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

Mody, Prerak; Huiskes, Merle; Chaves-de-Plaza, Nicolas F.; Onderwater, Alice; Lamsma, Rense; Hildebrandt, Klaus; Hoekstra, Nienke; Astreinidou, Eleftheria; Staring, Marius; More Authors

**DOI**
10.1016/j.phro.2024.100572

**Publication date**
2024

**Document Version**
Final published version

**Published in**
Physics and Imaging in Radiation Oncology

Physics and Imaging in Radiation Oncology 30 (2024) 100572

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

# 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 F. 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

ARTICLE INFO

*Keywords:*
Automated plan optimization
Auto contouring
Dose impact
Robot process automation
Automated plans

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.

## 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 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

* Corresponding author at: Division of Image Processing (LKEB), Department of Radiology, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands.
*E-mail address:* p.p.mody@lumc.nl (P. Mody).
*URL:* https://www.lkeb.nl (P. Mody).

https://doi.org/10.1016/j.phro.2024.100572
Received 26 November 2023; Received in revised form 21 March 2024; Accepted 21 March 2024
Available online 28 March 2024

[10,6], 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 ($\leqslant$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 [13,12,15], 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., 7000 cGy to the primary tumor, while others were prescribed 6600 cGy 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 5425 cGy/7000 cGy in 35 fractions. For post-operative patients, our clinic prescribed 5280 cGy/6600 cGy 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 $V_{95\%} \geqslant 98\%$ for DL1/DL2 and $D_{2\%} \leqslant 107\%$ for DL2 of the Clinical Target Volume (CTV) in a 21-scenario robust optimization with 3 mm 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 $V_{94\%} \geqslant 98\%$ [22] and voxel-wise maximum (vw-max) of $D_{2\%} \leqslant 107\%$.

### 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.

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 Fig. 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 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 3 mm 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:

$$\Delta D_x = D_{x,p1} - D_{x,p2}. \quad (1)$$

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:

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

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

For the above $\Delta D_x$ (dose) and $\Delta\text{NTCP}_d$ values, we performed a Wilcoxon signed-rank test ($p \leqslant 0.05$ is considered a significant difference) to evaluate if the differences between plans are significant.

