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

Mody, P.P.

Automated Quality Assurance of Deep Learning Contours in Head-and-Neck Radiotherapy

Prerak Mody

Colophon

About the cover:

The cover page was designed by Prerak Mody. On the front cover, you see an ensemble of contours in the middle represented by swirling strokes. The individual contours are depicted with various colors and represent the central theme of this thesis i.e., contouring of regions of interest within medical scans. These contours are made in collaboration by two parties: a human (clinician) and an AI. The AI (i.e., deep learning) has been anthromorphized and thus depicted via a human-like hand with a mesh design. The human and AI hands are collaborating on the contour which is the motivating factor for this thesis.

This inside cover envisions the future of radiotherapy, showcasing a clinician (in white coat) editing AI-generated contours of the head-and-neck area in real-time. As the patient is positioned on the linear accelerator gantry, the clinician uses automated quality assurance tools to quickly adjust the treatment plan based on the patient's latest anatomical scans, ensuring highly precise and adaptive treatment delivery. A technician (in blue overalls) helps the patient position themselves on the gantry.

Google Gemini, a visual language model (VLM) was used to assist with these designs.

Automated Quality Assurance
of Deep Learning Contours
in Head-and-Neck Radiotherapy
Prerak Mody

ISBN: 978-94-6522-967-6
Thesis layout & cover designed by Prerak Mody
Printed by Ridderprint, the Netherlands

Automated Quality Assurance of Deep Learning Contours in Head-and-Neck Radiotherapy

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van (waarnemend) rector magnificus,
volgens besluit van het college voor promoties
te verdedigen op donderdag 22 januari 2026
klokke 10:00 uur

door

Prerak Mody
geboren te Mumbai, Maharashtra, India
in 1993

Promotor:	Prof. dr. ir. M. Staring
	Prof. dr. ir. B.P.F. Lelieveldt
Leden promotiecommissie:	Prof. dr. C.R.N. Rasch
	Prof. dr. A. Mukhopadhyay
	Technical University, Darmstadt
	Prof. dr. J.P.W. Pluim
	Eindhoven University of Technology, Eindhoven
	Prof. dr. ir. C.A.T. van den Berg
	University Medical Center, Utrecht

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands.

The research in this thesis was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022)

The printing of this thesis was funded by:
Medis Medical Imaging
Library of Leiden University
Bontius Stichting/LUMC ResearchFoundation

Contents

List of abbreviations

QA	quality assessment
CT	computed tomography
PET	positron emission tomography
HN	head-and-neck
RT	radiotherapy
OAR	organ at risk
DL	deep learning
CTV	clinical target volume
RoI	region of interest
NTCP	normal tissue complication probability
AI	artificial intelligence
AvU	accuracy-vs-uncertainty
ECE	expected calibration error
R-AvU	region-based accuracy-vs-uncertainty
CE	cross entropy
PTV	planning target volume
DICE	Sørensen–Dice coefficient
D_{mean}	mean dose
D0.03cc	dose to 0.03 cubic centimeters of tissue
EUD	equivalent uniform dose
MaxEUD	maximum equivalent uniform dose
OOD	out of distribution

1

Introduction

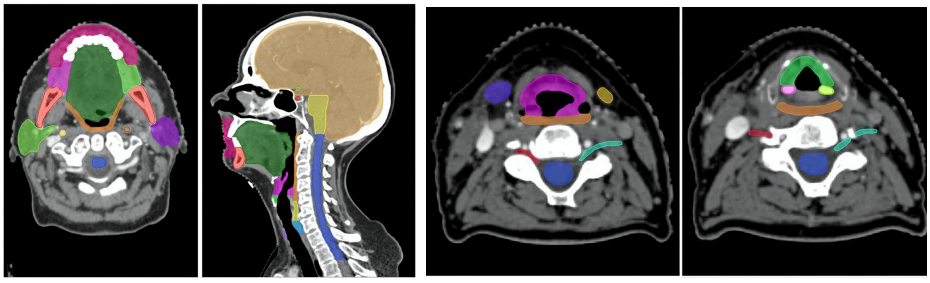
Deep learning, a form of a pattern recognition algorithm, has shown much promise in automating the contouring of anatomical structures for radiotherapy. However, such automation must be complemented by robust quality assurance (QA) mechanisms to ensure clinical reliability. This thesis addresses the 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)

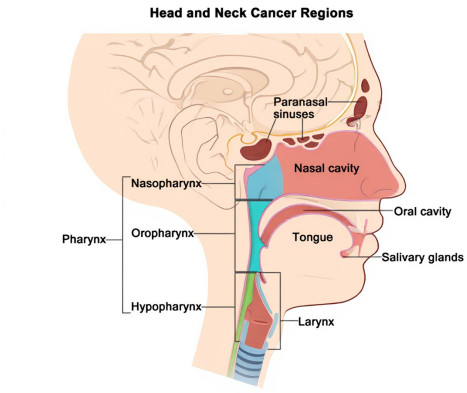


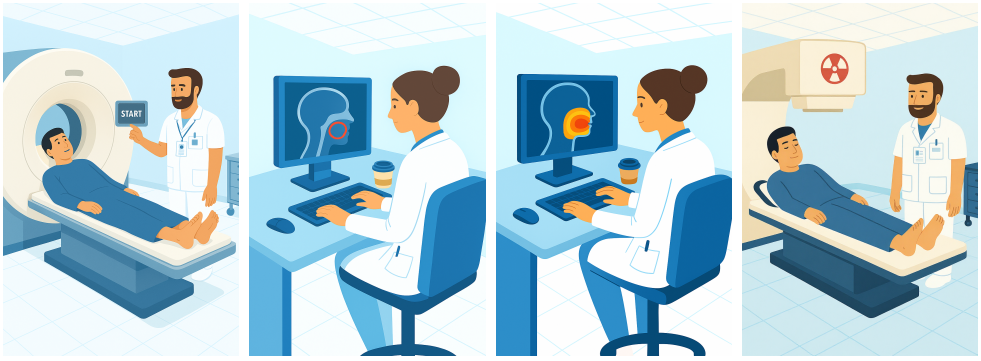
Figure 1.1: Contours for head-and-neck OARs on axial and sagittal views of a CT scan (a,b) [1] and sites where tumors are present (c) [2].

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.

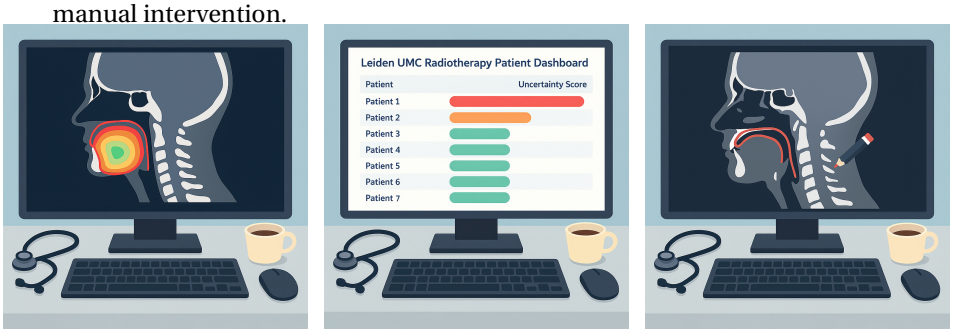
An alternative form of QA is to use post-commissioning image-based techniques which are also more broadly applicable to the whole field of medical image segmentation. For e.g., an issue with deep learning-based auto-contouring models is that they "fail silently" when operating in out-of-distribution (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.

2

Large-scale dose evaluation of deep learning organ contours in head-and-neck radiotherapy by leveraging existing plans

This chapter was adapted from:

Mody, Prerak, Merle Huiskes, Nicolas F. Chaves-de-Plaza, Alice Onderwater, Rense Lamsma, Klaus Hildebrandt, Nienke Hoekstra, Eleftheria Astreinidou, Marius Staring, and Frank Dankers. "Large-scale dose evaluation of deep learning organ contours in head-and-neck radiotherapy by leveraging existing plans." In *Physics and Imaging in Radiation Oncology* 30 (2024): 100572.

Abstract

Background and Purpose: Retrospective dose evaluation for organ-at-risk auto-contours has previously used small cohorts due to additional manual effort required for treatment planning on auto-contours. We aimed to do this at large scale, by a) proposing and assessing an automated plan optimization workflow that used existing clinical plan parameters and b) using it for head-and-neck auto-contour dose evaluation.

Materials and Methods: Our automated workflow emulated our clinic's treatment planning protocol and reused existing clinical plan optimization parameters. This workflow recreated the original clinical plan (P_{OG}) with manual contours (P_{MC}) and evaluated the dose effect ($P_{OG} - P_{MC}$) on 70 photon and 30 proton plans of head-and-neck patients. As a use-case, the same workflow (and parameters) created a plan using auto-contours (P_{AC}) of eight head-and-neck organs-at-risk from a commercial tool and evaluated their dose effect ($P_{MC} - P_{AC}$).

Results: For plan recreation ($P_{OG} - P_{MC}$), our workflow had a median impact of 1.0% and 1.5% across dose metrics of auto-contours, for photon and proton respectively. Computer time of automated planning was 25% (photon) and 42% (proton) of manual planning time. For auto-contour evaluation ($P_{MC} - P_{AC}$), we noticed an impact of 2.0% and 2.6% for photon and proton radiotherapy. All evaluations had a median ΔNTCP (Normal Tissue Complication Probability) less than 0.3%.

Conclusions: The plan replication capability of our automated program provides a blueprint for other clinics to perform auto-contour dose evaluation with large patient cohorts. Finally, despite geometric differences, auto-contours had a minimal median dose impact, hence inspiring confidence in their utility and facilitating their clinical adoption.

2.1 Introduction

Manual contouring of organs-at-risk (OAR) in radiotherapy is a time and resource-demanding task [5, 93, 94], especially in head-and-neck cancer due to a large OAR count [95]. Moreover, it is plagued by inter- and intra-annotator variability [10, 11, 96, 97] and hence there is a need for automation. In the last few years, availability of deep learning-based commercial tools have reduced the barriers for clinics to implement auto-contouring technology in daily practice. However, these tools may produce erroneous contours due to poor contrast, organ deformations, surgical removal of an organ or when tested on different patient cohorts [98]. Such cases may potentially lead to commercial providers providing updates to the underlying deep learning models. Thus, as deep learning auto-contouring tools are increasingly adopted in clinics, with the potential for future updates to models, there is a growing need to benchmark them, preferably at large-scale and in an automated manner.

As deep learning-based auto-contouring methods for head-and-neck OARs have been shown to offer satisfactory geometric performance [10, 99], the next step is to evaluate their dose impact [100]. However, we observed that dose-based studies on auto-contours tend to use either smaller ($\leq$ 20) [41, 45, 46, 48, 49, 101, 102] or medium-sized ($\leq$ 40) [50], rather than larger [47] datasets. Studies using larger datasets simply superimpose the automated contours on the clinical dose [47] which does not fully replicate the treatment planning process. Conversely, studies using smaller or medium-sized test datasets either made manual plans [41, 48–50], used knowledge-based planning [46], a template approach [45] or a priori multi-criteria optimization (MCO) [101, 102]. Since smaller datasets may be affected by sampling bias, there is a need to perform dose analysis with a larger patient cohort. However, a manual approach to plan optimization is simply not scalable. Moreover, existing automated approaches [45, 46, 101], if not already clinically implemented, require additional skills and resources. Therefore, there is a need for an automated approach to treatment planning that can be done at a large scale and also leverages existing clinical knowledge and work.

Thus, our contribution was to propose and assess a plan optimization method for retrospective studies that is scalable due to its automated nature and easily implementable due to the use of existing clinical resources (i.e., knowledge, tools and optimization parameters). We then used this approach in a use case to quantify auto-contour-induced dose effects for head-and-neck photon and proton radiotherapy.

2.2 Materials and methods

2.2.1 Data acquisition

Our dataset consists of 100 head-and-neck cancer patients, of which 70 had clinical plans made for photon therapy, while 30 had proton plans, at Leiden University Medical Center

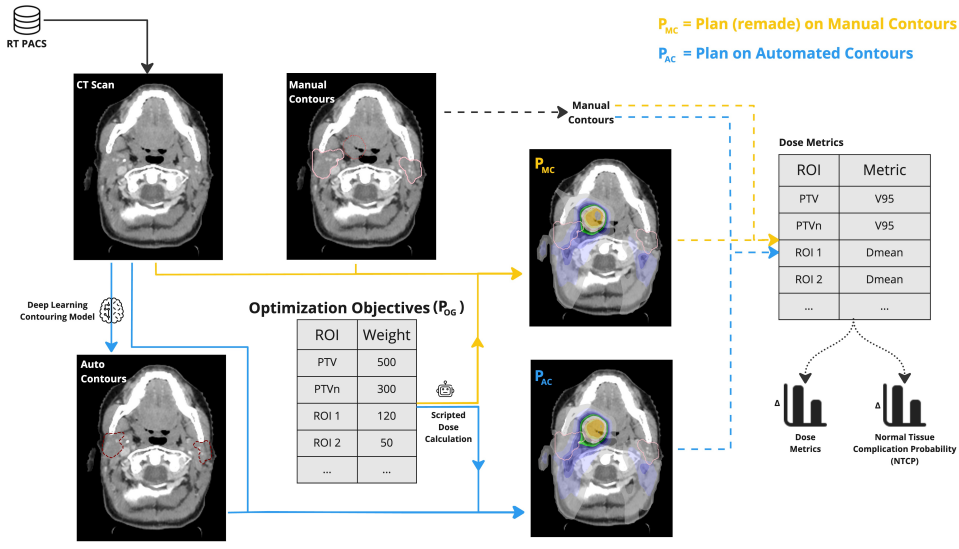


Figure 2.1: Workflow for automated plan optimization and use-case of evaluating the effect of automated contours on dose. By reusing original plan (P_{OG}) parameters, we made a plan for both the manual contours (P_{MC}) and automated contours(P_{AC}), shown with yellow and blue colors respectively. Dashed lines indicate the evaluation workflow where both doses were evaluated on the manual contours. Pink, maroon and orange contours are used to represent the manual, automated and PTV (DL1) contours respectively. Finally, we used manual contours to compute dose metrics and normal tissue complication probability (NTCP) [103] models and compare all plans.

(Leiden, The Netherlands) from 2021 to 2023. Patients were treated for either oropharyngeal (71) or hypopharyngeal (29) cancers with cancer stages T1-4, N0-3 and M0. 92 patients were treated with curative intent, i.e., 7000cGy to the primary tumor, while others were prescribed 6600cGy due to their post-operative nature. Details about CT scans used in planning are written in Section 2.6.1. The study was approved by the Medical Ethics Committee of Leiden, The Hague, Delft (G21.142, October 15, 2021). Patient consent was waived due to the retrospective nature of the study.

2.2.2 Automated Contours

For automated contouring, a commercial deep learning model from RayStation-10B (RaySearch Labs, Sweden) - “RSL Head and Neck CT" (v1.1.3) was used. A subset of the OARs which were used clinically for treatment planning were auto-contoured – Spinal Cord, Brainstem, Parotid (L/R), Submandibular (L/R), Oral Cavity, Esophagus, Mandible and Larynx (Supraglottic). See Section 2.6.2 for additional details.

2.2.3 Treatment Planning Protocol

We used volumetric modulated arc therapy (VMAT) to generate a photon plan using a 6MV dual arc beam. The elective and boost Planning Target Volumes (PTV), henceforth referred as DL1/DL2 (dose level 1/2) were prescribed 5425cGy/7000cGy in 35 fractions. For post-operative patients, our clinic prescribed 5280cGy/6600cGy in 33 fractions instead. Planning was done such that at least 98% of DL1 and DL2 volumes received 95% of the prescribed dose ($V_{95\%}$) and also by keeping $D_{0.03cc}$ for DL2 below 107% of the prescribed dose.

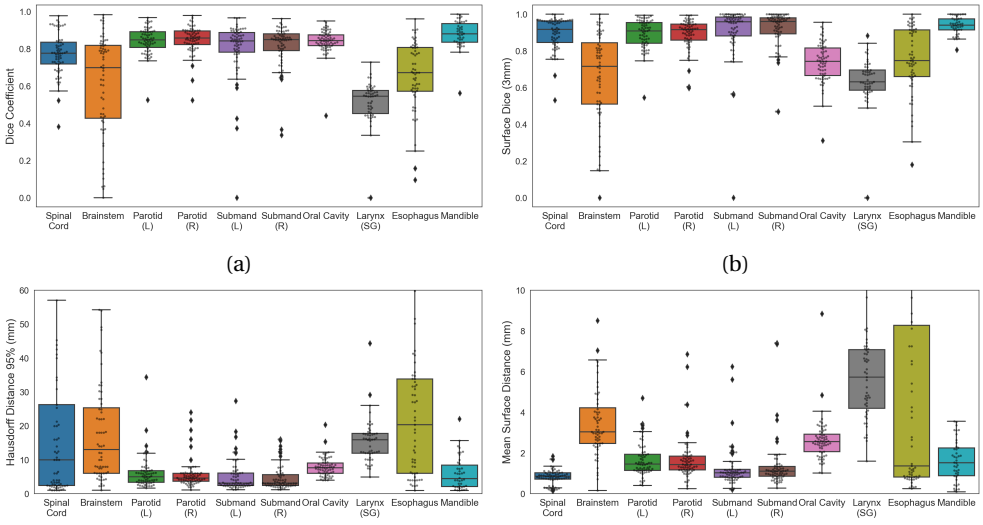
Proton plans consisted of six beam intensity modulated proton therapy (IMPT). Planning was done such that $V_{95\%} \geq 98\%$ for DL1/DL2 and $D_{2\%} \leq 107\%$ for DL2 of the Clinical Target Volume (CTV) in a 21-scenario robust optimization with 3mm setup and 3% proton range uncertainty. For robust evaluation of CTV DL1/DL2 we instead use 28-scenarios and test the voxel-wise minimum (vw-min) plan such that its $V_{94\%} \geq 98\%$ [104] and voxel-wise maximum (vw-max) of $D_{2\%} \leq 107\%$.

2.2.4 Automated Treatment Planning

To make our automated program, a four-step script [105–107] was created which uses manually defined beam settings and objective weights from the clinical plan (more details in Section 2.6.3). This approach is also referred as robot process automation (RPA) [108], a process wherein a program emulates a human.

In summary, for step 1, we began with an objective template i.e., a class solution with a standard set of weights that focuses on targets and the body contour. Step 2 then added dose-fall-off (DFO) objectives for organs which is the distance over which a specified high dose falls to a specified low dose. In step 3, we introduced equivalent uniform dose (EUD) objectives [109] on the OARs. Manual planning for the EUD objective involves iteratively fine-tuning its parameters. Since only the parameters of the last iteration were available to us, we instead followed a single-step optimization for this objective. Finally, in step 4, we used patient-specific control structure contours to reduce OAR dose or sculpt the dose to the targets. In the last step, we also updated any other weights the treatment planner might have changed compared to the objective template. Note, these final weight updates were asynchronous to manual planning, since we did not know when these weights were updated in the aforementioned process. Note that each of the above steps underwent four optimization cycles.

Using our automated program, we made two plans – 1) a plan optimized on manual contours (P_{MC}) and 2) a plan optimized on automated contours (P_{AC}) as shown in Figure 5.1. For the targets, elective lymph nodes, and OARs not available in the auto-contouring model we used manual contours which were used clinically for the original plan (P_{OG}). The plans were made using the Python 3.6 scripting interface of the Treatment Planning System (TPS) of RayStation. The scripts for this work are available at



(a) (b) (c) (d)

Figure 2.2: Box plots showing geometric (a) and surface metrics (b,c,d) for all our patients. The scatter points indicate the metric values for each patient.

https://github.com/prerakmody/dose-eval-via-existing-plan-parameters.

2.2.5 Geometric Evaluation

We used volumetric and surface distance metrics like Dice Coefficient, Hausdorff Distance 95% (HD95) and Mean Surface Distance (MSD) to evaluate our contours. Moreover, we also evaluated Surface DICE (SDC) with a margin of 3mm to gain insight into contour editing time requirements [110].

2.2.6 Dose and NTCP Evaluation

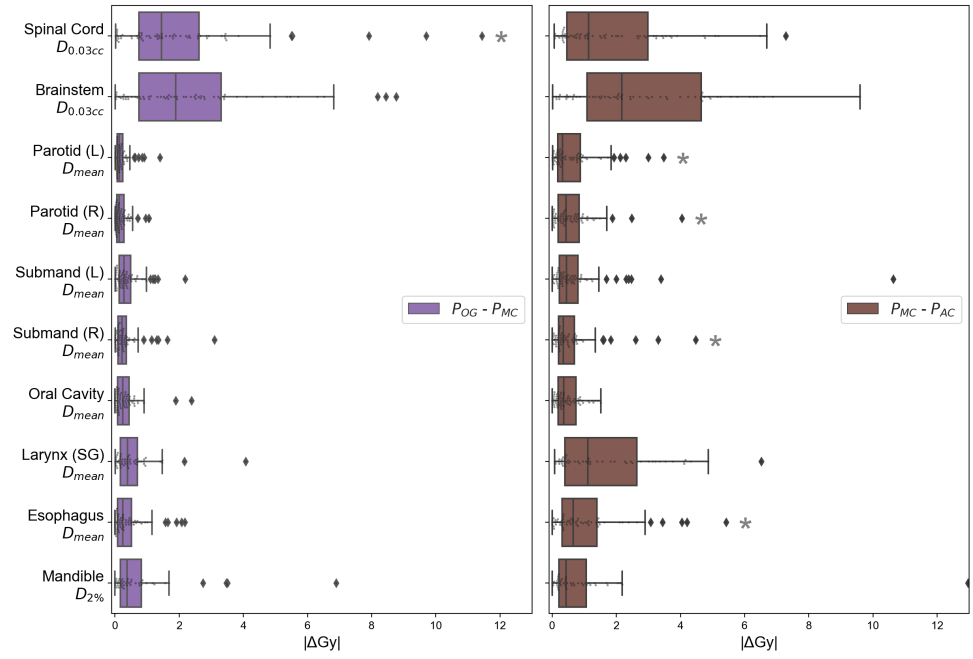
Given that our plans – P_{OG}, P_{MC} and P_{AC} have differences in the way they were created, we need to compare them. Metrics relevant to OARs were calculated and plans were compared in the following manner:

$$\Delta D_x = D_{x,p1} - D_{x,p2}. \tag{2.1}$$

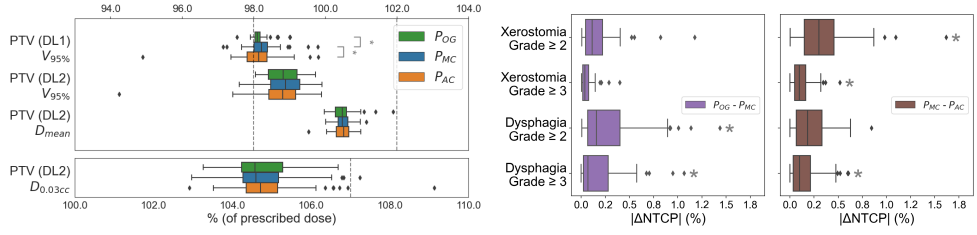
Here, x refers to the OAR for which we calculated a dose metric D and then compared it between any pair of plans $p1$ and $p2$. Here, D can refer to $D_{0.03cc}$ (Spinal Cord, Brainstem), D_{mean} (Parotid, Submandibular, Oral Cavity, Larynx (Supraglottic), Esophagus) or $D_{2\%}$ (Mandible).

For normal tissue complication (NTCP) probability [103] evaluation, we used a similar approach:

$$\Delta\text{NTCP}_d = \text{NTCP}_{d,p1} - \text{NTCP}_{d,p2}, \tag{2.2}$$



(a) Organs at Risk



(b) Targets

(c) NTCP

Figure 2.3: Dose metrics for the original (i.e., clinical) photon plans (P_{OG}) as well as plans (re)made on manual (P_{MC}) and automated (P_{AC}) contours using an automated program. $P_{OG}-P_{MC}$ shows the dose effect of the proposed planning process, while $P_{MC}-P_{AC}$ shows the effect of using auto-contours. Here * represents a p-value ≤ 0.05. In a) we see the difference in the dose metric of each OAR when comparing across plans. The plots in b) show us the metrics for the targets, while c) shows us the difference in NTCP values.

where d refers to either Xerostomia or Dysphagia with a grade ≥ 2 or ≥ 3.

For the above ΔD_x (dose) and ΔNTCP_d values, we performed a Wilcoxon signed-rank test ($p \leq 0.05$ is considered a significant difference) to evaluate if the differences between plans are significant.

2.3 Results

2.3.1 Geometric evaluation

Figure 2.2 shows five organs (Spinal Cord, Parotids, Submandibulars, Oral Cavity, Mandible) had a median DICE ≥ 0.78 (with additional summary measures tabulated in Section 2.6.2). In Figure 2.2b we observed that in general the surface DICE values for the OARs are higher than their DICE values, except for the oral cavity. Figure 2.2c and Figure 2.2d shows that HD95 and MSD had trends similar to DICE in Figure 2.2a. OARs with a median DICE ≥ 0.8 had their median HD95 less than 7.7mm and their median MSD less than 2.6mm. The spinal cord had DICE values that are better than brainstem, but its HD95 range was as long as brainstem.

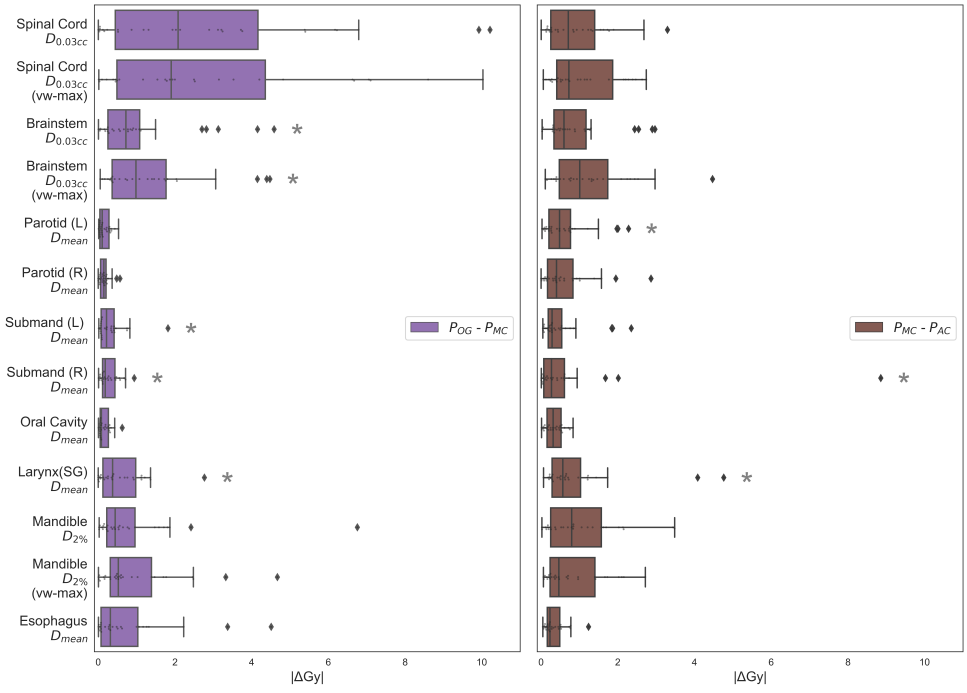
2.3.2 Dose evaluation

The median absolute value of P_{OG} (original plan) - P_{MC} (automated plan using manual contours) was 0.27Gy (1.0%), 1.66Gy (4.6%) and 0.21Gy (0.7%) for all, central nervous system (CNS), i.e., Brainstem and Spinal Cord and non-CNS organs, respectively. The same for P_{MC} - P_{AC} (automated plan using auto-contours) was 0.58Gy (2.0%), 1.86Gy (5.4%) and 0.46Gy (1.6%), with metrics of individual organs in Figure 2.3a listed in Section 2.6.4. Figure 2.3b shows dose metrics for targets where, for P_{MC} and P_{AC}, we achieved PTV (DL1) (V_{95}) ≥ 98.0% for 76% and 60% of plans. However, 96% and 93% of P_{MC} and P_{AC} plans achieved PTV (DL1) (V_{95}) ≥ 97.5%. For this metric, a statistically significant difference was observed between P_{OG} and P_{MC} as well as P_{MC} and P_{AC}. Finally, Figure 2.3c shows |ΔNTCP| results, where the maximum median across all toxicities was 0.3% (individual toxicity metrics in Section 2.6.5).

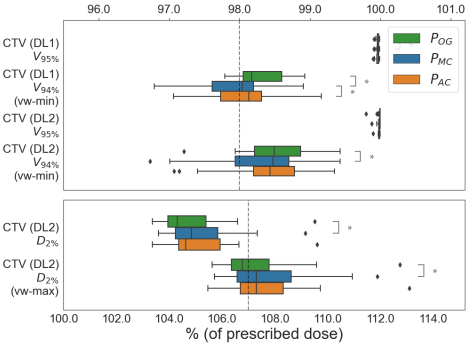
For proton, $|P_{OG} - P_{MC}|$ had a median value of 0.33Gy (1.5%), 1.13Gy (11.5%) and 0.22Gy (0.8%) for all, CNS and non-CNS organs, respectively. The same for $P_{MC} - P_{AC}$ was 0.48Gy (2.6%), 0.75Gy (6.9%) and 0.38Gy (1.8%). Figure 2.4b shows proton targets wherein 58% and 62% of P_{MC} and P_{AC} plans achieved PTV (DL1) (vw-min) (V_{94}) ≥ 98.0%, while 82% and 80% achieved PTV (DL1) (vw-min) (V_{94}) ≥ 97.5%. Similar to photon, a statistically significant difference was observed between P_{OG} and P_{MC} as well as P_{MC} and P_{AC}. For |ΔNTCP| (Figure 2.4c), the maximum median across all toxicities was 0.2%.

A weak Spearman correlation coefficient between DICE and dose differences ($|P_{MC} - P_{AC}|$) was observed for CNS organs ($|\rho_s| \leq 0.11$), across both photon and proton (Figure 2.5). Conversely, the Parotids, Submandibulars and Oral Cavity had relatively higher values ($-0.43 \leq \rho_s \leq -0.17$). The remaining organs did not have similar correlations across both radiotherapy treatments.

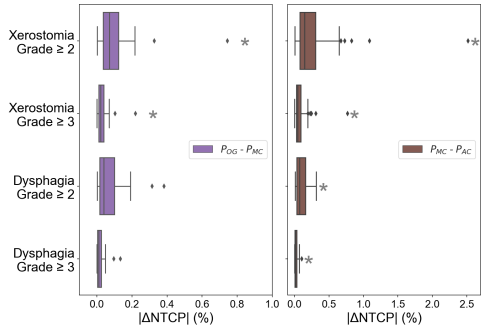
Finally, our automated plan optimization took 45 minutes and 2.5 hours of computer time, compared to 3 and 6 hours of manual time (on average, as estimated by our clinic's planners), for photon and proton, respectively.



(a) Organs at Risk

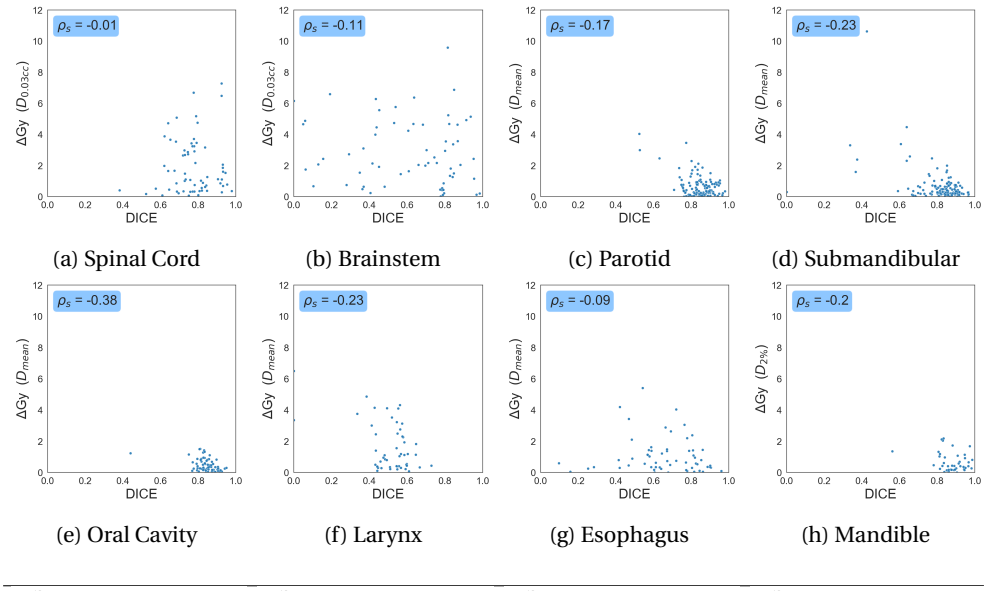


(b) Targets



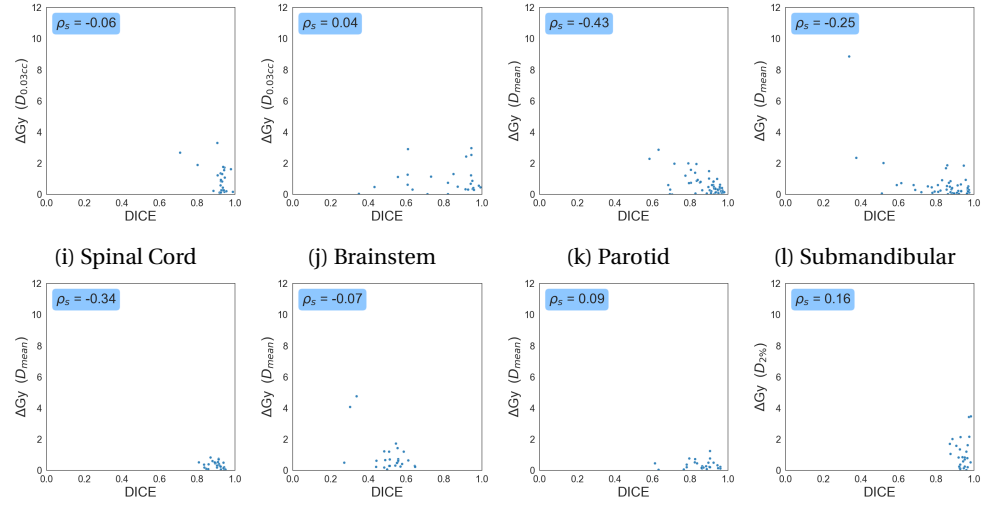
(c) NTCP

Figure 2.4: Dose metrics for the original proton plans (P_{OG}) as well as plans (re)made on manual (P_{MC}) and automated (P_{AC}) contours using an automated program. $P_{OG} - P_{MC}$ shows the dose effect of the proposed planning process, while $P_{MC} - P_{AC}$ shows the effect of using auto-contours. Here * represents a p-value ≤ 0.05. In a) we see the difference in the dose metric of each OAR when comparing across plans. The plots in b) show us the metrics for the targets, while c) shows us the difference in NTCP values.



(a) Spinal Cord (b) Brainstem (c) Parotid (d) Submandibular

(e) Oral Cavity (f) Larynx (g) Esophagus (h) Mandible



(i) Spinal Cord (j) Brainstem (k) Parotid (l) Submandibular

(m) Oral Cavity (n) Larynx (o) Esophagus (p) Mandible

Figure 2.5: Scatter plots for eight organs-at-risk from the auto-contouring module. Here we plot the DICE (x-axis) against each organs absolute dose metric differences, i.e., $|P_{MC} - P_{AC}|$ (y-axis) for photon (a-h) and proton (i-p) radiotherapy.

2.4 Discussion

This work aimed at proposing and assessing an automated plan optimization workflow for retrospective studies that can be easily implemented by clinics due to its use of existing clinical resources. Unlike previous works [41, 45, 46, 48, 49, 101, 102], we performed this

at large-scale and for both photon and proton radiotherapy. To replicate our approach, a clinic can simply use the scripting interface of their treatment planning system (TPS) and convert their planning process into a step-by-step approach. This requires minimal additional expertise (i.e., Python coding), for which many TPS solutions provide documentation. For head-and-neck radiotherapy, automated plans on manual contours (P_{MC}) showed a negligible difference (i.e., median impact of 1.0% and 1.5% across organs), when compared to the original clinical plan (P_{OG}) [111, 112]. Thus, the proposed evaluation process could serve as a springboard for clinics to validate an auto-contouring model, at large-scale, by simply reusing their existing plans. When using this program for the use case of head-and-neck auto-contour evaluation, the plan using auto-contours (P_{AC}) had a low dose impact when compared to the plan using using manual organ contours, for both photon (2.0%) and proton (2.6%) planning. Additionally, minuscule differences in NTCP values indicated that minor plan differences did not lead to large differences in long-term radiation-induced toxicity. This could potentially promote confidence in the community [113] to adopt auto-contouring to speed up clinical workflows.

