Advances in Radiation Oncology (2021) **6**, 100658

advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy

Ward van Rooij, MSc,* Wilko F. Verbakel, PhD, PDEng, Berend J. Slotman, MD, PhD, and Max Dahele, MBChB, PhD

Department of Radiation Oncology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, Amsterdam, The Netherlands

Received 27 August 2020; revised 14 December 2020; accepted 30 December 2020

Abstract

Purpose: Contouring organs at risk remains a largely manual task, which is time consuming and prone to variation. Deep learning-based delineation (DLD) shows promise both in terms of quality and speed, but it does not yet perform perfectly. Because of that, manual checking of DLD is still recommended. There are currently no commercial tools to focus attention on the areas of greatest uncertainty within a DLD contour. Therefore, we explore the use of spatial probability maps (SPMs) to help efficiency and reproducibility of DLD checking and correction, using the salivary glands as the paradigm.

Methods and Materials: A 3-dimensional fully convolutional network was trained with 315/264 parotid/submandibular glands. Subsequently, SPMs were created using Monte Carlo dropout (MCD). The method was boosted by placing a Gaussian distribution (GD) over the model's parameters during sampling (MCD + GD). MCD and MCD + GD were quantitatively compared and the SPMs were visually inspected.

Results: The addition of the GD appears to increase the method's ability to detect uncertainty. In general, this technique demonstrated uncertainty in areas that (1) have lower contrast, (2) are less consistently contoured by clinicians, and (3) deviate from the anatomic norm.

Conclusions: We believe the integration of uncertainty information into contours made using DLD is an important step in highlighting where a contour may be less reliable. We have shown how SPMs are one way to achieve this and how they may be integrated into the online adaptive radiation therapy workflow.

Sources of support: This research was supported by a grant from Varian Medical Systems, Palo Alto, CA.

Disclosures: The department of radiation oncology has a research collaboration with Varian Medical Systems, Palo Alto, CA, and Drs Slotman and Verbakel have received honoraria/travel support from Varian Medical Systems. Ward van Rooij has not received personal fees from Varian Medical Systems. Not during the conduct of the study nor outside the submitted work. Dr Verbakel reports grants from Varian Medical Systems during the conduct of the study and grants and personal fees from Varian Medical Systems outside the submitted work. Dr Dahele reports grants from Varian Medical Systems during the conduct of the study. Dr Slotman reports grants from Varian Medical Systems during the conduct of the study and grants and personal fees from ViewRay, Inc, outside the submitted work.

Medical imaging data used in this research are not available in accordance with privacy regulations under EU law (GDPR).

* Corresponding author: Ward van Rooij, MSc; E-mail: w.vanrooij@amsterdamumc.nl

https://doi.org/10.1016/j.adro.2021.100658

Introduction

Online adaptive radiation therapy (OART) accounts for anatomic changes over the course of treatment by reoptimizing the dose distribution according to the anatomy at that moment, improving the balance between target coverage and organ-at-risk (OAR) doses.[1,2] One prerequisite for the implementation of OART is a short time between imaging and delivery of the adapted treatment plan. In recent years, software improvements and increased computing power have enabled fast inverse-optimized intensity modulated treatment planning.[3]

Contouring OARs, however, remains a largely manual task, which is time consuming and prone to variation.[4-6] Automated deep learning-based delineation (DLD) shows promise both in terms of quality and speed.[7] Although DLD performs well, the average Sørensen-Dice similarity coefficient (SDC, described in the following sections), for e.g. a parotid or submandibular gland model, is rarely higher than 90%,[8] and outliers can drop as low as 40%.[7] Therefore, manual checking of DLD is still recommended.

Although DLD, even for multiple organs, can be as fast as a few seconds per patient, manually checking the generated structures is time consuming and can largely negate the potential time saved by DLD.[9] The need for manual checking due to less-than-optimal DLD is a barrier to its wider implementation. Speeding up manual checking therefore becomes relevant. One way to do this could be by highlighting parts of the DLD-generated structure that have a larger chance of being wrong. This may be done by showing the uncertainty in a DLD structure.