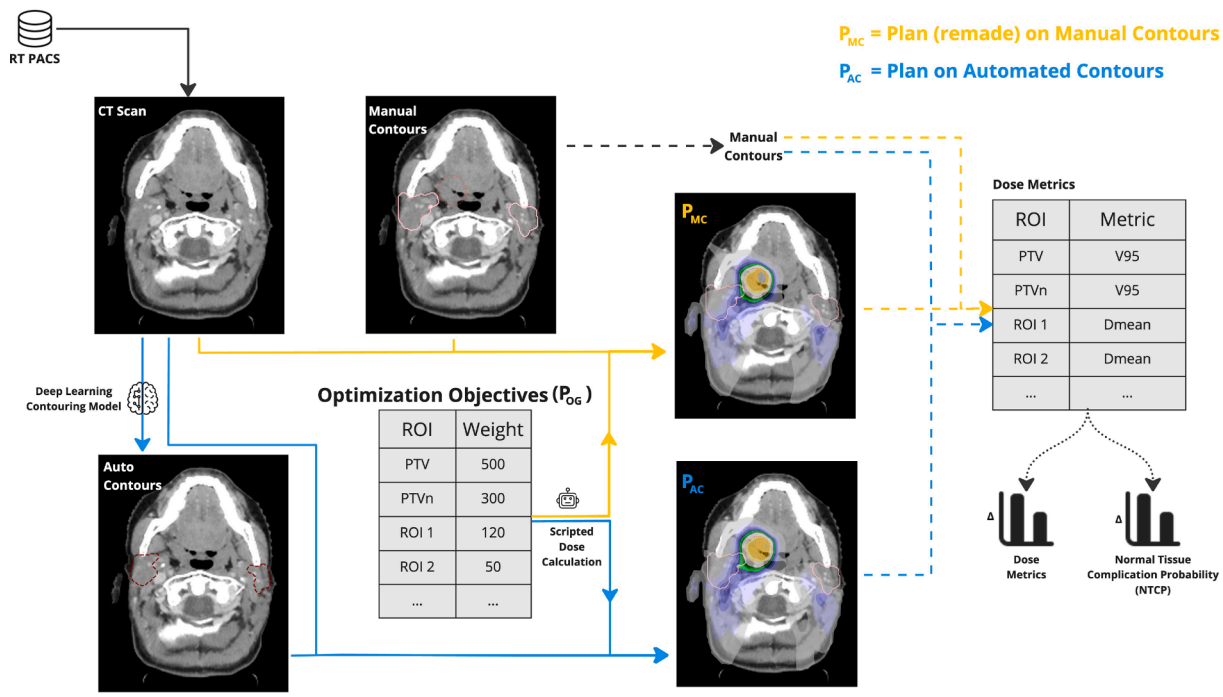



**Fig. 1.** Workflow for automated plan optimization and use-case of evaluating the effect of automated contours on dose. By reusing original plan ($P_{OG}$) parameters, we made a plan for both the manual contours ($P_{MC}$) and automated contours($P_{AC}$), shown with yellow and blue colors respectively. 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.

## 3. Results

### 3.1. Geometric evaluation

Fig. 2 shows five organs (Spinal Cord, Parotids, Submandibulars, Oral Cavity, Mandible) had a median DICE ⩾0.78 (with additional summary measures tabulated in Supplementary Material B). In Fig. 2b we observed that in general the surface DICE values for the OARs are higher than their DICE values, except for the oral cavity. Fig. 2c and Fig. 2d shows that HD95 and MSD had trends similar to DICE in Fig. 2a. OARs with a median DICE ⩾0.8 had their median HD95 less than 7.7 mm and their median MSD less than 2.6 mm. 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 0.27 Gy (1.0%), 1.66 Gy (4.6%) and 0.21 Gy (0.7%) for all, central nervous system (CNS), i.e., Brainstem and

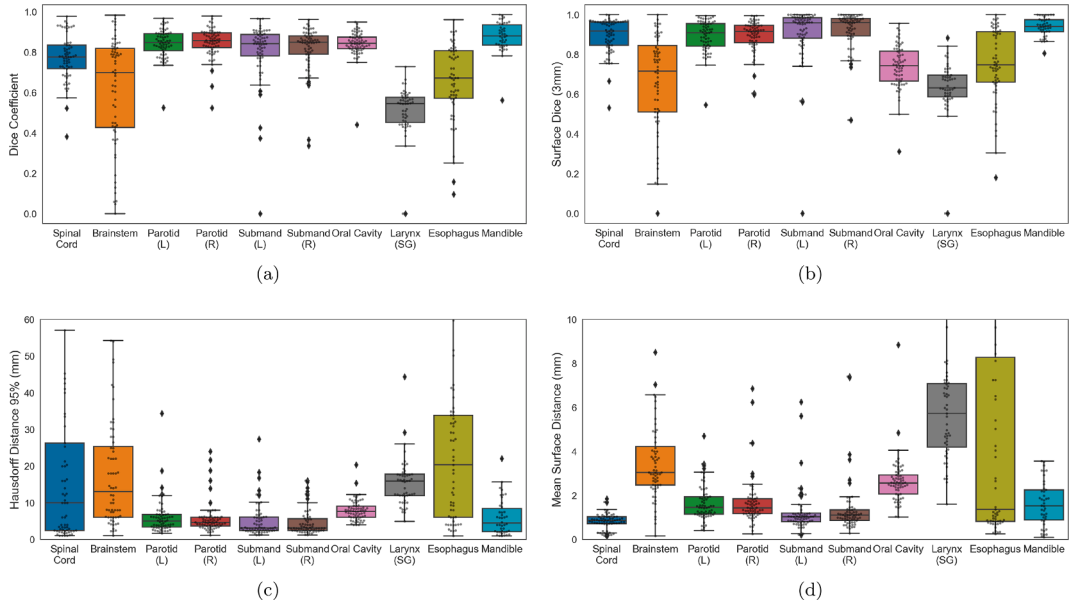



**Fig. 2.** Box plots showing geometric (a) and surface metrics (b–d) for all our patients. The scatter points indicate the metric values for each patient.