For five out of eight OARs (i.e., Spinal Cord, Parotid, Submandibular, Oral Cavity and Mandible), the average DICE scores may be considered on par with previous work (≈ 0.8) [10, 45, 99] (see Section 2.6.2). A visual inspection of the remaining auto-contours, i.e., Larynx (SG), Brainstem (and by extension the Spinal Cord) (Figure 2.6, Section 2.6.6) indicated that they had contouring protocols that differed from our clinic. Moreover, the auto-contouring model was trained on a different patient cohort, leading to additional contour differences with our clinical dataset. Finally, we chose to not perform any additional refinement on manual contours, since they were also used for making clinical plans (P_{OG}) delivered to patients. For e.g. in the first row of Figure 2.6, we see that only the caudal section of the Brainstem was annotated. Treatment planners find optimizing this section sufficient due to its potential for high dose from tumor proximity. The aforementioned reasons are why we noticed reduced measures for Larynx (SG), Brainstem and Spinal Cord in Figure 2.2.

A critique of using unmodified manual contours may be that a lack of “gold-standard" contours will not give accurate geometric measures. Since our primary goal however was dose evaluation using existing clinical resources (i.e., unmodified manual contours), we proceed without any refinement. Also, in an auto-contouring dose evaluation scenario, it is already sufficient to know that plans made on auto-contours are equivalent to plans made on manual contours as seen in Figure 2.3b (photon) and Figure 2.4b (proton). Thus, our approach of using existing manual contours improves the ease-of-implementation of auto-contour dose evaluation studies and enables evaluation at large-scale.

To evaluate the quality of our automated plans, we first assessed target dose metrics. We use PTV (DL1) ($V_{95\%}$) for photon and CTV (DL1) ($V_{94\%}$) (vw-min) for proton, since planners prioritize them due to their difficulty. Hence it serves as a good benchmark for

our automated plans. Results indicated that most of our plans (≥ 93% for photon and ≥ 80% for proton) were of near-clinical quality (i.e., ≥ 97.5%). Those plans that did not strictly achieve clinical quality (i.e., ≥ 98%) on the aforementioned metrics, had reduced dose coverage in either the most cranial or caudal slices. In a retrospective study for dose-evaluation of auto-contours, such a minor error will have a minimal effect on the dose metrics of organs we are interested in.

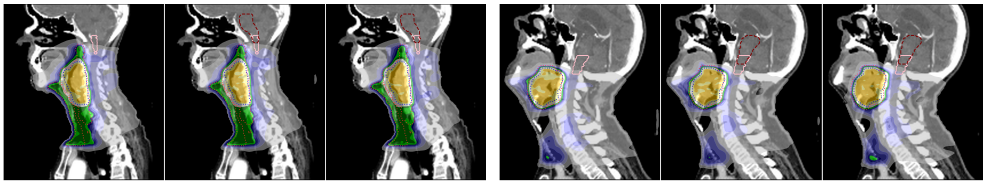
Figure 2.4b shows that most proton plans, including P_{OG}, tended to have hotspots, i.e., $D_{2\%}(vw-max) \geq 107\%$, unlike most photon plans which did not, i.e., $D_{0.03cc} \leq 107\%$ (Figure 2.3b). In our dataset, these proton plans were made for performing a plan comparison between photon and proton (via NTCP), according to the model-based selection [114]. If during proton treatment planning, the NTCP differences already indicated either a) high organ sparing or b) not sufficiently better organ sparing than photons, planners did not further optimize this plan. However, given that dose hotspots are quite small, they did not affect dose metrics for the auto-contoured organs in our study. Finally, differences in plans were also caused because the same plan optimization process when run twice, may lead to similar, but not exactly the same solution due to randomness in initialization.

Figure 2.3 shows that of all the organs the Spinal Cord and Brainstem had wider boxplots for both $P_{OG} - P_{MC}$ and $P_{MC} - P_{AC}$. This is because the $\Delta D_{0.03\text{cc}}$ metric is inherently more sensitive to dose changes than ΔD_{mean}. This is seen in the first row of Figure 2.6 where similar DICE values for the Brainstem output vastly different dose differences. For proton (Figure 2.4), we saw a similar trend for $P_{OG} - P_{MC}$, but not for $P_{MC} - P_{AC}$. This indicated that proton planning is more susceptible to workflow differences than contour differences of Brainstem and Spinal Cord, for our cohort of oro- and hypopharyngeal cancers, which are at a distance from these organs.

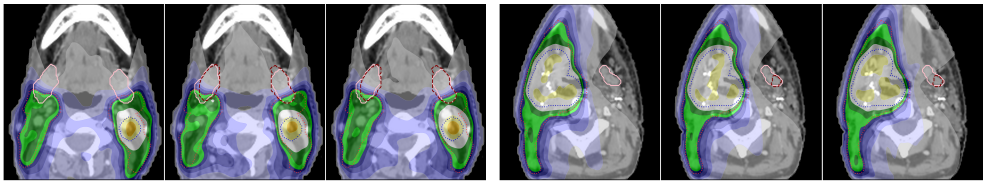
Figure 2.3a, 2.3c (photon) and Figure 2.4a, 2.4c (proton) show statistically significant differences, but from a clinical standpoint, the minor differences in organ dose metrics and ΔNTCP values may be clinically irrelevant.

Moving on to the effect of DICE on dose metric of organs (Figure 2.5), one would expect that a decrease in DICE would lead to higher ΔcGy values for organs. This was true for the Parotids, Submandibulars (Figure 2.6) and Oral Cavity across both photons and protons ($-0.43 \leq \rho_s \leq -0.17$). The Brainstem and Spinal Cord showed poor correlation scores for both forms of radiotherapy, primarily due to the sensitive nature of the $D_{0.03\text{cc}}$ metric. The Esophagus also showed low correlation, since, in many cases, it is caudally far away from the tumor regions for the patients in our cohort. The Larynx showed a high correlation for photon, but not for proton, which could be an effect of sample size. Finally, the Mandible, an organ with high DICE, showed opposite trends in photon and proton. Overall, we noticed that there was a low correlation between DICE and dose metrics.

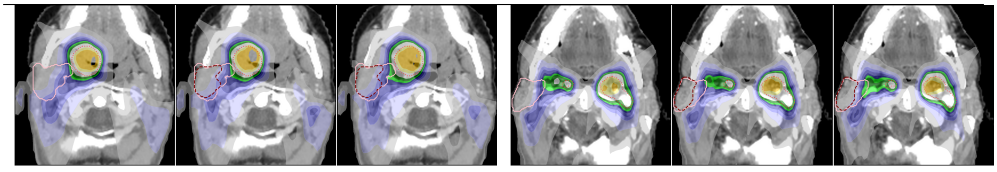
This work was inspired by prior research on treatment plan scripting [105, 106] to scale-up dose evaluation for auto-contours. However, some plans were still not of the



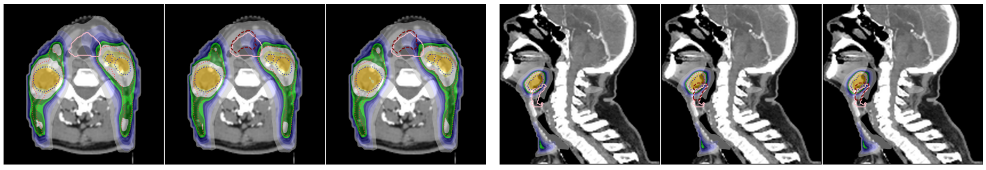
(a) Brainstem (DICE=0.13, $|\Delta D_{0.03cc}| = 6.0\%$) (b) Brainstem (DICE=0.19, $|\Delta D_{0.03cc}| = 27.2\%$)



(c) Submand (R) (DICE=0.82, $|\Delta D_{mean}| = 1.7\%$) (d) Submand (L) (DICE=0.42, $|\Delta D_{mean}| = 84.9\%$)



(e) Parotid (R) (DICE=0.85, $|\Delta D_{mean}| = 3.0\%$) (f) Parotid (R) (DICE=0.63, $|\Delta D_{mean}| = 20.5\%$)



(g) Larynx (SG) (DICE=0.64, $|\Delta D_{mean}| = 0.5\%$) (h) Larynx (SG) (DICE=0.55, $|\Delta D_{mean}| = 2.3\%$)

Figure 2.6: CT scans of photon (a-d) and proton (e-h) patients overlayed with a dose distribution as well as PTV (DL1) (orange), PTV (DL2) (blue), manual (pink) and automated (maroon) contours. Each example shows the P_{OG}, P_{MC} and P_{AC} plans from left to right. The dose metric in the sub-captions compares the absolute percentage difference of $P_{MC} - P_{AC}$.

highest possible quality since our four-step replication of the clinical process is a close, but imperfect emulation of a treatment planners approach. Non-iterative EUD optimization (step 3), lack of synchrony in weight updates between the manual and automated approach (step 4), and re-use of control structures from P_{OG} to P_{MC} and P_{AC} (step 4), led to small deviations from the original planning process. These limitations cause P_{MC} and P_{AC} dose metrics to be imprecise which could potentially impact our results. For future work we would like to more closely mimic the optimization steps as well as consider control structures specific to each plan, rather than simply copying them.

To conclude, we showed an automated approach to plan creation for retrospective

studies that was employed for the use-case of evaluating the dose impact of auto-contouring software, at scale. We hope our results showcasing low dose impact of auto-contours will inspire others to investigate and eventually use them in clinical settings.

2.5 Acknowledgement

The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022) and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate, The Netherlands.

2.6 Appendix

2.6.1 Data Acquisition

The CT scans of our dataset had a dimension of 512 x 512 pixels in the spatial plane with a pixel spacing in the range of [0.92-1.36] mm. Each CT slice was 2mm thick and each scan had between [128,199] slices. The scans were acquired from a Brilliance Big Bore (Philips Healthcare, Ohio, USA) with 120kV and 250mAs. Post acquisition, 64% of patients had Orthopedic Metal Artifact Reduction (O-MAR) processing done.

2.6.2 Automated Contours

The auto-contouring model of RayStation 10B (results in Table 2.1 and Table 2.2) first performed registration of the chosen CT scan using an atlas of CTs to narrow down CT size so it fits within the graphical processing unit (GPU) used for deep learning. Once registered, the mid-point of each OAR is detected and a 3D bounding box is cropped around that. This cropped area is then passed to a neural net trained for contouring that specific OAR. Each OAR-specific neural net is based on the UNet segmentation architecture whose output is a 3D probabilistic mask for that OAR. As a post-processing step, smoothing is performed on the surfaces of OARs. The model was trained using Tensorflow, an open-source deep neural net software package. During training, rotations, translations and elastic deformations were used to augment the training data. Details on patient cohort were not made public by the manufacturer.

RoI	DICE	SDC @ 3mm	HD95 (mm)	MSD (mm)
Spinal Cord ($D_{0.03cc}$)	0.78 [0.61,0.93]	0.92 [0.76,0.97]	10.0 [1.1,69.4]	0.9 [0.2,1.4]
Brainstem ($D_{0.03cc}$)	0.70 [0.07,0.95]	0.72 [0.18,0.95]	13.1 [2.5,49.0]	3.1 [1.1,8.3]
Parotid (L) (D_{mean})	0.85 [0.75,0.94]	0.91 [0.78,0.98]	5.0 [2.3,12.3]	1.5 [0.6,3.2]
Parotid (R) (D_{mean})	0.86 [0.74,0.94]	0.92 [0.75,0.98]	4.6 [2.2,15.7]	1.4 [0.6,4.2]
Submand (L) (D_{mean})	0.84 [0.59,0.93]	0.96 [0.74,1.00]	3.1 [1.7,16.3]	1.0 [0.5,5.3]
Submand (R) (D_{mean})	0.85 [0.68,0.92]	0.96 [0.75,1.00]	3.1 [1.7,16.3]	1.1 [0.6,3.5]
Oral Cavity (D_{mean})	0.84 [0.77,0.92]	0.74 [0.59,0.90]	7.7 [4.3,12.0]	2.6 [1.5,3.3]
Larynx (SG) (D_{mean})	0.54 [0.36,0.65]	0.63 [0.51,0.80]	15.9 [7.8,25.0]	5.7 [2.8,10.2]
Esophagus (D_{mean})	0.66 [0.28,0.90]	0.75 [0.41,0.97]	20.4 [2.5,63.9]	1.4 [0.3,18.8]
Mandible (D_{mean})	0.88 [0.81,0.97]	0.94 [0.87,1.00]	4.5 [1.1,14.0]	1.5 [0.2,3.4]

Table 2.1: Summary measures (median [5^{th} percentile, 95^{th} percentile]) for volumetric and surface metrics of auto-contours of RayStation 10B.

RoI	DICE	SDC @ 3mm	HD95 (mm)	MSD (mm)
Spinal Cord ($D_{0.03cc}$)	0.77 [0.74,0.80]	0.89 [0.87,0.91]	19.2 [13.6,24.7]	0.8 [0.7,0.9]
Brainstem ($D_{0.03cc}$)	0.61 [0.61,0.67]	0.66 [0.60,0.72]	18.0 [14.4,21.5]	3.8 [3.3,4.5]
Parotid (L) (D_{mean})	0.84 [0.84,0.86]	0.89 [0.87,0.91]	5.8 [4.8,6.8]	1.7 [1.5,1.8]
Parotid (R) (D_{mean})	0.85 [0.85,0.86]	0.89 [0.87,0.91]	5.8 [4.9,6.9]	1.7 [1.5,2.0]
Submand (L) (D_{mean})	0.80 [0.80,0.84]	0.90 [0.87,0.94]	6.2 [4.3,8.9]	2.3 [1.1,4.3]
Submand (R) (D_{mean})	0.82 [0.82,0.84]	0.92 [0.89,0.94]	4.8 [3.9,5.7]	1.4 [1.1,1.7]
Oral Cavity (D_{mean})	0.84 [0.82,0.86]	0.74 [0.71,0.76]	7.9 [7.2,8.6]	2.6 [2.4,2.9]
Larynx (SG) (D_{mean})	0.51 [0.47,0.54]	0.63 [0.58,0.67]	15.4 [13.7,17.3]	6.1 [5.3,7.0]
Esophagus (D_{mean})	0.66 [0.61,0.70]	0.75 [0.71,0.80]	23.8 [18.6,29.3]	5.8 [4.0,7.8]
Mandible (D_{mean})	0.88 [0.85,0.90]	0.94 [0.92,0.95]	6.1 [4.7,7.6]	1.6 [1.3,1.9]

Table 2.2: Summary measures (sample mean [bootstrapped 95% confidence interval]) for volumetric and surface metrics of auto-contours of RayStation 10B.

2.6.3 Automated Planning

For automated planning, we replicated the beam setup, OAR/target objectives for both photon and proton as per our institutions clinical head-and-neck protocol.

For photon (Table 2.3), our VMAT plans are made on an isotropic dose grid of 0.2cm The photon beams were commissioned on an Elekta Synergy system with Agility multi-leaf collimator.

For proton (Table 2.4), our IMPT plans are made on an isotropic dose grid of 0.3cm. This dose is delivered using pencil beam scanning (PBS) on a Varian ProBeam machine.

Step	RoI	Function	Description	Weight
1	PTV (DL1)	MinDose	100% of DL1 prescription	80.0 → *{VDT}*
1	PTV (DL1)	MaxDose	102% of DL1 prescription	50.0 → *{VDT}*
1	ring ≤ PTV (DL1)	MaxDose	96% of DL1 prescription	0.0 → *{VDT}*
1	PTV (DL2)	MinDose	100% of DL2 prescription	80.0 → *{VDT}*
1	PTV (DL2)	MaxDose	102% of DL2 prescription	50.0 → *{VDT}*
1	PTV (DL2)	UniformDose	100% of DL2 prescription	10.0
1	Body	DoseFallOff	From 100% to 0% of DL1 prescription over 5.0 cm	1.0
1	Body	DoseFallOff	From 100% to 26% of DL1 prescription over 2.0 cm	2.0
1	Body	DoseFallOff	From 100% to 64% of DL1 prescription over 0.5 cm	10.0
1	$\text{Ghost}_{\text{Cranial}}$	DoseFallOff	From 100% to 0% of DL1 prescription over 1.0 cm	0.5
1	$\text{Ghost}_{\text{Ear(L)}}$	DoseFallOff	From 100% to 46% of DL1 prescription over 2.0 cm	1.0
1	$\text{Ghost}_{\text{Ear(R)}}$	DoseFallOff	From 100% to 46% of DL1 prescription over 2.0 cm	1.0
1	Brainstem	MaxEUD	eudParameterA=50 (maxEUD=4000 cGy)	3.0
1	Brainstem (+3 cm)	MaxEUD	eudParameterA=50 (maxEUD=4400 cGy)	3.0
1	Spinal Cord	MaxEUD	eudParameterA=50 (maxEUD=4000 cGy)	3.0
1	Spinal Cord (+3 cm)	MaxEUD	eudParameterA=50 (maxEUD=4400 cGy)	3.0
2.1	Other Organs	DoseFallOff	From 100% to 20% of DL1 prescription over 2.0 cm	1.0
2.2	Other Organs	DoseFallOff	From 100% to 0% of DL1 prescription over 2.0 cm (*as determined by treatment planner*)	1.0
3	Other Organs	MaxEUD	eudParameterA=50, maxEUD={*VDT*}	1.0
4	Control Structures	{MinDose, MaxDose}	Dose=*{VDT}*	*{VDT}*

Table 2.3: Our 4-step emulation of the manual photon optimization process of our clinic. In each step, we also optimize for the objectives of the previous steps. We use *VDT* as an abbreviation for the phrase “value determined by treatment planner". The → indicates that the weight is modified at the end of Step 4.. Here DL1/DL2 stands for electives/boost regions of the tumor and prescription refers to a value of cGy that was assigned to a region-of-interest (RoI). Here “Other Organs" refers to Cochlea (L/R), Parotid (L/R). Submandibular (L/R), Muscle Constrictor (S/M/I), Cricopharyngeus, Larynx (SG), Glottic Area, Trachea, Esophagus and Oral Cavity. The rows shown here are created as objectives in our clinic's treatment planning solution.

Step	RoI	Function	Description	Weight	Robust
1	CTV (DL1)	MinDose	100% of DL1 prescription	800.0 → *{VDT}*	∗
1	CTV (DL1) - (CTV(DL2) + 3 mm)	MaxDose	102% of DL1 prescription	20.0 → *{VDT}*	∗
1	CTV (DL1) - (CTV(DL2) + 2 cm)	MaxDose	102% of DL1 prescription	80.0 → *{VDT}*	∗
1	CTV (DL2)	MinDose	100% of DL2 prescription	800.0 → *{VDT}*	∗
1	CTV (DL2)	MaxDose	100% of DL2 prescription	50.0 → *{VDT}*	∗
1	CTV (L)	MinDose	0 cGy and Beam={1,2,3}	0.0	
1	CTV (R)	MinDose	0 cGy and Beam={4,5,6}	0.0	
1	Body	DoseFallOff	From 101% to 0% of DL2 prescription over 2.0 cm	1.0	
1	Body	MaxDose	67% of DL2 prescription for each beam	10000.0	
1	Body	MaxDose	107% of DL2 prescription	100.0	∗
2	Mandible	MaxDose	107% of DL2 prescription	500.0 → *{VDT}*	∗
2	Organ Set 1	DoseFallOff	From 101% to 0% of DL2 prescription over 2.0 cm	1.0	
2	Organ Set 2	DoseFallOff	From 101% to 0% of DL2 prescription over 2.0 cm	1.0	
3.1	Organ Set 2	MaxEUD	eudParameterA=1, maxEUD={*VDT*}	1.0	
3.2	Organ Set 2 - (CTV (DL1) + 3 mm)	MaxEUD	eudParameterA=1, maxEUD={*VDT*}	1.0	
4	Control Structure	{MinDose, MaxDose}	Dose=*{VDT}*	*{VDT}*	{∗}

Table 2.4: Our 4-step emulation of the manual proton optimization process of our clinic. In each step, we also optimize for the objectives of the previous steps. We use *VDT* as an abbreviation for the phrase “value determined by treatment planner". The → indicates that the weight is modified at the end of Step 4.. Here DL1/DL2 stands for elective/boost regions of the CTV and prescription refers to a value in cGy that was assigned to a region-of-interest (RoI). “Organ Set 1" refers to Mandible, Brainstem, Spinal Cord, Esophagus, Trachea, Larynx (SG), Trachea and Glottic Area, while “Organ Set 2" refers to Parotid (L/R), Submandibular (L/R), Muscle Constrictor (S/M/I), and Oral Cavity. The ∗ mark is used to indicate those objectives which are robustly optimized. The rows shown here are created as objectives in our clinic’s treatment planning solution.

2.6.4 Organ Dose Metrics

We show dose metrics for organs available in the RayStation 10B auto-contouring module for photon (Table 2.5 and Table 2.6) and proton (Table 2.7 and Table 2.8). For the purpose of our study, we only included organs with available auto-contours, although additional organs-at-risk are evaluated clinically.

RoI	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Spinal Cord ($D_{0.03cc}$)	1.45 [0.06,5.51]	1.13 [0.18,5.16]
Brainstem ($D_{0.03cc}$)	1.88 [0.05,6.77]	2.17 [0.21,6.37]
Parotid (L) (D_{mean})	0.12 [0.02,0.72]	0.32 [0.02,2.10]
Parotid (R) (D_{mean})	0.13 [0.01,0.68]	0.42 [0.03,1.66]
Submand (L) (D_{mean})	0.27 [0.02,1.20]	0.45 [0.05,2.37]
Submand (R) (D_{mean})	0.21 [0.01,1.28]	0.35 [0.04,1.80]
Oral Cavity (D_{mean})	3.24 [0.01,0.86]	0.35 [0.05,1.32]
Larynx (SG) (D_{mean})	0.39 [0.03,1.47]	0.39 [0.21,4.24]
Esophagus (D_{mean})	0.24 [0.01,1.64]	0.65 [0.04,3.43]
Mandible ($D_{2\%}$)	0.37 [0.03,3.43]	0.43 [0.06,2.12]

Table 2.5: Median [5^{th} percentile, 95^{th}percentile] of the absolute dose metric values (in Gy) for $P_{OG} - P_{MC}$ and $P_{MC} - P_{AC}$ in photon radiotherapy.

RoI	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Spinal Cord ($D_{0.03cc}$)	2.01 [1.51,2.56]	1.90 [1.49,2.32]
Brainstem ($D_{0.03cc}$)	2.43 [1.90,3.01]	2.82 [2.36,3.34]
Parotid (L) (D_{mean})	0.21 [0.15,0.28]	0.66 [0.49,0.85]
Parotid (R) (D_{mean})	0.21 [0.15,0.27]	0.62 [0.48,0.80]
Submand (L) (D_{mean})	0.39 [0.30,0.49]	0.80 [0.52,1.22]
Submand (R) (D_{mean})	0.33 [0.23,0.45]	0.59 [0.42,0.80]
Oral Cavity (D_{mean})	0.32 [0.24,0.42]	0.49 [0.40,0.58]
Larynx (SG) (D_{mean})	0.55 [0.39,0.74]	1.65 [1.25,2.07]
Esophagus (D_{mean})	0.41 [0.29,0.54]	1.05 [0.80,1.38]
Mandible ($D_{2\%}$)	0.81 [0.48,1.22]	0.97 [0.54,1.60]

Table 2.6: Sample mean [bootstrapped 95% confidence interval] of the absolute dose metric values (in Gy) for $P_{OG} - P_{MC}$ and $P_{MC} - P_{AC}$ in photon radiotherapy.

RoI	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Spinal Cord ($D_{0.03cc}$)	2.08 [0.03,8.82]	0.70 [0.12,2.40]
Spinal Cord ($D_{0.03cc}$) (vw-max)	1.90 [0.05,8.07]	0.72 [0.15,2.57]
Brainstem ($D_{0.03cc}$)	0.72 [0.05,3.79]	0.59 [0.03,2.77]
Brainstem ($D_{0.03cc}$) (vw-max)	0.98 [0.13,4.30]	1.00 [0.19,2.81]
Parotid (L) (D_{mean})	0.10 [0.02,0.39]	0.48 [0.07,1.99]
Parotid (R) (D_{mean})	0.14 [0.01,0.43]	0.40 [0.03,1.80]
Submand (L) (D_{mean})	0.21 [0.06,0.79]	0.28 [0.05,1.85]
Submand (R) (D_{mean})	0.18 [0.03,0.70]	0.27 [0.01,1.89]
Oral Cavity (D_{mean})	0.08 [0.02,0.39]	0.31 [0.03,0.73]
Larynx (SG) (D_{mean})	0.37 [0.01,1.36]	0.56 [0.19,3.26]
Esophagus (D_{mean})	0.31 [0.01,3.03]	0.23 [0.07,0.77]
Mandible ($D_{2\%}$)	0.44 [0.01,2.19]	0.79 [0.06,2.92]
Mandible ($D_{2\%}$) (vw-max)	0.52 [0.01,2.98]	0.46 [0.08,2.13]

Table 2.7: Median [5^{th} percentile, 95^{th}percentile] of the absolute dose metric values (in Gy) for $P_{OG} - P_{MC}$ and $P_{MC} - P_{AC}$ in proton radiotherapy.

RoI	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Spinal Cord ($D_{0.03cc}$)	2.92 [1.93,4.00]	0.92 [0.65,1.20]
Spinal Cord ($D_{0.03cc}$) (vw-max)	2.93 [1.92,4.06]	1.08 [0.79,1.40]
Brainstem ($D_{0.03cc}$)	1.07 [0.67,1.54]	0.89 [0.60,1.20]
Brainstem ($D_{0.03cc}$) (vw-max)	1.35 [0.90,1.84]	1.27 [0.92,1.70]
Parotid (L) (D_{mean})	0.16 [0.11,0.21]	0.63 [0.43,0.87]
Parotid (R) (D_{mean})	0.15 [0.11,0.20]	0.62 [0.41,0.86]
Submand (L) (D_{mean})	0.32 [0.20,0.47]	0.51 [0.32,0.73]
Submand (R) (D_{mean})	0.27 [0.18,0.37]	0.71 [0.29,1.41]
Oral Cavity (D_{mean})	0.15 [0.10,0.21]	0.34 [0.26,0.42]
Larynx (SG) (D_{mean})	0.59 [0.39,0.83]	0.88 [0.54,1.30]
Esophagus (D_{mean})	0.75 [0.42,1.19]	0.34 [0.25,0.45]
Mandible ($D_{2\%}$)	0.88 [0.49,1.40]	1.00 [0.69,1.34]
Mandible ($D_{2\%}$) (vw-max)	0.95 [0.58,1.36]	0.79 [0.54,1.08]

Table 2.8: Sample mean [bootstrapped 95% confidence interval] of the absolute dose metric values (in Gy) for $P_{OG} - P_{MC}$ and $P_{MC} - P_{AC}$ in proton radiotherapy.

2.6.5 NTCP

For NTCP scores (Table 2.9 and Table 2.10), we used the formulae and parameters from the National Indication Protocol for Proton therapy (*Landelijk Indicatie Protocol Protonentherapie*) [103]. From this document, we referred to Section 3.3.3 and 3.3.4 for xerostomia and Section 3.4.3 and 3.4.4 for dysphagia. For all four toxicities, we used a baseline score of 0.

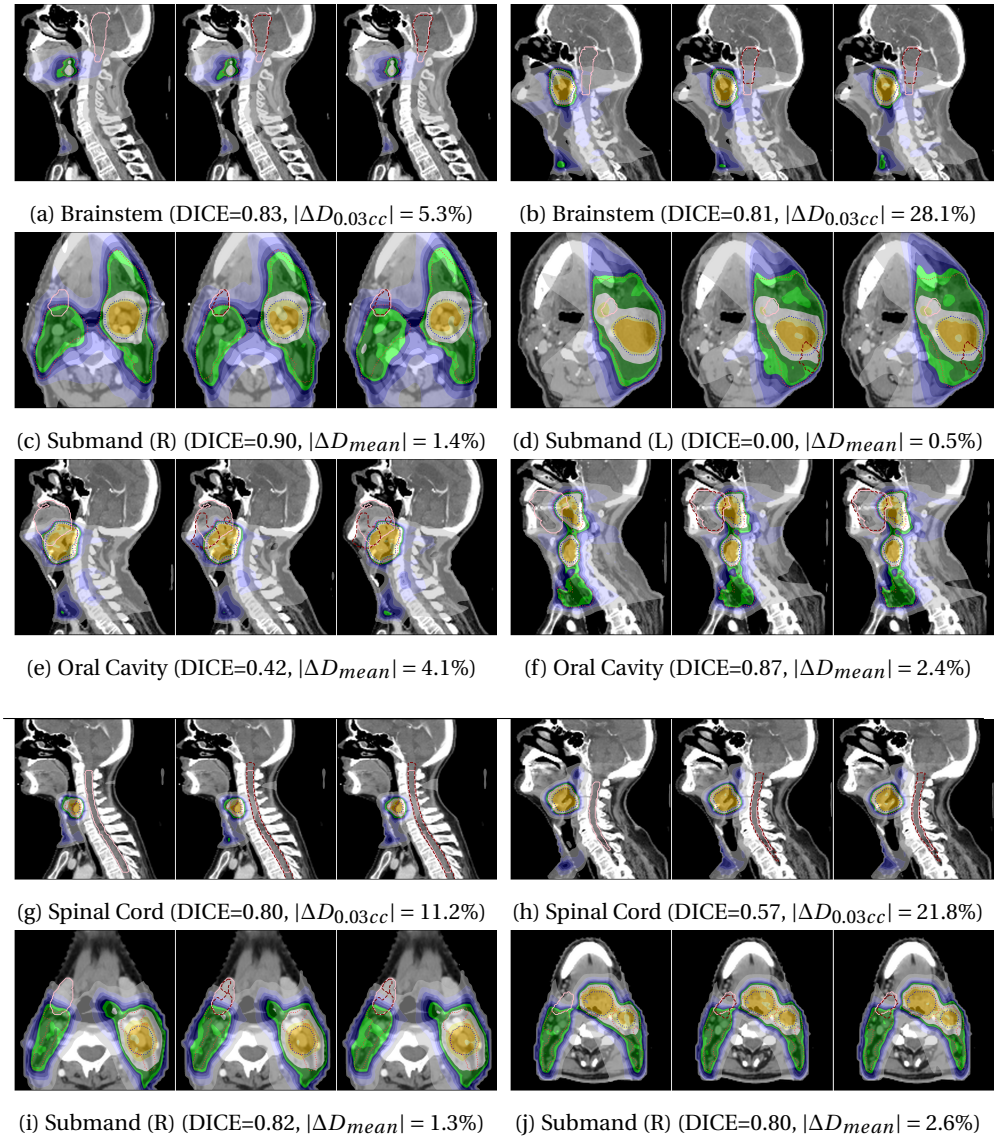
	Photon		Proton	
	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Xerostomia Grade ≥ 2	0.1 [0.0,0.5]	0.3 [0.0,0.9]	0.1 [0.0,0.3]	0.2 [0.0,1.0]
Xerostomia Grade ≥ 3	0.0 [0.0,0.2]	0.1 [0.0,0.3]	0.0 [0.0,0.1]	0.1 [0.0,0.3]
Dysphagia Grade ≥ 2	0.2 [0.0,0.9]	0.2 [0.0,0.6]	0.0 [0.0,0.3]	0.1 [0.0,0.3]
Dysphagia Grade ≥ 3	0.1 [0.0,0.7]	0.1 [0.0,0.5]	0.0 [0.0,0.1]	0.0 [0.0,0.1]

Table 2.9: Summary measures (median [5^{th} percentile, 95^{th} percentile]) for ΔNTCP (%) values in photon and proton radiotherapy for $\lvert P_{OG} - P_{MC} \rvert$ and $\lvert P_{MC} - P_{AC} \rvert$.

	Photon		Proton	
	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$	$\lvert P_{OG} - P_{MC} \rvert$	$\lvert P_{MC} - P_{AC} \rvert$
Xerostomia Grade ≥ 2	0.2 [0.1,0.2]	0.4 [0.3,0.4]	0.1 [0.1,0.2]	0.3 [0.2,0.5]
Xerostomia Grade ≥ 3	0.1 [0.0,0.1]	0.1 [0.1,0.2]	0.0 [0.0,0.1]	0.1 [0.1,0.2]
Dysphagia Grade ≥ 2	0.3 [0.2,0.4]	0.2 [0.2,0.3]	0.1 [0.1,0.1]	0.1 [0.1,0.1]
Dysphagia Grade ≥ 3	0.2 [0.1,0.3]	0.2 [0.1,0.2]	0.0 [0.0,0.0]	0.0 [0.0,0.0]

Table 2.10: Sample mean [bootstrapped 95% confidence interval]) for ΔNTCP (%) values in photon and proton radiotherapy for $\lvert P_{OG} - P_{MC} \rvert$ and $\lvert P_{MC} - P_{AC} \rvert$.

2.6.6 Visual Results



(a) Brainstem (DICE=0.83, $|\Delta D_{0.03cc}| = 5.3\%$)

(b) Brainstem (DICE=0.81, $|\Delta D_{0.03cc}| = 28.1\%$)

(c) Submand (R) (DICE=0.90, $|\Delta D_{mean}| = 1.4\%$)

(d) Submand (L) (DICE=0.00, $|\Delta D_{mean}| = 0.5\%$)

(e) Oral Cavity (DICE=0.42, $|\Delta D_{mean}| = 4.1\%$)

(f) Oral Cavity (DICE=0.87, $|\Delta D_{mean}| = 2.4\%$)

(g) Spinal Cord (DICE=0.80, $|\Delta D_{0.03cc}| = 11.2\%$)

(h) Spinal Cord (DICE=0.57, $|\Delta D_{0.03cc}| = 21.8\%$)

(i) Submand (R) (DICE=0.82, $|\Delta D_{mean}| = 1.3\%$)

(j) Submand (R) (DICE=0.80, $|\Delta D_{mean}| = 2.6\%$)

Figure 2.7: This figure shows CT scans of photon (a-f) and proton (g-j) patients overlayed with a dose distribution as well as PTV (DL1) (orange), PTV (DL2) (blue), manual (pink) and automated (maroon) contours. Each example shows the P_{OG}, P_{MC} and P_{AC} plans from left to right. The dose metric in the sub-captions compares the absolute percentage difference of $P_{MC} - P_{AC}$.

3

Comparing Bayesian Models for Organ Contouring in Head and Neck Radiotherapy

This chapter was adapted from:

Mody, Prerak P., Nicolas Chaves-de-Plaza, Klaus Hildebrandt, René van Egmond, Huib de Ridder, and Marius Staring. "Comparing Bayesian models for organ contouring in head and neck radiotherapy." In *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 100-109. SPIE, 2022.

Abstract

Deep learning models for organ contouring in radiotherapy are poised for clinical usage, but currently, there exist few tools for automated quality assessment (QA) of the predicted contours. Bayesian models and their associated uncertainty, can potentially automate the process of detecting inaccurate predictions. We investigate two Bayesian models for auto-contouring, DropOut and FlipOut, using a quantitative measure – expected calibration error (ECE) and a qualitative measure – region-based accuracy-vs-uncertainty (R-AvU) graphs. It is well understood that a model should have low ECE to be considered trustworthy. However, in a QA context, a model should also have high uncertainty in inaccurate regions and low uncertainty in accurate regions. Such behaviour could direct visual attention of expert users to potentially inaccurate regions, leading to a speed-up in the QA process. Using R-AvU graphs, we qualitatively compare the behaviour of different models in accurate and inaccurate regions. Experiments are conducted on the MICCAI2015 Head and Neck Segmentation Challenge and on the DeepMindTCIA CT dataset using three models: DropOut-DICE, Dropout-CE (Cross Entropy) and FlipOut-CE. Quantitative results show that DropOut-DICE has the highest ECE, while Dropout-CE and FlipOut-CE have the lowest ECE. To better understand the difference between DropOut-CE and FlipOut-CE, we use the R-AvU graph which shows that FlipOut-CE has better uncertainty coverage in inaccurate regions than DropOut-CE. Such a combination of quantitative and qualitative metrics explores a new approach that helps to select which model can be deployed as a QA tool in clinical settings.

3.1 Introduction

Radiotherapy is an important cancer treatment option due to its ability to treat cancerous tissue while simultaneously sparing healthy tissue [115]. During treatment planning there is a requirement to acquire diagnostic 3D images like CT, MR and PET scans and contour the healthy tissue or organs at risk (OAR) as well as tumorous tissue. This contouring task is time-consuming and is also subject to inter- and intra-annotator disagreement [6, 8]. As deep learning models have made great progress in this field [23, 24, 26, 28, 29] they are widely being considered as an automated technique to speed up and standardize the contouring process [34, 35]. However, to deploy such models in a clinical setting, a manual quality assessment (QA) of predicted contours needs to be performed before they can be used for radiation dosage calculation, which again, introduces a delay. This work investigates the potential usage of uncertainty heatmaps produced by Bayesian deep learning models to help speed up the manual QA process for OARs, by directing human attention to inaccurately segmented regions.

