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

Prerak Mody[a,b,*], Merle Huiskes[c], Nicolas Chaves de Plaza[b,d], Alice Onderwater[c], Rense Lamsma[c], Klaus Hildebrandt[d], Nienke Hoekstra[c], Eleftheria Astreinidou[c], Marius Staring[a,c], Frank Dankers[c]

[a]*Division of Image Processing (LKEB), Department of Radiology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands*
[b]*HollandPTC consortium – Erasmus Medical Center, Rotterdam, Holland Proton Therapy Centre, Delft, Leiden University Medical Center (LUMC), Leiden and Delft University of Technology, Delft, The Netherlands*
[c]*Department of Radiation Oncology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands*
[d]*Computer Graphics and Visualization Group, EEMCS, TU Delft, Delft, 2628 CD, The Netherlands*

## Abstract

(*249/250 words*)
*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 $\Delta$NTCP (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.

*Keywords:* Automated Plan Optimization, Auto Contouring, Dose Impact, Robot Process Automation, Automated Plans

(*2963/3000 words*)

## 1. Introduction

Manual contouring of organs-at-risk (OAR) in radiotherapy is a time and resource-demanding task [1–3], especially in head-and-neck cancer due to a large OAR count [4]. Moreover, it is plagued by inter- and

[*] Corresponding author
*Email address:* p.p.mody@lumc.nl (Prerak Mody)
*URL:* www.lkeb.nl (Prerak Mody)

intra-annotator variability [5–8] 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 [9]. 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 [6, 10], the next step is to evaluate their dose impact [11]. However, we observed that dose-based studies on auto-contours tend to use either smaller ($\leq 20$) [12–18] or medium-sized ($\leq 40$) [19], rather than larger [20] datasets. Studies using larger datasets simply superimpose the automated contours on the clinical dose [20] which does not fully replicate the treatment planning process. Conversely, studies using smaller or medium-sized test datasets either made manual plans [14, 17–19], used knowledge-based planning [13], a template approach [12] or a priori multi-criteria optimization (MCO) [15, 16]. 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 [12, 13, 15], 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. Materials and methods

### *2.1. Data acquisition*

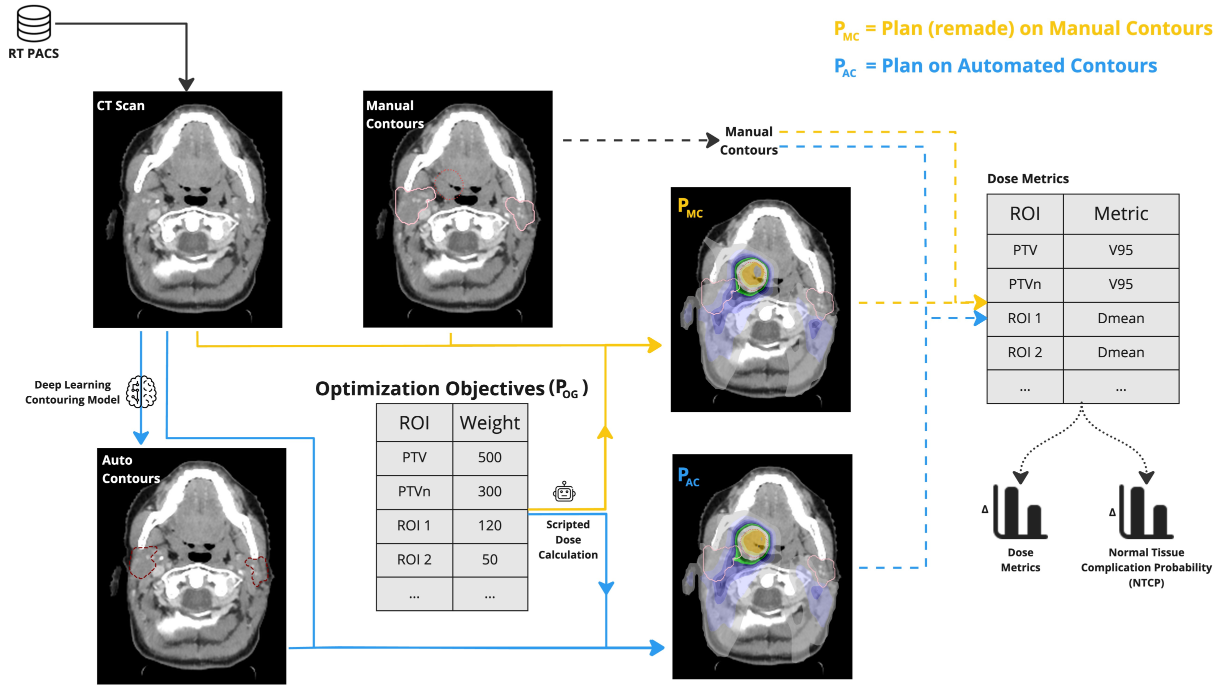
Our dataset consists of 100 head-and-neck cancer patients, of which 70 had clinical plans made for photon therapy, while 30 had proton plans, at Leiden University Medical Center (Leiden, The Netherlands) from 2021 to 2023. Patients were treated for either oropharyngeal (71) or hypopharyngeal (29) cancers 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 Supplementary Material A. 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. 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 Supplementary Material B for additional details.

