Implementation of delineation error detection systems in time-critical radiotherapy

Do AI-supported optimization and human preferences meet?

Chaves-de-Plaza, Nicolas F.; Mody, Prerak; Hildebrandt, Klaus; Staring, Marius; Astreinidou, Eleftheria; de Ridder, Mischa; de Ridder, Huib; Vilanova, Anna; van Egmond, René

DOI
10.1007/s10111-024-00784-4

Publication date
2024

Document Version
Final published version

Published in
Cognition, Technology and Work

Cognition, Technology & Work
https://doi.org/10.1007/s10111-024-00784-4

RESEARCH

Implementation of delineation error detection systems in time-critical radiotherapy: Do AI-supported optimization and human preferences meet?

Nicolas F. Chaves-de-Plaza[1,7] · Prerak Mody[3,7] · Klaus Hildebrandt[1] · Marius Staring[3,4] · Eleftheria Astreinidou[4] · Mischa de Ridder[5] · Huib de Ridder[2] · Anna Vilanova[6] · René van Egmond[2]

Received: 27 October 2023 / Accepted: 4 November 2024

Abstract
Artificial Intelligence (AI)-based auto-delineation technologies rapidly delineate multiple structures of interest like organs-at-risk and tumors in 3D medical images, reducing personnel load and facilitating time-critical therapies. Despite its accuracy, the AI may produce flawed delineations, requiring clinician attention. Quality assessment (QA) of these delineations is laborious and demanding. Delineation error detection systems (DEDS) aim to aid QA, yet questions linger about potential challenges to their adoption and time-saving potential. To address these queries, we first conducted a user study with two clinicians from Holland Proton Therapy Center, a Dutch cancer treatment center. Based on the study's findings about the clinicians' error detection workflows with and without DEDS assistance, we developed a simulation model of the QA process, which we used to assess different error detection workflows on a retrospective cohort of 42 head and neck cancer patients. Results suggest possible time savings, provided the per-slice analysis time stays close to the current baseline and trading-off delineation quality is acceptable. Our findings encourage the development of user-centric delineation error detection systems and provide a new way to model and evaluate these systems' potential clinical value.

Keywords Auto-delineation · Quality assessment · Process optimization · Information integration · Radiotherapy center · Time pressure

1 Introduction

External beam radiotherapy (EBRT) is a widely used cancer treatment that relies on the precise delineation of tumors and organs-at-risk (OARs) to optimize radiation dose delivery. Manual delineation is laborious and time-consuming, hindering the adoption of time-sensitive therapies like adaptive proton therapy (Albertini et al. 2020; Sonke et al. 2019; Castadot et al. 2010). AI technologies such as deep learning-based auto-delineation can swiftly generate delineations from CT or MRI scans, reducing clinician workload and enhancing consistency (Nikolov et al. 2021; Cardenas et al. 2019; Sonke et al. 2019). However, AI-generated delineations often contain inaccuracies requiring quality assessment (QA) by clinicians (Vandewinckele et al. 2020).

As Fig. 1 illustrates, the QA process involves clinicians navigating auto-delineated image slices to identify and correct errors, a particularly demanding task for anatomically complex regions like the head and neck. Recently, delineation error detection systems (DEDS) have been proposed