Organ contours are extracted by classifying the 3D voxels of a scan into different categories. It is well accepted that for a predictive classification model to be trusted, it should be calibrated. This means that its output confidence (i.e. probability value) should correspond to the likelihood of being accurate. In other words, in a calibrated model, voxels predicted to belong to an OAR with probability p, should have an accuracy equal to p. It has been previously shown that well-calibrated model confidences also produce uncertainty measures that correspond to inaccurate regions [76, 77]. Such a property may be useful in a radiotherapy QA context to direct visual attention of clinicians to inaccurate regions. Thus, this work further investigates this claim, for the purpose of choosing a model for clinical deployment, by analysing two deep Bayesian models - DropOut [116] and FlipOut [117]. Bayesian models were chosen as they offer a principled approach to capture uncertainty. We use a combination of a commonly used quantitative metric for model confidence calibration - expected calibration error (ECE) [118] and propose a new qualitative metric for uncertainty calibration - region-based accuracy-vs-uncertainty (R-AvU) graphs. Motivated by the observation that some models may provide us with similar ECE values, we use the R-AvU graphs to understand the differences in their uncertainty behavior. Previous uncertainty evaluation metrics like AvU [119] provide a single scalar value by performing an analysis on the accuracy and uncertainty of each voxel in a scan. To achieve a perfect AvU score, a model must have only accurate and certain or inaccurate and uncertain voxels, i.e. perfectly calibrated uncertainty. We believe this metric has the right motivation, but its formulation may not be sufficient from a QA perspective as it does not offer clear insight into the uncertainty calibration in accurate and inaccurate regions. Such region-specific insight is useful as high uncertainty in inaccurate regions and low uncertainty in accurate regions can provide heatmaps that could help direct visual

attention during QA. Hence, the R-AvU graph uses the building blocks of the AvU metric and plots the uncertainty probabilities in accurate and inaccurate regions across a range of uncertainty thresholds. We use entropy as an uncertainty metric in our experiments, which has been previously shown to capture both data and model uncertainty [120].

3.2 Method

3.2.1 Data

CT scans along with annotations for 9 organs at risk (OAR) in the head-and-neck area were used from the MICCAI 2015-Head and Neck Segmentation Challenge dataset [20]. This dataset provided 33 training and 10 test samples from the RTOG 0522 clinical trial [121]. Models trained on this dataset were also evaluated on a separate dataset titled DeepMindTCIA [22] which contains 15 patients. The DeepMindTCIA dataset also refers to the RTOG 0522 clinical trial along with the TCGA-HNSC [21] collection on The Cancer Imaging Archive (TCIA). Duplicate RTOG 0522 patients were removed from the DeepMind TCIA dataset if they were already present in the MICCAI dataset. Each CT volume is resampled to a resolution of (0.8, 0.8, 2.5) mm and cropped with a bounding box of dimensions (240,240,80) around the brainstem. The resampling and subsequent training was done at a fixed resolution so that it is convenient for the convolution kernels to learn anatomical feature extraction. The scans were cropped around the brainstem to reduce the computational complexity of patch extraction. The Hounsfield units were trimmed from -125 to +225 to better capture contrast for soft tissues. The models consumed random 3D patches of size (140,140,40) that were augmented with 3D translations, 3D rotations, 3D elastic deformations and Gaussian noise.

3.2.2 Neural Architecture

The base convolutional neural network (CNN) of our Bayesian models is inspired by FocusNet [24], a deterministic model. This model is a standard encoder-decoder architecture that uses Squeeze and Excitation [122] modules for improved feature extraction via channel attention, a DenseASPP [123] module to obtain sufficient receptive field and finally a supplementary network to prevent foreground-background imbalance for smaller organs at risk (OAR) like optic nerves and optic chiasm. Our implementation avoids the supplementary network for the sake of simplicity. We add Bayesian layers in the DenseASPP module which forms the middle layers of FocusNet.

A choice of either DropOut [124] or FlipOut [125] layers were used for Bayesian modelling. Bayesian modelling of a predictive model involves placing a prior over the models weights $p(W)$ and updating its posterior $p(W|D)$ via observations $D = (X, Y)$ where X and Y are training inputs and outputs respectively. Learning a distribution over the model weights, instead of simply learning fixed scalar values, helps us capture how much the output can vary when provided some input. Thus, Bayesian modelling helps us infer the

output distribution $p(y|x,D)$ where x is a test sample and y is its associated output by marginalizing over the posterior:

$$p(y|x,D) = \mathbb{E}_{W\sim p(W|D)}[p(y|x,W)]. \tag{3.1}$$

Theoretically, the DropOut model estimates the posterior distribution of a deep Gaussian process (*a Bayesian inference tool*) by placing a Bernoulli distribution with parameter p_d on the neural net weights. This was shown to be equivalent to performing dropout on the outputs of the layer that those weights belong to. Here output refers to the result of a convolution operation i.e. $w_h * x_h$, where w_h is the kernel weight and x_h is the input in some hidden layer and dropout refers to randomly setting this output to zero with probability p_d. FlipOut on the other hand assumes the weight distribution to be Gaussian. In practice, Monte-Carlo sampling via multiple forward passes is used to estimate or infer $p(y|x,D)$. Thus, in every forward pass, Dropout and FlipOut perform output space and weight space perturbations respectively. This is because during each forward pass the DropOut model drops outputs randomly while the FlipOut model samples new weights from a Gaussian distribution. Our DropOut model contains ~500k parameters, while the FlipOut model contains twice those parameters due to the Gaussian assumption. We chose a fixed probability of $p_d = 0.25$ for the Dropout model.

3.2.3 Training and Inference

During a single forward pass, the models produce 3D probability maps for each OAR, with each voxel being represented by a vector containing probability values for each OAR that sum to 1. An argmax operator is applied on each voxel's probability vector to assign it an OAR. For each OAR, we assume its 3D predicted probability map to be P_c and the corresponding ground truth probability map to be $Y_c = \{0,1\}$, where $c \in C$ stands for OAR class id. The models are trained using either soft-DICE [126] or cross-entropy (CE) loss, which is calculated for each OAR and then averaged to calculate the gradient for back propagation. During training, we perform only a single forward pass to calculate the loss. The DICE loss is calculated as follows:

$$DICE_c = \frac{2\sum_{i=1}^{N}(P_c^i Y_c^i)}{\sum_{i=1}^{N} P_c^i + \sum_{i=1}^{N} Y_c^i}, \tag{3.2}$$

$$L_{DICE} = 1 - \frac{1}{C}\left(w_c \sum_{c=1}^{C} DICE_c\right), \tag{3.3}$$

where P_c^i represents the predicted probability of one of N voxels, Y_c^i is its corresponding ground truth and w_c is the weight assigned to each class. We use a weighted approach since the OARs in the head and neck region suffer from an imbalanced class problem.

The weights are inversely proportional to the average voxel count of each OAR. Similar to DICE, the standard CE loss only penalizes the foreground of each organs probability map i.e. $\mathbb{1}_{\{Y_c=1\}}$. Our modified CE loss inspired by [127] also penalizes the background i.e. $\mathbb{1}_{\{(1-Y_c)=1\}}$ of these probability maps for additional supervision as follows:

$$CE_{foreground} = \sum_{i=1}^{N} (\mathbb{1}_{\{Y_c^i=1\}} \ln(P_c^i)) \tag{3.4}$$

$$CE_{background} = \sum_{i=1}^{N} (\mathbb{1}_{\{(1-Y_c^i)=1\}} \ln(1-P_c^i) \tag{3.5}$$

$$L_{CE} = \frac{1}{C}\left(w_c \sum_{c=1}^{C} (CE_{foreground} + CE_{background})\right), \tag{3.6}$$

which showed improved performance when compared to using the standard CE loss.

To train the FlipOut model, one minimizes the CE loss as well as the KL-Divergence term between the Gaussian prior $p(w)$ and the estimated posterior $p(w|D)$ [125]. For inferring the predictive distribution $p(y|x,D)$ from the model posterior $p(W|D)$, Monte Carlo sampling is performed. We perform $M = 30$ forward passes, each time sampling from the posterior to produce 3D activation maps $(P_c)_m$ for each OAR. These are then averaged $(\bar{P_c})$ and passed through the argmax operator to yield the output $\hat{Y}$ containing OAR ids.

$$\bar{P_c} = \frac{1}{M} \sum_{m=1}^{M} (P_c)_m \tag{3.7}$$

$$\hat{Y} = \arg\max_{c=1}^{C} [\bar{P_c}] \tag{3.8}$$

We train and evaluate 4 Bayesian models c.f. DropOut-CE-Basic, DropOut-DICE, DropOut-CE and FlipOut-CE along with some deterministic variants. Here Dropout-CE-Basic is the model trained with the foreground-only cross entropy loss while DropOut-CE is trained with the modified-CE loss described above. In the deterministic (i.e. non-Bayesian) variants c.f. DropOut-DICE-Det and DropOut-CE-Det, only a single forward pass (i.e $M = 1$) is performed. A deterministic analysis on FlipOut-CE is not done as its design leads to new weights being sampled in every forward pass. The models were trained for a 1000 epochs with the Adam optimizer and a fixed learning rate of 0.001, with one epoch looping over 33 patients in the MICCAI2015 training subset.

3.2.4 Uncertainty

Using the probability maps $(P_c)_m$ of each OAR, we compute the entropy maps and use them as uncertainty maps. Entropy is a term derived from information theory that captures the average amount of uncertainty present in a signal. Thus, if Monte Carlo (M) sampling in a Bayesian network leads to highly varying probability vectors for a voxel,

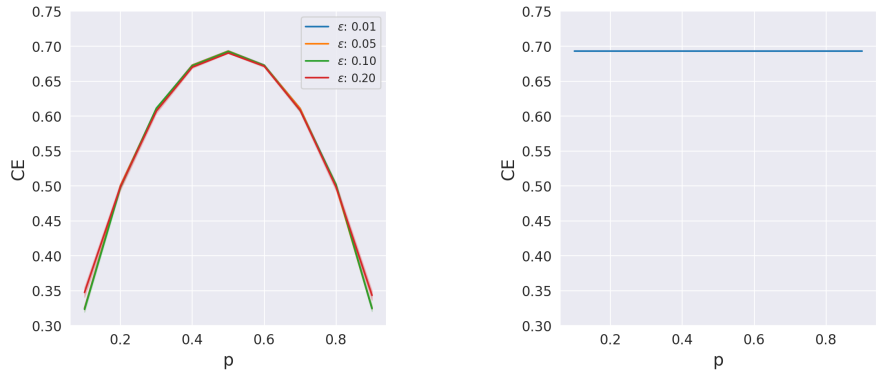


Figure 3.1: These figures show the behaviour of entropy for a simple binary classification problem of one voxel. Here p represents the foreground class probability and ϵ refers to the amount of output probability variability across Monte Carlo runs. The left figure shows uncertainty behavior when the output probability has some variability, while the right figure shows uncertainty behaviour in case of extreme probability changes.

it would have higher entropy. To calculate the 3D entropy map $H(y|x,D)$, we use the averaged probability heatmaps $\bar{P}_c$ of each OAR:

$$H(y|x,D) = -\sum_{i=1}^{C} \bar{P}_c \cdot \log(\bar{P}_c), \tag{3.9}$$

which has a maximum value when the average probability vector $\bar{P}_c^i$ for each voxel i has all its values as $\frac{1}{C}$. In our case of C=10 (9 OARs + background), the maximum entropy value is 2.3.

In Figure 3.1, we use a toy binary classification problem (e.g. foreground vs background classification for a single voxel) to understand the behavior of these metrics. In the left figure, we add uniform variability parameterized by ϵ to the foreground class probability p to replicate possible Monte Carlo outputs. Here, entropy is maximum at $p = 0.5$, i.e the model assigns equal probability to both foreground and background. It is lowest when the model is confident in its predictions i.e. $p = \{0,1\}$. Also, while increasing the amount of variation across different Monte Carlo outputs, there is no behavioral change in entropy as seen by the overlap of the curves. In the right figure, we investigate an extreme case wherein Monte Carlo sampling outputs probabilities such as $[p, 1-p, p, \ldots]$. This replicates extreme probability swings which might represent the case of a boundary voxel between an OAR and background where contrast is poor and hence the model is uncertain. Such outputs maximize the entropy across all probabilities.

3.2.5 Evaluation

For evaluation, we use two metrics: the expected calibration error (ECE) [118] for model confidence calibration and then region-based accuracy-vs-uncertainty (R-AvU) for uncertainty calibration. For e.g. in a foreground-background classification problem, if 100 voxels are assigned the foreground class with 70% probability, then we should expect that 70 of those voxels have been assigned the correct class. The error between the model confidence and its accuracy is considered as calibration error. When the same is averaged across multiple probability bins, we obtain the expected calibration error. Specifically, for each OAR, we calculate ECE_c by assigning the probability of each predicted OAR voxel i to one of B=10 equally spaced bins (B_p) between 0 and 1 as follows:

$$\mathrm{ECE}_c = \frac{1}{B}(\mathrm{acc}(B_p) - \mathrm{conf}(B_p)), \tag{3.10}$$

$$\mathrm{acc}(B_p) = \frac{1}{|B_p|} \sum_{i \in B_p} \mathbb{1}_{\hat{Y}_c = Y_c}, \tag{3.11}$$

$$\mathrm{conf}(B_p) = \frac{1}{|B_p|} \sum_{i \in B_p} (P_c)_i. \tag{3.12}$$

Here Y_c is the ground truth map, $\hat{Y}_c$ is the predicted map and P_c is the probability map belonging to a particular OAR. The lower the ECE values, the more calibrated a model is. Finally, to compute the R-AvU graphs we use uncertainty heatmaps to create line plots for the probability of uncertainty in inaccurate ($p(u|i)$) regions as well as the probability of uncertainty in accurate ($p(u|a, \sim a)$) regions. In the context of this graph, each voxel has two properties: its accuracy and uncertainty. Each voxel is then categorized as n_{ac}, n_{au}, n_{ic} and n_{iu} where n stands for the number of voxels, a for accurate, i for inaccurate, c for certain and u for uncertain. Using these terms, we find the two curves in the R-AvU graph

$$p(u|i) = \frac{n_{ui}}{n_{iu} + n_{ic}} \tag{3.13}$$

$$p(u|a, \sim a) = \frac{n_{au}}{n_{au} + n_{ac}} \tag{3.14}$$

We define accurate regions as those containing true positive (TP) voxels. We include the $\sim a$ term to denote *almost* TP voxels, as due to inter- and intra-observer variation, it is common to disregard false positive (FP) and false negative (FN) voxels very close to the ground truth contours. This is done by an erosion followed by a dilation on the inaccurate regions using a (3,3,1) filter which removes any small regions of error. The remaining FP and FN voxels are then considered as the inaccurate regions. Such an interpretation may be useful for radiotherapy QA, where smaller contouring errors may not have significant downstream effects on the calculated radiation dose for healthy tissue. Thus, such areas can be considered accurate enough and it is preferable from a visual attention standpoint that a model has lower uncertainty in these regions.

3.3 Results

3.3.1 Volumetric Performance

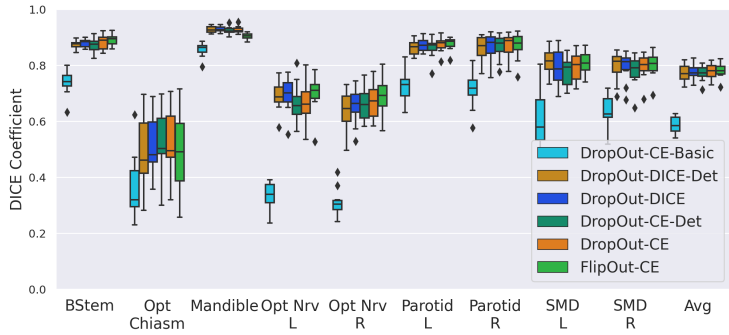
Figure 3.2 shows OAR DICE scores for the MICCAI 2015 test dataset on the left and for the DeepMindTCIA dataset on the right. For both datasets, the mandible and the brainstem (BStem) achieve the highest scores followed closely by the parotid and submandibular (SMD) glands while the optic organs (Opt Nrv L, Opt Nrv R and Opt Chiasm) have lower DICE scores overall. In the DeepMindTCIA dataset, we see various outliers for the right submandibular gland (SMD R). For the MICCAI 2015 test dataset, all models, except DropOut-CE-Basic have equivalent average performance in terms of standard medical segmentation metrics, i.e DICE (~0.77 - 0.78) and Hausdorff Distance 95% (~5mm - 7mm). We run a Wilcoxon signed-rank test on the Bayesian models and achieve p-values of 0.625 between DropOut-DICE and DropOut-CE, 0.275 between DropOut-DICE and FlipOut-CE and 1.0 between DropOut-CE and FlipOut-CE for the average DICE scores. For the average Hausdorff Distance 95% we achieve p-values of 0.027 between DropOut-DICE and DropOut-CE, 0.375 between DropOut-DICE and FlipOut-CE and 0.232 between DropOut-CE and FlipOut-CE. The results indicate that for the most part the models are not significantly different. Thus, we may compare these models using other metrics such as Expected Calibration Error (ECE) and Region-Accuracy vs Uncertainty (R-AvU). No statistical tests or additional metrics were used to study the DropOut-CE-Basic model due to its poor performance on average DICE (0.58) and average Hausdorff Distance 95% (15.95mm). Tensorflow [128] code to reproduce these results can be found at https://github.com/prerakmody/hansegmentation-uncertainty-qa.

3.3.2 Expected Calibration Error

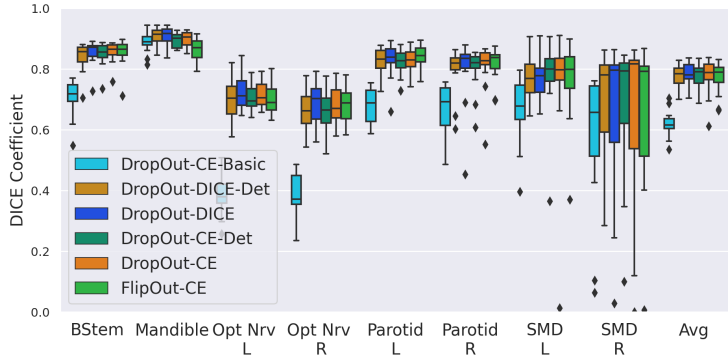
Figure 3.3 shows for both datasets that Dropout-DICE and DropOut-CE always have lower ECE than their deterministic counterparts Dropout-Dice-Det and DropOut-CE-Det. DropOut-CE on average has a lower ECE than DropOut-DICE, while FlipOut-Det and FlipOut-CE have similar ECE. The same holds for DropOut-CE and FlipOut-CE. For organs, we notice that the optic organs have the highest ECE compared to other organs for both datasets. The submandibular glands (SMD L and SMD R) and the right parotid gland have outliers in the DeepMindTCIA dataset as shown on the right side of Figure 3.3.

3.3.3 Region - Accuracy vs Uncertainty

Figure 3.4 represents $p(u|i)$ as a solid line plot and $p(u|a, \sim a)$ as a dotted line plot for entropy. A model for efficient QA would have high $p(u|i)$ and low $p(u|a, \sim a)$. The $p(u|i)$ and $p(u|a, \sim a)$ of the FlipOut-CE model is higher than that of the DropOut-CE model for the entire range of uncertainty thresholds. For entropy as the uncertainty metric, the DropOut-DICE model always has values lower than DropOut-CE and FlipOut-CE for both



(a) MICCAI2015 test dataset

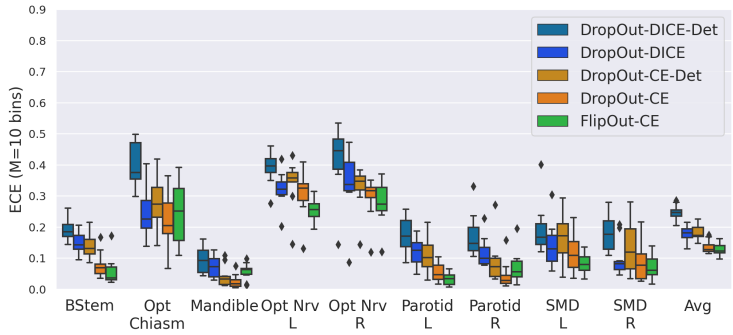


(b) DeepMindTCIA dataset

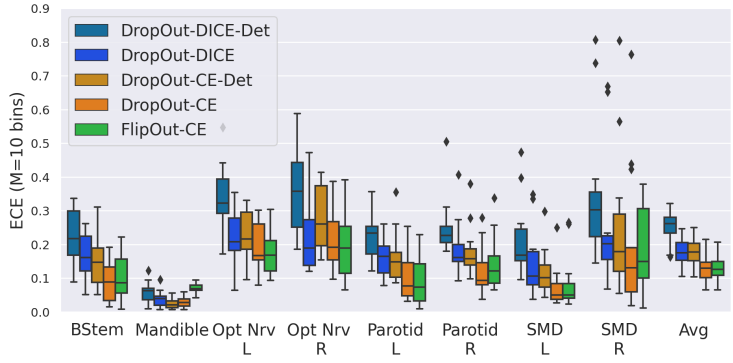
Figure 3.2: Boxplot depicting the DICE scores for the MICCAI2015 test dataset (a) and the DeepMindTCIA dataset (b). The x-axis shows the different organs and the average over all organs.

$p(u|i)$ and $p(u|a, \sim a)$. Similar trends are noticed for the DeepMindTCIA dataset, though the probability values are slightly reduced.

For visual results, we look at Figure 3.5 where the first column shows a CT slice and the second column shows the ground truth (GT) mask. The third, fourth and fifth columns are the deep learning predictions and the remaining columns are their corresponding uncertainty heatmaps. The first row in the figure shows a result from the MICCAI2015 test dataset representing a false positive prediction for the top slice of the brainstem. The second and third rows show predictions for the DeepMindTCIA dataset of the left parotid gland and mandible respectively. In these figures, red represents false positive, blue represents false negative and white represents true positive predictions.



(a) MICCAI2015 test dataset



(b) DeepMindTCIA dataset

Figure 3.3: Boxplot depicting the Expected Calibration Error (ECE) with M=10 bins for the MICCAI2015 test dataset (a) and the DeepMindTCIA dataset (b). The x-axis shows the different organs and the average.

3.4 Discussion and conclusion

This work exploited an existing deterministic model (i.e. FocusNet [24]) and investigated the model confidence calibration and uncertainty behavior of its Bayesian versions for efficient QA in a clinical radiotherapy setting. All Bayesian models, when averaged across organs at risk (OAR), performed equally well in terms of volumetric and surface distance measures, allowing us to compare across other metrics like expected calibration error (ECE) and region-based accuracy-vs-uncertainty (R-AvU). Using a modified cross entropy loss for our models improved their performance in comparison to its standard version as additional supervision is provided for both the foreground and background of each OAR. It was also important to use weights for each OAR to handle the problem of class imbalance. The right plot in Figure 3.2 shows low DICE scores for the right submandibular gland (SMD R) in the DeepMindTCIA dataset. This is because, in general, our models have reduced performance for the TCGA-HNSC patients when compared to the RTOG 0522 patients

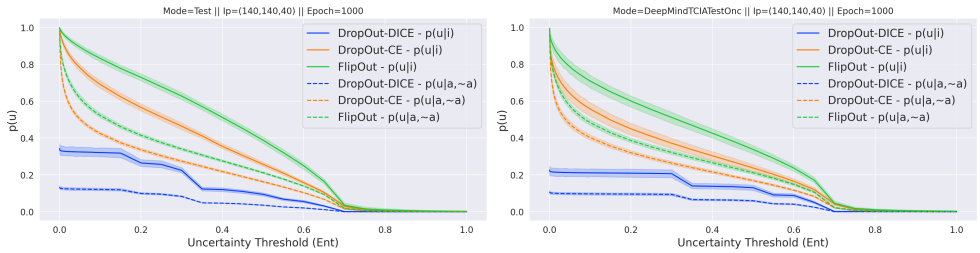


Figure 3.4: Line plots showing the uncertainty behaviour of different models in inaccurate ($p(u|i)$) and accurate ($p(u|a, \sim a)$) regions for the MICCAI2015 test set (left) and the DeepMindTCIA dataset (right).

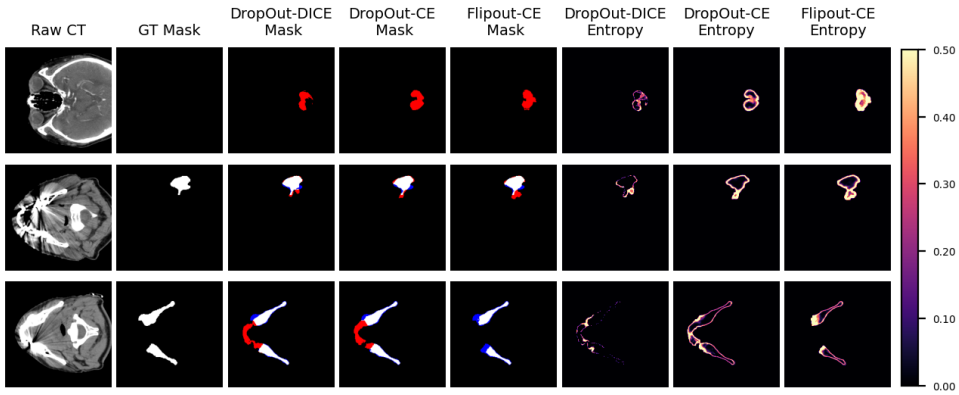


Figure 3.5: The first two columns depict the raw and ground truth data from the datasets, while the remaining columns show model predictions and their associated entropy heatmaps. In the predicted masks, white voxels are true positives, red voxels are false positives while blue voxels are false negatives.

due to poor contrast in the TCGA-HNSC CT scans.

Post model training, it is important to evaluate the ECE of a predictive model to check if it produces probability estimates that reflects its true underlying interpretation of a test sample. The boxplots in Figure 3.3 shows that performing Bayesian inference in neural networks always reduces or maintains calibration error (ECE). Thus, all subsequent model comparisons in this work only consider Bayesian models. It is also observed that CE as a loss function leads to reduced ECE compared to DICE, as also found by others[76]. This may be since CE is a strict scoring rule and hence achieves more reliable probability estimates. Also note that the modified CE achieved similar accuracy compared to DICE. This is an important result as most works in medical image segmentation rely on using the DICE loss. Once again, similar to DICE performance, the right submandibular gland (SMD R) in the DeepMindTCIA dataset has outlier ECE values. This is due to the fact the

models are highly confident but yet inaccurate, leading to large calibration errors.

Given that DropOut-CE and FlipOut-CE have similar ECE values, we refer to the R-AvU graphs to understand differences in their behavior in the context of output uncertainty. For entropy, the FlipOut-CE model has better uncertainty coverage than other models in inaccurate regions. This is reflected in Figure 3.4 where both its $p(u|i)$ and $p(u|a, \sim a)$ curves are higher than that of DropOut-CE. This means that FlipOut-CE misses less inaccurate regions than DropOut-CE, but also directs visual attention to areas that are accurate, more so than DropOut-CE, potentially slowing down QA. A possible reason for the behavior of FlipOut-CE could be that it uses a Gaussian distribution which might be more representative of the weight distribution than the Bernoulli distribution. Entropy for Dropout-DICE, which has the highest ECE, has uncertainty curves that do not sufficiently cover incorrect regions, thus reducing its potential as a contour QA candidate.

Focusing on the bright areas in Figure 3.5, the first and third row show that FlipOut-CE provides a better coverage of erroneous regions, while in the second row the bright areas of DropOut-DICE correspond to errors in the different lobes of the left parotid gland. In the third row of Figure 3.5 for CE-trained models, we see that there exists high uncertainty in the erroneous regions and low uncertainty along the borders of the mandible. The low uncertainty could be the effect of different annotation quality for different patients in the training data which leads to data-based uncertainty along the border regions of an OAR. A similar effect for CE-trained models is seen in row 2 for the left parotid gland, but in this case there is high uncertainty in both high and low error regions which does not satisfy our requirements for visual attention. It is due to this effect that the $p(u|a)$ curves have high probability values. Finally, uncertainty does not exactly correspond to voxel-wise error, so an additional visualization tool on top of the output uncertainty heatmaps may improve acceptability from clinical users.

To conclude, we show that considering both foreground and background regions in the probability maps of organs for the cross entropy (CE) loss improves model performance over the standard practice of only using the foreground regions. This is beneficial, as CE-trained models have better model confidence calibration than DICE trained models. We also explored how the combined use of a quantitative and qualitative measure can support the analysis and selection of Bayesian models for radiotherapy QA. It was observed, that on average, FlipOut-CE has more uncertainty coverage of both inaccurate and accurate regions than the DropOut models, possibly due to the Gaussian assumption in FlipOut compared to the Bernoulli assumption in DropOut. Future work may consider additional training objectives to push apart the $p(u|i)$ and $p(u|a, \sim a)$ curves with the $p(u|i)$ curve having high values and the $p(u|a, \sim a)$ curves having lower values. This will ensure visual attention in erroneous regions through the use of uncertainty heatmaps. One may also explore the use of uncertainty metrics like mutual information that only capture model uncertainty [120], unlike entropy that captures both data and model un-

certainty. It might be worthwhile to investigate which uncertainty metric is more useful within clinical workflows. Finally, this study could also be done for a contour propagation scenario in adaptive radiotherapy to observe if similar results are obtained.

3.5 Acknowledgements

The research for this work was funded by the HollandPTC-Varian Consortium (grant id 2019022).

4

Improving Uncertainty-Error Correspondence in Deep Bayesian Medical Image Segmentation

This chapter was adapted from:

Mody, Prerak, Nicolas F. Chaves-de-Plaza, Chinmay Rao, Eleftheria Astrenidou, Mischa de Ridder, Nienke Hoekstra, Klaus Hildebrandt, and Marius Staring. "Improving Uncertainty-Error Correspondence in Deep Bayesian Medical Image Segmentation." In *Machine Learning for Biomedical Imaging*, August 2024 issue (2024): 1048–82.
https://doi.org/10.59275/j.melba.2024-5gc8.

Abstract

Increased usage of automated tools like deep learning in medical image segmentation has alleviated the bottleneck of manual contouring. This has shifted manual labour to quality assessment (QA) of automated contours which involves detecting errors and correcting them. A potential solution to semi-automated QA is to use deep Bayesian uncertainty to recommend potentially erroneous regions, thus reducing time spent on error detection. Previous work has investigated the correspondence between uncertainty and error, however, no work has been done on improving the "utility" of Bayesian uncertainty maps such that it is only present in inaccurate regions and not in the accurate ones. Our work trains the FlipOut model with the Accuracy-vs-Uncertainty (AvU) loss which promotes uncertainty to be present only in inaccurate regions. We apply this method on datasets of two radiotherapy body sites, c.f. head-and-neck CT and prostate MR scans. Uncertainty heatmaps (i.e. predictive entropy) are evaluated against voxel inaccuracies using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. Numerical results show that when compared to the Bayesian baseline the proposed method successfully suppresses uncertainty for accurate voxels, with similar presence of uncertainty for inaccurate voxels. Code to reproduce experiments is available at https://github.com/prerakmody/bayesuncertainty-error-correspondence.

4.1 Introduction

In recent years, deep learning models are being widely used in radiotherapy for the task of medical image segmentation. Although these models have been shown to accelerate clinical workflows [129, 130], they still commit contouring errors [131]. Thus, a thorough quality assessment (QA) needs to be conducted, which places a higher time and manpower requirement on clinical resources. This creates a barrier to the adoption of such deep learning models [132]. Moreover, it also creates an obstacle for adaptive radiotherapy (ART) workflows, which have been shown to improve a patient's post-radiation quality-of-life [4]. This obstacle arises due to ART's need of regular contour updates. Currently, commercial auto-contouring tools do not have the ability to assist with quick identification and rectification of potentially erroneous predictions [131, 132].

Quality assessment (QA) of incorrect contours would require two steps – 1) error detection and 2) error correction [133]. Currently, errors are searched for by manual inspection and then rectified using existing contour editing tools. Error detection could be semi-automated by recommending either potentially erroneous slices of a 3D scan [63], or by highlighting portions of the predicted contours [58] or blobs [61]. Upon detection of the erroneous region, the contours could be rectified using point or scribble-based techniques [134, 135] in a manner that adjacent slices are also updated. For this work, we will focus on error detection.

Various approaches to error detection have suggested using Bayesian Deep Learning (BDL) and the uncertainty that it can produce as a method to capture potential errors in the predicted segmentation masks [56, 58, 61, 63, 64, 66, 69]. Although such works established the potential usage of uncertainty in the QA of predictions, it may not be sufficient in a clinical workflow that relies on pixel-wise uncertainty as a proxy for error detection. In our experiments with deep Bayesian models, we observed that the relationship between prediction errors and uncertainty is sub-optimal, and hence has low clinical "utility". Ideally, for semi-automated contour QA, the uncertainty should be present only in inaccurate regions and not in the accurate ones. At times, literature usually refers to this as uncertainty calibration [69, 136–139], but we find this term incorrect as historically, calibration is referred to in context of probabilities of a particular event [140]. Thus, we believe it is semantically incorrect to say uncertainty calibration and instead propose to use the term uncertainty-error correspondence.

To create a Bayesian model that is incentivized to produce uncertainty only in inaccurate regions, we use the Accuracy-vs-Uncertainty (AvU) metric [141] and its probabilistic loss version [137] during training of a UNet-based Bayesian model [142]. This loss promotes the presence of both **a**ccurate-if-**c**ertain (n_{ac}) as well as **i**naccurate-if-**u**ncertain (n_{iu}) voxels in the final prediction (Figure 4.1). With uncertainty present only around potentially inaccurate regions, one can achieve improved synergy between clinical experts

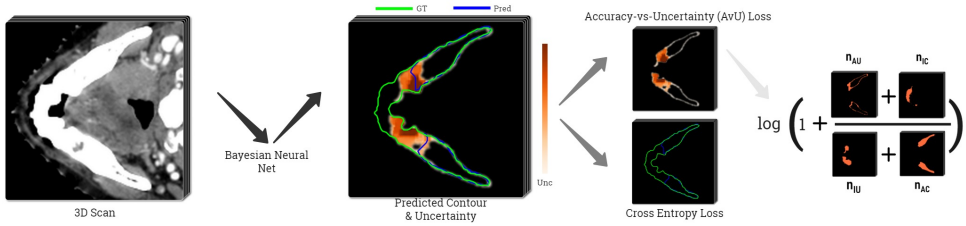


Figure 4.1: Method overview - A 3D medical scan (e.g. CT/MR) is input into a UNet-based Bayesian neural net to produce both predicted contours (*Pred*) and predictive uncertainty (*Unc*). While the cross-entropy loss is used to improve segmentation performance, the Accuracy-vs-Uncertainty (AvU) loss is used to improve uncertainty-error correspondence. The AvU loss is computed by comparing the prediction with the ground truth (*GT*) at a specific uncertainty threshold using four terms: count of accurate-and-certain (n_{AC}), accurate-and-uncertain (n_{AU}), inaccurate-and-certain (n_{IC}) and inaccurate-and-uncertain (n_{IU}) voxels.

and their deep learning tools during the QA stage. Our work is the first to use the AvU loss in a dense prediction task like medical image segmentation and also with datasets containing natural and not synthetic variations as was previously done [137]. This work extends our conference paper [143] with additional datasets, experiments and metrics. There, we adapt the original AvU loss by considering the full theoretical range of uncertainties in the loss, rather than one extracted from the validation dataset [137]. For our work we use the predictive entropy as an uncertainty metric [144].

Several other approaches have been considered in context of uncertainty, for e.g. ensembles, test time augmentation (TTA) and model calibration. While ensembles of models have good segmentation performance [62, 66], they are parameter heavy. TTA [60, 73] performs inference by modulating a models inputs, but does not perform additional training, so may be unable to transcend its limitations. Calibration techniques attempt to make predictions less overconfident [71, 145–149], however they do not explicitly align model errors with uncertainty. All the above methods are benchmarked on the truthfulness of their output probabilities (when compared against voxel accuracies) using metrics like expected calibration error (ECE). However, a model with lower ECE than its counterparts may not necessarily have higher uncertainty-error correspondence.

Finally, to evaluate calibrative and uncertainty-error correspondence metrics, one needs to compute the “true" inaccuracy map. Similar to our conference paper [143] and inspired by [58], we classify inaccuracies of predicted voxel maps into two categories: “errors" and “failures" (see Section 4.8.1). Segmentation “errors" are those inaccuracies which are considered an artifact similar to inter-observer variation, a phenomenon common in medical image segmentation [9, 150]. Thus, we consider these smaller inaccuracies to be accurate in our computations, under the assumption they do not require clinical

intervention. In the context of contour QA, such voxels should ideally be certain. Hence, only the segmentation “failures" are a part of the “true" inaccuracy map used to calculate the calibrative and uncertainty-error correspondence metrics.

To summarize, our contributions are as follows:

- For the purpose of semi-automated quality assessment of predicted contours, we aim to improve uncertainty-error correspondence (unc-err) in a Bayesian medical image segmentation setting, pioneering this in the context of radiation therapy. Specifically, we propose using the loss form of the Accuracy-vs-Uncertainty (AvU) metric while training a deep Bayesian segmentation model.
- We compare our Bayesian model with the AvU loss against an ensemble of deterministic models, five approaches employing calibration-based losses and also test time augmentation. We also perform an architectural comparison by comparing models with Bayesian convolutions placed in either the middle layers or decoder layers of a deep segmentation model.
- We benchmark unc-err of the segmentation models on both in- and out-of-distribution radiotherapy datasets containing head-and-neck CT and Prostate MR scans. Models are benchmarked on these datasets across discriminative, calibrative and uncertainty-error correspondence metrics.

4.2 Related Works

4.2.1 Epistemic and aleatoric uncertainty

Recent years have seen an increase in work that utilizes probabilistic modeling in deep medical image segmentation. The goal has been to account for uncertainty due to noise in the dataset (*aleatoric uncertainty*) as well as in the limitations of the predictive models learning capabilities (*epistemic uncertainty*). Noise in medical image segmentation refers to factors like inter- and intra- annotator contour variation [9, 150] due to factors such as poor contrast in medical scans. Works investigating aleatoric uncertainty model the contour diversity in a dataset by either placing Gaussian noise assumptions on their output [67] or by assuming a latent space in the hidden layers and training on datasets containing multiple annotations per scan [151]. A popular and easy-to-implement approach to model for aleatoric uncertainty is called test-time augmentation (TTA) [57]. Here, different transformations of the image are passed through a model, and the resulting outputs are combined to produce both an output and its associated uncertainty.