### *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.



RT PACS
CT Scan
Manual Contours
Manual Contours
$P_{MC}$ = Plan (remade) on Manual Contours
$P_{AC}$ = Plan on Automated Contours
$P_{MC}$
$P_{AC}$
Deep Learning Contouring Model
Auto Contours
Scripted Dose Calculation
Dose Metrics
Normal Tissue Complication Probability (NTCP)
Δ
Δ

Optimization Objectives ($P_{OG}$)

| ROI | Weight |
|---|---|
| PTV | 500 |
| PTVn | 300 |
| ROI 1 | 120 |
| ROI 2 | 50 |
| ... | ... |

Dose Metrics

| ROI | Metric |
|---|---|
| PTV | V95 |
| PTVn | V95 |
| ROI 1 | Dmean |
| ROI 2 | Dmean |
| ... | ... |

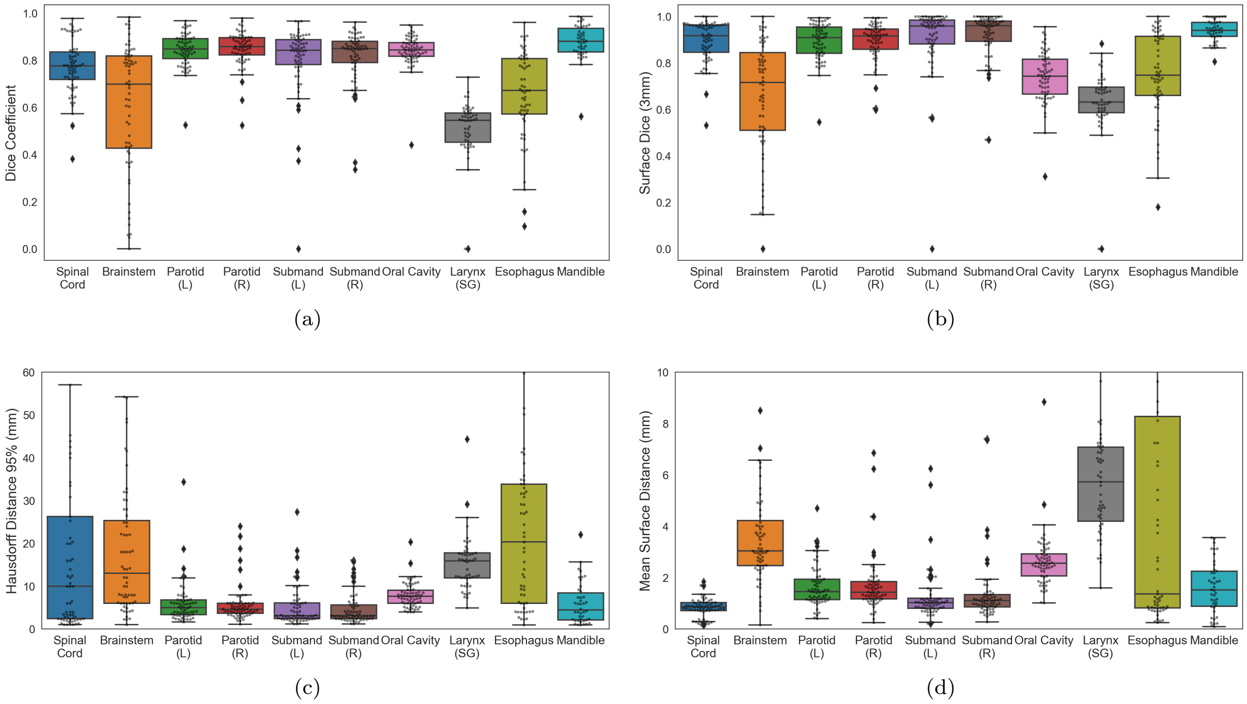
Figure 1: Workflow for automated plan optimization and use-case of evaluating the effect of automated contours on dose. By reusing original plan ($P_{OG}$) parameters, we made a plan for both the manual contours ($P_{MC}$) and automated contours($P_{AC}$), shown with yellow and blue colors respectively. 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) [21] models and compare all plans.

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 $V_{94\%} \geq 98\%$ [22] and voxel-wise maximum (vw-max) of $D_{2\%} \leq 107\%$.

### *2.4. Automated Treatment Planning*

To make our automated program, a four-step script [23–25] was created which uses manually defined beam settings and objective weights from the clinical plan (more details in Supplementary Material C). This approach is also referred as robot process automation (RPA) [26], 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 [27] 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.



(a)

(b)

(c)

(d)

Figure 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.

Using our automated program, we made two plans – 1) a plan optimized on manual contours ($\text{P}_{MC}$) and 2) a plan optimized on automated contours ($\text{P}_{AC}$) as shown in Figure 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 ($\text{P}_{OG}$). 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 https://github.com/prerakmody/dose-eval-via-existing-plan-parameters.

