Automated quality assurance of deep learning contours in head-and-neck radiotherapy

Mody, P.P.

6

Summary, discussion and future work

6.1 Thesis Summary

This thesis addresses the critical need for efficient and reliable quality assurance (QA) tools for automated organ and tumor contouring in radiotherapy. While deep learning models offer significant acceleration in contouring, the subsequent manual QA and refinement steps can be time-consuming and offset part of these gains, creating a bottleneck in clinical workflows. Two core themes of QA are explored in this thesis: *error detection* (identifying where contours are likely incorrect) and *error correction* (efficiently refining those errors) in either pre- or post-commissioning phases.

Specifically, this thesis explores: a) the development of an automated and scalable workflow for evaluating the pre-commissioning dosimetric impact of auto-contours (Chapter 2), b) the potential of Bayesian models and training losses to detect inaccurate predictions in the post-commissioning phase by leveraging their associated uncertainty (Chapter 3 & Chapter 4), and c) the improvement of error correction efficiency through AI-assisted refinement tools (Chapter 5). Thus, the overarching goal of this thesis is to explore different QA methodologies both pre- and post-commissioning of auto-contouring tools for head-and-neck radiotherapy.

6.2 Chapter Recapitulations

6.2.1 Chapter 2

This chapter addressed the need of large-scale pre-commissioning dosimetric evaluations of auto-contoured organs-at-risk (OARs). The main contribution was the development and assessment of an automated plan optimization workflow. This workflow was designed to emulate the clinic's treatment planning protocol by reusing existing clinical plan parameters (e.g., beam settings, objective weights). This approach, termed robot process automation (RPA), converts the complex manual planning process into a repeatable, step-by-step script using the Treatment Planning System's (TPS) scripting interface. This form of automated planning process is much faster compared to manual planning and allows one to scale pre-commissioning auto-contour error detection.

A study was conducted on a large cohort of 100 head-and-neck cancer patients (70 photon and 30 proton plans), allowing for robust statistical analysis. Results showed that using auto-contours resulted in minimal differences for dose coverage (e.g. D_{mean}, $D_{2\%}$)

and dose-related toxicities (i.e., NTCP) when compared to manual contours. Thus, this process of pre-commissioning QA showed that geometric differences introduced by auto-contouring had minimal clinical dosimetric consequences.

6.2.2 Chapter 3

Bayesian modeling choices can affect prediction uncertainty, which can potentially serve as a proxy for error in post-commissioning QA. Here, two Bayesian models (DropOut and FlipOut) were investigated and evaluated using expected calibration error (ECE) and a novel metric called region-based accuracy-vs-uncertainty (R-AvU). While ECE takes a more information theoretic approach to evaluate the models truthfulness, R-AvU takes a more visual approach to evaluate uncertainty utility. Experiments revealed that training with cross-entropy (CE) loss leads to better model calibration (i.e., ECE). Also, despite similar ECE values, FlipOut-CE demonstrated better uncertainty coverage in inaccurate regions than DropOut-CE when analyzed using R-AvU graphs. These results raise a question in context of translating research outputs to clinics: what metrics should one explore when evaluating for uncertainty as a proxy for contour error detection.

6.2.3 Chapter 4

While Bayesian models can produce uncertainty maps, their clinical utility depends on these maps aligning with true errors. Insights from Chapter 3 revealed that while Bayesian models produce uncertainty, its direct correspondence with prediction errors is often sub-optimal. This chapter introduced a differentiable loss formulation of the Accuracy-vs-Uncertainty (AvU) metric to explicitly encourage uncertainty where errors exist. Uncertainty heatmaps were evaluated against voxel inaccuracies using Receiver Operating Characteristic (ROC) curves (specifically, "uncertainty-ROC") and Precision-Recall (PR) curves. A key aspect of the evaluation was the distinction between segmentation "failures" (larger errors requiring intervention) and "errors" (smaller, acceptable inaccuracies akin to inter-observer variation), with only "failures" contributing to the "true" inaccuracy map.

Results showed that the AvU loss significantly improved calibrative (ECE) and uncertainty-error correspondence (ROC-AUC, PRC-AUC) metrics for both in-distribution (ID) and out-of-distribution (OOD) datasets. Compared to ensemble models (which use more parameters), the AvU model showed comparable or superior performance in uncertainty-error correspondence. Importantly, the study revealed that training for model calibration (e.g., using ECE-focused methods) does not necessarily translate to improved uncertainty outputs for error detection, emphasizing the unique advantage of the AvU loss. Thus, this chapter explored a novel technical approach to improve the utility of deep learning models for error detection in post-commissioning QA.

6.2.4 Chapter 5

Here the focus is shifted from post-commissioning error detection to post-commissioning error correction for auto-contour quality assurance. This chapter specifically aimed to compare the time-efficiency and contour quality of traditional manual brush tools against an AI-assisted "AI pencil" for auto-contour refinement. Many existing AI pencil methods in literature often lacked comprehensive human user evaluations, being limited to 2D settings or robotic users. A web-based interface was developed featuring an AI pencil capable of interpreting sparse 2D visual cues (scribbles) from users to generate 3D refinements of tumor contours on head-and-neck CT+PET scans.

