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

Mody, P.P.

1

Introduction

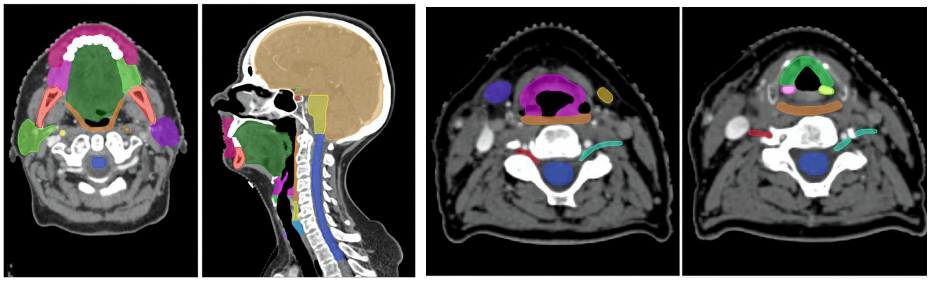
Deep learning, a form of a pattern recognition algorithm, has shown much promise in automating the contouring of anatomical structures for radiotherapy. However, such automation must be complemented by robust quality assurance (QA) mechanisms to ensure clinical reliability. This thesis addresses the growing need for automated QA for contours generated by deep learning models in head-and-neck radiotherapy. This chapter first outlines the anatomical complexity of the head-and-neck region and the multi-step radiotherapy (RT) workflow. Then the promise and limitations of deep learning-based automation in contouring is discussed. Finally, it concludes by motivating the need for automated tools that enable error detection and error correction of contours, forming the central theme of this thesis.

1.1 Head-And-Neck Anatomy

The head-and-neck area consists of 25 important organs from a radiotherapy perspective [1] as shown in Figure 1.1a, 1.1b. Note, that the head-and-neck region is considered separate from the brain region. Structures within the head-and-neck region are responsible for essential physiological functions — swallowing, breathing, salivation, taste, smell, speech and vision. Within this region, tumors are classified according to the site where they originate: laryngeal, pharyngeal (nasopharyngeal, oropharyngeal, hypopharyngeal), oral, salivary, nasal, or para-nasal, as shown in Figure 1.1.

Radiotherapy involves the delivery of radiation to tumors, termed the target in clinical terminology. This is done while simultaneously preserving healthy tissue, called organs-at-risk (OAR). Preservation of these OARs is paramount during cancer treatment to maintain a patients' quality of life post-treatment. However, the high density and proximity of these anatomical structures present a major challenge for automated contouring. Moreover, unlike organs, tumor shapes and sizes vary between patients and also during the course of treatment. Furthermore, factors like poor scan quality (e.g. CT, MR or PET) make detection of anatomical structures challenging.

Such a challenging task is ripe for automation, however, inaccurate automated delineations, especially those that go undetected, can lead to suboptimal radiation plans. This then leads to either healthy tissue toxicity, poor tumor reduction or both. Therefore, automated contouring must be accompanied by precise quality assessment (QA) mechanisms that can highlight potential inaccuracies and enable correction, without placing an



(a) (b)

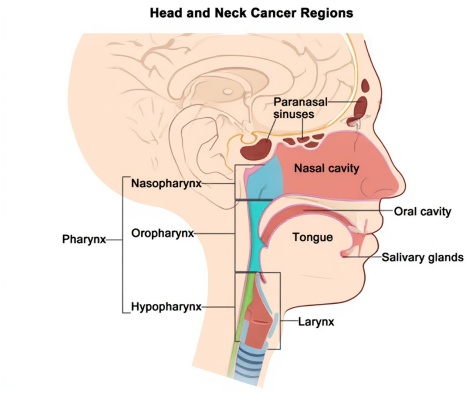


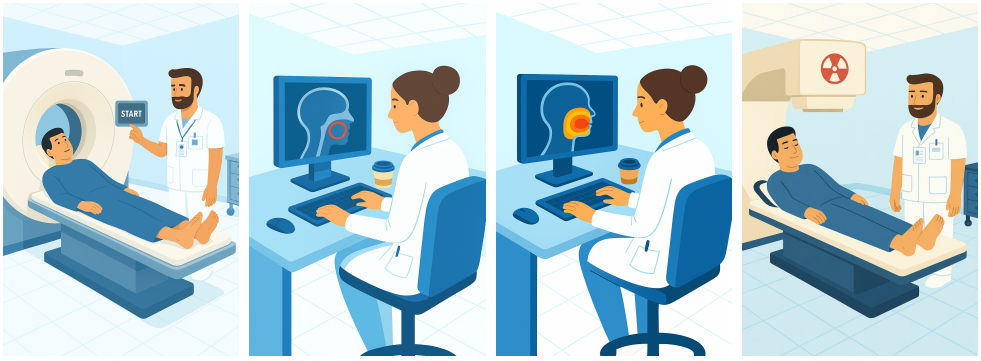
Figure 1.1: Contours for head-and-neck OARs on axial and sagittal views of a CT scan (a,b) [1] and sites where tumors are present (c) [2].

additional burden on clinicians.