In the case of medical image segmentation, uncertainty in a DLD structure translates to the probability that a certain voxel is part of that structure. It can be split into two parts: uncertainty inherent to the model and uncertainty inherent to the data[10] and may stem from, for example, inconsistent clinical training data or imaging data outside the range of the model. We are currently not aware of any commercially available radiation therapy tools that focus attention on the areas of greatest uncertainty in a DLD structure.

We explored the potential of spatial probability maps (SPMs) to increase the efficiency and reproducibility of DLD checking and correction, using the salivary glands as the paradigm. In so doing, our primary concern was not to develop or compare uncertainty quantification methods, but to take an established technique from the realms of research and explore how it can be applied to an area of current clinical need: automated OART.

Methods and Materials

Data

The potential of SPMs was retrospectively investigated for the left parotid and submandibular gland (PG/SMG) using 3-dimensional (3D) computed tomography (CT) based contours from head and neck cancer treatments. Whenever the right PG/SMG was available, it was flipped and added to the data set (assuming symmetry[11]), resulting in 315/264 PGs/SMGs. Inclusion of air/bone in the contour was corrected for by removing all voxels with a corresponding Hounsfield unit value of less than −300 or greater than 200 in the CT data. No additional curation was performed. The train set comprised 5/6 of the data, the test set 1/6 of the data, and the validation set 1/10 of the train set (all randomly selected). Cross-validation was not applied because verifying the geometric accuracy of the model was not the purpose of this study.

The preprocessing of the CT data consisted of cropping a region of interest ($64 \times 64 \times 32/96 \times 64 \times 64$ voxels for SMG/PG) centered on the OAR to limit memory usage, applying a Hounsfield unit window to remove extreme values and increase contrast, normalizing the data to [0,1], and subtracting the mean to center the data around 0 and in so doing facilitate training.

Model

The model was a fully convolutional network[12] based on the 3D U-net[13] with dropout[14] applied to all convolutional layers. Dropout turns off a random selection of parameters (in this case 255,289 in total) during each training instance. The number of parameters that are turned off depends on the dropout rate, which was 0.1 (~25,529 parameters turned off). Dropout is used to prevent overfitting. When a model has overfitted, it has been optimized on the train data too much because of which it does not generalize well to unseen data. The SDC was used as cost function. SDC is defined as

$$SDC = \frac{2tp}{2tp + fp + fn}$$

where *tp, fp,* and *fn* correspond to true positive, false positive, and false negative voxels, respectively. Early stopping was also applied to prevent overfitting; training was stopped when improvement for the validation set was <0.001 for at least 5 epochs. Adam[15] was the optimizer ($\beta_1 = 0.928$, $\beta_2 = 0.999$), allowing each parameter to have its own learning rate that can be adapted during training. Hyperparameter values were chosen based on prior non-exhaustive hyperparameter tuning.[7] The model

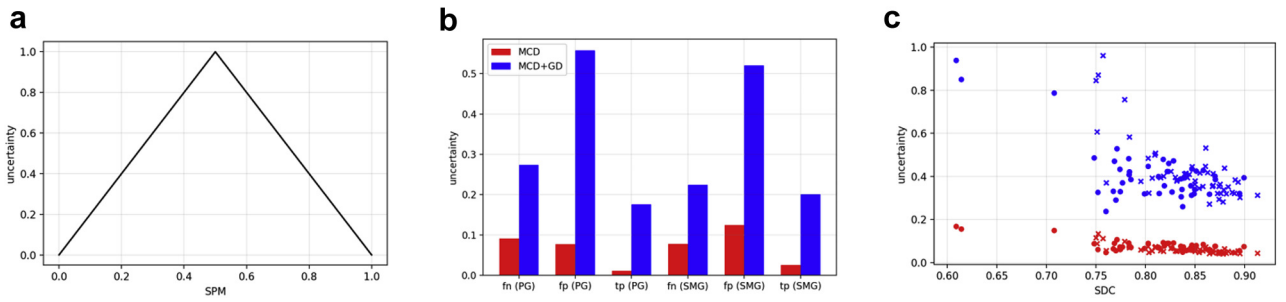

Figure 1 (a) Graph showing how the voxel-level uncertainty score (y-axis) was derived from the spatial probability map (SPM) voxel value (x-axis). (b) Graph showing the average uncertainty scores (y-axis) for Monte Carlo dropout (MCD) and MCD + Gaussian distribution (GD) for false negative (fn), false positive (fp), and true positive (tp) voxels of parotid and submandibular gland (PG/SMG). (c) Graph showing the relation between Sørensen-Dice similarity coefficient (SDC) (x-axis) and structure-level uncertainty score (y-axis) for MCD (red) and MCD + GD (blue); PGs are crosses, SMGs are dots. (A color version of this figure is available at https://doi.org/10.1016/j.adro.2021.100658.)

was built with Keras (https://keras.io/) on top of TensorFlow (https://www.tensorflow.org/) and trained with 1 GeForce RTX 2080ti.