✉ Nicolas F. Chaves-de-Plaza
n.f.chavesdeplaza@tudelft.nl

1 Computer Graphics and Visualization Lab, TU Delft, Delft, The Netherlands

2 Industrial Design Engineering, TU Delft, Delft, The Netherlands

3 Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

4 Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands

5 Department of Radiotherapy, University Medical Center Utrecht, Utrecht, The Netherlands

6 Department of Mathematics and Computer Science, TU Eindhoven, Eindhoven, The Netherlands

7 HollandPTC, Delft, The Netherlands

Published online: 26 November 2024

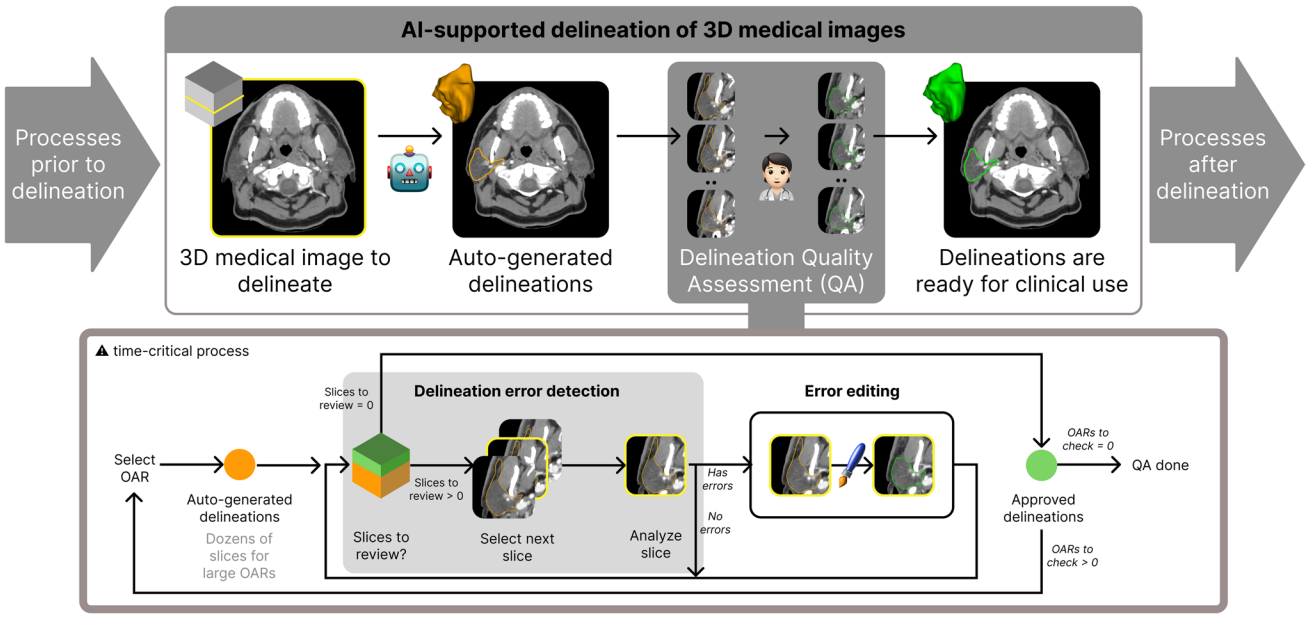

Fig. 1 Overview of the AI-infused delineation workflow. The input is a set of 3D image volumes to delineate, a computerized tomography (CT) in the example. After generating the initial delineations with the AI, the clinician proceeds to perform a quality assessment (QA). The process has two tasks that alternate until there are no more errors: delineation error detection and editing

to streamline QA by highlighting areas likely to contain errors (Sander et al. 2020; Zhou et al. 2023; Roberfroid et al. 2024). While these technologies promise to reduce QA time, their clinical implementation and impact on workflow efficiency remain underexplored.

This study aims to advance the clinical applicability of DEDS by addressing questions about the suitability of the DEDS workflow and its potential to expedite the QA process. We employed a mixed methods approach, starting with an observational user study involving a radiotherapy technologist and a radiation oncologist from Holland Proton Therapy Center (HollandPTC) to refine the DEDS workflow and validate several information sources for error detection and prioritization. This was followed by a simulation study that assessed the time-saving potential of various DEDS workflows across a diverse patient cohort with varying anatomies and error patterns.

The user study revealed a preference among the two clinicians for prioritizing errors based on clinical metrics, such as dose, over other forms of assistance with which they are less familiar. Further, DEDS assistance proved cumbersome, with the two clinicians expressing fatigue and confusion about the suggested slice orderings. These obstacles prompted the radiotherapy technologist to partially revert to a sequential slice-by-slice approach when navigating three-dimensional image volumes. Simulation results indicate that DEDS can improve the QA time-quality trade-off, although further refinement is needed for integration into clinical practice. This work sets a benchmark for DEDS evaluation and provides a simulation model that can be used to assess different error detection strategies.

2 Related work

Existing literature on user evaluation of radiotherapy software and workflows focuses on treatment planning process steps like delineation (Kalpathy-Cramer et al. 2014; Steenbakkers et al. 2005, 2006) and dose optimization (Mazur et al. 2014, 2013). Particular to the case of delineation, research has focused on understanding the delineation workflow (Aselmaa et al. 2017); and investigating the effect of alternative image modalities (Steenbakkers et al. 2006) and delineation uncertainty (Maruccio et al. 2024), and usability of semi-automatic editing tools (Aselmaa et al. 2017; Ramkumar et al. 2016, 2017). Recently introduced deep neural networks (DNNs) generating delineations of hundreds of OARs at once (Nikolov et al. 2021; Cardenas et al. 2019) prompt clinics to create clinician-centric delineation quality assessment (QA) processes to identify and rectify DNNs inaccuracies (Vandewinckele et al. 2020).

This paper focuses on the delineation error detection QA subprocess. Delineation error detection systems (DEDS) can identify errors at various levels, from voxels to anatomical structures (Altman et al. 2015; Hui et al. 2018; Rhee et al. 2019; Sandfort et al. 2021; Mody et al. 2022a). DEDS accelerate QA by directing attention to errors, reducing unnecessary scrutiny of clinically-acceptable delineations. For

instance, some DEDS employ AIs to predict errors within slices based on auto-generated delineations and their uncertainty (Sander et al. 2020). Recent developments even suggest a DEDS module that actively directs clinicians to the next slice for review based on predicted error extent (Zhou et al. 2023) or predicted dosimetric impact (Roberfroid et al. 2024). Despite advances in DEDS, their clinical implementation and associated user experience challenges remain largely unaddressed issues.