### 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 [28].

### 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:

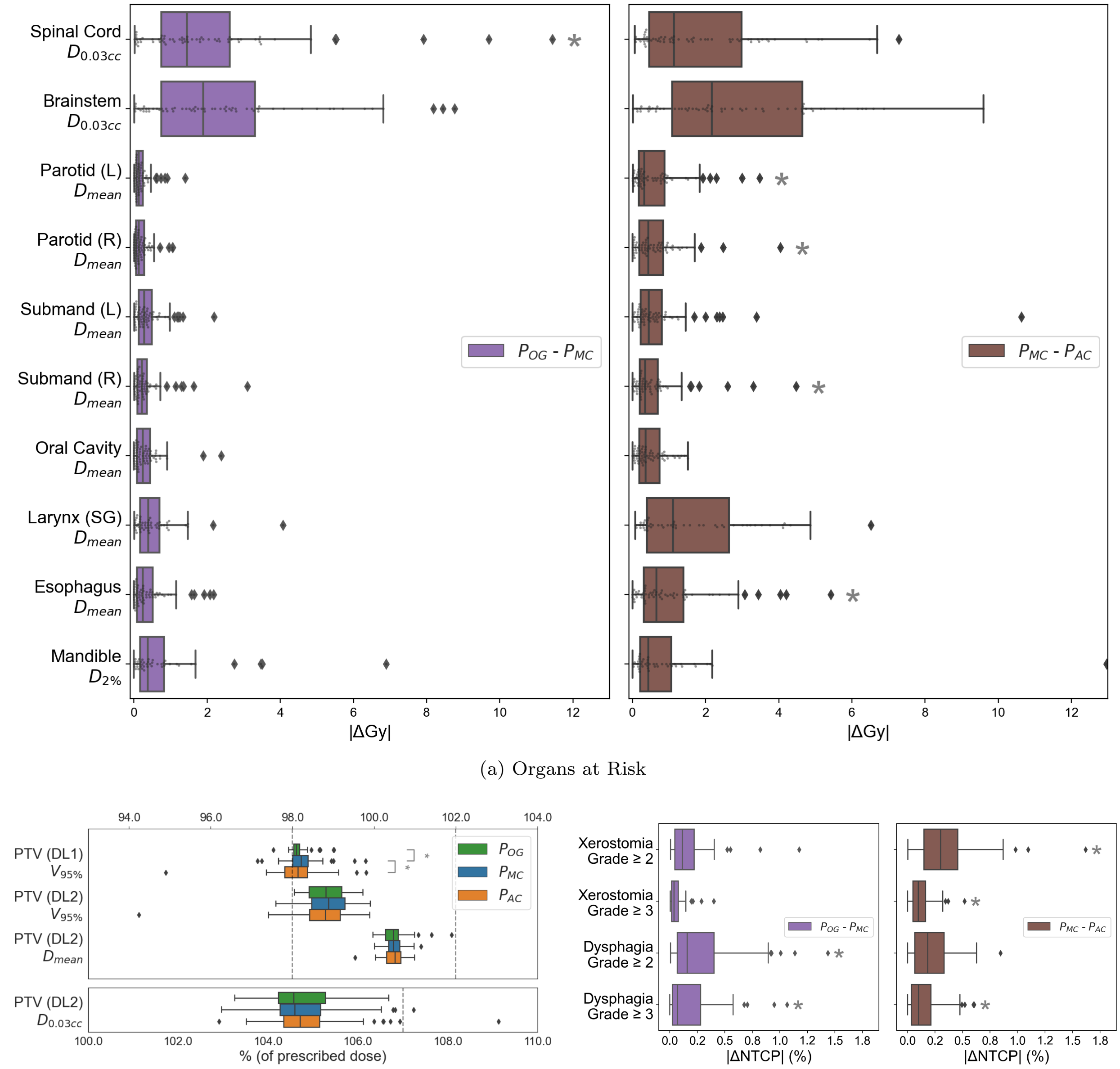
$$\Delta D_x = D_{x,p1} - D_{x,p2}. \tag{1}$$

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 [21] evaluation, we used a similar approach:

$$\Delta \text{NTCP}_d = \text{NTCP}_{d,p1} - \text{NTCP}_{d,p2}, \tag{2}$$

where $d$ refers to either Xerostomia or Dysphagia with a grade $\geq 2$ or $\geq 3$.



(a) Organs at Risk

(b) Targets

(c) NTCP

Figure 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 $\leq 0.05$. 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.

For the above $\Delta D_x$ (dose) and $\Delta \text{NTCP}_d$ 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.

## 3. Results

### 3.1. Geometric evaluation

Figure 2 shows five organs (Spinal Cord, Parotids, Submandibulars, Oral Cavity, Mandible) had a median DICE $\geq 0.78$ (with additional summary measures tabulated in Supplementary Material B). In Figure 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 2c and Figure 2d shows that HD95 and MSD had trends similar to DICE in Figure 2a. OARs with a median DICE $\geq 0.8$ 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.

### 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 3a listed in Supplementary Material D. Figure 3b shows dose metrics for targets where, for $P_{MC}$ and $P_{AC}$, we achieved PTV (DL1) ($V_{95}$) $\geq$ 98.0% for 76% and 60% of plans. However, 96% and 93% of $P_{MC}$ and $P_{AC}$ plans achieved PTV (DL1) ($V_{95}$) $\geq$ 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 3c shows $|\Delta\text{NTCP}|$ results, where the maximum median across all toxicities was 0.3% (individual toxicity metrics in Supplementary Material E).

