0.509
7	3.027	1.483	0.197
8	4.815	5.436	0.612
9	7.591	5.625	0.229
Mean	2.115	2.039	0.523

Sensitivity analysis: Table. 2 shows the difference in the mean dose for the reference plans (second column), predicted plans (third column) and the corresponding Dice Similarity Coefficient (DSC) (fourth column) between a reference contour and nine variations of the left optic nerve. The contours are indexed in ascending order of the difference in mean reference dose (column two). The difference in the predicted dose tracked the difference in reference well, while the DSC trends were harder to use for making contour quality decisions. For example, the contour with index 8 could be considered reasonable when looking only at DSC. But, the predicted dose showed that it is not as good dosimetrically, as index 1 having a lower DSC.

The average difference in the mean reference dose (ΔD:

$|R_i - R_0|$) was 2.115, while the same for predicted dose (ΔD: $|P_i - P_0|$) was 2.039. The correlation coefficient between reference and predicted dose differences across these contour variations was a strong 0.926, while that with the DSC was -0.471. Furthermore, Fig. 3 shows the similarity in DVH curves to qualitatively compare reference and predicted doses for four representative variations. These evaluations confirm that the prediction model reliably tracked dose changes across contour variations. This demonstrates the utility of dose prediction models during contouring so that edits are based on clinically relevant dosimetry rather than the current practice of using just image-based anatomy and geometry.

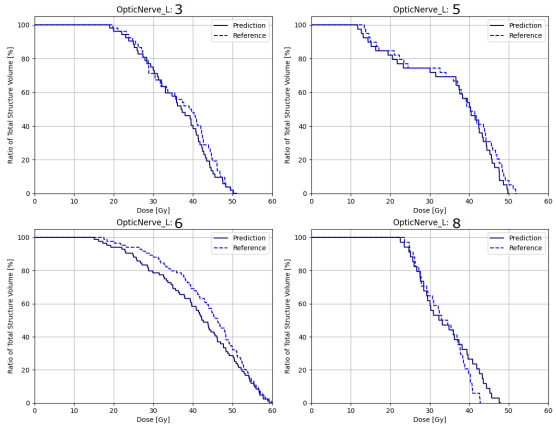

Fig. 3. Comparison of Dose Volume Histograms (DVH) for Optic Nerve Left - for four representative realistic contour variations (index matches those in Table. 2). A smaller gap between the two curves indicates better results.

Discussion: In this paper, we showed with experiments on a data set of 100 clinical glioblastoma cases that our dose prediction model has a mean dose error of less than 1 Gray on a test set of 20 clinical cases. This model was sufficiently sensitive to contour changes with a strong correlation while tracking dose changes, helping make better-informed contour editing decisions. However, a limitation is that separate models need to be trained for every tumor location, delivery machine, and planning software. We plan to build on this initial result to devise further experiments focusing on sensitivity.

Compliance with ethical standards: This study was conducted on retrospective human subject data from Inselspital (University Hospital Bern). Ethical approval was obtained from the regional ethics committee of the Canton of Bern.

Acknowledgements: We are funded by Swiss Cancer Research (KFS-5127-08-2020). We report no financial relationship or conflicts of interest.

4. REFERENCES

[1] J. R.McFaline-Figueroa and E. Q.Lee, "Brain tumors," *The American Journal of Medicine*, vol. 131, no. 8, pp. 874–882, 2018.

[2] R.Stupp, W. P.Mason, M. J.Van Den Bent, M.Weller, B.Fisher, M. J.Taphoorn, K.Belanger, A. A.Brandes, C.Marosi, U.Bogdahn, et al., "Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma," *New England Journal of Medicine*, vol. 352, no. 10, pp. 987–996, 2005.

[3] C.Scaringi, L.Agolli, and G.Minniti, "Technical advances in radiation therapy for brain tumors," *Anticancer research*, vol. 38, no. 11, pp. 6041–6045, 2018.

[4] I. J.Das, V.Moskvin, and P. A.Johnstone, "Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy," *Journal of the American College of Radiology*, vol. 6, no. 7, pp. 514–517, 2009.

[5] L. J.Peters, B.O'Sullivan, J.Giralt, T. J.Fitzgerald, A.Trotti, J.Bernier, J.Bourhis, K.Yuen, R.Fisher, and D.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*, vol. 28, no. 18, pp. 2996–3001, 2010.

[6] M.Niyazi, M.Brada, A. J.Chalmers, S. E.Combs, S. C.Erridge, A.Fiorentino, A. L.Grosu, F. J.Lagerwaard, G.Minniti, R.-O.Mirimanoff, et al., "ESTRO-ACROP guideline "target delineation of glioblastomas"," *Radiotherapy and Oncology*, vol. 118, no. 1, pp. 35–42, 2016.

[7] C.Guo, P.Huang, Y.Li, and J.Dai, "Accurate method for evaluating the duration of the entire radiotherapy process," *Journal of Applied Clinical Medical Physics*, vol. 21, no. 9, pp. 252–258, 2020.

[8] A.Babier, B.Zhang, R.Mahmood, K. L.Moore, T. G.Purdie, A. L.McNiven, and T. C.Chan, "Openkbp: The open-access knowledge-based planning grand challenge and dataset," *Medical Physics*, vol. 48, no. 9, pp. 5549–5561, 2021.

[9] S.Liu, J.Zhang, T.Li, H.Yan, and J.Liu, "Technical note: A cascade 3d u-net for dose prediction in radiotherapy," *Medical Physics, https://doi.org/10.1002/mp.15034*, 2021.

[10] A.Van Esch, L.Tillikainen, J.Pyykkonen, M.Tenhunen, H.Helminen, S.Siljamäki, J.Alakuijala, M.Paiusco, M.Iori, and D. P.Huyskens, "Testing of the analytical anisotropic algorithm for photon dose calculation," *Medical physics*, vol. 33, no. 11, pp. 4130–4148, 2006.