1.2 Radiotherapy Workflow and Automation Opportunities

To achieve the goal of targeting tumors while sparing OARs in radiotherapy, a complex multi-step workflow is followed. This involves image acquisition, image contouring, dose plan calculation, and finally dose plan delivery (Figure 1.2). Cancer treatment takes place over multiple sessions of radiation called fractions. Treatment is often given between 33-35 fractions [3]. Since the anatomy of a patient evolves over the course of treatment (e.g., tumor shrinkage or fat reduction), it would be ideal to rescan, recontour and replan to suit the radiation to the latest anatomy. This advanced form of radiotherapy called adaptive radiotherapy (ART) [4, 5], presents clinicians with challenges due to the added workload.

Each of the above steps present opportunities for automation. For example, contouring, the focus of this thesis, is a very time-consuming part of the radiotherapy workflow which is also beset with issues like inter- and intra-annotator variability [6–12]. While automation techniques like deep learning have shown to significantly accelerate OAR and



(a) Image Acquisition (b) Image Contouring (c) Radiation Planning (d) Radiation Delivery

Figure 1.2: A multi-step radiotherapy workflow.

tumor contour generation, errors in predicted contours can go undetected without proper quality assessment (QA) mechanisms, thereby introducing clinical risk. Automated QA for such contours, especially with the increased motivation in clinics to use ART, is thus critical to ensure treatment integrity over time [13, 14].

Radiotherapy dose planning is also time-consuming yet increasingly automatable, with extensive research [15, 16] validating their clinical efficacy. Scan quality improvement is also a hot topic of research [17–19], however both these topics fall outside the scope of the present thesis.

The integration of new tools into routine clinical workflows involves a process known as commissioning. This process ensures that these tools are safe, reliable, and effective before and after their deployment. Commissioning can be broadly categorized into two phases:

- Pre-commissioning validation: This phase occurs before a tool is introduced into clinical practice. For instance, understanding how an auto-contouring tool's outputs affect dose plans is crucial for patient safety during this stage.

- Post-commissioning quality assurance (QA): This phase focuses on ongoing QA to identify and correct any errors that may arise during the daily clinical use of the deployed tool. For e.g., this could include detecting potential auto-contouring errors.

The pre-commissioning phase is a particularly challenging task for clinics since they may not have experience with the new tool. Also, it involves curating sufficient dataset quantity and this is difficult for resource-constrained clinics. Thus, this task also offers possibilities for automation.

1.3 Deep Learning for Auto-Contouring and its Limitations

Driven by the accelerating demands of modern radiotherapy, many deep learning-based auto-contouring methods for radiotherapy have been explored [20–33] and tested for clinical practice [34–41]. Researchers have released large datasets [42–44], investigated novel preprocessing techniques, deep neural architectures, and loss functions to push the boundaries on medical image segmentation. Such extensive effort has resulted in improved performance of these models, a promising indication for reductions in manual workload and inter-annotator variability.

While the significant progress and growing confidence in deep learning models within the community have made their integration into clinical practice inevitable and already underway, these auto-contouring models still face limitations. They can struggle with factors like poor contrast regions, small structures (e.g., optic nerves, swallowing muscles), X-ray scattering in CT scans due to dental fillings, or handling conflicting information from multiple modalities (e.g., CT + PET scans for tumor contouring). Additionally, deep learning-based auto-contouring models often fail when used on scans from a different clinic or machine than the ones they were trained on, known as an out-of-distribution scenario.

Thus, although the adoption of these models has, to some extent, alleviated the contouring bottleneck in the radiotherapy workflow, it concurrently introduces a critical need for robust and efficient quality assurance (QA). Therefore, the next logical phase of research is to explore large-scale or human-interactive contour QA techniques to ensure the reliability and safety of automatically generated contours. This is the next phase to ensure further integration of deep learning solutions into routine clinical workflows.

1.4 Quality Assessment of Contour Automation

Auto-contouring tools can either be tested via clinically-oriented metrics or image-based metrics. In the field of radiotherapy, much work has been done to validate auto-contouring tools in the pre-commissioning phase using radiation dose-based metrics [41, 45–50]. The goal here is to check how different the radiation plans are when made via automated contours as compared to manual contours.

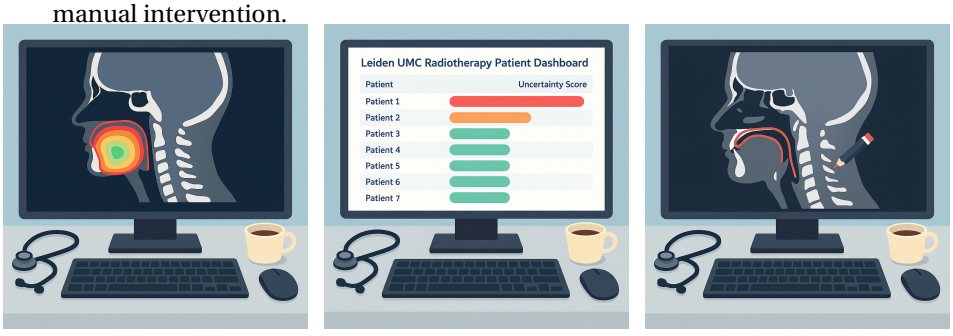
An alternative form of QA is to use post-commissioning image-based techniques which are also more broadly applicable to the whole field of medical image segmentation. For e.g., an issue with deep learning-based auto-contouring models is that they "fail silently" when operating in out-of-distribution (OOD) scenarios. To abate this, solutions have been proposed over the past 5 years to detect OOD samples [51–55]. While OOD techniques classify the whole image as either in-distribution (ID) or OOD, other works focus on extracting pixel-level uncertainty to guide QA activities [56–75]. Note that this uncertainty is often calculated using a deep learning models' output probabilities. Furthermore, to ensure that pixel-level uncertainty is trustworthy, the medical image segmentation com-