Spinal Cord and non-CNS organs, respectively. The same for $P_{MC}$ - $P_{AC}$ (automated plan using auto-contours) was 0.58 Gy (2.0%), 1.86 Gy (5.4%) and 0.46 Gy (1.6%), with metrics of individual organs in Fig. 3a listed in Supplementary Material D. Fig. 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, Fig. 3c shows |ΔNTCP| 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 0.33 Gy (1.5%), 1.13 Gy (11.5%) and 0.22 Gy (0.8%) for all, CNS and non-CNS organs, respectively. The same for $P_{MC} - P_{AC}$ was 0.48 Gy (2.6%), 0.75 Gy (6.9%) and 0.38 Gy (1.8%). Fig. 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| (Fig. 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 ($|\rho_s| \leqslant 0.11$), across both photon and proton (Fig. 5). Conversely, the Parotids, Submandibulars and Oral Cavity had relatively higher values ($-0.43 \leqslant \rho_s \leqslant -0.17$). The remaining organs did not have similar correlations across both radiotherapy treatments.

Finally, our automated plan optimization took 45 min and 2.5 h of computer time, compared to 3 and 6 h 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 ($P_{OG}$)

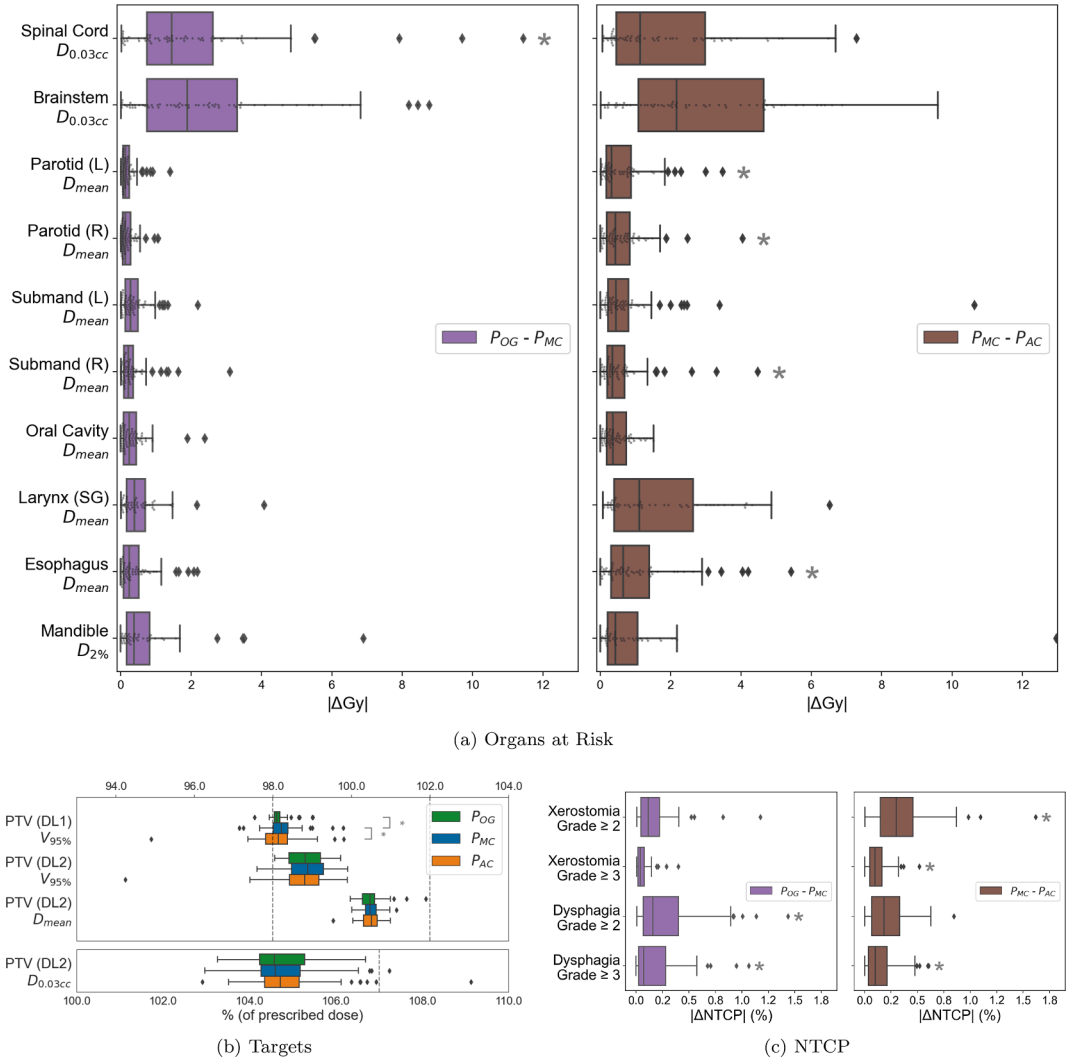



**Fig. 3.** Dose metrics for the original (i.e., clinical) photon plans ($P_{OG}$) as well as plans (re) made on manual ($P_{MC}$) and automated ($P_{AC}$) contours using an automated program. $P_{OG} - P_{MC}$ shows the dose effect of the proposed planning process, while $P_{MC} - P_{AC}$ shows the 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.