Spatial probability maps

SPMs were created using Monte Carlo dropout (MCD).[16] As explained earlier, dropout is used during training to prevent overfitting. However, it can also be used during testing to approximate a model's uncertainty. Because dropout turns off a random selection of the model's parameters, each pass-through is different. Prior experiments we ran showed the SPMs resulting from MCD were not very expressive. Therefore, MCD was boosted by sampling from a Gaussian distribution (GD; μ is equal to parameter value, σ = .015/.01 for PG/SMG) for each model parameter (n = 255289), from now on referred to as MCD + GD. Consequently, for each pass-through, parts of the model's parameters were turned off, whereas the remainder were slightly changed. In other words, the model's dimensional space and its point of convergence therein were varied. The σ values were maximized with 1 constraint: the SDC resulting from a majority vote among the generated models was not supposed to be lower than that of the base model. This method provided differing predictions for each forward propagation (n = 101). 101101 pass-throughs gave a well-calibrated uncertainty quantification and, having an odd amount, allowed for a majority vote. Each prediction was a 3D binary object, where 0 indicated the voxel was not part of the gland and 1 indicated the voxel was part of the gland. The predictions were then averaged to acquire the SPMs, where each voxel's value indicated the probability that that voxel was part of the gland.

Analysis

The SPMs were first analyzed in a quantitative manner. To do so, a voxel-level uncertainty score from 0 to 1 was defined (Fig 1a). The idea behind this score was as follows: when the value of a voxel in the SPM is 0.5, an uncertainty score of 1 is assigned because there is a lot of uncertainty about whether that voxel should be 0 or 1. Accordingly, when the value of a voxel in the SPM is either 0 or 1, the uncertainty score is 0 because there is very little uncertainty about whether that voxel should be 0 or 1.

The average uncertainty scores for the false negative, false positive, and true positive voxels were compared for MCD and MCD + GD. Like the model's predictions, the clinical structure was a 3D binary object, where 0 indicated the voxel was not part of the gland and 1 indicated the voxel was part of the gland. To retrieve an uncertainty score for the entire structure, the sum of the uncertainty score of all the voxels was divided by the sum of the object depicting the clinical structure because uncertainties tend to lie around the surface of structures, and larger structures have more surface area. The relation between SDC and the structure-level uncertainty score was then compared for MCD and MCD + GD. After the quantitative analysis, the SPMs resulting from MCD + GD were visually inspected to see whether they would be a valuable addition to the DLD checking and correction process.

Results

Generating the SPM for 1 gland took <2.5 seconds, which can be further optimized. Figure 1b,c shows the quantitative results of using MCD versus MCD + GD. The addition of a GD over the model's weights appears to increase the ability to detect uncertainties.

The difference in SPMs resulting from MCD and MCD + GD is illustrated by 2 examples in Figure 2. The dashed contours indicate that a certain percentage of the generated models agree that all voxels within that contour are part of the gland, so they are not generated by an individual model but illustrate a probability of the