In contrast to aleatoric uncertainty, epistemic uncertainty could be used to identify scans (or parts of the scan) that are very different from the training dataset. Here, the model is unable to make a proper interpolation from its existing knowledge. Methods such as ensembling [62] and Bayesian posterior inference (e.g., Monte-Carlo DropOut,

Stochastic Variational Inference) [56, 58, 61, 63, 64, 137, 152] are common methods to model epistemic uncertainty in neural nets. While Bayesian modeling is a more mathematically motivated and hence, principled approach to estimating uncertainty, ensembles have been motivated by the empirically-proven concept of bootstrapping. In contrast to Bayesian models where the perturbation is modelled by placing distributions on weights, ensembles use either different model weight initializations, or different subsets of the training data. In Bayesian inference techniques, perturbations are introduced in the models activation or weight space. Dropout [153] and DropConnect [154] are popular techniques that apply the Bernoulli distribution on these spaces. Stochastic variational inference (SVI) is another type of weight space perturbation that usually assumes the more expressive Gaussian distribution on the weights. Bayes by Backprop [155] and its resource-efficient variant such as FlipOut [142] are examples of SVI. For our work, we consider approaches that are designed for both epistemic uncertainty (Ensembles and SVI models) as well as aleatoric uncertainty (TTA).

4.2.2 Uncertainty use during training

Other works also use the uncertainty from a base segmentation network to automatically refine its output using a follow-up network. This refinement network can be graphical [156] or simply convolutional [58]. Uncertainty can also be used in an active learning scenario, either with [157] or without [68] interactive refinement. Shape-based features of uncertainty maps have also been shown to identify false positive predictions [72]. Similarly, we too use uncertainty in our training regime, but with the goal of promoting uncertainty only in those regions which are inaccurate, an objective not previously explored in medical image segmentation.

4.2.3 Model calibration

In context of segmentation, model calibration error is inversely proportional to the alignment of a models output probabilities with its pixel-wise accuracy. Currently there is no proof that reduction in model calibration error leads to improved uncertainty-error correspondence. However, a weak link can be assumed since both are derived from a models output probabilities. It is well known that the probabilities of deterministic models trained on the cross entropy (CE) loss are not well calibrated [145]. This means that they are overconfident on incorrect predictions and hence fail silently. This, which is an undesirable trait in context of segmentation QA and needs to be resolved.

To abate this overconfidence issue, methods such as post-training model calibration (or temperature scaling) [65, 145, 158], ensembles [62, 159], calibration-focused training losses [146, 148, 149, 160] and calibration-focused training targets [71, 147] have been shown to improve model calibration for deterministic models. Temperature scaling, a post-training model calibration technique, has been shown to perform poorly in out-of-domain (OOD) settings [159], relies wholly on an additional validation dataset and/or

needs explicit shape priors [65]. FinerLocal temperature scaling techniques have been proposed that calibrate on the image or pixel level [158], however they are still conceptually similar to the base method and are hence plagued by the same concerns. Others [65] used a shape prior module for out-of-domain robustness, but they only introduced synthetic textural variations in their work.

Another approach to model calibration is to regularize a model during train to promote uncertainty. For e.g. the ECP [146] technique explicitly adds the negative entropy to the training loss. Conversely, the Focal loss [148, 161] achieves thisattempts to calibrate a model implicitly by assigning lower weights (during training) to more confident predictions. Other methods smooth the hard targets of the ground truth towards a uniform distribution in the limit. For e.g. Label Smoothing [147, 162] does this by modifyingmodifies the class distribution of a pixel by calculating a weighted average (using parameter α) between the hard target and a uniform distribution. On the other hand, Spatial Varying Label Smoothing (SVLS) [71] modifies a pixel's class allocation by considering classes around it. To avoid excessively making the models predictions uniform, Margin-based Label Smoothing (MBLS) [149, 163] reformulates the above approaches by showing that they essentially perform loss optimization where an equality constraint is applied on a pixels logits. MBLS attempts to achieve the best discriminative-calibrative trade-off by softening this equality constraint. They subtract the max logit of a pixel with its other logits and only penalize those logit distances that are greater than a predetermined margin. Others extend thisthe MBLS framework by further tuningeither learning class-specific weights for the equality constraint [164] or reformulating SVLS to a similar formulation similar to MBLS [160]. Although these methods attempt to make models less overconfident, they do not explicitly align a model's error to its uncertainty.

There also exist other approaches to model calibration for e.g., multi-task learning [52], mixup augmentation [165] and shape priors [166]. Multi-task learning requires additional data that may not always be present, while mixup creates synthetic samples which are not representative of the real data distribution. Finally, shape priors may not be applicable to tumors with variable morphology.

Model calibration techniques are evaluated by metrics like Expected Calibration Error (ECE) and its variants [167], however others have also proposed terms like Uncertainty-Calibration Error (UCE) [168, 169]. While ECE evaluates the equivalency between accuracy and predicted probability, UCE compares inaccuracy and uncertainty. However, while it is semantically appropriate to expect an average probability of p $(0 \leq p \leq 1)$ to give the same average accuracy (i.e., the mathematical formulation of ECE), the same is not appropriate for inaccuracy and uncertainty u $(0 \leq u \leq 1)$. Hence, UCE is not applicable to our work.

To conclude, the issue with each of the aforementioned techniques for epistemic, aleatoric and calibrative modeling is that they do not explicitly train the model to develop an innate

sense of potential errors on a given segmentation task. Given that this is the primary requirement from a contour QA perspective, these models may be unable to have good uncertainty-error correspondence.

4.3 Methods

4.3.1 Neural Architecture

We adopt the OrganNet2.5D neural net architecture [170] which is a standard encoder-decoder model connected by four middle layers. It contains both 2D and 3D convolutions in the encoder and decoder as well as hybrid dilated convolutions (HDC) in the middle. This network performs fewer pooling steps to avoid losing image resolution and instead uses HDC to expand the receptive field. To obtain uncertainty corresponding to the output, we add stochasticity to the deterministic convolutional operations by replacing them with Bayesian convolutions [142, 155]. We experiment with replacing deterministic layers in both the HDC as well as the decoder layers to understand the effect of placement.

In a Bayesian model, a prior distribution is placed upon the weights and is then updated to a posterior distribution on the basis of the training data. During inference (Equation (4.1)), we sample from this posterior distribution $p(W|D)$ to estimate the output distribution $p(y|x,D)$ with x, y and W being the input, output and neural weight respectively:

$$p(y|x,D) = \mathbb{E}_{W\sim p(W|D)}\Big[p(y|x,W)\Big]. \tag{4.1}$$

This work uses a Bayesian posterior estimation technique called stochastic variational inference, where instead of finding the true, albeit intractable posterior, it finds a distribution close to it. We chose FlipOut-based [142] convolutions which assume the distribution over the neural weights to be a Gaussian and are factorizable over each hidden layer. Pure variational approaches would need to sample from this distribution for each element of the mini-batch [155]. However, the FlipOut technique only samples once and multiplies that random sample with a Rademacher matrix, making the forward pass less computationally expensive.

4.3.2 Training Objectives

In this section , we use a notation format, where capital letters denote arrays while non-capital letters denote scalar values.

4.3.2.1 Segmentation Objective

Upon being provided a 3D scan as input, our segmentation model predicts for each class $c \in C$, a 3D probability map $\hat{P}_c$ of the same size. Each voxel $i \in N$ has a predicted probability vector $\hat{P}^i$ containing values $\hat{p}_c^i$ for each class that sum to 1 (due to softmax). To calculate the predicted class of each voxel $\hat{y}^i$, we do:

$$\hat{y}^i = \underset{c\in C}{\operatorname{argmax}}\, \hat{p}_c^i. \tag{4.2}$$

To generate a training signal, the predicted probability vector $\hat{P}^i$ is compared to the corresponding one-hot vector Y^i in the gold standard 3D annotation mask. Y^i is composed of $y_c^i \in \{0,1\}$. Inspired by [171, 172], we re-frame the binary cross-entropy loss (Equation (4.3)), as penalizing both the foreground ($y_c^i = 1$) and background ($(1 - y_c^i) = 1$) voxels of the probability maps of each class with a weight w_c:

$$L_{\text{CE}} = -\frac{1}{|C|}\left(\sum_{c \in C} w_c \left[\sum_{i \in N}\left(y_c^i \ln(\hat{p}_c^i) + (1 - y_c^i)\ln(1 - \hat{p}_c^i)\right)\right]\right). \tag{4.3}$$

Note, we do not utilize the DICE loss for training as it has been shown to have lower model calibration metrics [173]. Also, since the CE loss is susceptible to fail during a class-imbalance, we use its weighted version.

4.3.2.2 Uncertainty Objective

In a Bayesian model, multiple forward passes ($m \in M$) are performed and the output 3D probability maps $(\hat{P_c})_m$ of each pass are averaged to output $\hat{P}_c$ (Equation (4.1)). Using $\hat{P}_c$, we can calculate a host of statistical measures like entropy, mutual information and variance. We chose entropy as it has been shown to capture both epistemic uncertainty, which we explicitly model in FlipOut layers, as well as aleatoric uncertainty, which is implicitly modeled due to training data [144]. We use the predicted class probability vector $\hat{P}^i$ for each voxel and calculate its (normalized) entropy u^i:

$$u^i = -\frac{1}{\ln(|C|)} \sum_{c \in C} \hat{p_c^i} \ln(\hat{p_c^i}). \tag{4.4}$$

Since we have access to the gold standard annotation mask, each voxel has two properties: accuracy and uncertainty. Accuracy is determined by comparing the gold standard class y^i to the predicted class $\hat{y^i}$. We use this to classify them in four different categories represented by n_{ac}, n_{au}, n_{ic} and n_{iu}, where n stands for the total voxel count and a, i, u, c represent the **a**ccurate, **i**naccurate, **u**ncertain and **c**ertain voxels. A visual representation of these terms can be seen in Figure 4.1. Here, a voxel is determined to be certain or uncertain on the basis of a chosen uncertainty threshold $t \in T$ where the maximum value in T is the maximum theoretical uncertainty threshold [143]. The aforementioned four terms are the building blocks of the Accuracy-vs-Uncertainty (AvU) metric [141] as shown in Equation (4.5) - Equation (4.7) and it has a range between [0,1]. A higher value indicates that uncertainty is present less in accurate regions and more in inaccurate regions, thus improving the “utility" of uncertainty as a proxy for error detection.

$$\mathrm{AvU}^{\mathrm{t}} = \frac{n_{\mathrm{ac}^{\mathrm{t}}} + n_{\mathrm{iu}^{\mathrm{t}}}}{n_{\mathrm{ac}^{\mathrm{t}}} + n_{\mathrm{au}^{\mathrm{t}}} + n_{\mathrm{ic}^{\mathrm{t}}} + n_{\mathrm{iu}^{\mathrm{t}}}} \tag{4.5}$$

$$n^t_{\mathrm{ac}} = \sum_{i \in \left\{ \substack{y_i = \hat{y}_i \ \& \\ u_i \leq t} \right\}} 1, \qquad n^t_{\mathrm{au}} = \sum_{i \in \left\{ \substack{y_i = \hat{y}_i \ \& \\ u_i > t} \right\}} 1 \tag{4.6}$$

$$n^t_{\mathrm{ic}} = \sum_{i \in \left\{ \substack{y_i \neq \hat{y}_i \ \& \\ u_i \leq t} \right\}} 1, \qquad n^t_{\mathrm{iu}} = \sum_{i \in \left\{ \substack{y_i \neq \hat{y}_i \ \& \\ u_i > t} \right\}} 1 \tag{4.7}$$

To maximize AvU for a neural net, one can turn it into a loss metric to be minimized. As done in [137] for an image classification setting, we minimize its negative logarithm (Equation (4.8)) to improve mathematical stability of gradient descent. However, the AvU metric, as defined above, is not differentiable with respect to the neural net's weights. This is due to all its constituent terms being produced either due to thresholding or max operations which introduce discontinuities that disrupt gradient flows.. This is because the model's outputs are simply used to create a mask and hence no backpropagation can take place. The AvU metric is made differentiable by instead using the uncertainty u^i derived from $\hat{P}^i$ (Equation (4.4)), thus allowing for gradient flows. Also, a smooth non-linear operation i.e., *tanh* is used to constrain its value (Equation (4.9)). The differentiable uncertainty term is multiplied by other scalar weighing terms c.f. the maximum probability ($\hat{p}^i = \max(\hat{P}^i)$) and accuracy/inaccuracy mask for a voxel. All these operations together allow us to calculate proxy values for n_{ac}, n_{au}, n_{ic} and n_{iu}. In addition, rather than evaluating the loss at a single uncertainty threshold, we integrate over the theoretical range of the uncertainty metric. Thresholding is done by once again multiplying the uncertainty value with a binary mask. The benefits of this thresholding were shown in our conference paper [143]:

$$\begin{aligned} L_{\mathrm{AvU}^{\mathrm{t}}} &= -\ln\left(1 + \frac{n^t_{\mathrm{au}} + n^t_{\mathrm{ic}}}{n^t_{\mathrm{ac}} + n^t_{\mathrm{iu}}}\right), \\ L_{\mathrm{AvU}} &= \frac{1}{T} \sum_{t \in T} L_{\mathrm{AvU}^{\mathrm{t}}}, \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} n^t_{\mathrm{ac}} &= \sum_{i \in \left\{ \substack{y_i = \hat{y}_i \ \& \\ u_i \leq t} \right\}} \hat{p}^i \cdot (1 - \tanh(u^i)), & n^t_{\mathrm{au}} &= \sum_{i \in \left\{ \substack{y_i = \hat{y}_i \ \& \\ u_i > t} \right\}} \hat{p}^i \cdot \tanh(u^i) \\ n^t_{\mathrm{ic}} &= \sum_{i \in \left\{ \substack{y_i \neq \hat{y}_i \ \& \\ u_i \leq t} \right\}} (1 - \hat{p}^i) \cdot (1 - \tanh(u^i)), & n^t_{\mathrm{iu}} &= \sum_{i \in \left\{ \substack{y_i \neq \hat{y}_i \ \& \\ u_i > t} \right\}} (1 - \hat{p}^i) \cdot \tanh(u^i). \end{aligned} \tag{4.9}$$

Finally, the total loss L combines the segmentation and uncertainty loss as:

$$L = L_{\mathrm{CE}} + \alpha \cdot L_{\mathrm{AvU}}. \tag{4.10}$$

4.3.3 Evaluation

4.3.3.1 Discriminative and Calibration Evaluation

We evaluate all models on the DICE coefficient for discriminative performance. Calibration is evaluated using the Expected Calibration Error (ECE) [145]. Numerical results are compared with the Wilcoxon signed-ranked test where a p-value ≤ 0.05 is considered significant.

4.3.3.2 Uncertainty Evaluation

As the model is trained on the Accuracy-vs-Uncertainty (AvU) metric, we calculate the AvU scores up to the maximum normalized uncertainty of the validation dataset. A curve with the AvU score on the y-axis and the uncertainty threshold on the x-axis is made and the area-under-the-curve (AUC) for each scan is calculated. AUC scores provide us with a summary of the model performance regardless of the uncertainty threshold, and hence we use it to compare all models.

The AvU metric outputs a single scalar value for the whole scan and does not offer much insight into the differences in uncertainty coverage between the accurate and inaccurate regions. To abate this, we compare the probability of **u**ncertainty in **i**naccurate regions $p(u|i)$ to the probability of **u**ncertainty in **a**ccurate regions $p(u|a)$. Let us plot $p(u|i)$ and $p(u|a)$ on the y-axis and x-axis of a graph respectively, and define n_{iu}, n_{au}, n_{ac} and n_{ic}, as the count of true positives, false positives, true negatives and false negatives respectively. Thus, $p(u|i)$ is the true positive rate and $p(u|a)$ is the false positive rate. Computing this at different uncertainty thresholds provides us with the Receiver Operating Characteristic (ROC) curve, which we call the uncertainty-ROC curve [154].

Given that ROC curves are biased in situations with class imbalances between positive (inaccurate voxels) and negative (accurate voxels) classes, we also compute the precision-recall curves [69].Here, precision is the probability of inaccuracy given uncertainty $p(i|u)$ and recall is the probability of uncertainty given inaccuracy $p(u|i)$. Note, that the precision-recall curves do not make use of n_{ac}, which can be high in count for a well-performing model.

Finally, to calculate the calibrative and uncertainty-correspondence metrics, we need an inaccuracy map. We use an inaccuracy map based on the concept of segmentation "failures" and "errors" (Section 4.8.1). To do this, we perform a morphological opening operation using a fixed kernel size of (3,3,1).

4.4 Experiments and Results

4.4.1 Datasets

4.4.1.1 Head-and-Neck CT

Our first dataset contained Head and Neck CT scans of patients from the RTOG 0522 clinical trial [174]. The annotated data, which had been collected from the MICCAI2015 Head and Neck Segmentation challenge, contained 33 CT scans for training, 5 for validation and 10 for testing [42]. We further expanded the test dataset with annotations of 8 patients belonging to the RTOG trial from the DeepMindTCIA dataset (DTCIA) [43]. This dataset included annotations for the mandible, parotid glands, submandibular glands and brainstem. Although there were annotations present for the optic organs, we ignored them for this analysis as they are smaller compared to other organs and require special architectural design choices. Since the train and test patients came from the same study, we considered this as an in-distribution dataset. We also tested our models on the STRUCTSeg (50 scans) dataset [175], hereby shortened as STRSeg. While the RTOG dataset contained American patients, the STRSeg dataset was made up of Chinese patients and hence considered out-of-distribution (OOD) in context of the training data. The uncertainties of this dataset were evaluated to a value of 0.4 since that is the maximum empirical normalized entropy.

4.4.1.2 Prostate MR

Our second dataset contained MR scans of the prostate for which we use the ProstateX repository [176] containing 66 scans as the training dataset. The Medical Decathlon (Prostate) dataset with 34 scans [177] and the PROMISE12 repository with 50 scans [178] served as our test dataset. The Medical Decathlon dataset (abbreviated as PrMedDec henceforth) contained scans from the same clinic as the ProstateX training dataset. We combined the Peripheral Zone (PZ) and Transition Zone (TZ) from the MedDec dataset into 1 segmentation mask. The PROMISE12 dataset (abbreviated as PR12) was chosen for testing since literature [62] has shown lower performance on it and hence it serves as a good candidate to evaluate the utility of uncertainty. This dataset is different from ProstateX due to the usage of an endo-rectal coil in many of its scans as well as the presence of gas pockets in the rectum and dark shadows due to the usage of older MR machines. Thus, although these datasets contained scans of the prostate region, there exists a substantial difference in their visual textures. The maximum empirical normalized entropy of this 2-class dataset is 1.0 and hence the uncertainty-error correspondence metrics were calculated till this value.

4.4.2 Experimental Settings

We tested the Accuracy-vs-Uncertainty (AvU) loss on four datasets containing scans of different modalities and body sites. We trained 11 models: *Det* (deterministic), *Det+AvU*,

Ensemble, *Focal*, *LS* (Label Smoothing), *SVLS* (Spatially Varying Label Smoothing), *MbLS* (Margin based Label Smoothing), *ECP (Explicit Confidence Penalty)*, TTA (Test-Time Augmentation), *Bayes* and *Bayes + AvU*. As the names suggest, *Bayes* and *Bayes + AvU* are Bayesian versions of the deterministic OrganNet2.5D model [170]. The baseline *Bayes* model contained Bayesian convolutions in its middle layers and was trained using only the cross-entropy (CE) loss. The *Bayes + AvU* was trained using both the CE and Accuracy-vs-Uncertainty (AvU) loss. Two additional Bayesian models were trained which tests if the placement of the Bayesian layers had any effect: *BayesH* and *BayesH + AvU*. Here, *BayesH* refers to the Bayesian model with Bayesian layers in the head of the model (i.e the decoder). Results for these models can be found in Section 4.8.7.

The *Ensemble* was made of $M = 5$ deterministic models with different initializations [159]. For TTA, we applied Gaussian noise and random pixel removals for $M = 5$ times each and then averaged their outputs. The hyperparameters of the other models were chosen on the basis of the best discriminative, calibrative and uncertainty-error correspondence metrics on the validation datasets (Section 4.8.3). For the calibration focused methods we used the following range of hyperparameters: Focal ($\gamma = 1, 2, 3$), MbLS ($m = 8, 10, 20, 30$) for head-and-neck CT, MbLS ($m = 3, 5, 8, 10$) for prostate MR, LS ($\alpha = 0.1, 0.05, 0.01$), SVLS ($\gamma = 1, 2, 3$) and ECP ($\lambda = 0.1, 1.0, 10.0, 100.0$) for head-and-neck CT and ECP ($\lambda = 0.1, 1.0, 10.0, 100.0, 1000.0$) for prostate MR. For the AvU loss, we evaluated weighting factors in the range [10,100,1000,10000] for the head-and-neck dataset, and [100,1000,10000] for the Prostate dataset.

We trained our models for 1000 epochs using the Adam optimizer with a fixed learning rate of 10^{-3}. The deterministic model contained $\approx$ 550K parameters and thus the *Ensemble* contained $\approx$ 2.75M parameters. Since the Bayesian models double the parameter count in their layers they incurred an additional parameter cost and ended up with a total of $\approx$ 900K parameters.

4.4.3 Results

In Section 4.4.3.1 and Section 4.4.3.2 we show discriminative (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (ROC-AUC, PRC-AUC) for the two datasets.

4.4.3.1 Head-and-neck CT

Results in Table 4.1 showed that the AvU loss on the *Bayes* model significantly improved calibrative and uncertainty-error correspondence (unc-err) metrics for both in-distribution (ID) and out-of-distribution (OOD) datasets. The *Bayes+AvU* model also always performed better than the *Det*, calibration-focused and *TTA* models for unc-err metrics. Also, its ECE scores were in most cases better than calibration-focused models. However, there was no clear distinction between the performance of the *Ensemble* and *Bayes+AvU* model for ECE and unc-err metrics across both datasets. Also, the AvU loss did not benefit the unc-err metrics for the *Det* model, in both datasets. Of all the calibration-focused models, *LS* had

Table 4.1: Volumetric (*DICE*), calibrative (*ECE*) and uncertainty-error correspondence (ROC-AUC, PRC-AUC) metrics for all models. Here, we evaluate head-and-neck (H&N) CT test datasets which are either in-distribution (ID) or out-of-distribution (OOD). The arrows in the table header indicate whether a metric should be high (↑) or low (↓). Here, † and **bold** are used to indicate a statistical significance and improved results upon comparing a Bayesian model and its AvU-loss version, while underlined numbers indicate the best value for a metric across a dataset.

Test Dataset	Model	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
ID H&N CT (RTOG)	Det	84.2 ± 2.7	9.0 ± 2.1	73.0 ± 5.7	21.0 ± 4.8
	Det + AvU	83.8 ± 2.9	8.6 ± 2.7	73.1 ± 6.0	20.8 ± 4.0
	Focal	84.3 ± 2.4	9.3 ± 1.5	70.3 ± 5.5	18.2 ± 3.2
	ECP	84.4 ± 2.3	9.0 ± 2.0	73.8 ± 5.4	21.0 ± 3.7
	LS	83.0 ± 3.0	7.5 ± 2.2	62.6 ± 3.3	17.5 ± 4.0
	SVLS	84.2 ± 2.6	9.0 ± 2.0	70.8 ± 7.1	18.1 ± 3.5
	MbLS	84.0 ± 2.6	9.2 ± 2.1	67.5 ± 5.7	19.5 ± 3.5
	TTA	84.1 ± 2.8	9.1 ± 2.1	72.9 ± 5.9	20.8 ± 3.9
	Ensemble	85.0 ± 2.6	7.8 ± 1.8	78.6 ± 4.7	25.7 ± 6.8
	Bayes	83.9 ± 2.6	8.6 ± 2.1	74.1 ± 5.4	22.1 ± 3.5
	Bayes+AvU	83.6 ± 2.5	**7.6 ± 2.5**†	**76.1 ± 5.6**†	**25.1 ± 5.3**†
OOD H&N CT (STRSeg)	Det	78.1 ± 4.6	12.9 ± 2.6	62.2 ± 4.5	24.1 ± 3.7
	Det + AvU	78.6 ± 4.7	12.7 ± 3.0	60.8 ± 4.7	22.4 ± 4.1
	Focal	77.2 ± 6.7	12.5 ± 2.9	57.0 ± 4.6	20.9 ± 4.2
	ECP	78.8 ± 4.3	12.5 ± 2.6	61.5 ± 4.8	23.2 ± 3.6
	LS	77.7 ± 6.0	10.3 ± 2.9	56.7 ± 3.3	20.6 ± 4.3
	SVLS	79.0 ± 6.0	11.3 ± 2.5	59.9 ± 5.4	21.6 ± 2.7
	MbLS	77.5 ± 6.3	13.4 ± 3.0	56.9 ± 5.0	21.5 ± 3.6
	TTA	78.1 ± 4.6	12.7 ± 2.6	62.7 ± 4.6	24.9 ± 4.1
	Ensemble	78.6 ± 5.2	10.6 ± 2.4	64.7 ± 4.9	28.2 ± 5.1
	Bayes	75.0 ± 9.9	12.4 ± 4.0	64.8 ± 5.0	27.7 ± 5.8
	Bayes+AvU	76.3 ± 7.7	**12.1 ± 3.7**	**65.8 ± 5.0**†	**30.1 ± 6.5**†

the lowest ECE and unc-err metrics, while the *ECP* model had the best unc-err metrics. When compared to *Det*, the *TTA* model improved calibrative and unc-err metrics for the OOD dataset, while maintaining it for the ID dataset.

Visually, the *Bayes+AvU* model was able to successfully suppress uncertainty in the true positive (TP) (Case 1/2 in Figure 4.2a) and true negative (TN) (Case 3 in Figure 4.2a) regions of the predicted contour. Moreover, it also showed uncertainty in false positive (FP) regions while also suppressing uncertainty in TP regions (Case 3 in Figure 4.2b). Calibrative models (e.g. *Focal*, *LS*, *SVLS*) tended to be quite uncertain in TP or TN regions, which may lead to additional QA time. Detailed descriptions are provided in Section 4.8.4.

Table 4.2: Volumetric (*DICE*), calibrative (*ECE*) and uncertainty-error correspondence (ROC-AUC, PRC-AUC) metrics for all models. Here, we evaluate Prostate MR test datasets which are either in-distribution (ID) or out-of-distribution (OOD). The arrows in the table header indicate whether a metric should be high (↑) or low (↓). Here, † and **bold** are used to indicate a statistical significance and improved results upon comparing a Bayesian model and its AvU-loss version, while underlined numbers indicate the best value for a metric across a dataset.

Test Dataset	Model	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
ID Prostate MR (PrMedDec)	Det	84.1 ± 5.6	12.9 ± 6.0	92.5 ± 5.7	28.0 ± 3.7
	Det + AvU	83.7 ± 6.8	16.9 ± 8.1	92.1 ± 6.8	28.2 ± 3.4
	Focal	81.1 ± 15.4	10.2 ± 5.0	93.2 ± 5.5	29.3 ± 3.4
	ECP	84.0 ± 5.5	16.7 ± 7.1	92.1 ± 6.0	27.6 ± 4.3
	LS	83.4 ± 7.2	15.1 ± 8.6	83.2 ± 7.8	25.1 ± 3.1
	SVLS	83.5 ± 6.7	14.0 ± 8.1	90.5 ± 7.9	21.7 ± 2.6
	MbLS	84.2 ± 4.9	17.9 ± 7.4	92.2 ± 5.6	26.9 ± 3.6
	TTA	83.8 ± 5.8	16.4 ± 7.1	92.7 ± 5.6	28.8 ± 3.9
	Ensemble	84.5 ± 5.7	11.3 ± 6.5	94.3 ± 4.3	30.0 ± 4.6
	Bayes	84.0 ± 5.8	8.6 ± 4.7	94.7 ± 3.1	29.1 ± 4.8
	Bayes+AvU	**84.9 ± 6.9**	8.9 ± 6.0	**95.7 ± 3.2**†	**30.5 ± 4.5**†
OOD Prostate MR (PR12)	Det	74.2 ± 12.6	15.6 ± 6.3	87.9 ± 7.5	22.1 ± 6.2
	Det + AvU	74.5 ± 13.0	27.6 ± 14.3	88.2 ± 7.6	22.0 ± 7.1
	Focal	71.2 ± 17.4	12.1 ± 5.8	89.0 ± 7.1	24.3 ± 6.7
	ECP	74.8 ± 12.5	22.3 ± 10.2	87.2 ± 8.1	20.6 ± 7.0
	LS	74.5 ± 13.0	21.7 ± 11.5	79.5 ± 8.9	19.1 ± 7.2
	SVLS	76.9 ± 11.5	17.9 ± 9.3	87.2 ± 7.2	16.4 ± 5.2
	MbLS	73.6 ± 12.5	19.9 ± 7.4	86.5 ± 7.2	21.8 ± 5.6
	TTA	74.0 ± 12.8	23.7 ± 11.4	88.6 ± 7.4	24.9 ± 5.8
	Ensemble	76.3 ± 12.2	9.7 ± 5.0	91.6 ± 5.2	28.4 ± 5.7
	Bayes	70.6 ± 16.6	11.8 ± 7.2	89.1 ± 7.4	25.7 ± 5.1
	Bayes+AvU	76.3 ± 12.6	11.4 ± 6.7	**90.6 ± 6.9**†	**26.2 ± 7.4**†

4.4.3.2 Prostate MR

Similar to the head-and-neck CT dataset, the use of the AvU loss on the baseline *Bayes* model significantly improved its uncertainty-error correspondence (unc-err) while maintaining calibration performance (Table 4.2). Moreover, it improved the DICE values such that its one of the most competitive amongst all models. Also, the *Bayes+AvU* had better performance in both unc-err and calibrative metrics when compared to the *Det*, calibration-focused and *TTA* models. When comparing to the *Ensemble*, the *Bayes+AvU* had similar DICE. While *Bayes+AvU* had better calibrative and unc-err performance in the in-distribution (ID) dataset, the *Ensemble* performed better in the out-of-distribution (OOD) setting. The AvU loss had no positive effect on the DICE and unc-err performance of the

Det model in both the ID and OOD setting, however there was an increase in ECE.

Visual results show that the *Bayes+AvU* successfully suppresses uncertainty in the true negative (Case 1 in Figure 4.3a, Case 2 in Figure 4.3b) and true positive (Case 2 in Figure 4.3a) regions of the predicted contour. It also shows uncertainty in the false positive regions (Case 2 in Figure 4.3a, Case 1/3 in Figure 4.3b)

4.5 Discussion

Although medical image segmentation using deep learning can now predict high quality contours which can be considered clinically acceptable, a manual quality assessment (QA) step is still required in a clinical setting. To truly make these models an integral part of clinical workflows, we need them to be able to express their uncertainty and for those uncertainties to be useful in a QA setting. To this end, we test 11 models which are either Bayesian, deterministic, calibration-focused or ensembled.

4.5.1 Discriminative and Calibrative Performance

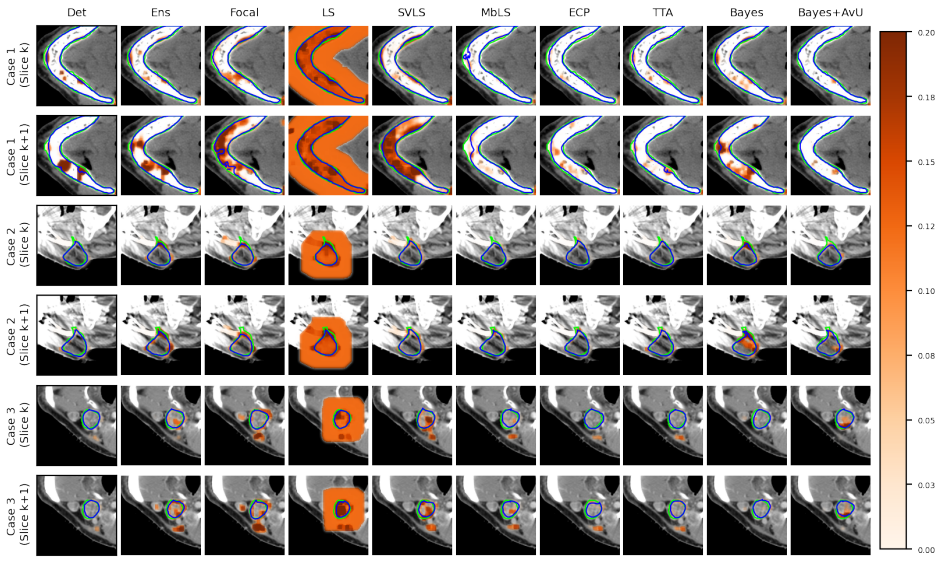
In context of DICE and ECE, the use of the AvU loss on the baseline *Bayes* model always showed results which have never statistically deteriorated. Moreover, the DICE results for the in-distribution (ID) head-and-neck dataset (RTOG) were on-par with existing state-of-the-art models (83.6 vs 84.7 for [43]). The same held for the ID Prostate dataset (PRMed-Dec) where results were better than advanced models (84.9 vs 83.0 for [177]). These results validate the use of our neural architecture [170], and training strategy.

Secondly, although the *Ensemble* model, in general, had better or equivalent DICE and ECE scores across all 4 datasets, it also required 3x more parameters than the *Bayes+AvU* model. Also, as expected, and due to 5x more parameters, the *Ensemble* model performed better than the *Det* model for DICE and ECE.

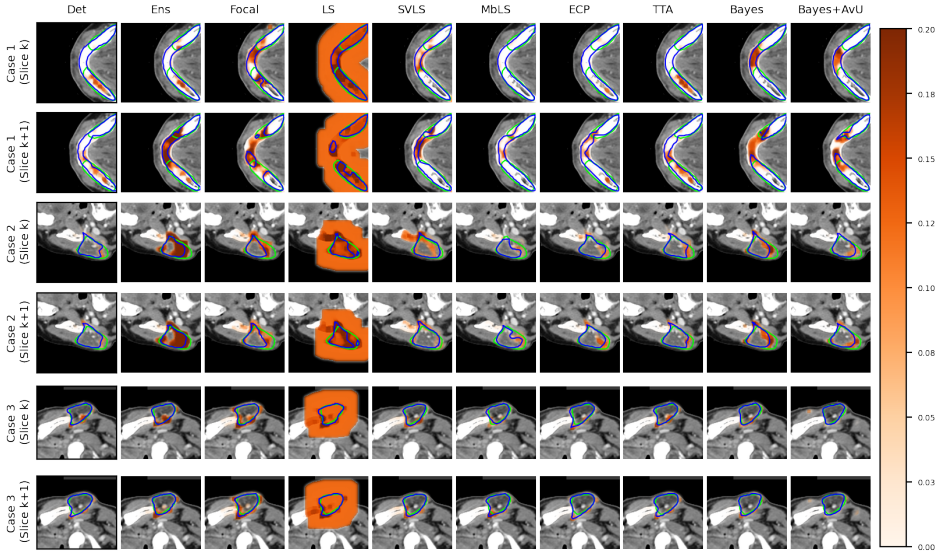
Finally, in the regime of segmentation “failures" as the inaccuracy map, the calibrative methods did not generally have improved calibration performance when compared to the *Det* model. In theory, these models regularize the model’s probabilities by making it more uncertain and hence avoid overconfidence. In practice however, this leads to the predicted contours being uncertain along their accurate boundaries, most evident in visual examples of the *Focal* and *SVLS* model (see Figure 4.2 and Figure 4.3). Also, visual image characteristics in different regions of the scan that are similar to the segmented organs may cause these models to showcase uncertainty in those areas (for e.g. patches of uncertainty in Case 3 of Figure 4.2a).