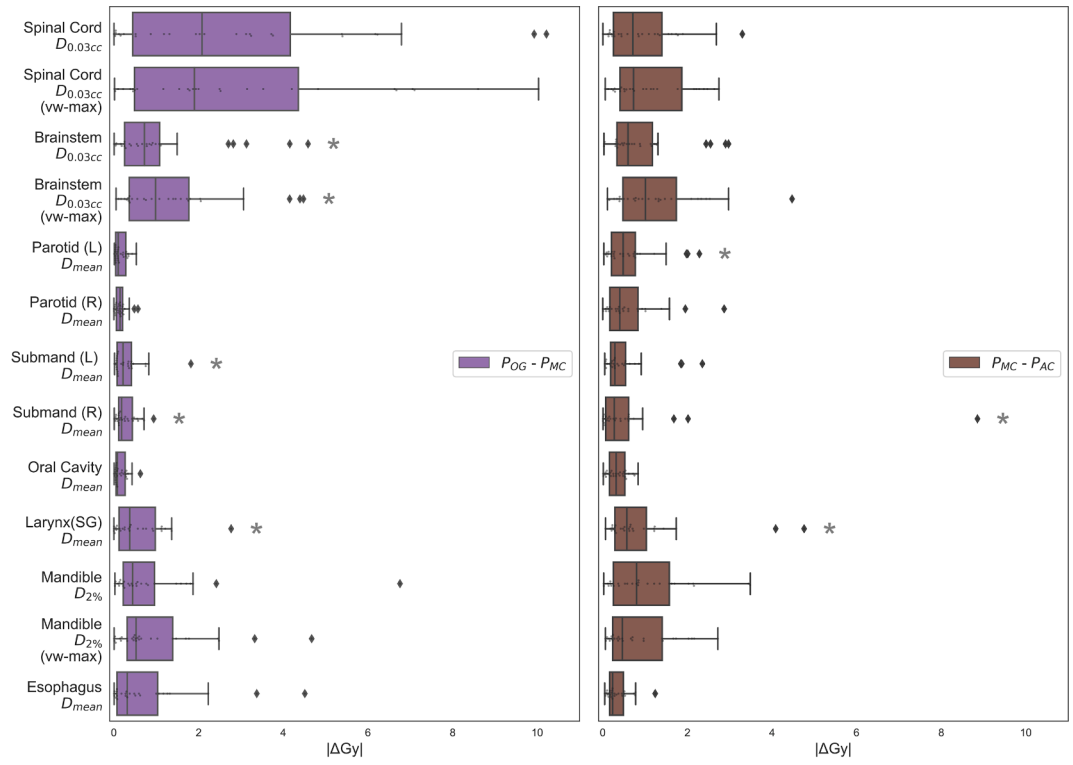



**Fig. 4.** Dose metrics for the original proton plans ($P_{OG}$) as well as plans (re) made on manual ($P_{MC}$) and automated ($P_{AC}$) contours using an automated program. $P_{OG} - P_{MC}$ shows the dose effect of the proposed planning process, while $P_{MC} - P_{AC}$ shows the effect of using auto-contours. Here $*$ represents a p-value $\leqslant 0.05$. 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.

[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 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 [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) (Fig. 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 Fig. 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 Fig. 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