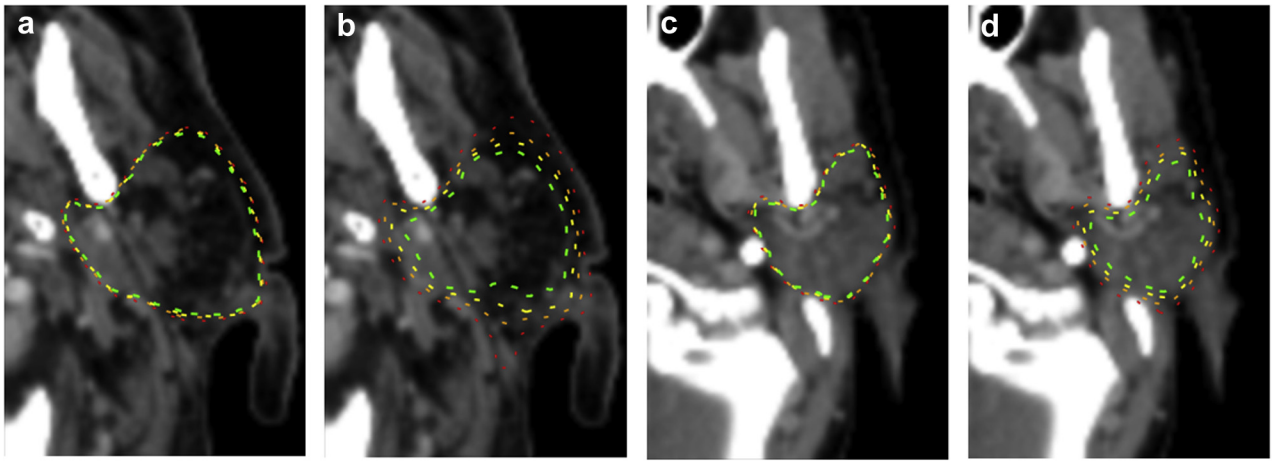

Figure 2 Difference in spatial probability maps (SPMs) for 2 slices (a,b and c,d) between Monte Carlo dropout (MCD) (a,c) and MCD + Gaussian distribution (GD) (b,d). Red contour means >10% of the generated models agree that the voxels within that contour are part of the structure. Orange >35%, yellow >60%, green >85%. (A color version of this figure is available at https://doi.org/10.1016/j.adro.2021.100658.)

underlying model. For visual inspection, we define the amount of spatial uncertainty for a particular area as the distance between the 4 lines, that is, there is more uncertainty when the dashed contours are farther apart.

Examples of PG/SMGs and their corresponding SPMs generated by MCD + GD can be seen in Figures 3 and 4. There is more uncertainty in areas that have a low amount of image contrast, as illustrated by the medial versus lateral parts in Figure 3a-c and Figure 4a versus 4b. Consistent with the low contrast in Figure 4c, there is considerable uncertainty. In a different slice from the same patient (Fig 4d), the gland is more clearly visible, and there is less uncertainty. In Figure 4e, there is more uncertainty compared with a different slice from the same patient (Fig 4f) where the structure is more visible. SPMs can draw attention to parts of an OAR that might be missed. An example of this is seen in Figure 3d where attention is drawn to the anterior PG.

We observed more uncertainty in areas that are less consistently contoured by clinicians. For the PG, these are the anterior and medial part of the gland (Fig 3e,f) and the cranial-most slices (Fig 3g,h). In Figure 4g, there is uncertainty surrounding the blood vessel. This may be due to the training data containing cases where the blood vessel is incorporated in the structure and cases where it is not.

Uncertain areas can be the result of unusual features in the data. An example is Figure 4h, where the SMG lies adjacent to a pathologic lymph node, degrading the model performance.

Discussion

We have analyzed the use of SPMs to highlight uncertainty in DLD contours. In general, the technique that was used demonstrated more uncertainty in areas that (1) have lower contrast, (2) are less consistently contoured by clinicians, and (3) deviate from the anatomic norm. This implies that the model is more sensitive to changes (dropout/GD) when it has to process data that contain more uncertainty (eg, low contrast). For obvious reasons, we could only show a selection of slices. Although the examples in Figure 1 and 2 were typical, there were also slices for which the observations we described did not hold. For instance, within Figure 4d, there is comparable uncertainty for both high- (anterior, lateral, posterior) and low-contrast (medial) areas within the image.

When comparing MCD to MCD + GD in a quantitative manner, the addition of the GD over the model's parameters appeared to increase the ability to detect uncertainty; in false positive, false negative, and true positive voxels, the MCD + GD showed more uncertainty. Ideally, these methods would only show uncertainty in the false voxels and not in the true voxels. The fact that they do not does not imply that the uncertainty method is failing, but may be due to a less-than-perfect model, which is to be expected with the limited amount of data that are available for training and the inherent error those data contain.[4-6] In fact, the sigma values were optimized to show the most uncertainty while not degrading the performance of the underlying model. If the sigma is lower, the method will show less uncertainty, and we will receive less information on where the contour may need to be checked. If the sigma is higher, the method will show an even higher amount of variance in the SPM, but will do so at the cost of degrading the underlying model. Future work could compare the SPM-derived uncertainty with the presence and magnitude of clinical edits by multiple observers to see if they concur.