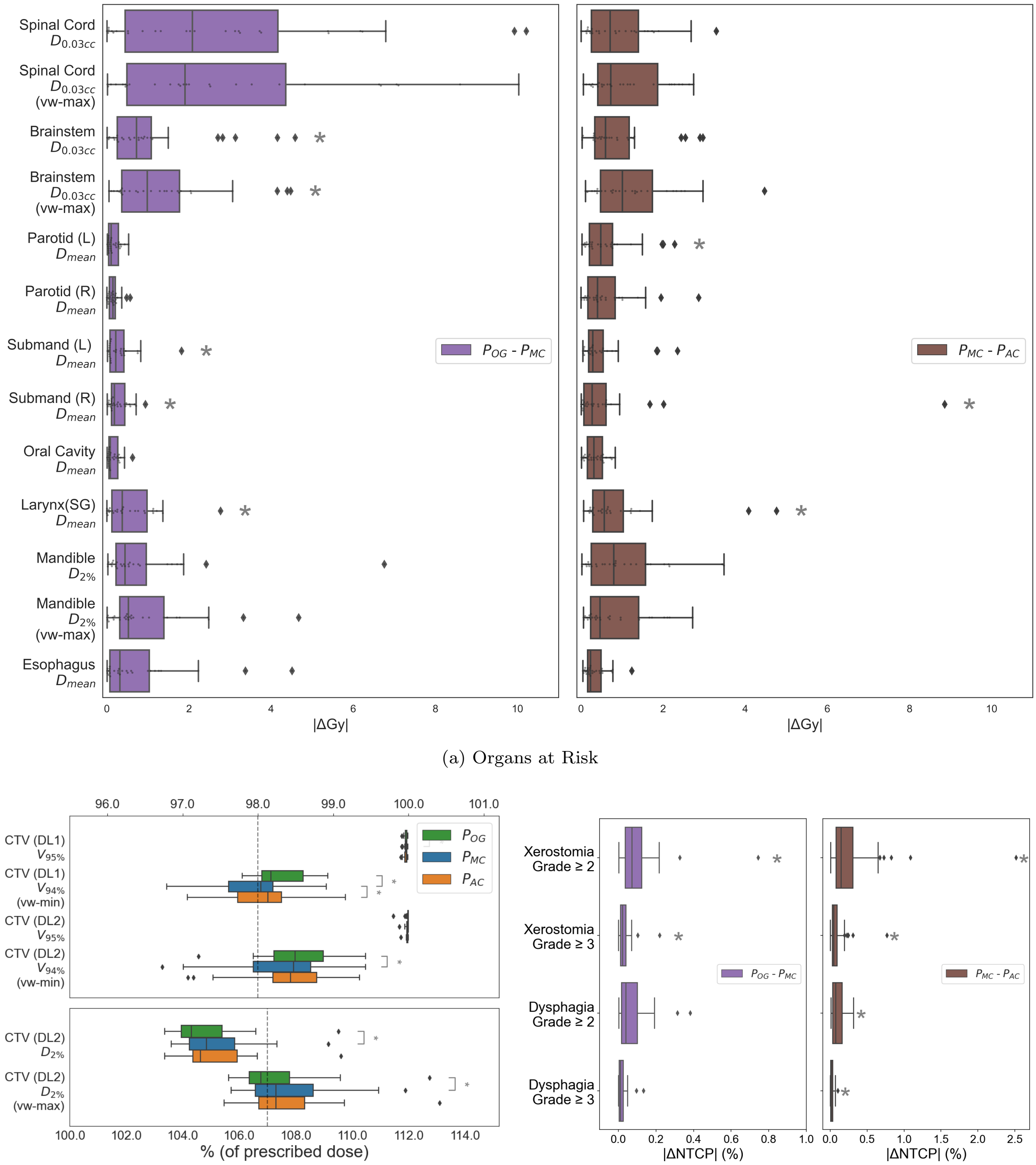
For proton, $|P_{OG} - P_{MC}|$ had a median value of 0.33Gy (1.5%), 1.13Gy (11.5%) and 0.22Gy (0.8%) for all, CNS and non-CNS organs, respectively. The same for $P_{MC} - P_{AC}$ was 0.48Gy (2.6%), 0.75Gy (6.9%) and 0.38Gy (1.8%). Figure 4b shows proton targets wherein 58% and 62% of $P_{MC}$ and $P_{AC}$ plans achieved PTV (DL1) (vw-min) ($V_{94}$) $\geq$ 98.0%, while 82% and 80% achieved PTV (DL1) (vw-min) ($V_{94}$) $\geq$ 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 $|\Delta\text{NTCP}|$ (Figure 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 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.

## 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 [12–18], 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 ($P_{OG}$) [29, 30]. 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