4.5.2 Uncertainty-Error Correspondence Performance

Although calibrative metrics are useful to compare the average truthfulness of a model’s probabilities, they may not be relevant to real-world usage in a pixel-wise segmentation QA scenario. Considering a clinical workflow in which uncertainty can be used as a proxy for error-detection, we evaluate the correspondence between them. Results showed that

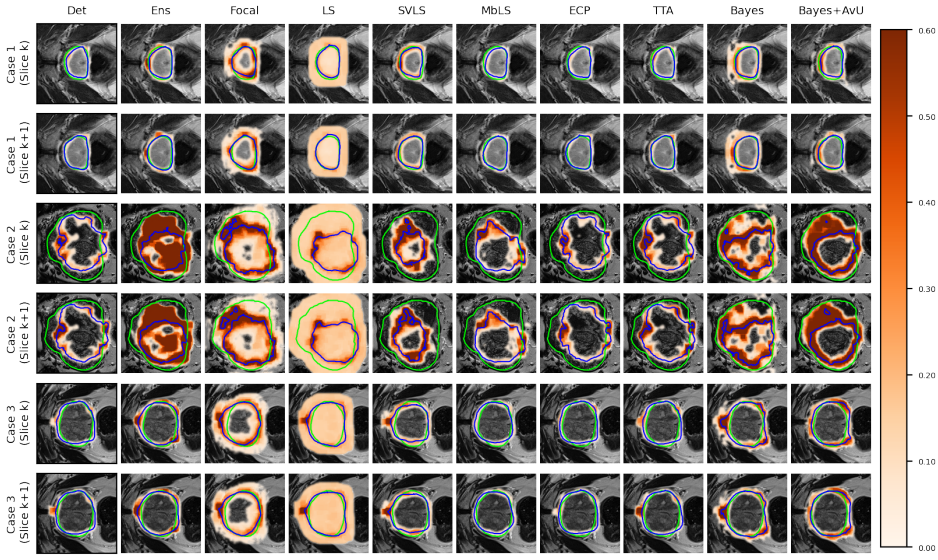


(a) H&N CT (RTOG) (in-distribution)

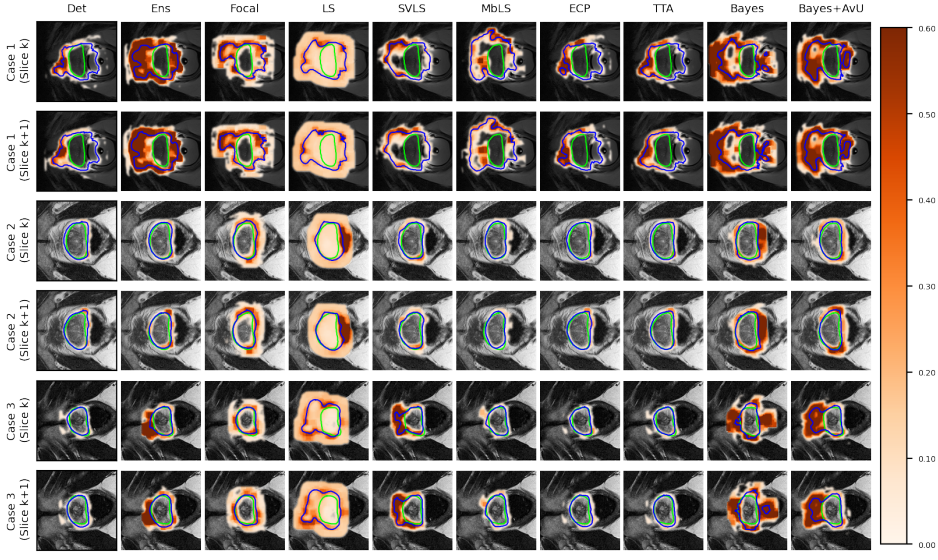


(b) H&N CT (STRSeg) (out-of-distribution)

Figure 4.2: Uncertainty-error correspondence for the head-and-neck (H&N) CT (a,b) dataset. Slices of the CT scans are shown in pairs to understand the 3D nature of segmentation uncertainty heatmaps. The color bar on the right depicts the range of uncertainty values while green and blue are used for ground truth and prediction contours respectively.

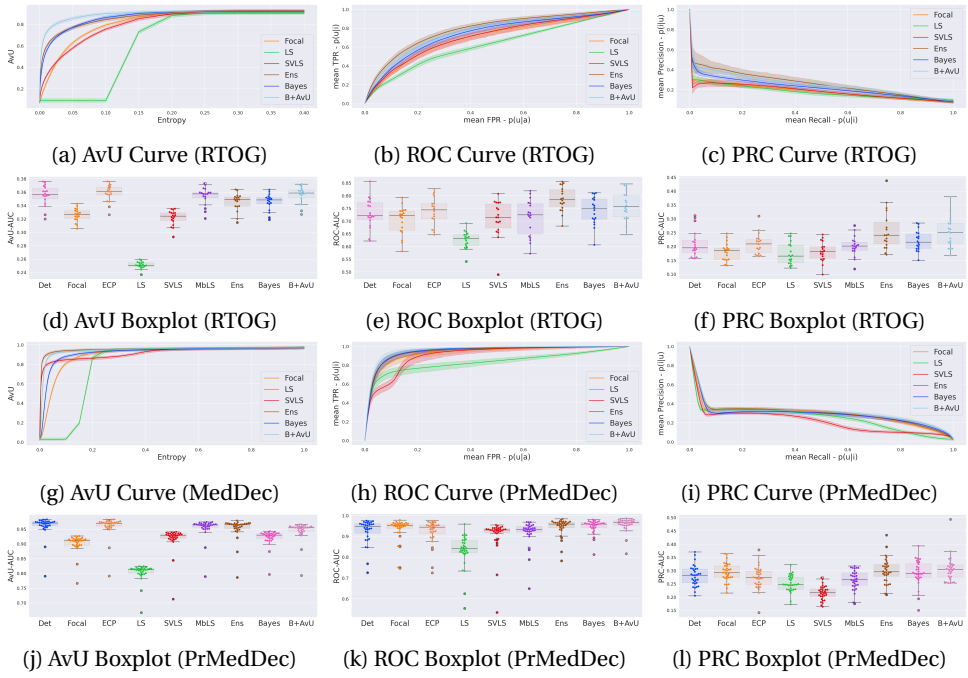


(a) Prostate MR (PrMedDec) (in-distribution)



(b) Prostate MR (PR12) (out-of-distribution)

Figure 4.3: Uncertainty-error correspondence for the Prostate MR (a,b) dataset. Slices of the MR scans are shown in pairs to understand the 3D nature of segmentation uncertainty heatmaps. The color bar on the right depicts the range of uncertainty values while green and blue are used for ground truth and prediction contours respectively.



(a) AvU Curve (RTOG) (b) ROC Curve (RTOG) (c) PRC Curve (RTOG)

(d) AvU Boxplot (RTOG) (e) ROC Boxplot (RTOG) (f) PRC Boxplot (RTOG)

(g) AvU Curve (MedDec) (h) ROC Curve (PrMedDec) (i) PRC Curve (PrMedDec)

(j) AvU Boxplot (PrMedDec) (k) ROC Boxplot (PrMedDec) (l) PRC Boxplot (PrMedDec)

Figure 4.4: The figures above show the distribution of the uncertainty-error correspondence metrics as curves and boxplots (with swarm plots) for patients from the RTOG clinical trial (a-f) as well as for the Medical Decathlon (Prostate) dataset (g-l). We only evaluate up to the maximum uncertainty of each dataset as the metrics do not change beyond that.

across both in- and out-of-distribution datasets, the *Bayes+AvU* model has one of the highest uncertainty-error correspondence metrics. Similar trends were observed for the *BayesH+AvU* (Section 4.8.7) model, however *Bayes+AvU* was better. We hypothesize that this is due to perturbations in the bottleneck of UNet-like models having a better understanding of semantic concepts (e.g., shape, size etc) than the decoder layers. However, the AvU loss did not offer benefit to the *Det* model on both datasets indicating that this loss may rely on the model to already exhibit some level of uncertainty.

An interesting case is shown in Figure 4.2b (Case 3) which showed uncertainty on the white blob (a vein) in the middle of the grey tissue of the organ. Many models showed uncertainty on the vein due to a difference in its texture from that of the organ. However, this information may be distracting to a clinician as they are using uncertainty for error detection. Given that there were no segmentation "failures", our *Bayes+AvU* model successfully suppressed all uncertainties. In another case (Figure 4.2a - Case 3), we saw that for 3D segmentation, uncertainty is also 3D in nature. Our *Bayes+AvU* model had an error in the second slice and correctly showed uncertainty there. However, this uncertainty

overflowed on the first slice and hence penalized the uncertainty-error correspondence metrics. Such results indicate that during contour QA, the clinician can potentially trust our AvU loss models more than other models as they are better indicative of potential errors. This reduces time wasted analyzing false positive regions (i.e., accurate but uncertain) and hence increases trust between an expert and deep learning-based contour QA tools. Also note that in general, the two-class prostate dataset visually showcased higher levels of uncertainty than the six-class head-and-neck dataset.

As seen in Table 4.1, Table 4.2 and Figure 4.4, there is no clear choice between the top two performing models i.e., *Bayes+AvU* and *Ensemble* for uncertainty-error correspondence. The visual results, however, indicate that the *Ensemble* model is more uncertain in accurate regions. Also, for all the datasets, the *Det* model has high AvU scores when compared to the *Bayes+AvU* model (Section 4.8.3). Here, it is important to consider that the AvU metric (Equation (4.7)) is essentially uncertainty accuracy, and thus, also comes with its own pitfalls. Given that all models had a DICE value which leads to more accurate terms and less inaccurate terms, the AvU metric got skewed due to the large count of n_{ac} terms. However, upon factoring the ROC and PRC curves, it becomes evident that the *Det* model is not the best performing for uncertainty-error correspondence.

Finally, all calibration-focused methods - *Focal, ECP, LS, SVLS* and *MBLS* had ROC and PRC metrics lower than the baseline *Bayes* model indicating that training for model calibration may not necessarily translate to uncertainty outputs useful for error detection.

4.5.3 Future Work

In a radiotherapy setting, the goal is to maximize radiation to tumorous regions and minimize it for healthy organs. This goal is often not optimally achieved due to imperfect contours caused by time constraints and amorphous region-of-interest boundaries on medical scans. Thus, an extension of our work could evaluate the contouring corrections made by clinicians in response to uncertainty-proposed errors in context of the dose changes to the different regions of interest. Such an experiment can better evaluate the clinical utility of an uncertainty-driven error correction workflow.

4.6 Conclusion

This work investigates the usage of the Accuracy-vs-Uncertainty (AvU) metric to improve clinical “utility" of deep Bayesian uncertainty as a proxy for error detection in segmentation settings. Experimental results indicate that using a differentiable AvU metric as an objective to train Bayesian segmentation models has a positive effect on uncertainty-error correspondence metrics. We show that our AvU-trained Bayesian models have equivalent or improved uncertainty-error correspondence metrics when compared to various calibrative and uncertainty-based methods. Given that our approach is a loss function, it can be used with other neural architectures capable of estimating uncertainty.

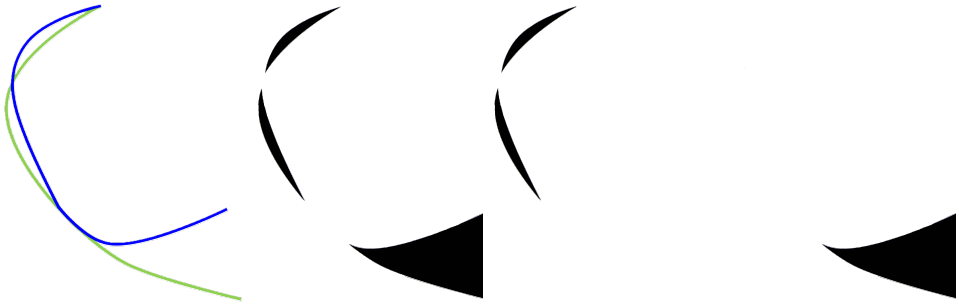
Given that deep learning models have shown the capability of reaching near expert-level performance in medical image segmentation, one of the next steps in their evolution is evaluating their clinical utility. Our work shows progress on this using a uncertainty-driven loss in a Bayesian setting. We do this for two radiotherapy body-sites and modalities as well in an out-of-distribution setting. Our hope is that the community is inspired by our positive results to further contribute to human-centric approaches to deep learning-based modeling.

4.7 Acknowledgement

The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022) and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate, The Netherlands.

4.8 Appendix

4.8.1 Segmentation “Failures" and “Errors"



(a) Contours (b) Inaccuracy Map (c) Segmentation Errors (d) Segmentation Failures

Figure 4.5: The green and blue contours in a) show the ground truth (GT) and predicted contours. In b) we see the inaccuracy map in black, while c) and d) show the smaller segmentation “errors” and larger segmentation “failures” respectively.

4.8.2 Weightage of AvU loss

The table below show the weights used for the AvU loss which were finetuned on the validation datasets of the head-and-neck CT and prostate MR. The final weightage was chosen by identifying the inflection point at which the *ROC-AUC* and *PRC-AUC* drop precipitously. Given that the AvU loss is a log term, its values are inherently small (≤ 1.0). This is then added to the cross-entropy term, which is a sum of logs (Eqn (3)) over all the voxels (=N) and all the classes (=C). Thus, we used a balancing term in the range of 10^1 to 10^4.

Table 4.3: Uncertainty-error correspondence results (higher is better) to select the weightage of the AvU loss. Underlined numbers indicate the maximum value for a metric.

Validation Dataset	Model	AvU-AUC (x10^{-2})	ROC-AUC (x10^{-2})	PRC-AUC (x10^{-2})
H&N CT (MICCAI2015)	Bayes	34.1 ± 0.7	79.1 ± 4.7	25.9 ± 2.9
	Bayes + 10AvU	34.5 ± 0.9	78.2 ± 6.0	26.1 ± 3.4
	Bayes + 100AvU	35.5 ± 0.6	79.6 ± 4.8	28.0 ± 3.5
	Bayes + 1000AvU	35.9 ± 6.9	76.4 ± 5.8	23.1 ± 1.7
Prostate MR (ProstateX)	Bayes	93.2 ± 1.8	95.3 ± 1.9	30.3 ± 2.9
	Bayes + 100AvU	94.9 ± 2.1	95.9 ± 2.0	31.5 ± 3.5
	Bayes + 1000AvU	95.5 ± 1.9	96.3 ± 2.4	32.0 ± 3.3
	Bayes + 10000AvU	96.1 ± 1.7	93.1 ± 2.1	29.3 ± 3.1

4.8.3 Hyperparameter selection

In the tables shown below, we report results for different hyperparameters of different model classes. If the DICE of a hyperparameter is 10.0 points lower than the class maximum, we ignore it. We also ignore models with large drops in ECE or AvU-AUC when compared to models in its own class. To choose the best hyperparameter, it has to perform as the best in four out of the five metrics, else we chose the middlemost hyperparameter.

Table 4.4: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on head-and-neck validation dataset for the purpose of hyperparameter selection. The experiment indicated as **bold** is the one with the best performance.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	83.6 ± 2.2	8.5 ± 1.6	35.1 ± 1.1	74.8 ± 5.0	24.5 ± 0.9
Det + 10AvU	83.4 ± 1.7	8.4 ± 1.6	35.4 ± 1.0	74.4 ± 4.8	23.2 ± 1.4
Det + 100AvU	83.6 ± 1.4	8.1 ± 1.5	36.1 ± 0.6	75.6 ± 2.9	23.4 ± 2.1
Det + 1000AvU	58.1 ± 6.9	14.7 ± 4.0	30.2 ± 1.6	78.8 ± 4.6	23.0 ± 10.6
Focal(γ=1)	84.1 ± 0.8	8.0 ± 0.6	32.4 ± 0.7	73.9 ± 4.1	21.5 ± 3.4
Focal(γ=2)	83.4 ± 1.3	9.6 ± 8.3	24.8 ± 1.1	73.7 ± 1.6	22.7 ± 4.1
Focal(γ=3)	84.1 ± 1.9	15.5 ± 1.9	17.5 ± 7.4	73.1 ± 3.2	12.6 ± 1.9
ECP(λ=0.1)	83.9 ± 1.3	8.3 ± 1.3	35.3 ± 0.8	75.1 ± 4.3	22.3 ± 0.8
ECP(λ=1.0)	84.0 ± 1.1	8.5 ± 0.8	35.4 ± 0.7	75.3 ± 3.2	23.4 ± 1.4
ECP(λ=10.0)	83.2 ± 2.1	8.7 ± 1.3	35.2 ± 0.8	74.9 ± 3.9	24.6 ± 2.1
ECP(λ=100.0)	81.2 ± 6.4	17.9 ± 1.5	28.7 ± 2.8	65.4 ± 5.6	17.5 ± 5.2
LS(α=0.01)	83.0 ± 2.1	8.1 ± 1.3	32.6 ± 0.9	70.9 ± 2.6	23.4 ± 2.7
LS(α=0.05)	83.6 ± 1.2	6.1 ± 1.0	24.9 ± 0.4	64.5 ± 3.3	18.1 ± 2.0
LS(α=0.1)	83.5 ± 1.2	7.9 ± 1.2	17.5 ± 0.1	63.9 ± 2.2	22.2 ± 1.1
SVLS(σ=1)	83.5 ± 1.3	7.7 ± 0.7	32.3 ± 0.8	71.5 ± 2.5	19.9 ± 0.4
SVLS(σ=2)	83.5 ± 1.7	8.1 ± 0.9	31.8 ± 1.0	70.5 ± 3.8	17.7 ± 1.5
SVLS(σ=3)	84.1 ± 2.0	7.7 ± 0.7	31.9 ± 1.0	71.3 ± 4.4	19.2 ± 3.2
MbLS(λ = 0.1,m=30)	82.7 ± 1.8	8.5 ± 0.6	34.9 ± 1.0	74.0 ± 4.3	23.1 ± 1.2
MbLS(λ = 0.1,m=20)	84.4 ± 1.4	8.0 ± 1.1	35.2 ± 0.7	72.3 ± 3.3	20.4 ± 1.0
MbLS(λ = 0.1,m=10)	82.7 ± 1.8	8.5 ± 0.6	32.9 ± 0.7	68.4 ± 3.0	21.7 ± 2.2
MbLS(λ = 0.1,m=8)	62.9 ± 7.6	18.75 ± 1.4	26.0 ± 0.4	74.9 ± 4.0	39.1 ± 2.7
MbLS(λ = 1,m=20)	83.2 ± 1.3	8.9 ± 1.5	35.0 ± 0.9	72.4 ± 4.4	22.5 ± 1.1
MbLS(λ = 10,m=20)	83.4 ± 1.4	8.5 ± 2.0	34.2 ± 1.1	72.2 ± 4.4	23.1 ± 2.0
MbLS(λ = 100,m=20)	81.8 ± 1.8	8.0 ± 1.1	32.1 ± 0.9	69.6 ± 4.8	21.0 ± 2.3
TTA	83.5 ± 2.2	8.5 ± 1.7	34.9 ± 1.1	75.3 ± 5.2	25.2 ± 1.7
Ens	84.9 ± 1.6	6.8 ± 0.9	34.1 ± 1.1	80.8 ± 3.2	28.2 ± 4.1
Bayes	84.2 ± 2.9	7.8 ± 1.3	34.1 ± 0.7	79.1 ± 4.7	25.9 ± 2.9
Bayes + 10AvU	83.1 ± 2.9	7.6 ± 2.0	34.5 ± 0.9	78.2 ± 6.0	26.1 ± 3.4
Bayes + 100AvU	83.2 ± 1.7	7.0 ± 1.9	35.5 ± 0.6	79.6 ± 4.8	28.0 ± 3.5
Bayes + 1000AvU	84.3 ± 1.0	7.5 ± 1.5	35.9 ± 6.9	76.4 ± 5.8	23.1 ± 1.7

Table 4.5: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on head-and-neck iD dataset. The experiment indicated as **bold** is the one with the best performance. * indicates hyperparameters chosen by the validation dataset.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	84.2 ± 2.7	9.0 ± 2.1	35.5 ± 1.5	73.0 ± 5.7	21.0 ± 4.8
Det + 10AvU	83.7 ± 2.3	9.3 ± 2.2	35.7 ± 1.3	70.6 ± 5.3	20.0 ± 3.6
Det + 100AvU*	83.8 ± 2.9	8.6 ± 2.7	36.2 ± 1.4	73.1 ± 6.0	20.8 ± 4.0
Det + 1000AvU	62.3 ± 5.6	12.1 ± 2.9	30.7 ± 1.2	78.0 ± 4.6	16.0 ± 9.0
Focal(γ=1)*	84.3 ± 2.4	9.3 ± 1.5	32.5 ± 0.9	70.3 ± 5.5	18.2 ± 3.2
Focal(γ=2)	84.2 ± 2.0	11.2 ± 1.6	25.1 ± 0.7	69.4 ± 4.9	17.2 ± 3.0
Focal(γ=3)	83.9 ± 2.5	15.7 ± 2.3	17.9 ± 5.3	70.5 ± 5.0	12.2 ± 2.9
ECP(λ=0.1)	84.4 ± 2.2	8.9 ± 2.1	35.7 ± 1.3	72.9 ± 6.3	20.1 ± 3.8
ECP(λ=1.0)*	84.4 ± 2.3	9.0 ± 2.0	35.9 ± 1.3	73.8 ± 5.4	21.0 ± 3.7
ECP(λ=10.0)	84.3 ± 2.7	9.2 ± 2.4	35.8 ± 1.4	73.5 ± 6.0	20.6 ± 4.3
ECP(λ=100.0)	70.8 ± 3.9	18.6 ± 2.8	21.4 ± 2.7	58.7 ± 2.6	28.9 ± 7.1
LS(α=0.01)	83.4 ± 2.8	9.0 ± 2.9	32.9 ± 0.1	66.1 ± 5.7	18.4 ± 3.6
LS(α=0.05)*	83.0 ± 3.0	7.5 ± 2.2	25.1 ± 0.5	62.6 ± 3.3	17.5 ± 4.0
LS(α=0.1)	84.1 ± 2.3	8.4 ± 2.9	17.5 ± 0.1	62.3 ± 2.5	18.5 ± 3.5
SVLS(σ=1)*	83.9 ± 2.5	9.0 ± 2.3	32.6 ± 1.1	69.6 ± 8.3	18.8 ± 2.8
SVLS(σ=2)	84.2 ± 2.6	9.0 ± 2.0	32.2 ± 1.1	70.8 ± 7.1	18.1 ± 3.5
SVLS(σ=3)	83.9 ± 2.7	9.0 ± 2.2	32.1 ± 1.2	69.3 ± 7.0	18.8 ± 2.8
MbLS(λ = 0.1,m=30)	83.7 ± 2.6	9.0 ± 2.0	35.4 ± 1.2	70.0 ± 5.6	19.7 ± 4.1
MbLS(λ = 0.1,m=20)*	84.0 ± 2.6	9.2 ± 2.1	35.3 ± 1.3	67.5 ± 5.7	19.5 ± 3.5
MbLS(λ = 0.1,m=10)	82.4 ± 2.6	9.8 ± 2.8	33.1 ± 1.2	64.1 ± 7.0	18.3 ± 3.1
MbLS(λ = 0.1,m=8)	62.4 ± 8.2	18.9 ± 1.6	26.3 ± 0.4	73.6 ± 6.6	38.3 ± 4.1
MbLS(λ = 1,m=20)	83.4 ± 2.5	9.2 ± 2.6	35.4 ± 1.3	71.3 ± 7.0	20.1 ± 3.6
MbLS(λ = 10,m=20)	83.0 ± 3.4	9.5 ± 2.8	34.6 ± 1.4	69.1 ± 6.0	20.1 ± 4.1
MbLS(λ = 100,m=20)	82.5 ± 3.2	9.1 ± 3.0	32.4 ± 1.4	68.2 ± 7.3	19.0 ± 2.9
TTA	84.1 ± 2.8	9.1 ± 2.1	35.5 ± 1.4	72.9 ± 5.9	20.8 ± 3.9
Ens	85.0 ± 2.7	7.8 ± 1.9	34.5 ± 1.2	78.6 ± 4.7	25.7 ± 6.8
Bayes	83.9 ± 2.6	8.7 ± 2.1	34.5 ± 1.2	74.1 ± 5.4	22.1 ± 3.5
Bayes + 10AvU	83.4 ± 2.8	8.7 ± 2.4	34.7 ± 1.3	74.7 ± 4.9	24.4 ± 4.1
Bayes + 100AvU*	83.6 ± 2.5	7.6 ± 2.5	35.6 ± 1.2	76.1 ± 5.6	25.1 ± 5.3
Bayes + 1000AvU	83.5 ± 3.0	8.5 ± 3.4	36.1 ± 1.5	77.2 ± 6.0	24.7 ± 4.5

Table 4.6: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on head-and-neck OOD dataset. The experiment indicated as **bold** is the one with the best performance. * indicates hyperparameters chosen by the validation dataset.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	78.1 ± 4.6	12.9 ± 2.6	33.4 ± 1.4	62.2 ± 4.5	24.1 ± 3.7
Det + 10AvU	76.3 ± 6.9	13.7 ± 3.5	33.3 ± 1.7	58.3 ± 4.6	23.3 ± 4.4
Det + 100AvU*	78.6 ± 4.7	12.7 ± 3.0	34.2 ± 1.5	60.8 ± 4.7	22.4 ± 4.1
Det + 1000AvU	42.5 ± 7.2	12.1 ± 2.1	28.9 ± 1.7	66.1 ± 5.8	19.0 ± 6.2
Focal(γ=1)*	77.2 ± 6.7	12.5 ± 2.9	30.6 ± 1.7	57.0 ± 4.6	20.9 ± 4.2
Focal(γ=2)	77.7 ± 5.2	12.2 ± 1.9	24.1 ± 0.9	57.5 ± 4.6	21.0 ± 4.1
Focal(γ=3)	79.0 ± 5.2	13.3 ± 1.6	18.6 ± 0.7	59.8 ± 4.9	16.6 ± 3.9
ECP(λ=0.1)	78.5 ± 4.9	12.6 ± 2.8	33.5 ± 1.6	59.8 ± 4.9	22.0 ± 3.8
ECP(λ=1.0)*	78.8 ± 4.3	12.5 ± 2.6	36.6 ± 1.5	61.5 ± 4.8	23.2 ± 3.6
ECP(λ=10.0)	78.9 ± 4.5	12.4 ± 2.5	33.8 ± 1.5	60.1 ± 4.7	22.1 ± 3.5
ECP(λ=100.0)	62.0 ± 6.1	20.0 ± 1.8	19.9 ± 2.9	56.0 ± 2.8	36.5 ± 9.7
LS(α=0.1)	77.7 ± 6.0	8.9 ± 2.7	17.9 ± 0.3	57.6 ± 1.9	23.9 ± 4.4
LS(α=0.05)*	77.7 ± 6.0	10.3 ± 2.9	24.3 ± 0.7	56.7 ± 3.3	20.6 ± 4.3
LS(α=0.01)	77.9 ± 5.4	13.3 ± 2.8	31.1 ± 1.5	58.6 ± 3.9	22.4 ± 3.7
SVLS(σ=1)*	78.3 ± 6.1	11.5 ± 3.0	31.4 ± 1.4	61.1 ± 4.9	23.3 ± 3.3
SVLS(σ=2)	79.0 ± 6.0	11.3 ± 2.5	31.4 ± 1.2	59.9 ± 5.4	21.6 ± 2.7
SVLS(σ=3)	78.6 ± 5.1	11.5 ± 2.9	31.1 ± 1.5	58.7 ± 5.0	22.5 ± 3.8
MbLS(λ = 0.1,m=30)	76.5 ± 7.1	13.6 ± 3.9	32.1 ± 2.9	58.9 ± 4.1	24.7 ± 7.7
MbLS(λ = 0.1,m=20)*	77.5 ± 6.3	13.4 ± 3.0	33.4 ± 1.5	56.9 ± 5.0	21.5 ± 3.6
MbLS(λ = 0.1,m=10)	76.8 ± 6.3	13.0 ± 3.2	31.7 ± 1.4	53.0 ± 4.5	20.6 ± 3.9
MbLS(λ = 0.1,m=8)	50.3 ± 10.6	20.1 ± 2.8	26.2 ± 0.9	61.1 ± 7.1	34.1 ± 3.7
MbLS(λ = 1,m=20)	77.3 ± 6.2	13.2 ± 2.8	33.3 ± 1.6	61.0 ± 4.5	23.4 ± 4.1
MbLS(λ = 10,m=20)	78.1 ± 5.3	13.0 ± 2.9	32.9 ± 1.5	57.0 ± 4.1	21.7 ± 3.5
MbLS(λ = 100,m=20)	78.2 ± 4.9	12.7 ± 2.5	31.6 ± 1.3	55.0 ± 5.1	19.7 ± 3.5
TTA	78.1 ± 4.6	12.7 ± 2.6	33.2 ± 1.5	62.7 ± 4.6	24.9 ± 4.1
Ens	78.6 ± 5.2	10.6 ± 2.4	32.1 ± 1.9	64.7 ± 4.9	28.2 ± 5.1
Bayes	75.0 ± 9.9	12.4 ± 4.0	32.2 ± 1.8	64.8 ± 5.0	27.7 ± 5.8
Bayes + 10AvU	74.9 ± 9.5	12.4 ± 4.0	32.1 ± 2.0	65.2 ± 4.6	29.1 ± 6.1
Bayes + 100AvU*	76.3 ± 7.7	12.1 ± 3.7	33.2 ± 1.7	65.8 ± 5.0	30.1 ± 6.5
Bayes + 1000AvU	75.5 ± 8.2	14.3 ± 4.1	33.5 ± 1.8	69.3 ± 5.6	32.9 ± 6.9

Table 4.7: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on prostate validation dataset for the purpose of hyperparameter selection. The experiment indicated as **bold** is the one with the best performance.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	85.9 ± 1.8	14.4 ± 3.2	96.5 ± 0.9	92.6 ± 4.1	26.5 ± 1.5
Det + 100AvU	84.8 ± 2.3	16.3 ± 3.9	96.1 ± 0.9	91.7 ± 4.2	27.9 ± 2.8
Det + 1000AvU	84.8 ± 1.9	16.0 ± 3.0	96.4 ± 0.9	93.6 ± 3.2	29.2 ± 1.0
Det + 10000AvU	84.9 ± 3.5	16.7 ± 5.1	96.5 ± 1.0	91.8 ± 2.7	25.9 ± 2.3
Ensemble	85.4 ± 1.7	13.4 ± 3.0	96.0 ± 1.0	94.8 ± 2.4	31.4 ± 1.6
Focal(γ=1)	84.5 ± 2.7	13.3 ± 4.3	90.7 ± 1.1	93.0 ± 4.1	29.4 ± 1.7
Focal(γ=2)	84.4 ± 2.1	9.8 ± 2.6	82.5 ± 1.0	93.8 ± 2.3	30.9 ± 2.1
Focal(γ=3)	84.5 ± 1.9	6.4 ± 1.5	58.9 ± 1.3	92.0 ± 4.3	30.5 ± 2.6
ECP(λ=0.1)	85.9 ± 1.8	14.6 ± 3.0	96.5 ± 0.9	91.9 ± 4.2	25.8 ± 1.7
ECP(λ=1.0)	85.7 ± 1.8	14.7 ± 3.0	96.4 ± 1.0	92.3 ± 3.9	26.4 ± 1.7
ECP(λ=10.0)	85.7 ± 1.7	14.8 ± 2.7	96.4 ± 1.0	91.9 ± 4.5	26.0 ± 1.8
ECP(λ=100.0)	85.7 ± 1.8	14.8 ± 2.8	96.4 ± 1.0	91.8 ± 4.3	25.8 ± 1.9
ECP(λ=1000.0)	86.0 ± 1.9	15.0 ± 3.0	85.0 ± 0.3	88.7 ± 2.1	26.7 ± 3.4
LS(α=0.01)	83.7 ± 2.5	17.2 ± 4.1	91.9 ± 0.9	85.8 ± 5.4	28.2 ± 2.9
LS(α=0.05)	85.1 ± 1.4	13.6 ± 2.3	80.8 ± 0.9	84.3 ± 5.7	25.4 ± 2.2
LS(α=0.1)	85.0 ± 2.1	11.1 ± 3.4	70.3 ± 0.6	85.1 ± 3.6	27.0 ± 2.2
SVLS(σ=1)	84.5 ± 1.9	14.0 ± 2.6	92.4 ± 1.0	91.8 ± 2.3	22.9 ± 1.8
SVLS(σ=2)	85.0 ± 1.8	12.9 ± 3.1	92.4 ± 0.9	91.4 ± 3.0	22.1 ± 1.4
SVLS(σ=3)	85.0 ± 1.6	13.1 ± 2.7	92.1 ± 0.9	91.2 ± 2.5	21.9 ± 1.4
MbLS(λ = 0.1,m=10)	84.8 ± 1.4	17.5 ± 5.1	95.7 ± 1.1	91.2 ± 4.1	31.1 ± 1.7
MbLS(λ = 0.1,m=8)	83.8 ± 1.3	16.0 ± 2.2	93.9 ± 0.9	90.5 ± 3.5	27.9 ± 2.1
MbLS(λ = 0.1,m=5)	84.3 ± 1.6	15.5 ± 2.8	90.4 ± 0.8	90.1 ± 5.6	28.2 ± 2.2
MbLS(λ = 0.1,m=3)	84.2 ± 2.1	12.8 ± 3.3	70.8 ± 0.4	82.1 ± 3.5	28.8 ± 5.6
MbLS(λ = 1.0,m=10)	83.7 ± 1.2	17.4 ± 4.4	96.0 ± 1.0	91.2 ± 3.9	30.5 ± 1.5
MbLS(λ = 10.0,m=10)	83.9 ± 1.5	17.8 ± 4.1	95.0 ± 1.2	90.8 ± 3.6	30.9 ± 1.6
TTA	85.6 ± 1.7	14.5 ± 3.1	96.3 ± 0.9	92.5 ± 4.0	27.2 ± 1.6
Bayes	85.7 ± 2.3	10.7 ± 3.0	93.2 ± 1.8	95.3 ± 1.9	30.3 ± 2.9
Bayes + 100AvU	86.1 ± 3.0	11.5 ± 3.8	94.9 ± 2.1	95.9 ± 2.0	31.5 ± 3.5
Bayes + 1000AvU	85.8 ± 2.8	12.0 ± 3.9	95.5 ± 1.9	96.3 ± 2.4	32.0 ± 3.3
Bayes + 10000AvU	86.0 ± 2.4	10.9 ± 3.0	96.1 ± 1.7	93.1 ± 2.1	29.3 ± 3.1

Table 4.8: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on prostate ID dataset. The experiment indicated as **bold** is the one with the best performance. * indicates hyperparameters chosen by the validation dataset.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	84.1 ± 5.6	12.9 ± 6.0	96.1 ± 3.4	92.5 ± 5.7	28.0 ± 3.7
Det + 100AvU	83.7 ± 6.7	16.6 ± 7.2	95.7 ± 3.3	91.6 ± 6.2	27.2 ± 2.9
Det + 1000AvU*	83.7 ± 6.8	16.9 ± 8.1	95.9 ± 3.8	92.1 ± 6.8	28.2 ± 3.4
Det + 10000AvU	83.4 ± 6.4	18.1 ± 7.9	96.1 ± 3.7	90.7 ± 5.6	26.1 ± 3.4
Focal (γ=1)*	81.1 ± 15.4	10.2 ± 5.0	90.3 ± 0.3	93.2 ± 5.5	29.3 ± 3.4
Focal (γ=2)	83.1 ± 6.2	10.4 ± 6.8	81.6 ± 2.5	92.9 ± 5.3	30.1 ± 3.7
Focal (γ=3)	82.3 ± 7.2	8.0 ± 6.4	58.7 ± 1.2	92.5 ± 5.4	31.8 ± 3.5
ECP (λ=0.1)	84.1 ± 5.4	16.5 ± 7.0	96.1 ± 3.4	92.3 ± 6.0	27.6 ± 3.9
ECP (λ=1.0)	84.1 ± 5.5	16.4 ± 7.0	96.1 ± 3.3	92.3 ± 6.0	27.8 ± 4.3
ECP (λ=10.0)*	84.0 ± 5.5	16.7 ± 7.1	96.1 ± 3.4	92.1 ± 6.0	27.6 ± 4.3
ECP (λ=100.0)	84.0 ± 5.5	16.6 ± 7.0	96.0 ± 3.0	92.1 ± 6.0	27.6 ± 4.1
ECP (λ=1000.0)	84.1 ± 5.7	16.6 ± 7.0	86.1 ± 3.2	92.2 ± 5.9	27.5 ± 4.3
LS (α=0.01)	82.5 ± 8.3	18.0 ± 9.4	91.3 ± 3.9	86.2 ± 7.9	27.0 ± 3.8
LS (α=0.05)*	83.4 ± 7.2	15.1 ± 8.6	80.4 ± 2.9	83.2 ± 7.8	25.1 ± 3.1
LS (α=0.1)	84.1 ± 5.6	11.6 ± 7.0	70.1 ± 1.8	84.7 ± 6.2	26.9 ± 3.3
SVLS (σ=1)	83.4 ± 7.1	14.7 ± 8.8	92.0 ± 3.7	90.9 ± 7.4	22.9 ± 2.9
SVLS (σ=2)*	83.5 ± 6.7	14.0 ± 8.1	91.9 ± 4.1	90.5 ± 7.9	21.7 ± 2.6
SVLS(σ=3)	83.2 ± 8.1	14.3 ± 9.7	91.5 ± 3.9	91.0 ± 6.8	23.1 ± 3.1
MbLS (λ = 1.0,m=3)	83.2 ± 6.3	13.3 ± 7.8	70.6 ± 1.7	82.2 ± 6.3	27.7 ± 3.4
MbLS (λ = 1.0,m=5)	82.8 ± 6.6	16.7 ± 8.0	89.9 ± 3.2	90.5 ± 7.2	27.2 ± 4.4
MbLS (λ = 1.0,m=8)	83.5 ± 5.8	17.1 ± 7.0	95.3 ± 3.6	93.0 ± 5.2	27.8 ± 4.1
MbLS (λ = 1.0,m=10)	84.2 ± 5.3	18.1 ± 6.1	95.5 ± 3.3	91.7 ± 6.1	26.5 ± 3.5
MbLS(λ = 1.0,m=10)*	84.2 ± 4.9	17.9 ± 7.4	95.6 ± 2.9	92.2 ± 5.6	26.9 ± 3.6
MbLS(λ = 10.0,m=10)	83.9 ± 5.2	17.9 ± 8.0	95.1 ± 3.2	91.9 ± 5.9	26.2 ± 4.1
TTA	83.8 ± 5.8	16.4 ± 7.1	96.0 ± 3.5	92.7 ± 5.6	28.8 ± 3.9
Ensemble	84.5 ± 5.7	11.3 ± 6.5	95.2 ± 3.5	94.3 ± 4.3	30.0 ± 4.6
Bayes	84.0 ± 5.8	8.6 ± 4.7	92.1 ± 2.6	94.7 ± 3.1	29.1 ± 4.8
Bayes + 100AvU	84.1 ± 6.4	12.0 ± 6.2	94.4 ± 3.1	95.5 ± 2.9	28.9 ± 5.0
Bayes + 1000AvU*	84.9 ± 6.9	8.9 ± 6.0	94.5 ± 3.2	95.7 ± 3.2	30.5 ± 4.5
Bayes + 10000AvU	85.2 ± 5.9	11.0 ± 6.3	94.2 ± 3.6	95.9 ± 3.5	30.2 ± 4.0

Table 4.9: Volumetric (DICE), calibrative (ECE) and uncertainty-error correspondence metrics (AvU-AUC, ROC-AUC, PRC-AUC) on prostate OOD dataset. The experiment indicated as **bold** is the one with the best performance. * indicates hyperparameters chosen by the validation dataset.