There are several ways in which SPMs could be applied in a clinical setting. When all the OARs for a

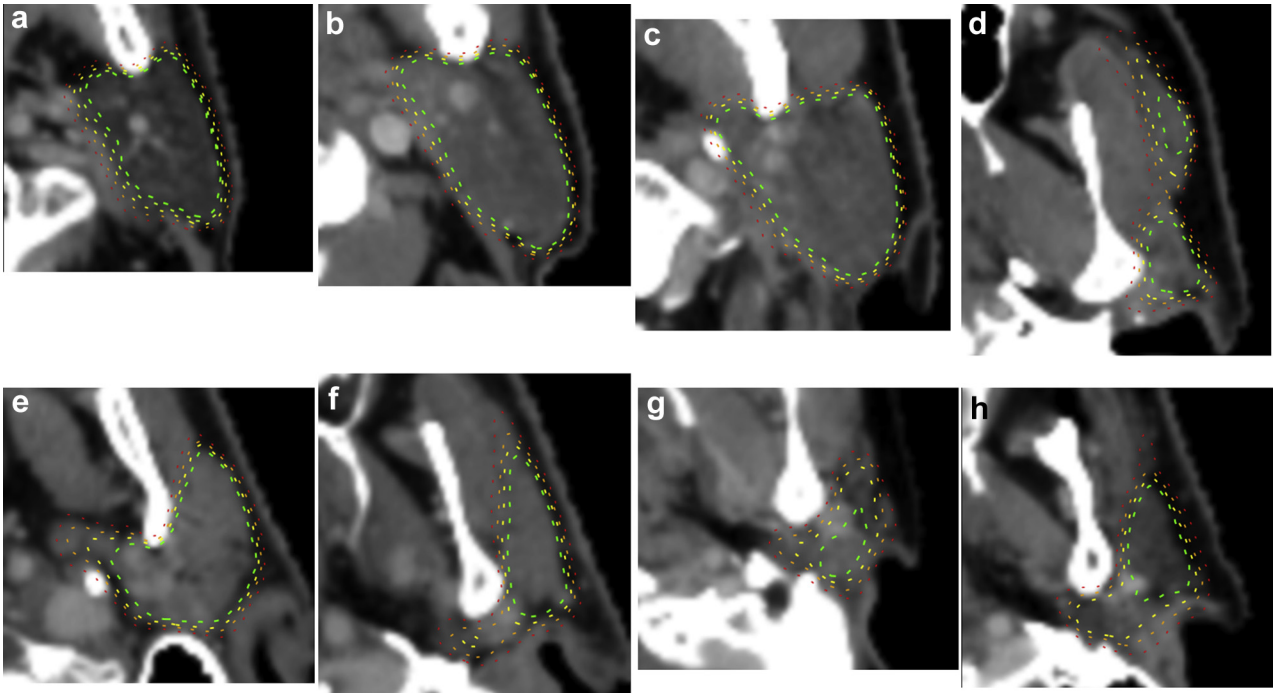

Figure 3 Illustrative examples of parotid gland slices with the spatial probability maps (SPMs) generated by Monte Carlo dropout (MCD) + Gaussian distribution (GD) overlaid. Red contour means >10% of the generated models agree that the voxels within that contour are part of the structure. Orange >35%, yellow >60%, green >85%. (A color version of this figure is available at https://doi.org/10.1016/j.adro.2021.100658.)