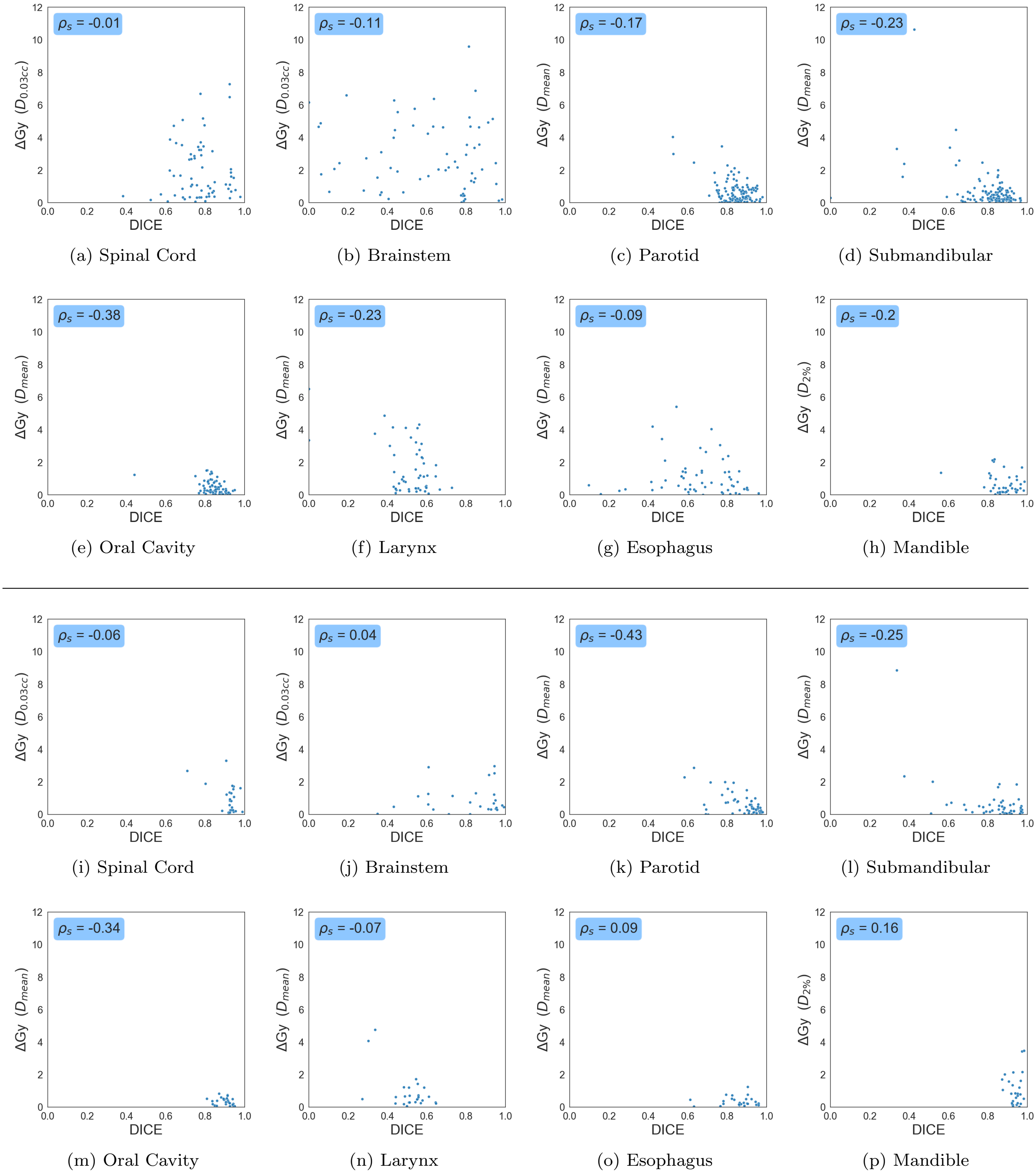


(a) Organs at Risk

(b) Targets

(c) NTCP

Figure 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 $\leq 0.05$. 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 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.

in long-term radiation-induced toxicity. This could potentially promote confidence in the community [31] 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 ($\approx 0.8$) [6, 10, 12] (see Supplementary Material B). A visual inspection of the remaining auto-contours, i.e., Larynx (SG), Brainstem (and by extension the Spinal Cord) (Figure 6, Supplementary Material F) 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 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.

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 3b (photon) and Figure 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.

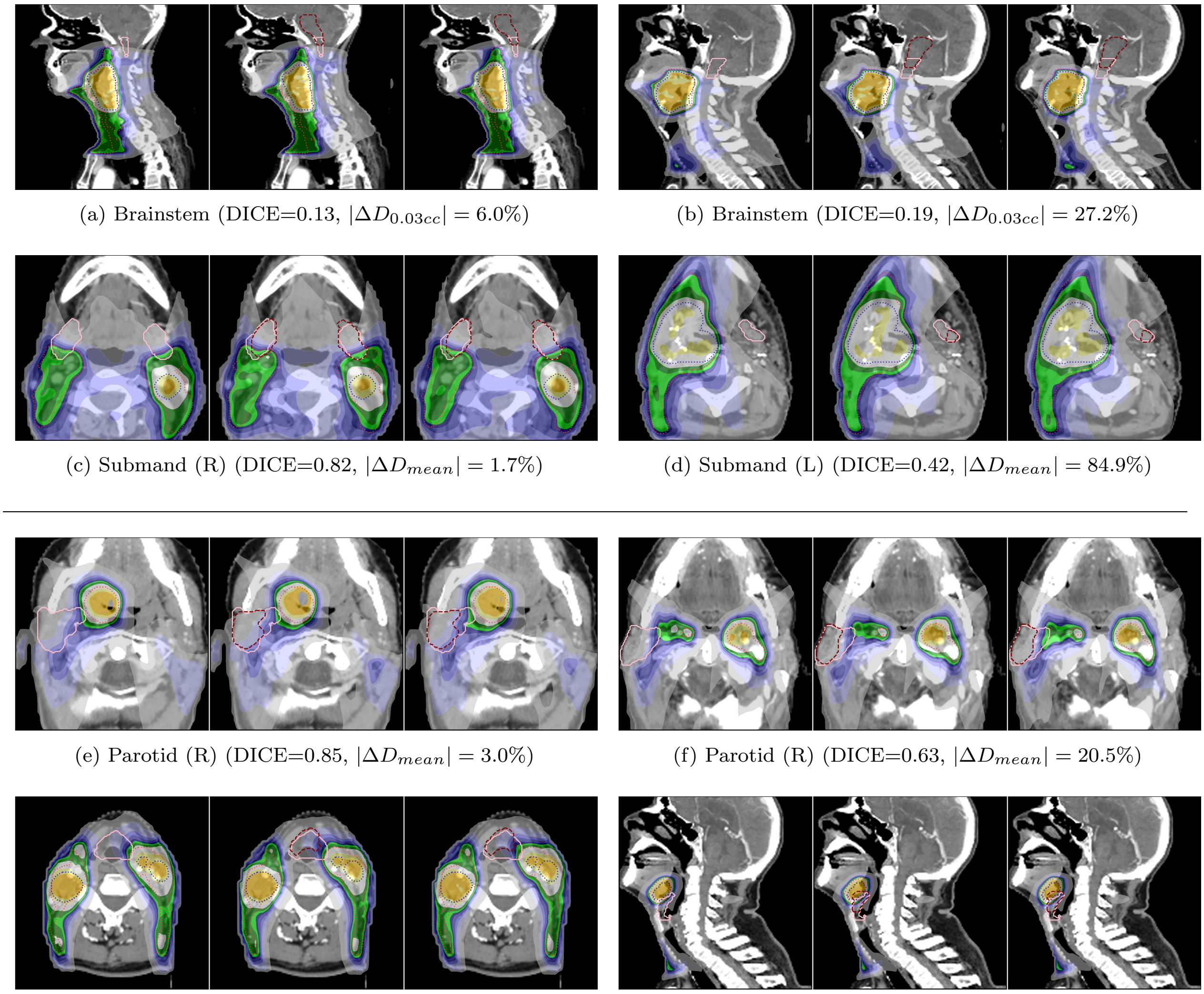
To evaluate the quality of our automated plans, we first assessed target dose metrics. We use PTV (DL1) ($V_{95\%}$) for photon and CTV (DL1) ($V_{94\%}$) (vw-min) for proton, since planners prioritize them due to their difficulty. Hence it serves as a good benchmark for our automated plans. Results indicated that most of our plans ($\geq 93\%$ for photon and $\geq 80\%$ for proton) were of near-clinical quality (i.e., $\geq 97.5\%$). Those plans that did not strictly achieve clinical quality (i.e., $\geq 98\%$) 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 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 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 [32]. 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 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 $\Delta D_{\text{mean}}$. This is seen in the first row of Figure 6 where similar DICE values for the Brainstem output vastly different dose differences. For proton (Figure 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 3a, 3c (photon) and Figure 4a, 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 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 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.