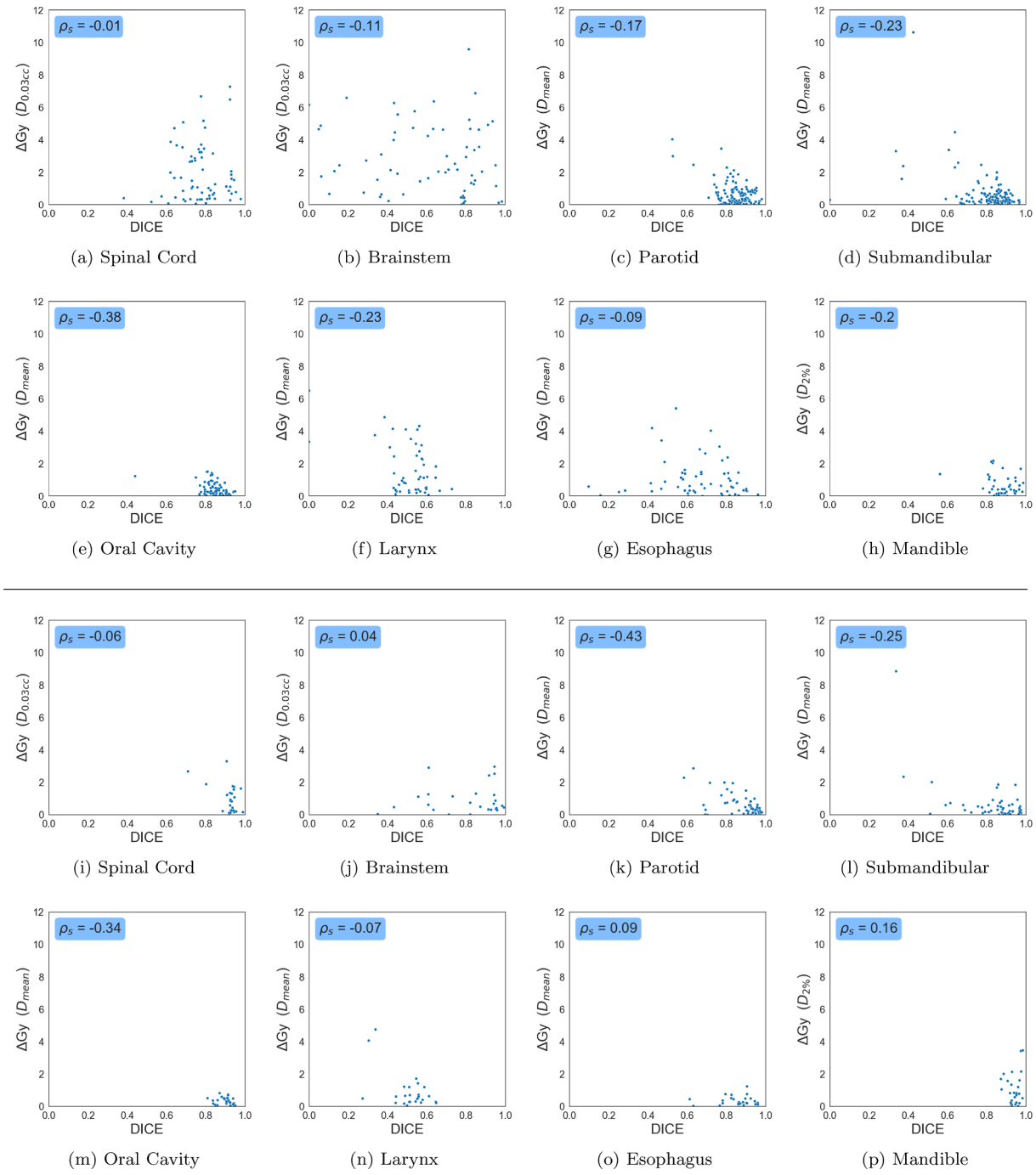