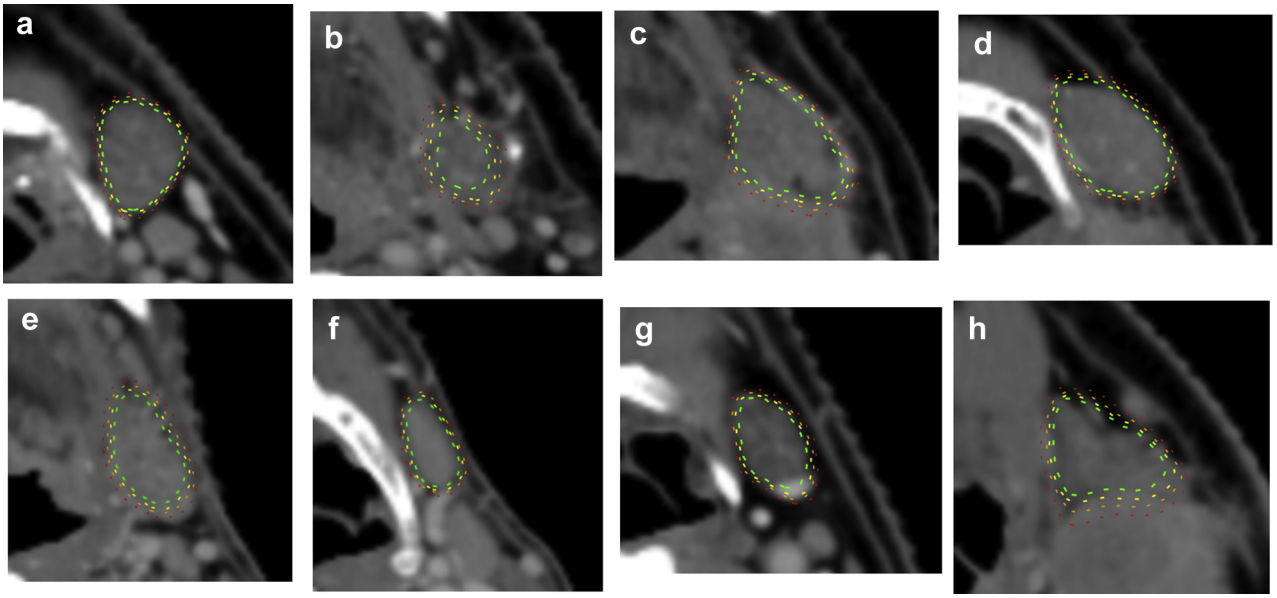

Figure 4 Illustrative examples of submandibular gland (SMG) slices with the spatial probability maps (SPMs) generated by Monte Carlo dropout (MCD) + Gaussian distribution (GD) overlaid. Red contour means >10% of the generated models agree that the voxels within that contour are part of the structure. Orange >35%, yellow >60%, green >85%. (A color version of this figure is available at https://doi.org/10.1016/j.adro.2021.100658.)

specific treatment are contoured by DLD, the clinician would be able to see the SPM of each OAR. One option is that the clinician is immediately presented with the SPMs of all OARs. However, in cases where there are a lot of OARs, like head and neck cancer, checking all SPMs may be too time-consuming. Therefore, other options include that (1) the clinician selects the OARs for which he/she would like to see the SPM, for instance based on