Experiment	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	AvU-AUC ↑ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC ↑ (x10^{-2})
Det	74.2 ± 12.6	15.6 ± 6.3	92.3 ± 5.4	87.9 ± 7.5	22.1 ± 6.2
Det + 100AvU	74.2 ± 13.3	23.6 ± 11.2	93.0 ± 4.2	87.1 ± 6.2	22.2 ± 5.7
Det + 1000AvU*	74.5 ± 13.0	27.6 ± 14.3	92.2 ± 5.7	88.2 ± 7.6	22.0 ± 7.1
Det + 10000AvU	72.7 ± 15.1	27.6 ± 14.3	92.4 ± 5.2	82.3 ± 9.4	19.6 ± 6.2
Focal(γ=1)*	71.2 ± 17.4	12.1 ± 5.8	85.4 ± 6.1	89.0 ± 7.1	24.3 ± 6.7
Focal(γ=2)	76.7 ± 10.8	12.8 ± 8.2	72.0 ± 9.3	87.2 ± 7.6	22.4 ± 6.4
Focal(γ=3)	73.2 ± 13.7	11.6 ± 7.7	49.7 ± 9.4	87.1 ± 8.5	27.0 ± 7.2
ECP(λ=0.1)	74.6 ± 12.5	22.8 ± 10.5	92.1 ± 5.5	87.6 ± 7.6	21.3 ± 6.6
ECP(λ=1.0)	73.9 ± 13.1	23.2 ± 10.7	91.9 ± 5.6	87.2 ± 7.2	21.2 ± 6.4
ECP(λ=10.0)*	74.8 ± 12.5	22.3 ± 10.2	91.6 ± 6.3	87.2 ± 8.1	20.6 ± 7.0
ECP(λ=100.0)	74.9 ± 12.3	22.7 ± 10.5	92.1 ± 5.5	87.7 ± 8.0	21.5 ± 7.2
ECP(λ=1000.0)	74.6 ± 12.5	22.7 ± 10.3	92.2 ± 5.6	87.6 ± 7.7	21.5 ± 6.7
LS(α=0.01)	71.6 ± 15.1	24.6 ± 11.6	87.9 ± 5.3	84.3 ± 7.5	22.7 ± 6.2
LS(α=0.05)*	74.5 ± 13.0	21.7 ± 11.5	77.2 ± 4.6	79.5 ± 8.9	19.1 ± 7.2
LS(α=0.1)	75.2 ± 12.2	18.1 ± 10.1	67.4 ± 3.8	79.0 ± 8.4	19.9 ± 6.4
SVLS(σ=1)	74.9 ± 11.7	19.7 ± 9.1	88.5 ± 5.5	87.2 ± 7.4	18.7 ± 5.1
SVLS(σ=2)*	76.9 ± 11.5	17.9 ± 9.3	88.3 ± 5.2	87.2 ± 7.2	16.4 ± 5.2
SVLS(σ=3)	74.3 ± 13.5	21.4 ± 12.6	88.4 ± 5.1	86.3 ± 8.2	19.4 ± 5.0
MbLS(λ = 0.1, m=10)	72.3 ± 15.9	20.9 ± 7.9	91.4 ± 5.7	87.9 ± 6.9	22.2 ± 6.7
MbLS(λ = 0.1, m=8)	74.1 ± 13.5	20.7 ± 8.7	88.3 ± 8.2	85.0 ± 10.4	18.8 ± 8.8
MbLS(λ = 0.1, m=5)	74.7 ± 13.3	22.0 ± 11.3	86.9 ± 5.0	87.1 ± 7.9	22.0 ± 6.4
MbLS(λ = 0.1, m=3)	74.0 ± 13.3	20.5 ± 11.7	68.6 ± 2.9	78.0 ± 7.2	21.5 ± 6.7
MbLS(λ = 1.0, m=10)*	73.6 ± 12.5	19.9 ± 7.4	91.8 ± 3.4	86.5 ± 7.2	21.8 ± 5.6
MbLS(λ = 10.0, m=10)	72.1 ± 16.1	20.2 ± 6.7	91.4 ± 5.5	86.5 ± 9.0	22.2 ± 6.7
TTA	74.0 ± 12.8	23.7 ± 11.4	92.8 ± 4.8	88.6 ± 7.4	24.9 ± 5.8
Ensemble	76.3 ± 12.2	9.7 ± 5.0	89.9 ± 6.6	91.6 ± 5.2	28.4 ± 5.7
Bayes	70.6 ± 16.6	11.8 ± 7.2	86.2 ± 6.0	89.1 ± 7.4	25.7 ± 5.1
Bayes + 100AvU	72.1 ± 14.4	20.0 ± 11.8	91.0 ± 3.8	92.7 ± 4.0	30.2 ± 6.5
Bayes + 1000AvU*	76.3 ± 12.6	11.4 ± 6.7	89.5 ± 6.2	90.6 ± 6.9	26.2 ± 7.4
Bayes + 10000AvU	76.6 ± 12.7	17.1 ± 10.1	88.6 ± 6.5	90.4 ± 6.3	23.3 ± 7.4

4.8.4 Visual Results

Visual results in Figure 4.2 and Figure 4.3 show pairs of consecutive CT/MR slices to better understand the 3D nature of the output uncertainty across all models. We show examples with both high and low DICE to investigate the presence and absence of uncertainty in different regions of the model prediction.

4.8.5 Head-And-Neck CT

The first two rows of Figure 4.2a and Figure 4.2b show the mandible (i.e. lower jaw bone) with only the *Bayes+AvU* model having overall low uncertainty in accurate regions and high uncertainty in (or close to) inaccurate regions.

In the next set of rows for head-and-necks CTs, we observe the parotid gland, a salivary organ, with (Figure 4.2a - Case 2) and without (Figure 4.2b - Case 2, Case 3) a dental scattering issue. In both cases, while the *Det* model shows low uncertainty, the baseline *Bayes* model shows high uncertainty in accurate regions. Usage of the AvU loss lowers uncertainty in these regions, while still exhibiting uncertainty in the erroneous regions, for e.g. the medial (i.e. internal) portion of the organ in Figure 4.2a (Case 2).

Moving on to our last case, we see the submandibular gland, another salivary gland in Figure 4.2a (Case 3). The *Ensemble*, *Focal*, *SVLS* and *MBLS* models all display high uncertainty in the core of the organ, which are also accurately predicted. On the other hand, the AvU loss minimizes the uncertainty and shows uncertainty in the erroneous region on the second slice.

4.8.6 Prostate MR

For the prostate datasets, we see two cases with high DICE in Figure 4.3a (Case 1) and Figure 4.3b (Case 2) where the use of the AvU loss reduces uncertainty for the baseline *Bayes* model.

We also see cases with low DICE in Figure 4.3a (Case 2) and Figure 4.3b (Case 1). Due to their low DICE all models display high uncertainty, but the *Bayes+AvU* model shows high overlap between its uncertain and erroneous regions. The same is also observed in Figure 4.3b (Case 3).

Finally, in Figure 4.3a (Case 3), we do not see any clear benefit of using the AvU loss on the *Bayes* model.

4.8.7 BayesH model

Table 4.10: Volumetric (*DICE*) , calibrative (*ECE*) and uncertainty-error correspondence metrics (ROC-AUC, PRC-AUC) for different Bayesian models. We evaluate head-and-neck (H&N) CT and Prostate MR test datasets which are either in-distribution (ID) or out-of-distribution (OOD). The arrows in the table header indicate whether a metric should be high (↑) or low (↓). Here, † and **bold** are used to indicate a statistical significance and improved results upon comparing a Bayesian model and its AvU-loss version, while underlined numbers indicate the best value for a metric across a dataset.

Test Dataset	Model	DICE ↑ (x10^{-2})	ECE ↓ (x10^{-2})	ROC-AUC ↑ (x10^{-2})	PRC-AUC↑ (x10^{-2})
ID H&N CT (RTOG)	Det	84.2 ± 2.7	9.0 ± 2.1	73.0 ± 5.7	21.0 ± 4.8
	Ensemble	85.0 ± 2.6	8.6 ± 2.1	78.6 ± 4.7	25.7 ± 6.8
	Bayes	83.9 ± 2.6	8.6 ± 2.1	74.1 ± 5.4	22.1 ± 3.5
	Bayes+AvU	83.6 ± 2.5	**7.6 ± 2.5**†	**76.1 ± 5.6**†	**25.1 ± 5.3**†
	BayesH	83.6 ± 2.9	9.2 ± 2.6	70.4 ± 7.0	20.1 ± 3.8
	BayesH+AvU	**84.1 ± 2.7**	**8.4 ± 2.4**†	**74.1 ± 5.4**†	**21.3 ± 4.6**†
OOD H&N CT (STRSeg)	Det	78.1 ± 4.6	12.9 ± 2.6	62.2 ± 4.5	24.1 ± 3.7
	Ensemble	78.6 ± 5.2	10.6 ± 2.4	64.7 ± 4.9	28.2 ± 5.1
	Bayes	75.0 ± 9.9	12.4 ± 4.0	64.8 ± 5.0	27.7 ± 5.8
	Bayes+AvU	**76.3 ± 7.6**†	**12.1 ± 3.7**	**65.8 ± 5.0**†	**30.1 ± 6.5**†
	BayesH	77.5 ± 6.6	12.6 ± 3.3	61.1 ± 4.1	23.5 ± 4.7
	BayesH+AvU	**78.8 ± 5.1**†	**12.1 ± 3.2**	**64.8 ± 3.8**†	**23.8 ± 4.0**†
ID Prostate MR (PrMedDec)	Det	84.1 ± 5.6	12.9 ± 6.0	92.5 ± 5.7	28.0 ± 3.7
	Ensemble	84.5 ± 5.7	11.3 ± 6.5	94.3 ± 4.3	30.0 ± 4.6
	Bayes	84.0 ± 5.8	8.6 ± 4.7	94.7 ± 3.1	29.1 ± 4.8
	Bayes+AvU	**84.9 ± 6.9**	8.9 ± 6.0	**95.7 ± 3.2**†	**30.5 ± 4.5**†
	BayesH	82.3 ± 5.2	9.3 ± 4.3	93.6 ± 2.9	28.4 ± 4.2
	BayesH+AvU	**84.5 ± 6.3**†	9.4 ± 6.5	**94.9 ± 3.1**†	**30.1 ± 4.9**†
OOD Prostate MR (PR12)	Det	74.2 ± 12.6	15.6 ± 6.3	87.9 ± 7.5	22.1 ± 6.2
	Ensemble	76.3 ± 12.2	9.7 ± 5.0	91.6 ± 5.2	28.4 ± 5.7
	Bayes	70.6 ± 16.6	11.8 ± 7.2	89.1 ± 7.4	25.7 ± 5.1
	Bayes+AvU	**76.3 ± 12.6**†	**11.4 ± 6.7**†	**90.6 ± 6.9**†	**26.2 ± 7.4**†
	BayesH	71.3 ± 14.4	12.1 ± 6.7	88.9 ± 6.3	25.1 ± 4.9
	BayesH+AvU	**74.1 ± 13.8**†	**11.9 ± 6.2**†	**89.9 ± 6.2**†	**25.9 ± 5.4**†

5

Manual Brush vs AI Pencil: Evaluating tools for auto-contour refinement of head-and-neck tumors on CT+PET scans

This chapter was adapted from:

Mody, Prerak, Nicolas Chaves de Plaza, Mark Gooding, Martin de Jong, Mischa de Ridder, Niels den Hans, Jos Elbers, Klaus Hildebrandt, Marius Staring. "Manual Brush vs AI Pencil: Evaluating tools for auto-contour refinement of head-and-neck tumors on CT+PET scans." (*submitted*)

Abstract

Background and Purpose: To resolve errors in auto-contours, clinicians currently use manual brush-like tools. These can be inefficient, especially for larger errors since one needs to rectify each incorrect pixel. An alternative is AI-assisted contour refinement using sparse visual cues like pencil strokes (or scribbles) drawn within false-positive and false-negative regions. However, existing AI pencil methods are limited to evaluations using either robot users or contour refinements being propagated only in 2D. We bridge these gaps and compare the time-efficiency and contour quality of the manual brush against the AI pencil for auto-contour refinement.

Materials and Methods: We designed a web-based interface and an AI pencil to conduct auto-contour refinement sessions with both tumor contouring experts (x4) and non-experts (x7) across 6 patients. Our AI pencil supports 2D interactions to refine 3D tumor contours on head-and-neck CT + PET scans. We compared the efficiency (time) and effectiveness (DICE / surface DICE @ 2mm) of the manual brush and AI pencil.

Results: For tumor auto-contour refinement, the AI pencil was [5%-78%] faster across 42 non-expert sessions and [16%-97%] faster across 24 expert sessions. The average inter-observer variability (calculated by DICE / surface DICE@2mm) across 6 patients was equivalent between the manual brush (0.89/0.90) and AI pencil (0.90/0.92) for the expert sessions.

Conclusions: The AI pencil offers a promising alternative to traditional manual brushes in auto-contouring based radiotherapy workflows. It improves the time efficiency while maintaining final contour quality for auto-contour refinement.

5.1 Introduction

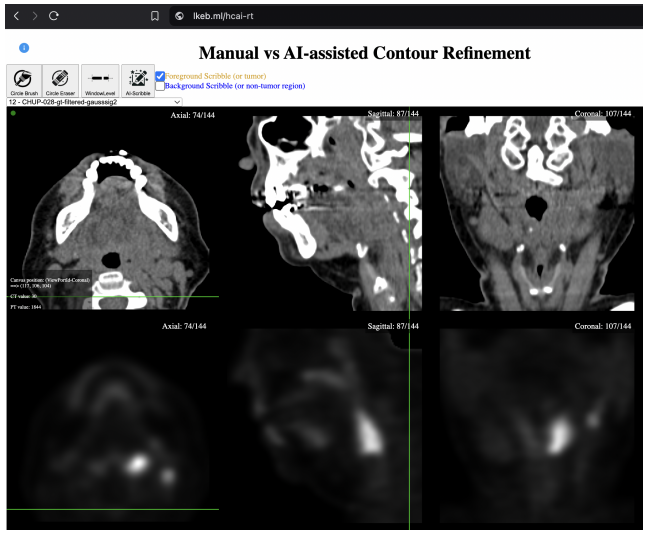
Auto-contouring in radiotherapy has made great progress over the last 5 years with improvements in AI (i.e., deep learning) models and a proliferation of clinically-available commercial tools [179–182]. Widespread use of these AI-based auto-contouring tools can be attributed to the time gains they provide. However, as these tools are still imperfect, clinicians currently perform a time-consuming manual quality assessment (QA) and refinement step [citations]. This bottleneck offsets some of the time gains provided by auto-contouring.

A few automated techniques have been proposed to reduce the auto-contour refinement bottleneck by either error-detection [80, 81, 183] or error-correction [84, 91, 92, 184]. This work focuses on the kind of error-correction wherein a user provides sparse feedback iteratively to improve an imperfect auto-contour. This feedback is usually sparse visual cues like pencil strokes (in the form of dots or scribbles) in the erroneous regions to rectify them. Literature on contouring with sparse user input has mostly reported on single-step auto-contouring with 2D [85–87] or 3D [88–90, 185–187] models. Few works report on results of iterative contour refinement [84, 91, 92, 184]. Two of the iterative contour refinement studies conducted a study with clinical users and reported time savings [92, 184] with models that takes a single modality as input. Since time-based evaluation with real clinical users is important, we build on this trend with a lightweight and multi-modal 3D model and also track the evolution of contouring metrics as more interactions were provided by the user.

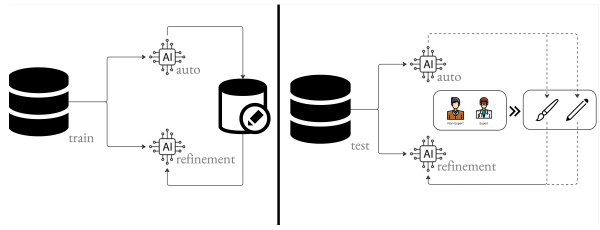
Thus, our main aim was to compare the time efficiency and contour quality of auto-contour refinement with human users across two tools. Non-experts as well as experts participated in our study on head-and-neck tumor contour refinement using both our proposed AI pencil (capable of using sparse 2D inputs to make 3D improvements) or the traditional manual brush. Additionally, we report on interaction dynamics like pixels drawn during refinement as well as inter observer variablity [188], for the multi-step refinement process. To accomplish the above, we designed and open-sourced a web-based contouring interface (Figure 5.1a)
(`https://github.com/prerakmody/interactive-autocontour-refinement`)

5.2 Materials and methods

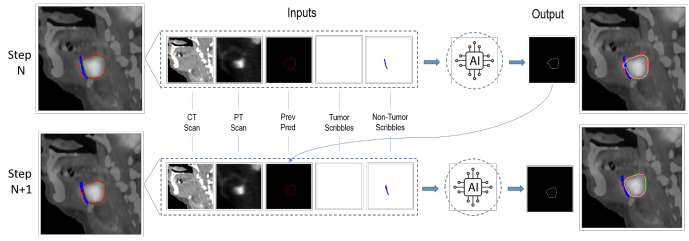
To compare the two contour refinement tools, we used a head-and-neck tumor dataset (Section 5.2.1) and trained both an auto-contouring and contour-refinement model (i.e., AI pencil) on it (Section 5.2.2). The contour refinement tools were then evaluated (Section 5.2.3 on a web interface (Section 5.2.4) by our users (Section 5.2.5).



(a) Web interface for contouring



(b) Training & testing workflow



(c) Neural network design

Figure 5.1: a) Web interface to perform contour refinement which shows axial/sagittal/coronal views for both PET + CT scans. On the top of the interface, the user can select contour editing tools (manual brush or AI pencil) and a patient from a list. b) Training workflow (left) showing the same database used to train the auto-contouring and contour-refinement models. Testing workflow (right) showing how auto-contours are modified by (non) experts using brushes (manual) or scribbles (AI-assistance). c) Inputs used within the neural net to make contour refinements with ground-truth (green), prediction (red) and refinement (pink).

5.2.1 Dataset

A head-and-neck tumor dataset from the Hecktor2022 challenge was used [44] which contains 524 pairs of CT and PET scans from seven clinics. The data originated from four countries; we used data from three of them (Canada, Switzerland, United States of America; 452 pairs) for training and validation, and from the remaining country (France; 72 pairs) for testing. More details can be found in Section 5.7.1.

5.2.2 Auto-contour and contour-refinement model training

A standard UNet architecture (~1.2M parameters) implemented in the MONAI framework [189] was chosen for both the auto-contouring and the contour-refinement model. Each model was trained using the standard cross-entropy loss. The goal here was not to achieve the best contouring performance, but rather to provide an initial segmentation for contour refinement. More details can be found in Section 5.7.2.

The auto-contouring model took as input the CT and PET scans and outputted a tumor mask, and was trained using the ground truth annotations. The contour-refinement (i.e., AI pencil) model took five inputs: CT, PET, the previously predicted contour, tumor scribbles and non-tumor scribbles and outputs a refined contour (Figure 5.1c). This model was trained using the mask predictions of the auto-contouring model as input. During training we simulated human scribbles by generating logic-based 2D scribbles in the false positive (FP) and false negative regions (FN) of the auto-contour models' predictions [84]. Depending on the region (i.e. FP or FN), the scribble was passed either as a tumor (for FN) or non-tumor (for FP) scribble. The model was then trained by comparing the refined predictions against the gold-standard.

5.2.3 Model (auto-contour and contour-refinement) validation

The outputs of both models were evaluated using the DICE metric and the surface DICE (@2mm) metric. The 2mm threshold was motivated by the HD50 results in [190]. The single-step auto-contouring model produced only one value for these metrics per patient. However, the contour refinement tools - the manual brush and AI pencil (i.e., contour-refinement model) were applied iteratively and hence evaluated at each interaction with the above metrics.

To verify whether both tools produced similar contours, we compute the inter-tool variability per patient by comparing the final contours from the manual brush and AI pencil. Moreover, we computed inter-observer-variability (IoV) [188] for each tool to determine if automation tools lead to standardization. The IoV computes metrics between the contours of multiple observers (using the same tool) and reports the median. A higher value indicates more agreement between the observers.

Finally, we logged the time taken for both refinement tools and compared them. User interaction count, pixels drawn, and slices scrolled were also logged as they directly influ-

enced the total time taken. An interaction was defined as a complete mouse click (press and release). Savings provided by the AI pencil when compared to the manual brush were also shown in percentages as:

$$\Delta M(\%) = \Big(M_{\text{manual}} - M_{\text{AI}}\Big) / M_{\text{manual}}, \tag{5.1}$$

where M can be either total time, total interactions, total pixels drawn, or total slices scrolled.

5.2.4 Web-based tool

A web-based user interface (Figure 5.1a) was developed using off-the-shelf libraries for the frontend (cornerstone3D [191]), backend (FastAPI [192]) and DICOM database (Orthanc [193]). This interface provided the manual brush, AI pencil as well as panning, zooming and scrolling capabilities for both the registered PET and CT scans. Shortcuts were provided to show/unshow contours, to change size of the manual brush as well as to switch between the foreground (tumor) and background (non-tumor) scribbles of the AI pencil. For more details check Section 5.7.3.

5.2.5 User Cohort

To compare the manual brush against the AI pencil, four head-and-neck tumor contouring experts (radiation oncologists with 2/4/11/21 years of experience) and seven non-experts (PhD candidates on AI in medical imaging) participated in this study. To support the non-experts, the ground truth tumor contour was given to them as a reference, and they were tasked to refine towards it. The experts were not shown the reference contours, and they were tasked to refine based on their expert opinion.

To compute the evaluation metrics, for the non-experts we compared each interaction against the reference contour, while for the experts we compared against their personal final contour. Consequently, the final metric value for the experts would always be the maximal (1.0).

We conducted our study in sessions, where in each session, the user is assigned a patient and a contour-refinement tool (manual brush or AI pencil, see Figure 5.1b). All users initially underwent training sessions with four patients from the validation dataset, before the start of the study. For each user, there is a gap of at least one week between the manual brush and AI pencil sessions of a patient. This reduces potential learning effects on the anatomy of that patient.

5.3 Results

The auto-contouring model achieved an average DICE score of 0.76 and average surface DICE score (@2mm) of 0.72 on the test set. From this set, we selected 5 patients with a surface DICE in the $[0.65, 0.7]$ range as cases in need of QA, and 1 patient with a surface DICE of 0.88 as a high-quality case.

	P1	P2	P3	P4	P5	P6
NE1	536 (78%)	1018 (65%)	322 (29%)	328 (46%)	96 (20%)	722 (52%)
NE2	326 (68%)	403 (57%)	440 (64%)	167 (38%)	112 (31%)	514 (56%)
NE3	319 (57%)	402 (55%)	391 (57%)	197 (47%)	223 (36%)	676 (58%)
NE4	90 (22%)	125 (24%)	194 (40%)	180 (54%)	130 (49%)	110 (19%)
NE5	112 (25%)	324 (50%)	96 (29%)	271 (54%)	201 (61%)	52 (10%)
NE6	440 (56%)	494 (78%)	229 (51%)	189 (38%)	340 (70%)	826 (61%)
NE7	199 (33%)	121 (21%)	289 (41%)	17 (5%)	246 (49%)	305 (43%)
Range (min, max)	[90,536] ([22%,78%])	[121,1018] ([21%,78%])	[96,440] ([29%,64%])	[17,328] ([5%,54%])	[96,340] ([20%,70%])	[52,826] ([10%,61%])

Table 5.1: Time savings in non-expert (NE) sessions.

	P1	P2	P3	P4	P5	P6
E1	276 (54%)	135 (52%)	52 (20%)	676 (86%)	408 (88%)	775 (97%)
E2	504 (70%)	307 (70%)	138 (42%)	304 (63%)	361 (85%)	666 (76%)
E3	230 (50%)	151 (73%)	29 (16%)	251 (85%)	238 (75%)	223 (68%)
E4	118 (40%)	71 (47%)	46 (34%)	222 (78%)	55 (83%)	292 (78%)
Range (min, max)	[118,504] ([40%,70%])	[71,307] ([47%,73%])	[29,138] ([16%,42%])	[222,676] ([63%,86%])	[55,408] ([75%,88%])	[223,775] ([68%,97%])

Table 5.2: Time savings in expert (E) sessions.

	P1	P2	P3	P4	P5	P6
NE1	97 (59%)	103 (61%)	136 (69%)	88 (67%)	49 (56%)	202 (73%)
NE2	56 (58%)	60 (66%)	78 (75%)	69 (72%)	43 (67%)	128 (81%)
NE3	105 (74%)	26 (30%)	102 (69%)	73 (70%)	73 (82%)	116 (65%)
NE4	98 (65%)	46 (42%)	170 (80%)	88 (83%)	95 (84%)	178 (78%)
NE5	55 (56%)	8 (14%)	62 (65%)	67 (67%)	64 (81%)	62 (52%)
NE6	130 (73%)	19 (19%)	111 (70%)	65 (65%)	118 (87%)	121 (66%)
NE7	152 (64)	49 (44%)	109 (71%)	47 (55%)	87 (76%)	158 (75%)
Range (min, max)	[55,130] ([56%,74%])	[8,103] ([14%,66%])	[62,170] ([65%,80%])	[47,88] ([55%,83%])	[43,118] ([56%,87%])	[62,202] ([52%,81%])

Table 5.3: Interaction count savings in non-expert (NE) sessions.

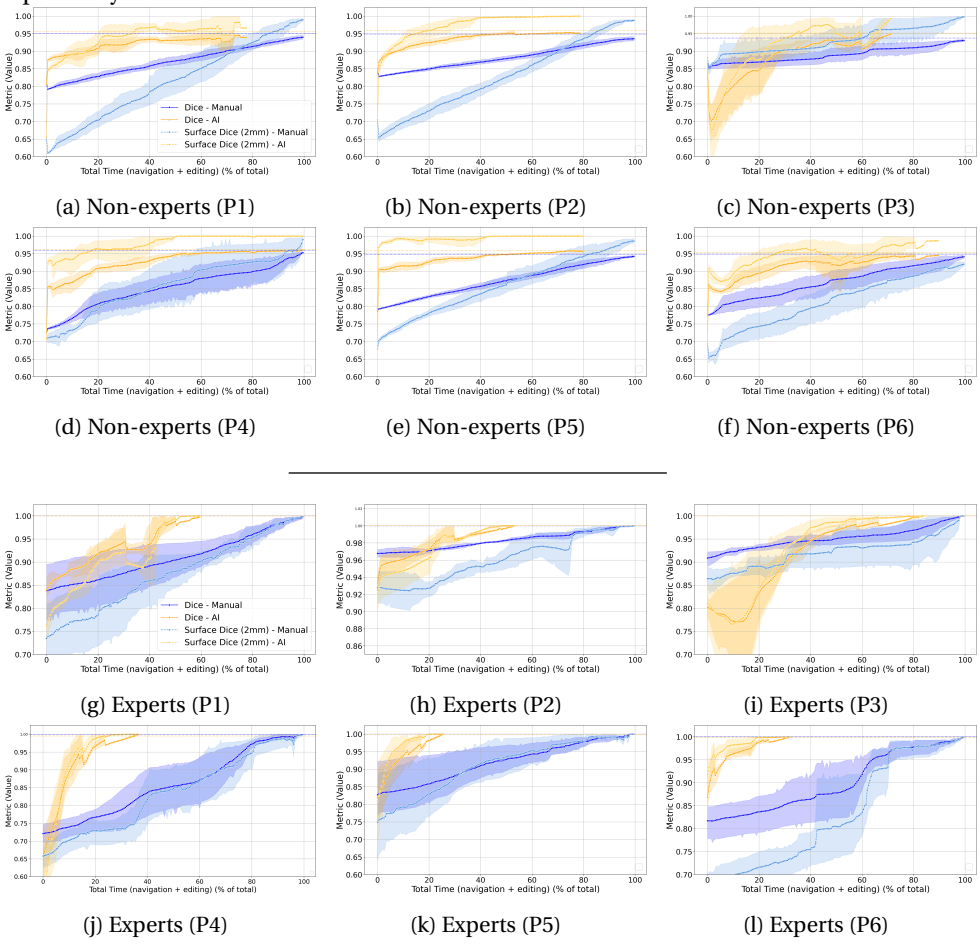
	P1	P2	P3	P4	P5	P6
E1	99 (77%)	19 (55%)	41 (67%)	117 (88%)	117 (95%)	190 (97%)
E2	119 (85%)	40 (78%)	66 (80%)	76 (80%)	69 (86%)	143 (88%)
E3	91 (81%)	18 (72%)	33 (66%)	78 (90%)	57 (90%)	87 (93%)
E4	72 (74%)	38 (84%)	43 (62%)	79 (90%)	26 (92%)	104 (95%)
Range (min, max)	[72,119] ([74%,85%])	[19,40] ([55%,84%])	[33,66] ([62%,80%])	[76,117] ([80%,90%])	[26,117] ([86%,95%])	[87,190] ([88%,97%])

Table 5.4: Interaction count savings in expert (E) sessions.

Table 5.5: Time (a,b) and interaction count (c,d) savings provided by the AI pencil when compared to manual brush for experts (E) and non-experts (NE) across patients (P). Time savings are in seconds and the percentages indicate how fast the AI pencil is when compared to the manual brush (i.e., $(T_{\text{Manual}} - T_{\text{AI}}) / T_{\text{Manual}}$.

Upon comparing the tools for the refinement of auto-contours, the AI pencil was [5% – 78%] faster for non-expert sessions (Table 5.1) and [16% – 97%] for expert sessions (Table 5.2). The same can be seen in line plots of Figure 5.2 depicting DICE (surface DICE @

2mm) performance across time. This is because the AI pencil required [14% – 87%] and [55% – 97%] fewer interactions for non-expert (Table 5.3) and expert (Table 5.4) sessions, respectively.



(a) Non-experts (P1) (b) Non-experts (P2) (c) Non-experts (P3)

(d) Non-experts (P4) (e) Non-experts (P5) (f) Non-experts (P6)

(g) Experts (P1) (h) Experts (P2) (i) Experts (P3)

(j) Experts (P4) (k) Experts (P5) (l) Experts (P6)

Figure 5.2: Line plots (with 95% CI) comparing contour refinement sessions for manual brush (in blue) and AI pencil (in orange) tools across 7 non-experts (a-f) and 4 experts (g-l). Here each session corresponds to a user (non-expert/expert) refining one patient (P) using a specific tool. The timing of each session is normalized into the [0,100] range by taking the max time across manual brush and AI pencil sessions and assigning it a value of 100. The normalized time on the x-axis is a combination of the slice navigation and contour editing time for each session.

Additionally, Figure 5.3 shows a histogram plot of pixels drawn during contour refinement which is [63% – 93%] and [81% – 99%] less for the AI pencil than the manual brush for non-expert and expert sessions, respectively (Section 5.7.4). Figure 5.4 shows visual examples of scribbles and the contours they produce. AI pencil scribbles can be used to deal with both false positives (Figure 5.4a, 5.4c, 5.4d, 5.4f, 5.4g, 5.4h, 5.4i) and false

negatives (Figure 5.4b, 5.4e, 5.4j). While most scribbles were shorter in length others were lengthier and explicit with their feedback (Figure 5.4f, 5.4i). Regardless of the style of the scribble, a sparse scribble in 2D (on slice *s*) propagates its changes in 3D. This is seen when one observes the updated contour on slices *s-1* and *s+1*. While some interactions align the refined contour with the reference contour with a single scribble (Figure 5.4g, 5.4h), others still require more interaction (Figure 5.4i, 5.4j)

For inter-tool variability, we compared the experts final contours obtained via the manual brush and the AI pencil, and noticed DICE (surface DICE @ 2mm) of patients in the range of $[0.72, 0.87]([0.68, 0.83])$ (Table 5.6). Finally, the IoV metric [188] was in the range of $[0.81, 0.96](0.74, 0.94)$ for the manual brush and $[0.88, 0.95](0.85, 0.95)$ for the AI pencil Table 5.7.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
E1	0.78 (0.66)	0.85 (0.74)	0.81 (0.77)	0.67 (0.58)	0.87 (0.92)	0.80 (0.74)
E2	0.85 (0.85)	0.87 (0.82)	0.80 (0.83)	0.71 (0.62)	0.81 (0.80)	0.81 (0.81)
E3	0.87 (0.86)	0.89 (0.85)	0.83 (0.84)	0.74 (0.74)	0.81 (0.85)	0.76 (0.66)
E4	0.84 (0.77)	0.88 (0.83)	0.80 (0.78)	0.78 (0.76)	0.81 (0.74)	0.77 (0.65)
Avg Patient Dice	0.84 (0.79)	0.87 (0.81)	0.81 (0.81)	0.72 (0.68)	0.83 (0.83)	0.79 (0.71)

Table 5.6: Metrics between the final contours of the tools.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
Manual Brush IoV	0.81 (0.74)	0.96 (0.91)	0.94 (0.94)	0.89 (0.93)	0.82 (0.85)	0.89 (0.92)
AI Pencil IoV	0.88 (0.85)	0.95 (0.94)	0.87 (0.89)	0.88 (0.95)	0.87 (0.93)	0.93 (0.95)

Table 5.7: Interobserver Variability (IoV) across patients.

Table 5.8: Dice (Surface Dice @ 2mm) when comparing a patients final contours across manual brush and AI pencil (a). The same metrics were also used to show inter-observer variability (IoV) for both tools as computed in [188].

5.4 Discussion

The widespread adoption of auto-contouring has reduced the contouring bottleneck in radiotherapy. Speeding up quality assessment (QA) of these auto-contours will further diminish this bottleneck. We investigated the use of an AI pencil for this purpose and compare its time-effectiveness, user load and capability for contour standardization against the standard manual brush. Our AI pencil could understand sparse visual cues and was able to propagate the 2D cue to 3D updates (Figure 5.4). This reduced the total effort to

QA the auto-contours as seen in Table 5.5 and Figure 5.2.

5.4.1 Time for auto-contour refinement

As a proof of principle, we first conducted our auto-contour refinement session with non-experts. They were shown the reference contour since they need to be provided a target contour to achieve. Since both the manual brush and AI pencil sessions aim to refine the predicted auto-contour to the same reference, their timing curves can be directly compared. The results (Table 5.1, Figure 5.2a - 5.2f) show that the use of the AI pencil speeds up auto-contour refinement with early and obvious advantage in a majority of cases (Figure 5.2a, 5.2b, 5.2d, 5.2e, 5.2f). For case Figure 5.2c, which was the case with the high initial DICE, we observed an early drop in performance of the AI pencil, recovering from this after 20% of the manual brush interactions. A potential explanation of this behavior is the fact that the two AI models were not internally aligned, meaning that the first iteration of the AI pencil, i.e. still having little manual input, may default to its own prediction of the contour.

Having established a proof-of-principle with the non-experts, we then tested our AI pencil in a real-world setting where the experts did not see the reference contour. In half the cases (Figure 5.2j, 5.2k, 5.2l), we could see a clear and early improvement due to the use of the AI pencil. In other cases (Figure 5.2g, 5.2h) we also saw smaller or later improvements over the manual brush. And finally, similar to the non-experts, we saw a drop and eventual rise for case Figure 5.2i. Note that for experts, their final contour served as the reference standard for each of their refinement steps, as it reflected their internal judgment on the true tumor contour.

Finally, it can be seen from Figure 5.2 that the manual brush slowly but steadily increased contour quality, while the AI pencil increased more sharply and then plateaued. This is because the manual brush could only edit one slice at a time while the AI pencil had the capability of using sparse 2D scribble inputs to refine contours in 3D.

5.4.2 Contour Consistency

The inter-tool variability of the tumor refinement sessions (Table 5.6) indicated that the experts produced similar contours regardless of the tool. These numbers provided a sense of validity to the final contours submitted in this study. In daily clinical practice, our experts also expected to receive MRI scans, physician notes, and endoscopy videos. However, since we worked with an open dataset (Hecktor2022 [44]), we did not have access to these resources. This could be a potential factor behind the aforementioned inter-tool variability. The introduction of additional resources in future studies could reduce it.

The inter-observer variability (Table 5.7) indicated the AI pencil leads to slightly better standardization of final tumor contours between experts. Thus, the AI pencil could offer both speed and quality for contour refinement.

5.4.3 Tooling

Ideally, auto-contour refinement tools like AI pencils should be embedded and tested in high-end contouring platforms offered by commercial radiotherapy software. However, none of the widely used commercial software's provide capabilities to access their contouring tools via a programmatic interface. Thus, we chose to build upon open-source libraries to create a web-hosted and open-source interaction platform for this contouring study. The AI pencil can be also readily integrated in open-source platforms like 3DSlicer [194] or Napari [195], however, the web-based framework was instrumental for conducting this study with experts from different institutes.