(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 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 $P_{MC} - P_{AC}$.

This work was inspired by prior research on treatment plan scripting [23, 24] to scale-up dose evaluation for auto-contours. However, some plans were still not of the 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.

## Funding

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.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Chaves-de Plaza NF, Mody P, Hildebrandt K, Staring M, Astreinidou E, de Ridder M, 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, p. 111–131.

[2] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:1–13. https://doi.org/10.1118/1.4871620.

[3] Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online Adaptive Radiation Therapy. Int J Radiat Oncol Biol Phys 2017;99:994–1003. https://doi.org/10.1016/j.ijrobp.2017.04.023.

[4] Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother Oncol 2015;117:83–90. https://doi.org/10.1016/j.radonc.2015.07.041.

[5] Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D Variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7(1). https://doi.org/10.1186/1748-717X-7-32.

[6] Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol 2020;144:152–158. https://doi.org/10.1016/j.radonc.2019.10.019.

[7] van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiat Oncol 2021;16:1–11. https://doi.org/10.1186/s13014-020-01677-2.

[8] Stelmes JJ, Vu E, Grégoire V, Simon C, Clementel E, Kazmierska J, 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. Radiat Oncol 2021;16:1–10. https://doi.org/10.1186/s13014-021-01809-2.

[9] Brunenberg EJ, Steinseifer IK, van den Bosch S, Kaanders JH, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imaging Radiat Oncol 2020;15:8–15. https://doi.org/10.1016/j.phro.2020.06.006.

[10] Ng CK, Leung VW, Hung RH. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring for Head and Neck Radiation Therapy. Appl Sci 2022;12. https://doi.org/10.3390/app122211681.

[11] Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–191. https://doi.org/10.1016/j.radonc.2021.05.003.

[12] Kieselmann JP, Kamerling CP, Burgos N, Menten MJ, Fuller CD, Nill S, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. Phys Med Biol 2018;63:aacb65. https://doi.org/10.1088/1361-6560/aacb65.

[13] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. Int J Radiat Oncol Biol Phys 2019;104:677–684. https://doi.org/10.1016/j.ijrobp.2019.02.040.

[14] Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiat Oncol 2021;16:1–14. https://doi.org/10.1186/s13014-021-01837-y.

[15] Costea M, Zlate A, Durand M, Baudier T, Grégoire V, Sarrut D, 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. Radiother Oncol 2022;177:61–70. https://doi.org/10.1016/j.radonc.2022.10.029.

[16] Costea M, Zlate A, Serre AA, Racadot S, Baudier T, Chabaud S, et al. Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images. Radiother Oncol 2023;188:109870. https://doi.org/10.1016/j.radonc.2023.109870.

[17] Lucido JJ, DeWees TA, Leavitt TR, Anand A, Beltran CJ, Brooke MD, et al. Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. Front Oncol 2023;13. https://doi.org/10.3389/fonc.2023.1137803.

[18] Smolders AJ, Choulilitsa E, Czerska K, Bizzocchi N, Krcek R, Lomax AJ, et al. Dosimetric comparison of auto-contouring techniques for online adaptive proton therapy. Phys Med Biol 2023;68:175006. https://doi.org/10.1088/1361-6560/ace307.

[19] Koo J, Caudell J, Feygelman V, Latifi K, Moros EG. Essentially unedited deep-learning-based OARs are suitable for rigorous oropharyngeal and laryngeal cancer treatment planning. J Appl Clin Med Phys 2023;:1–10 https://doi.org/10.1002/acm2.14202.

[20] van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, Steenbakkers Roel JH, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiother Oncol 2020;142:115–123. https://doi.org/10.1016/j.radonc.2019.09.022.

[21] Landelijk Platform Protonentherapie (LPPT) Landelijk Platform Radiotherapie Hoofd-halstumoren (LPRHHT). 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.