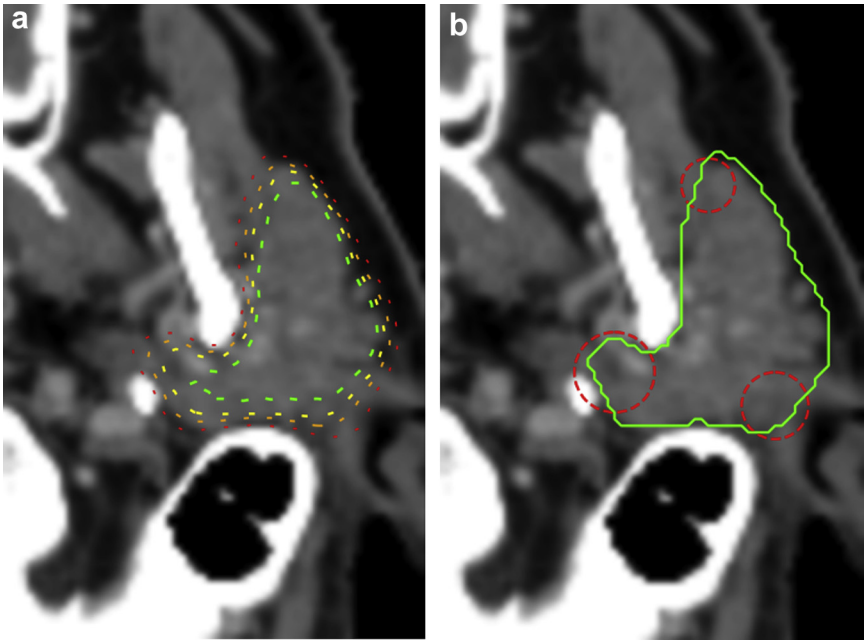

Figure 5 Example of how spatial probability maps (SPMs) could be used in a simplified manner in clinical practice. (a) An image with the SPM overlaid. (b) The same image with the contour generated by the model without distribution over the parameters (green, Sørensen-Dice similarity coefficient [SDC] = 0.86) and red dotted circles indicating uncertain areas based on the SPM in (a). (A color version of this figure is available at https://doi.org/10.1016/j.adro.2021.100658.)

proximity to the planning target volume or other a priori knowledge; (2) only particular areas are highlighted based on the distance between uncertainty lines of the SPMs (Fig 5); (3) a separate model is used to predict the performance of the DLD model[17] and flag OARs that are likely to have been poorly contoured, showing an SPM only for those OARs.[4] A standardized score indicates the amount of uncertainty in the contour, and SPMs are shown depending on whether or not the score exceeds a particular threshold. One could also think of functionalities to enable the fast-paced workflow of OART, like being able to select one of the uncertainty lines as the contour. Alternatively, SPMs could be exploited when a model is being trained by giving more weight to uncertain areas when updating the model's parameters, similar to active learning principles.[18]

In the case of OART, this could result in the following workflow (Fig 6): imaging data are acquired that are passed through a DLD model, resulting in a contour for each OAR. Next, some method is used to generate the corresponding SPMs. Based on the SPMs and other variables (e.g., image characteristics like amount of contrast) a performance estimator can be used, together with prior knowledge, to select those structures for which the SPM should be presented. Subsequently, the selected contours can be adjusted and the entire array of OAR contours can be used as input for an automated treatment planning system.

In this analysis, we only explored two (related) methods of generating uncertainty information, with the purpose of demonstrating the use of SPMs in a clinical setting. A considerable body of research has looked into ways of quantifying DLD uncertainty, using various methods, the most prevalent of which by far is MCD.[16,19-22] Another method is to train an ensemble of multiple models and average their predictions.[23,24] Both MCD and ensembles only tend to capture the uncertainty that is inherent to the model; uncertainty that can be explained away by having an infinite amount of data samples.[10] To capture the uncertainty that is inherent to the data, other methods have been investigated, like using a heteroscedastic noise model[10] or performing data augmentation during testing.[25] These methods may be useful for certain radiation therapy purposes, where there is known to be a lot of variance in contouring.[4-6] When multiple classes need to be segmented in a single image, these methods are not suitable to capture the relations between voxels of the same class. To tackle that problem, more advanced models have been designed.[26-28] Because our model only has to output a single class, we did not need such complex models. Furthermore, our aim was specifically to demonstrate how a relatively simple technique can be of use in a clinical setting. Future work should focus on systematically comparing various methods to quantify spatial uncertainty. Such systematic comparisons should tackle both data uncertainty and model uncertainty and should include validated evaluation metrics and identical data across methods.

In summary, we believe the integration of uncertainty information into contours made using DLD is an

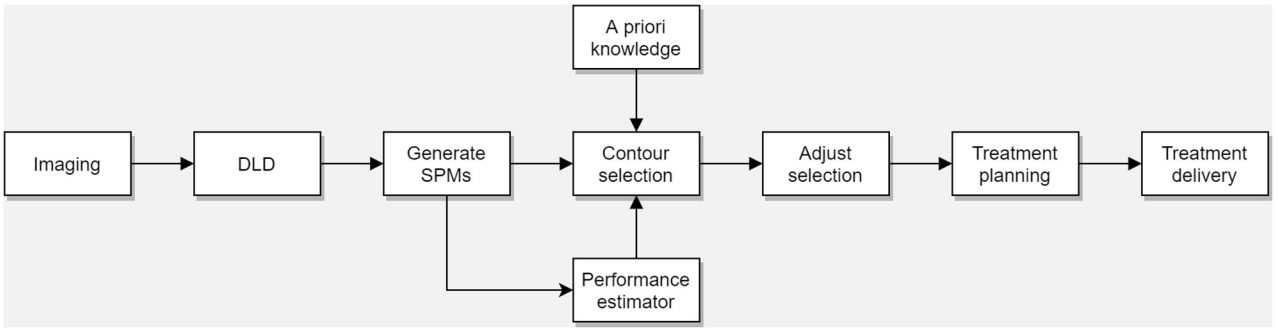

Figure 6 Diagram depicting the online adaptive radiation therapy (OART) workflow with the incorporation of spatial probability maps (SPMs).