The study enlisted the help of both non-clinical and clinical users to participate in refinement sessions of a patients auto-contour. The AI pencil consistently demonstrated superior time efficiency, being 5%-78% faster in non-expert sessions and 16%-97% faster in expert sessions compared to the manual brush. This remarkable speed-up is primarily attributed to the AI pencil's ability to propagate sparse 2D scribble inputs into comprehensive 3D contour refinements, obviating the need for tedious slice-by-slice editing. And despite the significant speed advantage, the final contour quality achieved with the AI pencil was equivalent to that of the manual brush. The AI pencil typically achieved a sharp increase in contour quality early in the refinement process before plateauing, contrasting with the manual brush's more gradual improvement. By demonstrating its effectiveness with human users in a 3D context, this work significantly contributes to alleviating the QA bottleneck and enhancing the overall efficiency of radiotherapy workflows.

6.3 Discussion and future work

The research presented in this thesis collectively addresses critical challenges in the safe, efficient and trustworthy integration of QA tools for deep learning-based auto-contouring models in clinical radiotherapy. By tackling both error detection and error correction within the QA workflow in both pre- and post-commissioning scenarios, this thesis contributes to advancing human-centric AI applications in medical image segmentation.

Building upon the foundations laid in the aforementioned chapters, several discussion points and promising avenues for future research emerge:

- Clinical buy-in – Often technical research tries to optimize on certain prespecified metrics and does not translate this into the clinic. This lack of *bench-to-bedside* attitude is often caused due to the structure of research projects. A missing factor is often sufficient clinical buy-in/involvement which leads to research being left on dusty shelves. Researchers should consider structuring their teams and mentors that involve multi-disciplinary skills to understand the full breadth and depth of the problem at hand.

- Renewing contouring guidelines – Chapter 2 showed both correlations and non-correlations between DICE and dose differences. Larger studies could redefine contouring guidelines, potentially evolving fixed anatomical guidelines into those with margins that could accommodate inter- and intra-observer variability.

- Understanding the utility of uncertainty in clinical settings – Uncertainty is a mathematical concept that has the potential to offer insights into the confidence of data-driven techniques like deep learning. However, often the community uses pure mathematical concepts like ECE (with its grouping mechanism) to evaluate the utility of a models uncertainty. Such metrics dont capture uncertainty in a pixel-wise (or granular manner). Thus, pushing the boundaries of existing metrics, although important, is not sufficient to adapt research innovations to daily clinical practice.

- Pixel-vs-Slice-vs-Region Uncertainty – It is possible that there is a practical limit to how much "uncertainty tuning" clinicians can benefit from before it becomes cognitive overload. On the one hand, too much uncertainty-driven decision making (e.g., pixel-wise) can be cognitively taxing. However, on the other hand, averaged uncertainty (e.g., on the slice or organ/tumor level) may not effectively guide contour refinement actions. Thus, researchers need to ponder on the granularity of uncertainty that we need in medical image segmentation applications.

- Connecting loss functions to clinical usability – The DICE loss is a geometry-based loss as it looks at the overall structure and shape of the ground truth and prediction. However, surprisingly a pixel-based approach i.e., the cross-entropy loss performed better at being truthful about its confidence in its predictions. Thus, makers of auto-contouring tools need to think deeper on how their loss functions affect the end users experiences.

- Analysing dataset requirements – One of the barriers to translating research into clinical practise is the high amount of training data required. However, literature shows similar performance with varying sizes of datasets. More work with tools like learning curves can inform the community better on the minimal dataset requirements to achieve clinical standards for contouring of organs and targets.

- Frameworks for real world clinical validation – Tools for robust experimentation and evaluation are what drive any field forward as it lowers the barriers for newcomers to contribute to the field. This can be seen with programming languages like Python and deep learning frameworks like Tensorflow and PyTorch. A similar example for medical image segmentation is the `grand-challenge.org` platform. Thus, as deep learning tools become more common in the field of medical imaging, the community needs to focus on how to build similar frameworks for uncertainty as a proxy for error detection and also for interactive segmentation.

- Trust in AI-driven actions – For the case of interactive contour refinement, how do we ensure clinicians trust AI-generated refinements enough to avoid reverting to manual corrections? And can such tools adapt to the diverse ways different clinicians approach contour editing? Thus, there may be a need for metrics that track how reliable is the model in local regions where the user makes their scribbles. And does the model secretly make any spurious predictions in regions far away from the users interaction.

- Role of regulatory bodies – Healthcare systems need to be regulated by governmental bodies due to the critical nature of the service they provide. However, research innovations often outpace regulatory bodies and in the meantime there is a possibility that innovations not rigorously or accurately tested can be used by clinicians. For e.g., in the case of deep learning-based auto-contouring there is very little discussion on the need for country/demographic-based benchmark datasets. Thus, it is very cumbersome for clinical innovators to determine how to evaluate commercial solutions since they need to be the ones to curate their own internal dataset which often tend to be messy due to the busy workload of clinicians. We implore the reader of this thesis to ponder upon this point and fill the aforementioned gap.

6.4 General conclusions

In an era of growing cancer incidence and limited clinical resources, this thesis contributes essential tools for ensuring safe, effective integration of deep learning auto-contouring into radiotherapy workflow. By offering practical, human-centric methods for both precise error detection and efficient error correction, this work helps bridge the gap between advanced deep learning models and their safe and effective quality assessment for integration into daily clinical radiotherapy practice. We hope to inspire others to pursue work that bridges the gap between mathematical uncertainty metrics and practical clinical trust. Likewise, interactive AI tools must evolve to reflect the diverse ways clinicians work.

Ultimately, this research aims to safeguard high-quality patient care and enhance workflow efficiency, with the positive results intended to inform and advance human-centric deep learning for medical imaging.