5.4.4 Future Work

Previous work has established proof on the viability of interactive contouring with 2D [84–87], 2.5D [184] and 3D [88–92, 185–187] models. However, depending on the contouring task, the imaging input could be either 2D (e.g., X-ray, ultrasound, histopathology, fundus scans) or 3D (e.g., CT, MR, PET). As the field of interactive contouring matures, future work should consider releasing models that are both 2D and 3D capable. Also, progress in computer vision often occurs because of the presence of open benchmark datasets. Previous 2D [87] and 3D [196, 197] datasets consist of unimodal scans (e.g. CT, MR, PET, fundus, X-ray). However, medical image contouring, especially in radiotherapy, is usually done in a multi-modal manner, like in our study. Future research will benefit from the curation of such multi-modal datasets where interactive tools like the AI pencil can be tested. Finally, many of state-of-the-art 3D models capable of iterative contour refinement [88, 90, 92] are large models and could affect inference time. Hence, neural net parameter count is an important factor and should be considered as a factor when running clinical trials on these models.

5.5 Conclusion

With a projection for increased occurrence of cancer cases and a shortage of radiotherapy clinicians [198, 199], there is an increasing need for automating the radiotherapy workflow [200]. Since human supervision is still paramount [201], human-centric AI techniques like the proposed AI pencil for contour quality refinement will be critical in achieving safety and efficiency standards for high quality radiotherapy care.

5.6 Acknowledgement

The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022) and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate, The Netherlands. Web hosting was partially funded by research credits from the Google Cloud Platform.

5.7 Appendix

5.7.1 Dataset

The clinics within the Hecktor dataset [44] are abbreviated by the challenge organizers as: Canada (CHGJ, CHUS, CHMR, CHUM), Switzerland (CHUV), United States of America (MDA) and France (CHUP). Our auto-contouring model was trained on Canadian (CHGJ, CHMR, CHUM), Swiss (CHUM) and American (MDA) patients. We kept aside data from one Canadian clinic (i.e., CHMR as in-distribution data) for validation and one French clinic (i.e., CHUP as out-of-distribution data) for testing purposes. We use CHMR for also training our users on our interface as well as the use of the AI pencil.

We resampled all scans to an isotropic voxel size of 1mm using B-spline interpolation, and contours using nearest-neighbor interpolation. For the sake of simplicity, we cropped an area of (144,144,144) around the primary head-and-neck tumor and used that during training and testing of both the auto-contouring and contour-refinement (i.e., AI pencil) models. All scans were normalized using Hounsfield Unit (HU) windowing ([-250,250]) for the CT scan and SUV windowing ([0,25]) for the PET scan. Finally, we performed z-normalization of the scans as is the standard practice for AI models.

5.7.2 Auto-contouring and contour-refinement models

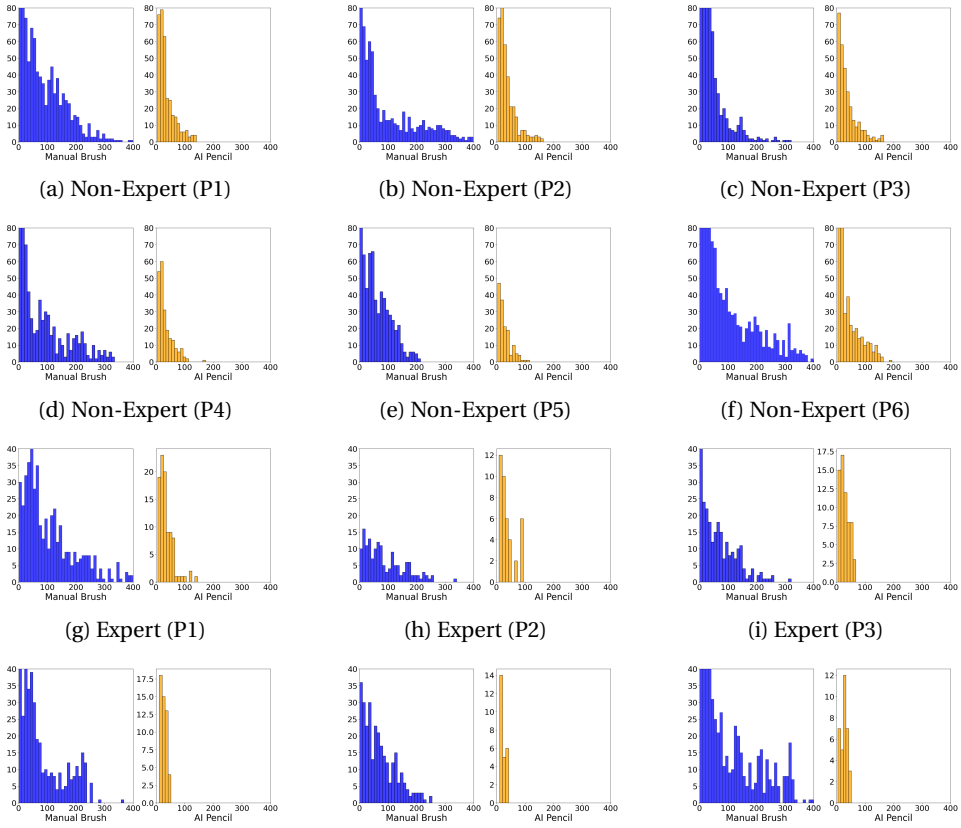
Both the auto-contour and contour-refinement models in this work were setup using the MONAI framework [189]. While the auto-contouring model has only a two-channel input (i.e. CT + PET), the contour-refinement model had a five-channel input (i.e., CT + PET + Auto-Contour + Tumor Scribble + Non-Tumor Scribble). The four internal residual convolutional blocks contained 16, 32, 64 and 128 layers. The models were trained using a standard cross-entropy loss and with the Adam optimizer (fixed learning rate of 0.001).

During training, we generate synthetic scribbles to simulate human scribbles. We generated two types of scribbles: contour-based or morphology-based (via medial axis or skeletonization) similar to previous work [87]. For training data augmentation, we sampled a portion of the scribble pixels and performed deformations on it to simulate human scribble randomness.

5.7.3 Web interface

For the user interface (Figure 5.1a), we used cornerstone3D [191], a medical imaging library written in the Javascript programming language. They provide software components for rendering DICOM images and contours along with contouring tools like brushes and pencils. For the database, we used the Orthanc DICOM server [193] that hosted the DICOM files of CT and PET scans, reference contours and also the auto-contouring predictions. The scribbles (of the AI pencil) provided by the user on the frontend were then sent to a backend FastAPI server [192] that used the Python programming language. Here,

we performed AI inference to refine the contour on the basis of the AI pencils' scribbles and then sent the refined contour back to the frontend.



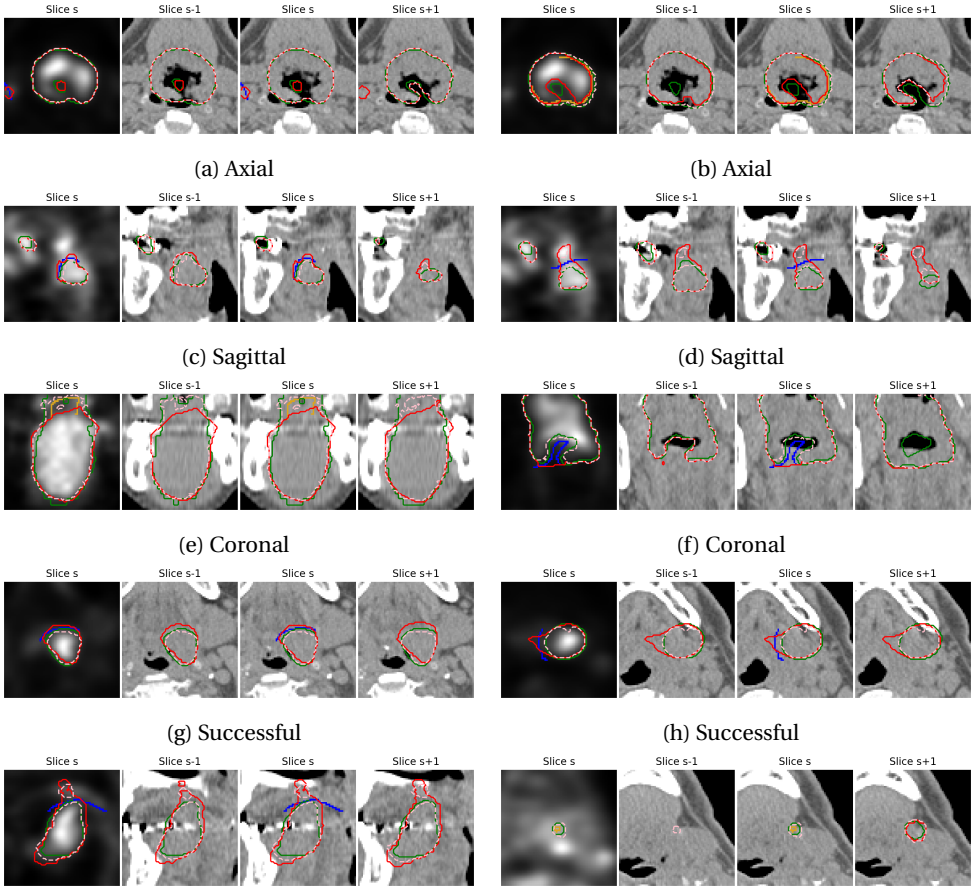
(a) Non-Expert (P1) (b) Non-Expert (P2) (c) Non-Expert (P3)

(d) Non-Expert (P4) (e) Non-Expert (P5) (f) Non-Expert (P6)

(g) Expert (P1) (h) Expert (P2) (i) Expert (P3)

(j) Expert (P4) (k) Expert (P5) (l) Expert (P6)

Figure 5.3: Histogram plots showing pixels drawn (x-axis) during auto-contour refinement of a patient (P) tumor by non-experts (a-f) and experts (g-l).



(a) Axial (b) Axial

(c) Sagittal (d) Sagittal

(e) Coronal (f) Coronal

(g) Successful (h) Successful

(i) (Partially) Successful (j) (Partially) Successful

Figure 5.4: Visual results for AI pencil refinement sessions with PET and CT scans showing previous reference contour (green), predicted contour (red) and refined contour (pink) along with scribble (yellow=tumor, blue=non-tumor). Results are shown for axial (a,b), sagittal (c,d) and coronal (e,f) views. While some scribbles successfully produce a refined contour that matches the reference (g,h), others are only partially successful(i,j).

5.7.4 Additional results on user effort

The total time taken during an auto-contour refinement session is a combination of the total user interactions, total pixels drawn (Table 5.9, 5.10) and the total slices scrolled (Table 5.11, 5.12,). Users also spent time on analysing the output of their previous interaction (either manual brush or AI-pencil), however, we do not capture these pauses in interaction.

In the tables below we show the savings in pixels drawn and slices scrolled with the AI pencil when compared to the manual brush.

	P1	P2	P3	P4	P5	P6
NE1	11035 (87%)	8925 (77%)	3801 (68%)	9308 (90%)	4885 (85%)	17987 (88%)
NE2	9465 (84%)	9057 (88%)	3566 (75%)	8102 (75%)	4770 (88%)	15753 (90%)
NE3	13501 (90%)	8832 (82%)	3326 (63%)	7568 (88%)	5787 (93%)	13993 (79%)
NE4	11039 (84%)	7918 (77%)	4756 (72%)	8294 (92%)	5941 (91%)	15319 (84%)
NE5	9926 (84%)	7704 (84%)	2803 (68%)	7207 (86%)	4302 (88%)	12542 (81%)
NE6	10803 (88%)	8991 (78%)	3383 (63%)	7044 (89%)	5510 (90%)	15799 (88%)
NE7	8339 (77)	6802 (75%)	3053 (68%)	6062 (79%)	4199 (84%)	12739 (81%)
Range	[8339,13501] ([84%,90%])	[6802,9057] ([75%,88%])	[2803,4756] ([63%,75%])	[6062,9308] ([75%,92%])	[4199,5941] ([84%,93%])	[12542,17987] ([79%,90%])

Table 5.9: Pixels drawn savings in non-expert (E) sessions.

	P1	P2	P3	P4	P5	P6
E1	9017 (89%)	3230 (81%)	2578 (81%)	9595 (94%)	8252 (98%)	16293 (98%)
E2	16253 (97%)	2981 (92%)	5539 (92%)	7203 (95%)	4569 (95%)	18516 (97%)
E3	14691 (92%)	2595 (89%)	3518 (86%)	8759 (97%)	5991 (97%)	12848 (98%)
E4	10417 (94%)	3406 (94%)	4328 (89%)	7139 (96%)	1103 (96%)	14310 (99%)
Range	[9017,16253] ([89%,97%])	[2595,3406] ([81%,94%])	[2578,5539] ([81%,92%])	[7203,9595] ([94%,97%])	[1103,8252] ([95%,98%])	[12848,18516] ([97%,99%])

Table 5.10: Pixels drawn savings in expert (E) sessions.

	P1	P2	P3	P4	P5	P6
NE1	55 (21%)	-239 (-79%)	-130 (-45%)	-78 (-11%)	-520 (-137%)	131 (13%)
NE2	-487 (-123%)	-170 (-33%)	-317 (-85%)	86 (16%)	45 (8%)	-237 (-47%)
NE3	230 (20%)	-127 (-45%)	87 (12%)	-390 (-167%)	32 (7%)	-687 (-470%)
NE4	12 (3%)	-401 (-98%)	-109 (-38%)	-493 (-109%)	-25 (-3%)	-803 (-117%)
NE5	-135 (-23%)	207 (19%)	154 (48%)	-237 (-100%)	-107 (-23%)	-95 (-18%)
NE6	-402 (-230%)	230 (36%)	10 (3%)	-367 (-67%)	52 (17%)	239 (56%)
NE7	-210 (-31%)	-810 (-447%)	391 (37%)	-326 (-51%)	-261 (-46%)	-211 (-26%)
Range	[-487,230] ([-230%,21%])	[-810,230] ([-447%,36%])	[-317,391] ([-85%,37%])	[-493,86] ([-167%,16%])	[-520,52] ([-137%,52%])	[-803,239] ([-470%,56%])

Table 5.11: Slices scrolled savings in non-expert (E) sessions.

	P1	P2	P3	P4	P5	P6
E1	-452 (-178%)	234 (34%)	-323 (-101%)	126 (32%)	266 (72%)	443 (62%)
E2	-37 (-9%)	-91 (-26%)	-56 (-12%)	-328 (-100%)	141 (25%)	300 (39%)
E3	-112 (-20%)	340 (50%)	-190 (-47%)	463 (61%)	95 (24%)	537 (48%)
E4	440 (50%)	251 (38%)	200 (24%)	954 (59%)	246 (64%)	292 (24%)
Range	[-452,440] ([-178%,50%])	[-91,340] ([-26%,50%])	[-323,200] ([-101%,24%])	[-328,954] ([-100%,61%])	[95,266] ([24%,72%])	[292,537] ([24%,62%])

Table 5.12: Slices scrolled savings in expert (E) sessions.

Table 5.13: Savings in pixels drawn (a,b) and slices scrolled (c,d) when using AI pencil as compared to manual brush. Savings in percentages are shown in brackets.

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.

7

Samenvatting, discussie en toekomstig werk

7.1 Samenvatting van de dissertatie

Deze dissertatie richt zich op de dringende behoefte aan efficiënte en betrouwbare kwaliteitscontrole (QA)-tools voor geautomatiseerde contourbepaling van organen en tumoren bij radiotherapie. Hoewel deep learning modellen een aanzienlijke versnelling bieden, kunnen de daaropvolgende handmatige QA- en verfijningsstappen tijdrovend zijn en zo de winst gedeeltelijk tenietdoen, wat leidt tot een knelpunt in klinische workflows. Twee hoofdthema's binnen QA worden onderzocht: *foutdetectie* (waar zijn contouren waarschijnlijk fout) en *foutcorrectie* (hoe corrigeer je die efficiënt), in zowel pre- als post-commissioning fasen.

Dit proefschrift onderzoekt specifiek: a) de ontwikkeling van een geautomatiseerde en schaalbare workflow voor het evalueren van de dosimetrische impact van autocontouren vóór ingebruikname (Chapter 2), b) het potentieel van Bayesiaanse modellen en trainingsverliezen om onnauwkeurige voorspellingen te detecteren in de post-ingebruiknamefase door gebruik te maken van de bijbehorende onzekerheid (Chapter 3 & Chapter 4), en c) de verbetering van de efficiëntie van foutcorrectie met behulp van AI-ondersteunde verfijningstools (Chapter 5). Het overkoepelende doel van dit proefschrift is dan ook om verschillende QA-methodologieën te verkennen, zowel vóór als na ingebruikname van autocontouringtools voor hoofd-halsradiotherapie.

7.2 Hoofdstuk Samenvattingen

7.2.1 Hoofdstuk 2

Dit hoofdstuk richtte zich op de noodzaak van grootschalige pre-commissioning dosimetrische evaluaties van automatisch gecontourde organen-at-risk (OARs). Het belangrijkste resultaat was een geautomatiseerde workflow voor planningsoptimalisatie, gebaseerd op bestaande klinische instellingen, bekend als robotic process automation (RPA). Via scripting in het Treatment Planning System (TPS) werd een handmatige planningsprocedure geautomatiseerd.

Een studie werd uitgevoerd op 100 hoofd-hals patiënten (70 fotonen, 30 protonen). Resultaten toonden minimale dosimetrische verschillen tussen automatische en handmatige contouren. Dit wijst erop dat geometrische afwijkingen veroorzaakt door automa-

tische contouren beperkte klinische impact hebben, en bevestigt de bruikbaarheid van de voorgestelde QA-aanpak.

7.2.2 Hoofdstuk 3

Dit hoofdstuk onderzocht hoe modelkeuzes de onzekerheid beïnvloeden, die als proxy kan dienen voor fouten in post-commissioning QA. Twee Bayesiaanse modellen (DropOut en FlipOut) werden geëvalueerd met behulp van Expected Calibration Error (ECE) en een nieuwe metriek, Region-based Accuracy-vs-Uncertainty (R-AvU). Waar ECE een informatietheoretische benadering is, biedt R-AvU een meer visuele evaluatie. Training met cross-entropy verlies (CE) gaf betere calibratie (lagere ECE). FlipOut-CE toonde betere onzekerheidsdekking in foutieve regio's dan DropOut-CE volgens de R-AvU grafieken. Deze resultaten roepen de vraag op: welke metriek moet men gebruiken bij onzekerheidsevaluatie voor foutdetectie?

7.2.3 Hoofdstuk 4

Hoewel Bayesiaanse modellen onzekerheidskaarten kunnen produceren, is hun klinisch nut afhankelijk van de mate waarin deze kaarten overeenkomen met echte fouten. Hoofdstuk 2 toonde aan dat deze overeenstemming vaak suboptimaal is. Dit hoofdstuk introduceerde een differentieerbaar verlies op basis van de Accuracy-vs-Uncertainty (AvU) metriek, die expliciet onzekerheid stimuleert waar fouten voorkomen. De kaarten werden geëvalueerd via ROC-curves ("uncertainty-ROC") en Precision-Recall-curves. Een belangrijk aspect was het onderscheid tussen "fouten" (klein, acceptabel) en "falen" (groter, vereisen interventie).

De AvU-verliesfunctie verbeterde significante calibratie (ECE) en de overeenkomst tussen onzekerheid en fouten (ROC-AUC, PRC-AUC), voor zowel in-distributie (ID) als out-of-distributie (OOD) datasets. AvU presteerde zelfs beter dan ensemble-modellen. Dit toont dat optimalisatie op ECE niet voldoende is om bruikbare onzekerheidskaarten te produceren — AvU biedt een unieke meerwaarde.

7.2.4 Hoofdstuk 5

De focus ligt hier op foutcorrectie, post-commissioning. In dit hoofdstuk werd de efficiëntie en kwaliteit van handmatige borstels vergeleken met een AI-ondersteunde "AI potlood"-tool. Bestaande AI-potloodtools missen vaak evaluatie met menselijke gebruikers en werken enkel in 2D.

Een webinterface werd ontwikkeld waarin gebruikers 2D-aanduidingen (scribbles) konden geven, waarna de AI potlood 3D-refinements uitvoerde op CT+PET scans van hoofd-hals tumoren. Zowel klinische als niet-klinische gebruikers namen deel. De AI-potlood was 5–78% sneller bij niet-experts en 16–97% sneller bij experts, terwijl de uiteindelijke kwaliteit gelijkwaardig bleef. De AI potlood bereikte snel een hoge kwaliteit, in tegen-

stelling tot de geleidelijke verbetering bij de handmatige tool. Dit toont de kracht van AI-geassisteerde QA bij radiotherapie.

7.3 Discussie en toekomstig werk

Deze dissertatie adresseert belangrijke uitdagingen in veilige, efficiënte en betrouwbare integratie van QA-tools voor deep learning-gebaseerde auto-contouring in de kliniek. Zowel foutdetectie als foutcorrectie, in pre- en post-commissioning scenario's, worden aangepakt, met nadruk op menselijke bruikbaarheid.

Toekomstige onderzoekslijnen zijn:

- Klinische betrokkenheid - Technisch onderzoek probeert vaak te optimaliseren op basis van bepaalde, vooraf gespecificeerde criteria, maar vertaalt dit niet naar de klinische praktijk. Dit gebrek aan 'van tafel tot bed'-mentaliteit wordt vaak veroorzaakt door de structuur van onderzoeksprojecten. Een ontbrekende factor is vaak voldoende klinische betrokkenheid, waardoor onderzoek op een stoffige plank blijft liggen. Onderzoekers zouden moeten overwegen hun teams en mentoren zo in te richten dat ze multidisciplinaire vaardigheden inzetten om de volledige breedte en diepte van het probleem te begrijpen.

- Vernieuwing van contourrichtlijnen – Chapter 2 toonde zowel correlaties als non-correlaties tussen DICE en dosisverschillen. Grotere studies zouden de contourrichtlijnen opnieuw kunnen definiëren, en mogelijk vaste anatomische richtlijnen laten evolueren naar richtlijnen met marges die rekening houden met inter- en intra-observatorvariabiliteit.

- Het nut van onzekerheid in klinische settings begrijpen – Onzekerheid is een wiskundig concept dat de potentie heeft om inzicht te bieden in de betrouwbaarheid van datagestuurde technieken zoals deep learning. De community gebruikt echter vaak puur wiskundige concepten zoals ECE (met zijn groeperingsmechanisme) om het nut van de onzekerheid van een model te evalueren. Dergelijke statistieken leggen onzekerheid niet pixelgewijs (of op een gedetailleerde manier) vast. Het verleggen van de grenzen van bestaande statistieken, hoe belangrijk ook, is dus niet voldoende om onderzoeksinnovaties aan te passen aan de dagelijkse klinische praktijk.

- Pixel- vs. slice- vs. regio-onzekerheid: Het is mogelijk dat er een praktische limiet is aan de mate waarin clinici kunnen profiteren van 'onzekerheidsafstemming' voordat het leidt tot cognitieve overbelasting. Enerzijds kan te veel op onzekerheid gebaseerde besluitvorming (bijv. per pixel) cognitief veeleisend zijn. Anderzijds biedt gemiddelde onzekerheid (bijv. op het niveau van een plak of orgaan/tumor) mogelijk niet effectief houvast voor het verfijnen van contouren. Daarom moeten

onderzoekers nadenken over de granulariteit van onzekerheid die we nodig hebben in medische beeldsegmentatietoepassingen.

- Verliesfuncties en klinische bruikbaarheid: De DICE-loss is een geometrie-gebaseerde verliesfunctie, omdat het kijkt naar de algehele structuur en vorm van de 'ground truth' en de voorspelling. Verrassend genoeg presteerde een pixel-gebaseerde aanpak, namelijk de cross-entropy loss, echter beter in het weergeven van de werkelijke vertrouwelijkheid van zijn voorspellingen. Daarom moeten makers van autocontouring-tools dieper nadenken over de manier waarop hun verliesfuncties de ervaring van de eindgebruiker beïnvloeden.

- De eisen voor de dataset analyseren – Een van de belemmeringen voor het vertalen van onderzoek naar de klinische praktijk is de grote hoeveelheid benodigde trainingsdata. De literatuur laat echter vergelijkbare prestaties zien met datasets van verschillende groottes. Meer werk met hulpmiddelen zoals leercruves kan de gemeenschap beter informeren over de minimale vereisten voor datasets om te voldoen aan klinische normen voor het contouren van organen en doelwitten.

- Frameworks voor klinische validatie in de praktijk – Frameworks voor klinische validatie in de praktijk – Tools voor robuuste experimentatie en evaluatie zijn wat elk vakgebied vooruithelpt, aangezien ze de drempels voor nieuwkomers verlagen om bij te dragen. Dit is te zien bij programmeertalen zoals Python en deep learning frameworks zoals Tensorflow en PyTorch. Een vergelijkbaar voorbeeld voor medische beeldsegmentatie is het grand-challenge.org-platform. Nu deep learning tools steeds gangbaarder worden in de medische beeldvorming, moet de gemeenschap zich richten op de ontwikkeling van vergelijkbare frameworks voor onzekerheid als een proxy voor foutdetectie en voor interactieve segmentatie.

- Vertrouwen in AI-gedreven acties – Voor de context van interactieve contourverfijning, hoe kunnen we ervoor zorgen dat clinici de door AI gegenereerde verfijningen voldoende vertrouwen om niet terug te vallen op handmatige correcties? En kunnen dergelijke tools zich aanpassen aan de diverse manieren waarop verschillende clinici contourbewerking benaderen? Het kan dus nodig zijn om statistieken te gebruiken die bijhouden hoe betrouwbaar het model is in lokale gebieden waar de gebruiker zijn krabbels maakt. En doet het model stiekem onterechte voorspellingen in gebieden ver weg van de interactie van de gebruiker?.

- Rol van regelgevende instanties: Gezondheidszorgsystemen moeten worden gereguleerd door overheidsinstanties vanwege de kritieke aard van de dienst die ze leveren. Onderzoeks-innovaties overtreffen echter vaak de regulerende instanties en in de tussentijd bestaat de mogelijkheid dat innovaties die niet rigoureus of nauwkeurig

zijn getest, door artsen kunnen worden gebruikt. In het geval van op deep learning gebaseerde auto-contouring is er bijvoorbeeld heel weinig discussie over de noodzaak van land-/demografisch-gebaseerde benchmark datasets. Dit maakt het voor klinische innovators zeer omslachtig om te bepalen hoe ze commerciële oplossingen moeten evalueren, aangezien zij degene moeten zijn die hun eigen interne dataset samenstellen, die vaak rommelig is vanwege de drukke werkdruk van clinici. We smeken de lezer van dit proefschrift om over dit punt na te denken en de bovengenoemde lacune op te vullen..

7.4 Algemene conclusies

In een tijdperk van toenemende incidentie van kanker en beperkte klinische middelen, levert dit proefschrift essentiële hulpmiddelen om de veilige, effectieve integratie van deep learning auto-contouring in de radiotherapieworkflow te waarborgen. Door praktische, mensgerichte methoden te bieden voor zowel precieze foutdetectie als efficiënte foutcorrectie, helpt dit werk de kloof te overbruggen tussen geavanceerde deep learning-modellen en hun veilige en effectieve kwaliteitsbeoordeling voor integratie in de dagelijkse klinische radiotherapiepraktijk. We hopen anderen te inspireren om werk na te streven dat de kloof overbrugt tussen wiskundige onzekerheidsmetrieken en praktisch klinisch vertrouwen. Evenzo moeten interactieve AI-hulpmiddelen evolueren om de diverse manieren waarop clinici werken te weerspiegelen.

Uiteindelijk streeft dit onderzoek ernaar om patiëntenzorg van hoge kwaliteit te waarborgen en de workflowefficiëntie te verbeteren, waarbij de positieve resultaten bedoeld zijn om mensgerichte deep learning voor medische beeldvorming te informeren en te bevorderen.

References

[1] Charlotte L Brouwer 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”. In: *Radiotherapy and Oncology* 117.1 (2015), pp. 83–90.

[2] United States government. *Head and Neck Cancers*. 2021. URL: https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet.

[3] Foteini Simopoulou et al. “Does adaptive radiotherapy for head and neck cancer favorably impact dosimetric, clinical, and toxicity outcomes?: A review”. In: *Medicine* 103.26 (2024), e38529.

[4] Jakub Grepl et al. “MRI-based adaptive radiotherapy has the potential to reduce dysphagia in patients with head and neck cancer”. In: *Physica Medica* 105 (2023), p. 102511.

[5] Stephanie Lim-Reinders et al. “Online Adaptive Radiation Therapy”. In: *Int J Radiat Oncol Biol Phys* 99 (2017), pp. 994–1003. URL: https://doi.org/10.1016/j.ijrobp.2017.04.023.

[6] Charlotte L Brouwer et al. “3D variation in delineation of head and neck organs at risk”. In: *Radiation Oncology* 7.1 (2012), pp. 1–10.

[7] Shivakumar Gudi et al. “Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site”. In: *Journal of medical imaging and radiation sciences* 48.2 (2017), pp. 184–192.

[8] Julie van der Veen, Akos Gulyban, and Sandra Nuyts. “Interobserver variability in delineation of target volumes in head and neck cancer”. In: *Radiotherapy and Oncology* 137 (2019), pp. 9–15.

[9] Julie van der Veen, Akos Gulyban, and Sandra Nuyts. “Interobserver variability in delineation of target volumes in head and neck cancer”. In: *Radiotherapy and Oncology* 137 (2019), pp. 9–15.

[10] Jordan Wong et al. “Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning”. In: *Radiother Oncol* 144 (2020), pp. 152–158. URL: https://doi.org/10.1016/j.radonc.2019.10.019.

[11] J. van der Veen et al. “Interobserver variability in organ at risk delineation in head and neck cancer”. In: *Radiat Oncol* 16 (2021), pp. 1–11. URL: https://doi.org/10.1186/s13014-020-01677-2.

[12] Ruta Zukauskaite et al. “Delineation uncertainties of tumour volumes on MRI of head and neck cancer patients”. In: *Clinical and Translational Radiation Oncology* 36 (2022), pp. 121–126.

[13] Michaël Claessens et al. “Quality assurance for AI-based applications in radiation therapy”. In: *Seminars in radiation oncology*. Vol. 32. 4. Elsevier. 2022, pp. 421–431.

[14] Sandra Nuyts et al. “Adaptive radiotherapy for head and neck cancer: Pitfalls and possibilities from the radiation oncologist’s point of view”. In: *Cancer Medicine* 13.8 (2024), e7192.

[15] Dan Nguyen et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. In: *Physics in medicine & Biology* 64.6 (2019), p. 065020.

[16] Masahide Saito et al. “Evaluation of deep learning based dose prediction in head and neck cancer patients using two different types of input contours”. In: *Journal of Applied Clinical Medical Physics* 25.12 (2024), e14519.

[17] Yanxia Liu et al. “CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy”. In: *Computerized medical imaging and graphics* 91 (2021), p. 101953.

[18] Wen Chen et al. “Clinical enhancement in AI-based post-processed fast-scan low-dose CBCT for head and neck adaptive radiotherapy”. In: *Frontiers in artificial intelligence* 3 (2021), p. 614384.

[19] Adrian Thummerer et al. “SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy”. In: *Medical physics* 50.7 (2023), pp. 4664–4674.

[20] Patrik F Raudaschl et al. “Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015”. In: *Medical physics* 44.5 (2017), pp. 2020–2036.

[21] Margarita L Zuley et al. “Radiology data from the cancer genome atlas head-neck squamous cell carcinoma [TCGA-HNSC] collection”. In: *Cancer Imaging Arch* 10 (2016), K9.

[22] Stanislav Nikolov et al. “Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study”. In: *J Med Internet Res* 23.7 (July 2021), e26151. ISSN: 1438-8871. DOI: 10.2196/26151. URL: http://www.ncbi.nlm.nih.gov/pubmed/34255661.

[23] Wentao Zhu et al. “AnatomyNet: Deep Learning for Fast and Fully Automated Whole-volume Segmentation of Head and Neck Anatomy”. In: *Medical physics* 46.2 (2019), pp. 576–589.

[24] Yunhe Gao et al. “FocusNet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck CT images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 829–838.

[25] Ozan Oktay et al. “Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers”. In: *JAMA network open* 3.11 (2020), e2027426.

[26] Dazhou Guo et al. “Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4223–4232.

[27] Christina Hague et al. “An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy”. In: *Radiotherapy and Oncology* 158 (2021), pp. 112–117.

[28] Yunhe Gao et al. “FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images”. In: *Medical Image Analysis* 67 (2021), p. 101831.

[29] Zijie Chen et al. “A Novel Hybrid Convolutional Neural Network for Accurate Organ Segmentation in 3D Head and Neck CT Images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 569–578.

[30] John C Asbach et al. “Deep learning tools for the cancer clinic: an open-source framework with head and neck contour validation”. In: *Radiation Oncology* 17.1 (2022), p. 28.

[31] Lucía Cubero et al. “Deep learning-based segmentation of head and neck organs-at-risk with clinical partially labeled data”. In: *Entropy* 24.11 (2022), p. 1661.

[32] Peiru Liu et al. “Deep learning algorithm performance in contouring head and neck organs at risk: a systematic review and single-arm meta-analysis”. In: *BioMedical Engineering OnLine* 22.1 (2023), p. 104.

[33] Skylar S Gay et al. “Fully-automated, CT-only GTV contouring for palliative head and neck radiotherapy”. In: *Scientific reports* 13.1 (2023), p. 21797.

[34] Lisanne V van Dijk et al. “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring”. In: *Radiotherapy and Oncology* 142 (2020), pp. 115–123.

[35] Charlotte L Brouwer et al. “Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy”. In: *Physics and imaging in radiation oncology* 16 (2020), pp. 54–60.

[36] Jordan Wong et al. “Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers”. In: *Radiation Oncology* 16.1 (2021), p. 101.

[37] Yang Zhong et al. “A preliminary experience of implementing deep-learning based auto-segmentation in head and neck cancer: a study on real-world clinical cases”. In: *Frontiers in oncology* 11 (2021), p. 638197.

[38] Andrea D’Aviero et al. “Clinical validation of a deep-learning segmentation software in head and neck: an early analysis in a developing radiation oncology center”. In: *International Journal of Environmental Research and Public Health* 19.15 (2022), p. 9057.

[39] Yasmin McQuinlan et al. “An investigation into the risk of population bias in deep learning autocontouring”. In: *Radiotherapy and Oncology* 186 (2023), p. 109747.

[40] Yunfei Hu et al. “Clinical assessment of a novel machine-learning automated contouring tool for radiotherapy planning”. In: *Journal of Applied Clinical Medical Physics* 24.7 (2023), e13949.

[41] J. John Lucido et al. “Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning”. In: *Front Oncol* 13 (2023). URL: https://doi.org/10.3389/fonc.2023.1137803.

[42] Patrik F Raudaschl et al. “Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015”. In: *Medical physics* 44.5 (2017), pp. 2020–2036.

[43] Stanislav Nikolov et al. “Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study”. In: *Journal of Medical Internet Research* 23.7 (2021), e26151.

[44] Valentin Oreiller et al. “Head and neck tumor segmentation in PET/CT: the HECKTOR challenge”. In: *Medical image analysis* 77 (2022), p. 102336.

[45] J. P. Kieselmann et al. “Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region”. In: *Phys Med Biol* 63 (2018), aacb65. URL: https://doi.org/10.1088/1361-6560/aacb65.

[46] Ward van Rooij et al. “Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation”. In: *Int J Radiat Oncol Biol Phys* 104 (2019), pp. 677–684. URL: https://doi.org/10.1016/j.ijrobp.2019.02.040.

[47] Lisanne V. van Dijk et al. “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring”. In: *Radiother Oncol* 142 (2020), pp. 115–123. URL: https://doi.org/10.1016/j.radonc.2019.09.022.

[48] Hongbo Guo et al. “The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer”. In: *Radiat Oncol* 16 (2021), pp. 1–14. URL: https://doi.org/10.1186/s13014-021-01837-y.

[49] Andreas Johan Smolders et al. “Dosimetric comparison of autocontouring techniques for online adaptive proton therapy”. In: *Phys Med Biol* 68 (2023), p. 175006. URL: https://doi.org/10.1088/1361-6560/ace307.

[50] Jihye Koo et al. “Essentially unedited deep-learning-based OARs are suitable for rigorous oropharyngeal and laryngeal cancer treatment planning”. In: *J Appl Clin Med Phys* (2023), pp. 1–10. URL: https://doi.org/10.1002/acm2.14202.

[51] Camila González et al. “Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation”. In: *Medical image analysis* 82 (2022), p. 102596.

[52] Davood Karimi and Ali Gholipour. “Improving calibration and out-of-distribution detection in deep models for medical image segmentation”. In: *IEEE Transactions on Artificial Intelligence* 4.2 (2022), pp. 383–397.

[53] Lyndon Boone et al. “ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI”. In: *NeuroImage* 278 (2023), p. 120289.

[54] Anton Vasiliuk et al. “Limitations of out-of-distribution detection in 3d medical image segmentation”. In: *Journal of Imaging* 9.9 (2023), p. 191.

[55] Zesheng Hong et al. “Out-of-distribution detection in medical image analysis: A survey”. In: *arXiv preprint arXiv:2404.18279* (2024).

[56] Felix JS Bragman et al. “Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 3–11.

[57] Guotai Wang et al. “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks”. In: *Neurocomputing* 338 (2019), pp. 34–45.