important step in highlighting where a contour may be less reliable. We have shown how SPMs are one way to achieve this and how they may be integrated into the OART workflow.

References

1. Lim-Reinders S, Keller BM, Al-Ward S, et al. Online adaptive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2017;99:994-1003.
2. Sonke J, Aznar M, Rasch C. Adaptive radiotherapy for anatomical changes. *Semin Radiat Oncol*. 2019;29:245-257.
3. Tol J, Delaney AR, Dahele M, et al. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys*. 2015;91:612-620.
4. Brouwer CL, Steenbakkers RJ, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol*. 2012;7:32.
5. Nelms BE, Tomé WA, Robinson G, et al. Variations in the contouring of organs-at-risk: Test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys*. 2012;82:368-378.
6. Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117:83-90.
7. van Rooij W, Dahele M, Ribeiro Brandao H, et al. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *Int J Radiat Oncol Biol Phys*. 2019;104:677-684.
8. Nikolov S, Blackwell S, Mendes R, et al. *Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy*. 2018;arXiv:1809.04430v1.
9. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*. 2018;126:312-317.
10. Kendall A, Gal Y. *What uncertainties do we need in Bayesian deep learning for computer vision*. 2017;arXiv:1703.04977.
11. Stimec B, Nikolic S, Rakocevic Z, et al. Symmetry of the submandibular glands in humans: A postmortem study assessing the linear morphometric parameters. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 2006;102:391-394.
12. Long J, Shelhamer E, Darrell T. *Fully convolutional networks for semantic segmentation*. 2014;arXiv:1411.4038.
13. Cicek O, Abdulkadir A, Lienkamp SS, et al. *3D U-Net: Learning dense volumetric segmentation from sparse annotation*. 2016;arXiv:1606.06650v1.
14. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
15. Kingma DP, Ba J. *Adam: A method for stochastic optimization*. 2015;arXiv:1412.6980v9.
16. Gal Y, Ghahramani Z. *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*. 2015;arXiv:1506.02142.
17. DeVries T, Taylor GW. *Leveraging uncertainty estimates for predicting segmentation quality*. 2018;arXiv:1807.00502.
18. Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. *JAIR*. 1996;4:129-145.
19. Roy AG, Conjeti S, Navab N, et al. *Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling*. 2018;arXiv:1804.07046.
20. Nair T, Precup D, Arnold DL, et al. *Exploring uncertainty in deep networks for multiple sclerosis lesion detection and segmentation*. 2018;arXiv:1808.01200.
21. Orlando JI, Seeböck P, Bogunovic H, et al. *U2-Net: A Bayesian U-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans*. 2019;arXiv:1901.07929.
22. Hiasa Y, Otake Y, Takao M, et al. *Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling*. 2019;arXiv:1907.08915.
23. Lakshminarayanan B, Pritzel A, Blundell C. *Simple and scalable predictive uncertainty estimation using deep ensembles*. 2017;arXiv:1612.01474.
24. Karimi D, Zeng Q, Mathur P, et al. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Med Image Anal*. 2019;57:186-196.
25. Wang G, Li W, Aertsen M, et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. 2019;338:34-45.
26. Kohl S, Romera-Paredes B, Meyer C, et al. *A probabilistic U-net for segmentation of ambiguous images*. 2018;arXiv:1806.05034.
27. Kohl S, Romera-Paredes B, Maier-Hein KH, et al. *A hierarchical probabilistic U-net for modeling multiscale ambiguities*. 2019;arXiv:1905.13077.
28. Baumgartner CF, Tezcan KC, Chaitanya K, et al. *PHiSeg: Capturing uncertainty in medical image segmentation*. 2019;arXiv:1906.04045.