[22] Korevaar EW, Habraken SJM, Scandurra D, Kierkels RGJ, Unipan M, Eenink MGC, et al. Practical robustness evaluation in radiotherapy – A photon and proton-proof alternative to PTV-based plan evaluation. Radiother Oncol 2019;141:267–274. https://doi.org/10.1016/j.radonc.2019.08.005.

[23] Xhaferllari I, Wong E, Bzdusek K, Lock M, Chen JZ. Automated IMRT planning with regional optimization using planning scripts. J Appl Clin Med Phys 2013;14:176–191. https://doi.org/10.1120/jacmp.v14i1.4052.

[24] Speer S, Klein A, Kober L, Weiss A, Yohannes I, Bert C. Automation of radiation treatment planning. Strahlentherapie Und Onkol 2017;193:656–665. https://doi.org/10.1007/s00066-017-1150-9.

[25] Teruel JR, Malin M, Liu EK, Mccarthy A, Hu K, Cooper BT, et al. Full automation of spinal stereotactic radiosurgery and stereotactic body radiation therapy treatment planning using Varian Eclipse scripting. J Appl Clin Med Phys 2020;21:122–131. https://doi.org/10.1002/acm2.13017.

[26] Aalst WMPVD, Bichler M, Heinzl A. Robotic Process Automation. Business & Information Systems Engineering 2018;60:269–272. https://doi.org/10.1007/s12599-018-0542-4.

[27] Niemierko A. Reporting and analyzing dose distributions: A concept of equivalent uniform dose. Med Phys 1997;24:103–110. https://doi.org/10.1118/1.598063.

[28] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res 2021;23:e26151. https://doi.org/10.2196/26151.

[29] Gu X, Strijbis VIJ, Slotman BJ, Dahele MR, Verbakel WFAR. Dose distribution prediction for head-and-neck cancer radiotherapy using a generative adversarial network: influence of input data. Front Oncol 2023;13:1251132. https://doi.org/10.3389/fonc.2023.1251132.

[30] Jaworski EM, Mierzwa ML, Vineberg KA, Yao J, Shah JL, Schonewolf CA, et al. Development and Clinical Implementation of an Automated Virtual Integrative Planner for Radiation Therapy of Head and Neck Cancer. Adv Radiat Oncol 2023;8:101029. https://doi.org/10.1016/j.adro.2022.101029.

[31] Petragallo R, Bardach N, Ramirez E, Lamb JM. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation : A survey study of medical dosimetrists. Journal of Applied Clinical Medical Physics 2022;23:1–10. https://doi.org/10.1002/acm2.13568.

[32] Langendijk JA, Hoebers FJ, De Jong MA, Doornaert P, Terhaard CH, Steenbakkers RJ, et al. National protocol for model-based selection for proton therapy in head and neck cancer. Int J Part Ther 2021;8:354–365. https://doi.org/10.14338/IJPT-20-00089.1.

[33] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer; 2015, p. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

## Supplementary Material A. 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, 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.

## Supplementary Material B. Automated Contours

The auto-contouring model of RayStation 10B 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 [33] 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 B.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 B.2: Summary measures (sample mean [bootstrapped 95% confidence interval]) for volumetric and surface metrics of auto-contours of RayStation 10B.

## Supplementary Material C. 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, 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, 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 | $80.0 \rightarrow \{VDT\}$ |
| 1 | PTV (DL1) | MaxDose | 102% of DL1 prescription | $50.0 \rightarrow \{VDT\}$ |
| 1 | ring $\leq$ PTV (DL1) | MaxDose | 96% of DL1 prescription | $0.0 \rightarrow \{VDT\}$ |
| 1 | PTV (DL2) | MinDose | 100% of DL2 prescription | $80.0 \rightarrow \{VDT\}$ |
| 1 | PTV (DL2) | MaxDose | 102% of DL2 prescription | $50.0 \rightarrow \{VDT\}$ |
| 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 | Ghost$_{\text{Cranial}}$ | DoseFallOff | From 100% to 0% of DL1 prescription over 1.0 cm | 0.5 |
| 1 | Ghost$_{\text{Ear(L)}}$ | DoseFallOff | From 100% to 46% of DL1 prescription over 2.0 cm | 1.0 |
| 1 | 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<br>(*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 C.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 | $800.0 \rightarrow \{VDT\}$ | $*$ |
| 1 | CTV (DL1) - (CTV(DL2) + 3 mm) | MaxDose | 102% of DL1 prescription | $20.0 \rightarrow \{VDT\}$ | $*$ |
| 1 | CTV (DL1) - (CTV(DL2) + 2 cm) | MaxDose | 102% of DL1 prescription | $80.0 \rightarrow \{VDT\}$ | $*$ |
| 1 | CTV (DL2) | MinDose | 100% of DL2 prescription | $800.0 \rightarrow \{VDT\}$ | $*$ |
| 1 | CTV (DL2) | MaxDose | 100% of DL2 prescription | $50.0 \rightarrow \{VDT\}$ | $*$ |
| 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 | $500.0 \rightarrow \{VDT\}$ | $*$ |
| 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 C.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.

## Supplementary Material D. Organ Dose Metrics

In Table D.5 and Table D.7, we show dose metrics for organs available in the RayStation 10B auto-contouring module. 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 D.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 D.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 D.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 D.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.

## Supplementary Material E. NTCP

For NTCP scores, we used the formulae and parameters from the National Indication Protocol for Proton therapy (*Landelijk Indicatie Protocol Protonentherapie*) [21]. 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.