munity has explored the concept of calibration of a models output probabilities [76–79]. Yet other works train an additional deep learning model to classify auto-contours within predefined categories such as acceptable/non-acceptable [80–83] on a slice-level.

Further work has also attempted to tackle the problem of how to quickly deal with errors once detected using interactive segmentation techniques. Interactive segmentation aims to efficiently refine faulty contours with minimal manual intervention. Very little work existed prior to the start of this thesis [84] however, some work has been published since then [85–92].

Inspired by previous work, this thesis proposes categorizing QA approaches as follows:

- Error Detection: Identifying regions where the model may have failed, using uncertainty estimation or dose impact analysis.
- Error Correction: Providing efficient tools for refining faulty contours with minimal manual intervention.



(a) Error Detection (via dose)

(b) Error Detection (via uncertainty)

(c) Error Correction

Figure 1.3: Potential applications of the output of this thesis.

The themes of error-detection and error-correction can also be understood through Figure 1.3 which is an illustration on potential user interfaces for these themes. Figure 1.3a shows a clinicians desktop where the radiation dose of the patient along with the contour is shown to determine whether contour QA shall have a significant impact on the patient. Figure 1.3b shows another error detection view wherein the patients to consider for QA are organized by the underlying models contouring uncertainty. Finally, in Figure 1.3c, the clinician can do semi-automated editing of auto-contours via an AI pencil.

Previous works established techniques to perform clinical validation or proposed technical improvements to improve on widely accepted metrics for OOD detection or calibration. However, this thesis aims to extend the aforementioned works with a focus on the clinician. It asks how one can define measures, explore novel techniques or QA at scale so that research keeps the clinician at the center. For error detection, it focuses on how dose-based clinical evaluations can be done at scale to compare various auto-contouring tools or how one can use image-based uncertainty which actually aligns with the true error.

And finally for error correction, what are the tools needed and which are the metrics that inform the utility for an interactive contour refinement technique in real world clinical settings.

1.5 Thesis outline

The aim of this thesis is to develop and evaluate automated methods for both error detection and error correction of contours generated by deep learning-based auto-contouring tools for head-and-neck radiotherapy. Beyond their technical contributions, the proposed methods can also be viewed from a commissioning perspective for integrating such tools into clinical workflows. This includes pre-commissioning validation, which involves assessing whether an auto-contouring tool is safe and reliable enough for clinical introduction. It also encompasses post-commissioning quality assurance (QA), focused on identifying and correcting errors that may arise during daily clinical use once the tool has been deployed.

This thesis is organized as follows:

Chapter 2 addresses pre-commissioning validation, exploring large-scale retrospective dose evaluations to quantify the clinical impact of auto-contouring errors. Before introducing an auto-contouring tool into clinical practice, understanding how its contours affect dose plans helps ensure patient safety. The proposed workflow emulates existing clinical treatment planning protocols and reuses optimization parameters, functioning as a form of robot process automation (RPA).

Chapter 3 focuses on post-commissioning error detection, investigating Bayesian models for automatically flagging potentially inaccurate regions in auto-generated contours. The goal is to help clinicians quickly identify problematic areas requiring manual review during routine clinical use. We also investigate loss functions and uncertainty metrics and their role in evaluating uncertainty. This work establishes a new approach combining quantitative and qualitative metrics for selecting appropriate models for clinical QA deployment.

Chapter 4 also relates to post-commissioning error detection, improving how well uncertainty maps from Bayesian models correspond to true contouring errors. Better uncertainty-error correspondence improves the utility of these maps for clinicians in real-time QA of patient scans. To improve uncertainty-error correspondence we utilize a differentiable version of an uncertainty metric and then evaluate on a per-pixel basis.

Chapter 5 transitions to post-commissioning error correction, evaluating an AI-assisted contour refinement tool ("AI pencil") that enables efficient correction of identified contouring errors. The tool's speed and effectiveness are compared to the traditional manual brush. A web-based interface and an AI pencil were developed, that supports 2D interactions to refine 3D auto-contours. User experiments with both experts and non-experts were done to compare the time-efficiency and contour quality of both tools.

Chapter 6 summarizes the thesis contributions and discusses future research directions on the topics of clinical validation, uncertainty in medical image segmentation and human-centric AI techniques to speed up contour QA.