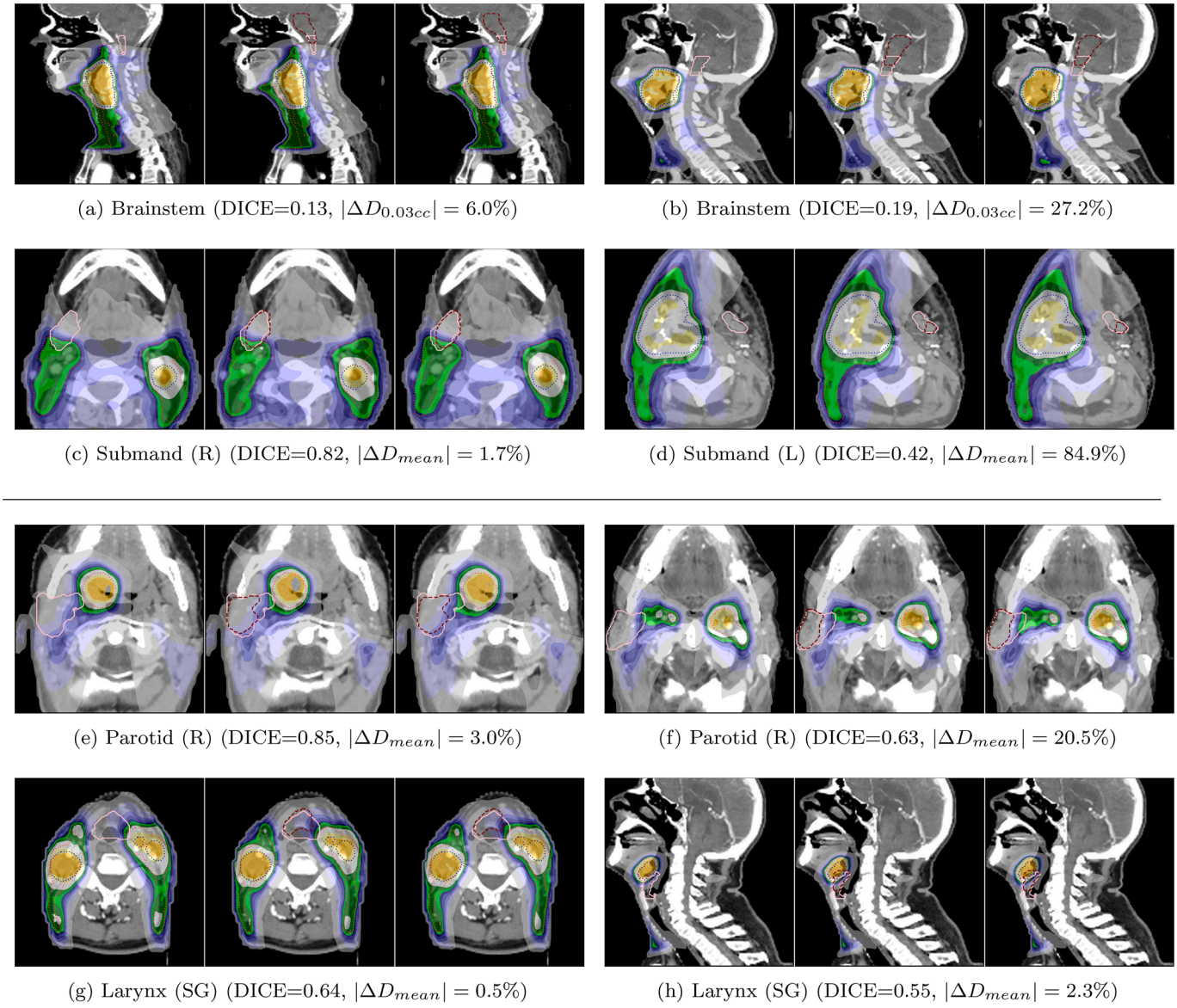