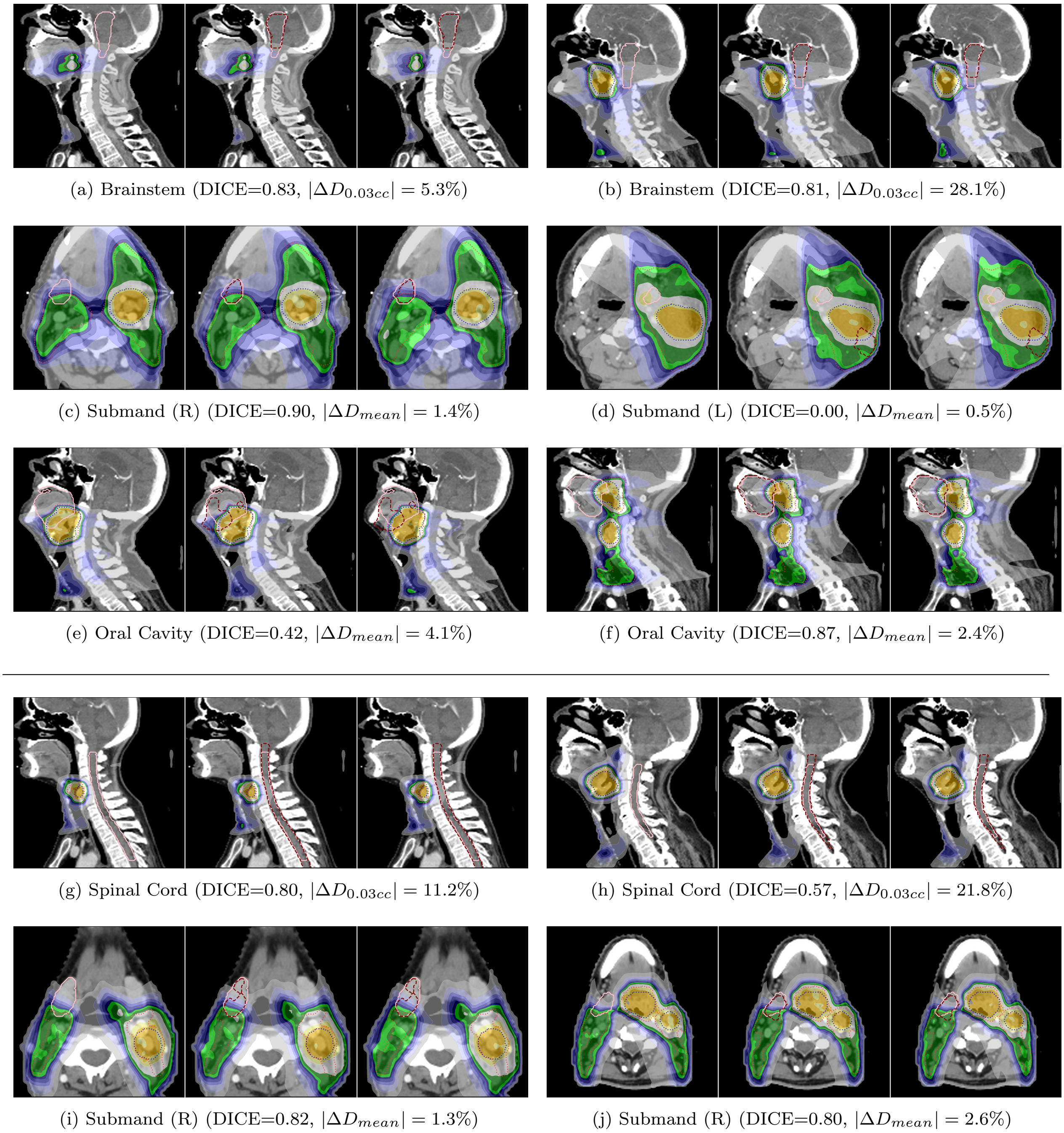
| | Photon | | Proton | |
|---|---|---|---|---|
| | $\|P_{OG} - P_{MC}\|$ | $\|P_{MC} - P_{AC}\|$ | $\|P_{OG} - P_{MC}\|$ | $\|P_{MC} - P_{AC}\|$ |
| Xerostomia Grade $\geq 2$ | 0.1 [0.0,0.5] | 0.3 [0.0,0.9] | 0.1 [0.0,0.3] | 0.2 [0.0,1.0] |
| Xerostomia Grade $\geq 3$ | 0.0 [0.0,0.2] | 0.1 [0.0,0.3] | 0.0 [0.0,0.1] | 0.1 [0.0,0.3] |
| Dysphagia Grade $\geq 2$ | 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 $\geq 3$ | 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 E.9: Summary measures (median [$5^{th}$ percentile, $95^{th}$ percentile]) for ΔNTCP (%) values in photon and proton radiotherapy for $|P_{OG} - P_{MC}|$ and $|P_{MC} - P_{AC}|$.

| | Photon | | Proton | |
|---|---|---|---|---|
| | $\|P_{OG} - P_{MC}\|$ | $\|P_{MC} - P_{AC}\|$ | $\|P_{OG} - P_{MC}\|$ | $\|P_{MC} - P_{AC}\|$ |
| Xerostomia Grade $\geq 2$ | 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 $\geq 3$ | 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 $\geq 2$ | 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 $\geq 3$ | 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 E.10: Sample mean [bootstrapped 95% confidence interval]) for ΔNTCP (%) values in photon and proton radiotherapy for $|P_{OG} - P_{MC}|$ and $|P_{MC} - P_{AC}|$.

# Supplementary Material F. 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 F.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 $P_{MC} - P_{AC}$.