[58] Jörg Sander, Bob D de Vos, and Ivana Išgum. “Automatic segmentation with detection of local segmentation failures in cardiac MRI”. In: *Scientific Reports* 10.1 (2020), p. 21769.

[59] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. “Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation”. In: *Frontiers in neuroscience* 14 (2020), p. 282.

[60] Guotai Wang et al. “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks”. In: *Neurocomputing* 338 (2019), pp. 34–45.

[61] Tanya Nair et al. “Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation”. In: *Medical image analysis* 59 (2020), p. 101557.

[62] Alireza Mehrtash et al. “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation”. In: *IEEE transactions on medical imaging* 39.12 (2020), pp. 3868–3878.

[63] Guotai Wang et al. “Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer. 2020, pp. 279–288.

[64] Azat Garifullin, Lasse Lensu, and Hannu Uusitalo. “Deep Bayesian baseline for segmenting diabetic retinopathy lesions: Advances and challenges”. In: *Computers in Biology and Medicine* 136 (2021), p. 104725.

[65] Cheng Ouyang et al. “Improved post-hoc probability calibration for out-of-domain MRI segmentation”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022*. Springer. 2022, pp. 59–69.

[66] Matthew Ng et al. “Estimating Uncertainty in Neural Networks for Cardiac MRI Segmentation: A Benchmark Study”. In: *IEEE Transactions on Biomedical Engineering* (2022), pp. 1–12. DOI: 10.1109/TBME.2022.3232730.

[67] Miguel Monteiro et al. “Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12756–12767.

[68] Sora Iwamoto et al. “Improving the Reliability of Semantic Segmentation of Medical Images by Uncertainty Modeling with Bayesian Deep Networks and Curriculum Learning”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021.

[69] Robin Camarasa et al. “A Quantitative Comparison of Epistemic Uncertainty Maps Applied to Multi-Class Segmentation”. In: *Machine Learning for Biomedical Imaging* 1.UNSURE2020 special issue (2021), pp. 1–10.

[70] Tewodros Weldebirhan Arega, Stéphanie Bricq, and Fabrice Meriaudeau. “Leveraging Uncertainty Estimates to Improve Segmentation Performance in Cardiac MR”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021, pp. 24–33.

[71] Mobarakol Islam and Ben Glocker. “Spatially varying label smoothing: Capturing uncertainty from expert annotations”. In: *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*. Springer. 2021, pp. 677–688.

[72] Ishaan Bhat et al. “Influence of uncertainty estimation techniques on false-positive reduction in liver lesion detection”. In: *Machine Learning for Biomedical Imaging* 1 (December 2022 issue 2022), pp. 1–33. ISSN: 2766-905X. DOI: 10.59275/j.melba.2022-5937. URL: https://melba-journal.org/2022:030.

[73] Achim Hekler, Titus J Brinker, and Florian Buettner. “Test time augmentation meets post-hoc calibration: uncertainty quantification under real-world conditions”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 14856–14864.

[74] Hongwei Bran Li et al. “QUBIQ: Uncertainty Quantification for Biomedical Image Segmentation Challenge”. In: *arXiv preprint arXiv:2405.18435* (2024).

[75] Kareem A Wahid et al. “Artificial intelligence uncertainty quantification in radiotherapy applications-A scoping review”. In: *Radiotherapy and Oncology* (2024), p. 110542.

[76] Jörg Sander et al. “Towards increased trustworthiness of deep learning segmentation methods on cardiac MRIg”. In: *Medical Imaging 2019: Image Processing*. Vol. 10949. International Society for Optics and Photonics. 2019, p. 1094919.

[77] Alireza Mehrtash et al. “Confidence calibration and predictive uncertainty estimation for deep medical image segmentation”. In: *IEEE transactions on medical imaging* 39.12 (2020), pp. 3868–3878.

[78] Thomas Buddenkotte et al. “Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation”. In: *Computers in Biology and Medicine* 163 (2023), p. 107096.

[79] Yidong Zhao et al. “Bayesian uncertainty estimation by hamiltonian monte carlo: Applications to cardiac mri segmentation”. In: *arXiv preprint arXiv:2403.02311* (2024).

[80] Dong Joo Rhee et al. “Automatic detection of contouring errors using convolutional neural networks”. In: *Medical physics* 46.11 (2019), pp. 5086–5097.

[81] Edward GA Henderson et al. “Automatic identification of segmentation errors for radiotherapy using geometric learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 319–329.

[82] Lars Johannes Isaksson et al. “Quality assurance for automatically generated contours with additional deep learning”. In: *Insights into Imaging* 13.1 (2022), p. 137.

[83] Biling Wang et al. “AI-Assisted Decision-Making for Clinical Assessment of Auto-Segmented Contour Quality”. In: *arXiv preprint arXiv:2505.00308* (2025).

[84] Tomas Sakinis et al. “Interactive segmentation of medical images through fully convolutional neural networks”. In: *arXiv preprint arXiv:1903.08205* (2019).

[85] Zixiang Wei et al. “Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy”. In: *Physics and Imaging in Radiation Oncology* 25 (2023), p. 100408.

[86] Jun Ma et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (2024), p. 654.

[87] Hallee E. Wong et al. “ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image”. In: *European Conference on Computer Vision (ECCV)* (2024).

[88] Yufan He et al. “VISTA3D: Versatile Imaging SegmenTation and Annotation model for 3D Computed Tomography”. In: *arXiv preprint arXiv:2406.05285* (2024).

[89] Yuxin Du et al. “Segvol: Universal and interactive volumetric medical image segmentation”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 110746–110783.

[90] Haoyu Wang et al. “Sam-med3d: towards general-purpose segmentation models for volumetric medical images”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 51–67.

[91] Andres Diaz-Pinto et al. “DeepEdit: Deep editable learning for interactive segmentation of 3D medical images”. In: *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*. Springer. 2022, pp. 11–21.

[92] Fabian Isensee et al. “nnInteractive: Redefining 3D Promptable Segmentation”. In: *arXiv preprint arXiv:2503.08373* (2025).

[93] Nicolas F Chaves-de-Plaza et al. “Towards fast human-centred contouring workflows for adaptive external beam radiotherapy”. In: *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference*. 2022, pp. 111–131.

[94] Gregory Sharp et al. “Vision 20/20: Perspectives on automated image segmentation for radiotherapy”. In: *Med Phys* 41 (2014), pp. 1–13. URL: https://doi.org/10.1118/1.4871620..

[95] Charlotte L. Brouwer 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”. In: *Radiother Oncol* 117 (2015), pp. 83–90. URL: https://doi.org/10.1016/j.radonc.2015.07.041.

[96] Charlotte L. Brouwer et al. “3D Variation in delineation of head and neck organs at risk”. In: *Radiat Oncol* 7.1 (2012). URL: https://doi.org/10.1186/1748-717X-7-32.

[97] J. J. Stelmes et al. “Quality assurance of radiotherapy in the ongoing EORTC 1420 “Best of” trial for early stage oropharyngeal, supraglottic and hypopharyngeal carcinoma: results of the benchmark case procedure”. In: *Radiat Oncol* 16 (2021), pp. 1–10. URL: https://doi.org/10.1186/s13014-021-01809-2.

[98] Ellen J.L. Brunenberg et al. “External validation of deep learning-based contouring of head and neck organs at risk”. In: *Phys Imaging Radiat Oncol* 15 (2020), pp. 8–15. URL: https://doi.org/10.1016/j.phro.2020.06.006.

[99] Curtise K.C. Ng, Vincent W.S. Leung, and Rico H.M. Hung. “Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring for Head and Neck Radiation Therapy”. In: *Appl Sci* 12 (2022). URL: https://doi.org/10.3390/app122211681.

[100] Michael V. Sherer et al. “Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review”. In: *Radiother Oncol* 160 (2021), pp. 185–191. URL: https://doi.org/10.1016/j.radonc.2021.05.003.

[101] Madalina Costea et al. “Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system”. In: *Radiother Oncol* 177 (2022), pp. 61–70. URL: https://doi.org/10.1016/j.radonc.2022.10.029.

[102] Madalina Costea et al. “Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images”. In: *Radiother Oncol* 188 (2023), p. 109870. URL: https://doi.org/10.1016/j.radonc.2023.109870.

[103] Landelijk Platform Radiotherapie Hoofd-halstumoren (LPRHHT) Landelijk Platform Protonentherapie (LPPT). *Landelijk Indicatie Protocol Protonentherapie (versie 2.2) (LIPPv2.2)*. https://nvro.nl/images/documenten/rapporten/2019-08-15__Landelijk_Indicatieprotocol_Protonentherapie_Hoofdhals_v2.2.pdf. 2019.

[104] Erik W Korevaar et al. “Practical robustness evaluation in radiotherapy – A photon and proton-proof alternative to PTV-based plan evaluation”. In: *Radiother Oncol* 141 (2019), pp. 267–274. URL: https://doi.org/10.1016/j.radonc.2019.08.005.

[105] Ilma Xhaferllari et al. “Automated IMRT planning with regional optimization using planning scripts”. In: *J Appl Clin Med Phys* 14 (2013), pp. 176–191. URL: https://doi.org/10.1120/jacmp.v14i1.4052.

[106] Stefan Speer et al. “Automation of radiation treatment planning”. In: *Strahlentherapie Und Onkol* 193 (2017), pp. 656–665. URL: https://doi.org/10.1007/s00066-017-1150-9.

[107] Jose R Teruel et al. “Full automation of spinal stereotactic radiosurgery and stereotactic body radiation therapy treatment planning using Varian Eclipse scripting”. In: *J Appl Clin Med Phys* 21 (2020), pp. 122–131. URL: https://doi.org/10.1002/acm2.13017.

[108] Wil M P Van Der Aalst, Martin Bichler, and Armin Heinzl. “Robotic Process Automation”. In: *Business & Information Systems Engineering* 60 (2018), pp. 269–272. URL: https://doi.org/10.1007/s12599-018-0542-4.

[109] Andrzej Niemierko. “Reporting and analyzing dose distributions: A concept of equivalent uniform dose”. In: *Med Phys* 24 (1997), pp. 103–110. URL: https://doi.org/10.1118/1.598063.

[110] Stanislav Nikolov et al. “Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study”. In: *J Med Internet Res* 23 (2021), e26151. URL: https://doi.org/10.2196/26151.

[111] Xiaojin Gu et al. “Dose distribution prediction for head-and-neck cancer radiotherapy using a generative adversarial network: influence of input data”. In: *Front. Oncol* 13 (2023), p. 1251132. URL: https://doi.org/10.3389/fonc.2023.1251132.

[112] Elizabeth M Jaworski et al. “Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer”. In: *Adv Radiat Oncol* 8 (2023), p. 101029. URL: https://doi.org/10.1016/j.adro.2022.101029.

[113] Rachel Petragallo et al. “Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation : A survey study of medical dosimetrists”. In: *Journal of Applied Clinical Medical Physics* 23 (2022), pp. 1–10. URL: https://doi.org/10.1002/acm2.13568.

[114] Johannes A. Langendijk et al. “National protocol for model-based selection for proton therapy in head and neck cancer”. In: *Int J Part Ther* 8 (2021), pp. 354–365. URL: https://doi.org/10.14338/IJPT-20-00089.1.

[115] Mei Ling Yap et al. “Global access to radiotherapy services: have we made progress during the past decade?” In: *Journal of global oncology* 2.4 (2016), pp. 207–215.

[116] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[117] Tyler LaBonte, Carianne Martinez, and Scott A Roberts. “We Know Where We Don’t Know: 3D Bayesian CNNs for Credible Geometric Uncertainty”. In: *arXiv preprint arXiv:1910.10793* (2019).

[118] Chuan Guo et al. “On calibration of modern neural networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[119] Jishnu Mukhoti and Yarin Gal. “Evaluating Bayesian Deep Learning Methods for Semantic Segmentation”. In: *CoRR* abs/1811.12709 (2018). arXiv: 1811.12709. URL: http://arxiv.org/abs/1811.12709.

[120] Yarin Gal. “Uncertainty in deep learning”. In: (2016).

[121] K Kian Ang et al. “Randomized phase III trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage III to IV head and neck carcinoma: RTOG 0522”. In: *Journal of clinical oncology* 32.27 (2014), p. 2940.

[122] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[123] Maoke Yang et al. “Denseaspp for semantic segmentation in street scenes”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3684–3692.

[124] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[125] Yeming Wen et al. “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”. In: *International Conference on Learning Representations (ICLR)*. 2017. eprint: 1803.04386 (cs.LG).

[126] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.

[127] Saeid Asgari Taghanaki et al. “Combo loss: Handling input and output imbalance in multi-organ segmentation”. In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 24–33.

[128] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[129] W Jeffrey Zabel et al. “Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy”. In: *Practical Radiation Oncology* 11.1 (2021), e80–e89.

[130] Lisanne V Van Dijk et al. “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring”. In: *Radiotherapy and Oncology* 142 (2020), pp. 115–123.

[131] Charlotte L Brouwer et al. “Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy”. In: *Physics and imaging in radiation oncology* 16 (2020), pp. 54–60.

[132] Rachel Petragallo et al. “Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists”. In: *Journal of Applied Clinical Medical Physics* 23.5 (2022), e13568.

[133] Nicolas F Chaves-de-Plaza et al. “Towards fast human-centred contouring workflows for adaptive external beam radiotherapy”. In: *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference*. 2022.

[134] Wenhui Lei et al. “DeepIGeoS-V2: deep interactive segmentation of multiple organs from head and neck images with lightweight CNNs”. In: *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention: International Workshops, LABELS 2019, HAL-MICCAI 2019, and CuRIOUS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 4*. Springer. 2019, pp. 61–69.

[135] Bhavani Sambaturu et al. “ScribbleNet: Efficient interactive annotation of urban city scenes for semantic segmentation”. In: *Pattern Recognition* 133 (2023), p. 109011.

[136] Ananya Kumar, Percy S Liang, and Tengyu Ma. “Verified uncertainty calibration”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[137] Ranganath Krishnan and Omesh Tickoo. “Improving model calibration with accuracy versus uncertainty optimization”. In: *Advances in Neural Information Processing Systems* (2020).

[138] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. “Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning”. In: *International conference on machine learning*. PMLR. 2020, pp. 11117–11128.

[139] Sebastian Gruber and Florian Buettner. “Better uncertainty calibration via proper scores for classification and beyond”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 8618–8632.

[140] A Philip Dawid. “The well-calibrated Bayesian”. In: *Journal of the American Statistical Association* 77.379 (1982), pp. 605–610.

[141] Jishnu Mukhoti and Yarin Gal. “Evaluating Bayesian Deep Learning Methods for Semantic Segmentation”. In: *CoRR* abs/1811.12709 (2018). arXiv: 1811.12709.

[142] Yeming Wen et al. “Flipout: Efficient pseudo- independent weight perturbations on mini-batches”. In: *Proceedings of the 6th International Conference on Learning Representations*. 2018.

[143] Prerak Mody et al. “Improving Error Detection in Deep Learning Based Radiotherapy Autocontouring Using Bayesian Uncertainty”. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022*. 2022. ISBN: 978-3-031-16748-5.

[144] Yarin Gal. “Uncertainty in Deep Learning”. In: *PhD Thesis, University of Cambridge* (2016).

[145] Chuan Guo et al. “On calibration of modern neural networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[146] Gabriel Pereyra et al. “Regularizing neural networks by penalizing confident output distributions”. In: *arXiv preprint arXiv:1701.06548* (2017).

[147] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. “When does label smoothing help?” In: *Advances in neural information processing systems* 32 (2019).

[148] Jishnu Mukhoti et al. “Calibrating deep neural networks using focal loss”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15288–15299.

[149] Balamurali Murugesan et al. “Calibrating segmentation networks with margin-based label smoothing”. In: *Medical Image Analysis* 87 (2023), p. 102826.

[150] Charlotte L Brouwer et al. “3D variation in delineation of head and neck organs at risk”. In: *Radiation Oncology* 7.1 (2012), pp. 1–10.

[151] Shi Hu et al. “Supervised uncertainty quantification for segmentation with multiple annotations”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 137–145.

[152] Eli Gibson et al. “Artificial Intelligence with Statistical Confidence Scores for Detection of Acute or Subacute Hemorrhage on Noncontrast CT Head Scans”. In: *Radiology: Artificial Intelligence* 4.3 (2022), e210115.

[153] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *International conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[154] Li Wan et al. “Regularization of neural networks using dropconnect”. In: *International conference on machine learning*. PMLR. 2013, pp. 1058–1066.

[155] Charles Blundell et al. “Weight uncertainty in neural network”. In: *International conference on machine learning*. PMLR. 2015, pp. 1613–1622.

[156] Roger D Soberanis-Mukul, Nassir Navab, and Shadi Albarqouni. “Uncertainty-based graph convolutional networks for organ segmentation refinement”. In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 755–769.

[157] Andres Diaz-Pinto et al. “DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images”. In: *Data Augmentation, Labelling, and Imperfections: Second MICCAI Workshop, DALI 2022, Held in Conjunction with MICCAI 2022*. Springer. 2022, pp. 11–21.

[158] Zhipeng Ding et al. “Local temperature scaling for probability calibration”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6889–6899.

[159] Yaniv Ovadia et al. “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift”. In: *Advances in neural information processing systems* 32 (2019).

[160] Balamurali Murugesan et al. “Trust your neighbours: Penalty-based constraints for model calibration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 572–581.

[161] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[162] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[163] Bingyuan Liu et al. “The devil is in the margin: Margin-based label smoothing for network calibration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 80–88.

[164] Bingyuan Liu et al. “Class adaptive network calibration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16070–16079.

[165] Sunil Thulasidasan et al. “On mixup training: Improved calibration and predictive uncertainty for deep neural networks”. In: *Advances in neural information processing systems* 32 (2019).

[166] Davood Karimi et al. “Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images”. In: *Medical image analysis* 57 (2019), pp. 186–196.

[167] Jeremy Nixon et al. “Measuring Calibration in Deep Learning.” In: *CVPR workshops*. Vol. 2. 7. 2019.

[168] Max-Heinrich Laves et al. “Well-calibrated model uncertainty with temperature scaling for dropout variational inference”. In: *arXiv preprint arXiv:1909.13550* (2019).

[169] Biraja Ghoshal and Allan Tucker. *On Calibrated Model Uncertainty in Deep Learning*. 2022. URL: `https://europepmc.org/article/PPR/PPR517139`.

[170] Zijie Chen et al. “A novel hybrid convolutional neural network for accurate organ segmentation in 3D head and neck CT images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 569–578.

[171] Saeid Asgari Taghanaki et al. “Combo loss: Handling input and output imbalance in multi-organ segmentation”. In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 24–33.

[172] Michael Yeung et al. “Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation”. In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102026.

[173] Prerak P Mody et al. “Comparing Bayesian models for organ contouring in head and neck radiotherapy”. In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 100–109.

[174] K Kian Ang et al. “Randomized phase III trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage III to IV head and neck carcinoma: RTOG 0522”. In: *Journal of clinical oncology* 32.27 (2014), p. 2940.

[175] Xianghua Ye et al. “Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study”. In: *Nature Communications* 13.1 (2022), p. 6137.

[176] Anneke Meyer et al. “Anisotropic 3D multi-stream CNN for accurate prostate segmentation from multi-planar MRI”. In: *Computer Methods and Programs in Biomedicine* 200 (2021), p. 105821.

[177] Michela Antonelli et al. “The medical segmentation decathlon”. In: *Nature communications* 13.1 (2022), p. 4128.

[178] Geert Litjens et al. “Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge”. In: *Medical image analysis* 18.2 (2014), pp. 359–373.

[179] Paul J Doolan et al. “A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy”. In: *Frontiers in oncology* 13 (2023), p. 1213068.

[180] Gerd Heilemann et al. “Clinical implementation and evaluation of auto-segmentation tools for multi-site contouring in radiotherapy”. In: *Physics and Imaging in Radiation Oncology* 28 (2023), p. 100515.

[181] Young Woo Kim, Simon Biggs, and Elizabeth Claridge Mackonis. “Investigation on performance of multiple ai-based auto-contouring systems in organs at risks (oars) delineation”. In: *Physical and Engineering Sciences in Medicine* 47.3 (2024), pp. 1123–1140.

[182] Lee Goddard et al. “Evaluation of multiple-vendor AI autocontouring solutions”. In: *Radiation Oncology* 19.1 (2024), p. 69.

[183] Prerak Mody et al. “Improving Uncertainty-Error Correspondence in Deep Bayesian Medical Image Segmentation”. In: *Machine Learning for Biomedical Imaging* 2 (August 2024 issue 2024), pp. 1048–1082. ISSN: 2766-905X. DOI: https://doi.org/10.59275/j.melba.2024-5gc8. URL: https://melba-journal.org/2024:018.

[184] Michael J Trimpl et al. “Deep learning-assisted interactive contouring of lung cancer: Impact on contouring time and consistency”. In: *Radiotherapy and Oncology* 200 (2024), p. 110500.

[185] Douwe J Spaanderman et al. “Minimally interactive segmentation of soft-tissue tumors on CT and MRI using deep learning”. In: *European Radiology* (2024), pp. 1–10.

[186] Mathis Ersted Rasmussen et al. “A simple single-cycle interactive strategy to improve deep learning-based segmentation of organs-at-risk in head-and-neck cancer”. In: *Physics and Imaging in Radiation Oncology* 26 (2023), p. 100426.

[187] Zixiang Wei et al. “An Interactive Deep-Learning Workflow for Head and Neck Gross Tumour Volume Segmentation”. In: *Available at SSRN 5219763* (2025).

[188] Julie van der Veen, Akos Gulyban, and Sandra Nuyts. “Interobserver variability in delineation of target volumes in head and neck cancer”. In: *Radiotherapy and Oncology* 137 (2019), pp. 9–15.

[189] M Jorge Cardoso et al. “Monai: An open-source framework for deep learning in healthcare”. In: *arXiv preprint arXiv:2211.02701* (2022).

[190] Heleen Bollen, Akos Gulyban, and Sandra Nuyts. “Impact of consensus guidelines on delineation of primary tumor clinical target volume (CTVp) for head and neck cancer: Results of a national review project”. In: *Radiotherapy and Oncology* 189 (2023), p. 109915.

[191] Hugo Pereira, Luis Romero, and Pedro Miguel Faria. “Web-Based DICOM Viewers: A Survey and a Performance Classification”. In: *Journal of Imaging Informatics in Medicine* (2024), pp. 1–19.

[192] Bill Lubanovic. *FastAPI*. " O’Reilly Media, Inc.", 2023.

[193] Sébastien Jodogne. “The Orthanc ecosystem for medical imaging”. In: *Journal of digital imaging* 31.3 (2018), pp. 341–352.

[194] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh. “3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support”. In: *Intraoperative imaging and image-guided therapy*. Springer, 2013, pp. 277–289.

[195] napari contributors. *napari: a multi-dimensional image viewer for Python*. urlhttps://doi.org/10.5281/zenodo.3555620. 2019. DOI: 10.5281/zenodo.3555620.

[196] Constantin Ulrich et al. “RadioActive: 3D Radiological Interactive Segmentation Benchmark”. In: *CoRR* (2024).

[197] Junlong Cheng et al. “Interactive medical image segmentation: A benchmark dataset and baseline”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 20841–20851.

[198] May Abdel-Wahab et al. “Global radiotherapy: current status and future directions—white paper”. In: *JCO global oncology* 7 (2021), pp. 827–842.

[199] Hongcheng Zhu et al. “Global radiotherapy demands and corresponding radiotherapy-professional workforce requirements in 2022 and predicted to 2050: a population-based study”. In: *The Lancet Global Health* 12.12 (2024), e1945–e1953.

[200] Mark J Gooding et al. “Fully automated radiotherapy treatment planning: A scan to plan challenge”. In: *Radiotherapy and Oncology* 200 (2024), p. 110513.

[201] Dylan Callens et al. “Is full-automation in radiotherapy treatment planning ready for take off?” In: *Radiotherapy and Oncology* (2024), p. 110546.

[202] Stanislav Nikolov et al. “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy”. In: *ArXiv e-prints* (2018). arXiv: 1809.04430 [cs.CV]. URL: https://arxiv.org/abs/1809.04430.

[203] Young Woo Kim, Simon Biggs, and Elizabeth Claridge Mackonis. “Investigation on performance of multiple ai-based auto-contouring systems in organs at risks (oars) delineation”. In: *Physical and Engineering Sciences in Medicine* 47.3 (2024), pp. 1123–1140.

[204] Prerak Mody et al. “Large-scale dose evaluation of deep learning organ contours in head-and-neck radiotherapy by leveraging existing plans”. In: *Physics and Imaging in Radiation Oncology* 30 (2024), p. 100572.

[205] Jishnu Mukhoti and Yarin Gal. “Evaluating bayesian deep learning methods for semantic segmentation”. In: *arXiv preprint arXiv:1811.12709* (2018).

[206] Alex Kendall and Yarin Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[207] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. “Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding”. In: *arXiv preprint arXiv:1511.02680* (2015).

[208] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. “Being bayesian, even just a bit, fixes overconfidence in relu networks”. In: *International conference on machine learning*. PMLR. 2020, pp. 5436–5446.

[209] Andrew YK Foong et al. “‘In-Between’Uncertainty in Bayesian Neural Networks”. In: *arXiv preprint arXiv:1906.11537* (2019).

[210] Aryan Mobiny et al. “Dropconnect is effective in modeling uncertainty of bayesian deep networks”. In: *Scientific reports* 11.1 (2021), p. 5458.

[211] Alex Kendall and Yarin Gal. “What uncertainties do we need in Bayesian deep learning for computer vision?” In: *Advances in neural information processing systems* 30 (2017).

[212] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.

[213] Tomaž Vrtovec et al. “Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods”. In: *Medical physics* 47.9 (2020), e929–e950.

[214] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer. 2015, pp. 234–241. URL: https://doi.org/10.1007/978-3-319-24574-4_28.

[215] Wentao Zhu et al. “Anatomynet: Deep 3d squeeze-and-excitation u-nets for fast and fully automated whole-volume anatomical segmentation”. In: *BioRxiv* (2018), p. 392969.

[216] Carlos E Cardenas et al. “Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations”. In: *Medical physics* 47.5 (2020), pp. 2317–2322.

[217] Valentin Oreiller et al. “Head and neck tumor segmentation in PET/CT: the HECKTOR challenge”. In: *Medical image analysis* 77 (2022), p. 102336.

[218] Johannes A Langendijk et al. “Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach”. In: *Radiotherapy and Oncology* 107.3 (2013), pp. 267–273.

[219] Lisa Van den Bosch et al. “Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: A new concept for individually optimised treatment”. In: *Radiotherapy and Oncology* 157 (2021), pp. 147–154.

[220] *Scripting in Raystation*. `https://www.raysearchlabs.com/siteassets/about-overview/media-center/wp-re-ev-n-pdfs/white-papers/white-paper-5---scripting-aug-20152.pdf`.

[221] *Scripting in Raystation*. URL: `https://www.raysearchlabs.com/siteassets/about-overview/media-center/wp-re-ev-n-pdfs/white-papers/white-paper-5---scripting-aug-20152.pdf`.

[222] Michael J Trimpl et al. “Interactive contouring through contextual deep learning”. In: *Medical Physics* 48.6 (2021), pp. 2951–2959.

[223] Alexander Kirillov et al. “Segment anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[224] Yunyang Xiong et al. “Efficientsam: Leveraged masked image pretraining for efficient segment anything”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 16111–16121.

[225] Douwe J Spaanderman et al. “Minimally interactive segmentation of soft-tissue tumors on CT and MRI using deep learning”. In: *European Radiology* (2024), pp. 1–10.

[226] Hallee E Wong et al. “Scribbleprompt: fast and flexible interactive segmentation for any biomedical image”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 207–229.

[227] Sarbani Ghosh Laskar et al. “Access to radiation therapy: from local to global and equality to equity”. In: *JCO global oncology* 8 (2022), e2100358.

List of publications

Journal articles

Mody, Prerak, Nicolas F. Chaves-de-Plaza, Chinmay Rao, Eleftheria Astrenidou, Mischa de Ridder, Nienke Hoekstra, Klaus Hildebrandt, and Marius Staring. "Improving Uncertainty-Error Correspondence in Deep Bayesian Medical Image Segmentation." *Machine Learning for Biomedical Imaging*, August 2024 issue (2024): 1048–82. https://doi.org/10.59275/j.melba.2024-5gc8.

Mody, Prerak, Merle Huiskes, Nicolas F. Chaves-de-Plaza, Alice Onderwater, Rense Lamsma, Klaus Hildebrandt, Nienke Hoekstra, Eleftheria Astreinidou, Marius Staring, and Frank Dankers. "Large-scale dose evaluation of deep learning organ contours in head-and-neck radiotherapy by leveraging existing plans." *Physics and Imaging in Radiation Oncology* 30 (2024): 100572.

Chaves-de-Plaza, Nicolas F., **Prerak Mody**, Marius Staring, René van Egmond, Anna Vilanova, and Klaus Hildebrandt. "Inclusion depth for contour ensembles." *IEEE Transactions on Visualization and Computer Graphics* 30, no. 9 (2024): 6560-6571.

Chaves-de-Plaza, Nicolas F., Mathijs Molenaar, **Prerak Mody**, Marius Staring, René van Egmond, Elmar Eisemann, Anna Vilanova, and Klaus Hildebrandt. "Depth for Multi-Modal Contour Ensembles." *Computer Graphics Forum*, vol. 43, no. 3, p. e15083. 2024.

Jia, Jingnan, Bo Yu, **Prerak Mody**, Maarten K. Ninaber, Anne A. Schouffoer, Jeska K. de Vries-Bouwstra, Lucia JM Kroft, Marius Staring, and Berend C. Stoel. "Using 3D point cloud and graph-based neural networks to improve the estimation of pulmonary function tests from chest CT." *Computers in Biology and Medicine* 182 (2024): 109192.

Chaves-de-Plaza, Nicolas F., **Prerak Mody**, Klaus Hildebrandt, Marius Staring, Eleftheria Astreinidou, Mischa de Ridder, Huib de Ridder, Anna Vilanova, and René van Egmond. "Implementation of delineation error detection systems in time-critical radiotherapy: Do AI-supported optimization and human preferences meet?." *Cognition, Technology & Work* (2024): 1-17.

Mody, Prerak, Nicolas Chaves de Plaza, Mark Gooding, Martin de Jong, Mischa de Ridder, Niels den Hans, Jos Elbers, Klaus Hildebrandt, Marius Staring. "Manual Brush vs AI Pencil: Evaluating tools for auto-contour refinement of head-and-neck tumors on CT+PET". *Sumbitted* (2025).

Gao, Ruochen, **Prerak Mody**, Chinmay Rao, Frank Dankers, and Marius Staring. "On factors that influence deep learning-based dose prediction of head and neck tumors." *Physics in Medicine & Biology* 70, no. 11 (2025): 115006.

International conference proceedings

Mody, Prerak, Nicolas F. Chaves-de-Plaza, Klaus Hildebrandt, and Marius Staring. "Improving error detection in deep learning based radiotherapy autocontouring using Bayesian uncertainty." *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 70-79. Cham: Springer Nature Switzerland, 2022.

Chaves-de-Plaza, Nicolas F., **Prerak Mody**, Klaus Hildebrandt, Marius Staring, Eleftheria Astreinidou, Mischa de Ridder, Huib de Ridder, and René van Egmond. "Towards fast human-centred contouring workflows for adaptive external beam radiotherapy." *Proceedings of the Human Factors and Ergonomics Society Europe* (2022): 111-31.

Mody, Prerak P., Nicolas Chaves-de-Plaza, Klaus Hildebrandt, René van Egmond, Huib de Ridder, and Marius Staring. "Comparing Bayesian models for organ contouring in head and neck radiotherapy." *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 100-109. SPIE, 2022.

Tan, Yicong, **Prerak Mody**, Viktor van der Valk, Marius Staring, and Jan van Gemert. "Analyzing components of a transformer under different dataset scales in 3D prostate CT segmentation." *Medical Imaging 2023: Image Processing*, vol. 12464, pp. 49-60. SPIE, 2023.

Open source software

Mody, Prerak, Koning, Patrick. "A terminal user interface (TUI) to view the status of your SLURM cluster." *online at https://pypi.org/project/slurm-viewer/* (2024).

Acknowledgements

If I have seen further it is by standing on the shoulders of giants.

SIR ISAAC NEWTON

A Ph.D. journey is a challenging endeavor, filled with both triumphs and setbacks. Navigating this multi-year project would have been impossible without the support, motivation, and guidance of colleagues, friends, and family.

First and foremost, I extend my gratitude to my Ph.D. advisor, Marius Staring. Marius was a mentor who has an incredibly positive outlook and was thus able to always provide me motivation in the face of experimental setbacks. Recognizing my passion for sharing research, Marius consistently motivated and supported my ambition to present my ideas to various research labs. I also want to express my appreciation to Boudewijn Lelieveldt, who, alongside Marius, interviewed me for this Ph.D. position and also fostered a culture of openness at LKEB. I would also like to thank the funding agencies: Varian (Stockholm, Sweden) and HollandPTC (Delft, The Netherlands).

I also extend my gratitude to my closest academic collaborator: Nicolas Chaves de Plaza and his PhD advisors Klaus Hildebrandt and René van Egmond. Despite the challenges faced in a PhD, Nicolas always found a way to give it a positive spin when compared to other jobs. Klaus always had actionable feedback in our meetings and as my MSc thesis advisor was also the one who recommended me for this PhD!

Among my peers, I am especially thankful for the camaraderie and intellectual exchange shared with Viktor van der Valk, Jingnan Jia, Chinmay Rao, Ruochen Gao, Li-Hsin Cheng, Cedric Rodriguez and Vangelis Kostoulas. Viktor was the person I could always go to when things got tough, and Jingnan's sunny disposition never failed to cheer me up. I also thank them for organizing LKEB's Thursdays meetup with me where we read and debated over deep learning methodologies. Cedric, Vangelis and I enjoyed a memorable conference experience in the US which turned into a friendship that still lasts today. Cedric and I also collaborated on organizing LUMC-wide sessions on the application of AI in clinical settings. Ruochen and Chinmay were always enthusiastic about engaging in technical discussions, which ultimately lead to fruitful research collaborations. I particularly thank Chinmay for his unwavering support during the stressful review period of my first journal publication.

Other colleagues I would like to thank are Mohammed Elhmady, Hessam Sokooti and Pieter Kitslaar. While Mohammed and Hessam were helped me navigate work-from-home challenges during the start of my PhD (2020), Pieter, my current manager at Medis Medical Imaging provided me with time to write this thesis. Michèle Huijberts always provided

the IT support I needed and also a good late evening chat. I would also like to thank Patrick de Koning, for his partnership on my first open-source software: *slurm-viewer*. From the LUMC radiotherapy lab, I am grateful to Frank Dankers for providing support during the publication of my third paper and Eleftheria Astrenidou for my first. I also appreciate the participation of Alex Vieth, Yauhenia Makarevich, Faeze Gholamiankhah as well as Chinmay, Ruochen, Patrick and Frank for my fourth study. Finally, I thank all LKEB members for their part in creating a wonderful research culture at the lab: Rob, Jouke, Berend, Oleh, Niels, Baldur, Berend, Alex, YanLi, Silvia, Xiatong, Simon, Yunjie, Soumyadeep, Laurens, Donghang, ChangLi and Efe.

In a foreign land, my close friends have been a constant source of support throughout this journey. I am grateful to Avinash Kini and Siddarth Bharteeya. Our friendship, forged during the CoVid pandemic, has enjoyed both intellectual discourse and joyous laughter. Siddarth was always available for a chat (and a new business idea!), while Avinash proved to be an exceptional listener and a skilled arbitrator of lively debates. My master's degree friends, Ravi Autar, Shirani Bisnayak, Arthur Hoyesan, and Jesse Hangenaars, also played a significant role in my life during my Ph.D. journey. I cherish Ravi's exceptional hosting skills, Shirani's ever expanding knowledge on pop culture, Arthur's wonderful sense of humor, and my consistent quarterly lunches with Jesse.

Of course, my deepest gratitude goes to my family: my mother Savita Mody, for instilling in me the virtue of patience; my father Pradeep Mody, for always being a great listener; and my brother Prakhar Mody, for his steadfast commitment to logical thinking. I also thank my uncle Viren Radia and my aunt Sonal Radia for always being available to have a chat with me.

Last, but certainly not least, I wish to express my profound appreciation to my partner, Sailee Sansgiri. Her constant presence in my life made every effort worthwhile. Sharing my successes, failures, apprehensions, and fears with her was the bedrock that carried me through this Ph.D. I aspire to spend the rest of my life with her to eventually grow old together. Her wit, humor, charm, smile, and creativity are an everlasting source of inspiration for me.

Finally, to everyone who has been a part of this journey, thank you.

Curriculum Vitae

Prerak was born in Mumbai, Maharashtra province, India in 1993. He finished his high school studies at Arya Vidya Mandir, Mumbai, India in 2011. He obtained a Bachelors of Technology (B.Tech) degree in Computer Science and Enginering at Vellore Institute of Technology (VIT), India in 2015. His masters studies (MSc) in Computer Science was done at Technical University (TU) Delft, Netherlands from 2018-2020. The topic of the masters thesis at Philips, Netherlands was on 3D human pose estimation in privacy preserving settings (for e.g. using point cloud images).

Shortly after he started his PhD at Division of Image Processing (LKEB), Radiology, Leiden University Medical Center with a focus on human-centered techniques for contouring in radiotherapy. Since 2025 he works at Medis Medical Imaging building software solutions to detect coronary structures within CT scans using deep learning.