**Fig. 5.** Scatter plots for eight organs-at-risk from the auto-contouring module. Here we plot the DICE (x-axis) against each organs absolute dose metric differences, i.e., $|P_{MC} - P_{AC}|$ (y-axis) for photon (a–h) and proton (i–p) radiotherapy.

scenario, it is already sufficient to know that plans made on auto-contours are equivalent to plans made on manual contours as seen in Fig. 3b (photon) and Fig. 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.

Fig. 4b shows that most proton plans, including $P_{OG}$, tended to have hotspots, i.e., $D_{2\%}(vw-max) \geqslant 107\%$, unlike most photon plans which did not, i.e., $D_{0.03cc} \leqslant 107\%$ (Fig. 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,



(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, $|\Delta D_{mean}| = 2.3\%$)

**Fig. 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 $P_{MC} - P_{AC}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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.

Fig. 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 $\Delta D_{0.03cc}$ metric is inherently more sensitive to dose changes than $\Delta D_{mean}$. This is seen in the first row of Fig. 6 where similar DICE values for the Brainstem output vastly different dose differences. For proton (Fig. 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.

Fig. 3a, 3c (photon) and Fig. 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 (Fig. 5), one would expect that a decrease in DICE would lead to higher ΔcGy values for organs. This was true for the Parotids, Submandibulars (Fig. 6) and Oral Cavity across both photons and protons ($-0.43 \leqslant \rho_s \leqslant -0.17$). The Brainstem and Spinal Cord showed poor correlation scores for both forms of radiotherapy, primarily due to the sensitive nature of the $D_{0.03cc}$ 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 [24,23] 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.

## CRediT authorship contribution statement

**Prerak Mody:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Merle Huiskes:** Methodology, Writing – review & editing. **Nicolas F. Chaves-de-Plaza:** Writing – review & editing. **Alice Onderwater:** Methodology. **Rense Lamsma:** Methodology, Writing – review & editing. **Klaus Hildebrandt:** Writing – review & editing. **Nienke Hoekstra:** Conceptualization, Methodology, Writing – review & editing. **Eleftheria Astreinidou:** Conceptualization, Methodology, Writing – review & editing. **Marius Staring:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Frank Dankers:** Conceptualization, Methodology, Software, Data curation, Writing – review & editing, Supervision.

## 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.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.phro.2024.100572.

## 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. 111–31.
[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;7:1–10. 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;144:152–8. 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;160:185–91. 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. https://doi.org/10.1088/1361-6560/aacb65. 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;104:677–84. 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 autocontouring 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;142:115–23. 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;141:267–74. 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;14:176–91. 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;193:656–65. 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;21:122–31. https://doi.org/10.1002/acm2.13017.
[26] Aalst WMPVD, Bichler M, Heinzl A. Robotic process automation. Business Inf. Syst. Eng. 2018;60:269–72. 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;24:103–10. 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. J Appl Clin Med Phys 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;8:354–65. https://doi.org/10.14338/IJPT-20-00089.1.