In adaptive radiotherapy, clinicians prioritize areas based on dose distribution and patient malignancies (Chaves-de-Plaza et al. 2022). Various studies explore the dosimetric impact of delineation errors (Guo et al. 2021; Mövik et al. 2023; van Rooij et al. 2019). Recent work introduces a DEDS that utilizes deformations of auto-generated delineations and dose prediction technologies to identify dosimetrically relevant areas for inspection (Roberfroid et al. 2024). We incorporate dose as a clinically relevant priority measure and discuss alternatives with the two clinicians in the study when dose information is unavailable.

3 Materials and methods

We used imaging data associated with a retrospective cohort of 42 head and neck cancer patients treated at Holland Proton Therapy Center (HollandPTC) between 2018 and 2020. The study from which the patient data was taken received IRB approval from Holland Proton Therapy Center (HollandPTC), and all patients provided informed consent. Data from three patients were employed for the user study and the complete cohort for the simulation study.

Figure 2 presents an overview of the different types of three-dimensional images available per patient plus the additional ones we derived, like AI delineations and their uncertainty. In the remainder of this paper, we distinguish three-dimensional images, or image volumes, using a `monospace` font. Unless stated otherwise, operations on pairs of volumes are applied voxel-wise, yielding a new volume (i.e., $\texttt{vol3} = \texttt{vol1} + \texttt{vol2}$). We use subscripts on the volume to index slices or voxels, which we specify in the text. For instance, $\texttt{vol}_s$ in the figure refers to the s^{th} 2D axial slice of `vol`.

3.1 Imaging data

The top section of Fig. 2 displays slices of the patient's CT scan (`image`) and organ-at-risk (OAR) delineations ($\texttt{del}^*$) used for the original treatment planning. We define $\texttt{del}^*$ as delineation ground truth in our studies. In the user study, participants did not have access to $\texttt{del}^*$ while performing the error detection tasks. $\texttt{del}^*(\texttt{OAR})$ represents the delineation of a specific OAR, which is a binary image with ones where the OAR lies and zeros otherwise. `image` and $\texttt{del}^*$ have width, height and slice dimensions of sizes $512 \times 512 \times 195$ voxels and spacing of $0.98 \times 0.98 \times 2$ mm.

Each patient file also included the treatment `dose` distribution volume, representing radiation deposition in space. In Fig. 2, brighter yellow and darker purple colors mean higher and lower dose values, respectively. We resampled the `dose` to match the dimension sizes of `image` and $\texttt{del}^*$. We include the `dose` in our studies because the participants have an adaptive radiotherapy background, where the dose is used as a heuristic to determine which slices need more attention (Roberfroid et al. 2024). In certain situations, metrics such as the distance to the target volumes may be more appropriate than the dose. Deciding to prioritize one over the other would necessitate rearranging the slices and consequently altering the workflow, which constitutes the primary focus of our paper.

For preprocessing, we cropped all three volumes using a bounding box centered at the brain stem with dimensions $240 \times 240 \times 80$ voxel and spacing of $0.8 \times 0.8 \times 2.5$ mm. Linear interpolation was applied to `image` and `dose`, while nearest-neighbor interpolation was used for $\texttt{del}^*$. These preprocessing steps aligned the data with the input format expected by the AI.

3.2 AI delineations, uncertainty and error

We fed the patient's `image` in the HollandPTC dataset to a pre-trained state-of-the-art Bayesian deep neural network (the AI in this work), to generate ten candidate delineations for each input image. For this, we used the FlipOut model described in Mody et al. (2022b), which is based on the FocusNet architecture, employing a modified cross-entropy loss. The model generates delineation candidates by running ten times, each with a different set of weights sampled from a learned distribution. The network was trained on a subset of 33 patients of the MICCAI2015 head and neck dataset (Raudaschl et al. 2017). For each patient, there are delineations for nine OARs of which we used six: BrainStem, Mandible, parotid glands (Parotid_L and Parotid_R), and submandibular glands (Submand_L and Submand_R). We refer the reader to the original publication for more details about the network architecture and training.

Each AI-generated candidate $\texttt{cdel}^i$ with $i \in \{1, \ldots, 10\}$ is a label map volume, with each voxel having the ID of the OAR it belongs to (or zero if background). To aggregate the candidates into the predicted delineation `del`, we computed the voxel-wise median label:

$$\texttt{del} = \mathbb{M}(\texttt{cdel}^1, \ldots, \texttt{cdel}^{10}), \tag{1}$$

where $\mathbb{M}$ denotes the voxel-wise median function. `del` is also a label map with the same dimensions and spacing as `image`. To obtain an OAR's predicted segmentation $\texttt{del}(\texttt{OAR})$, it suffices to set voxels matching match a given OAR ID to

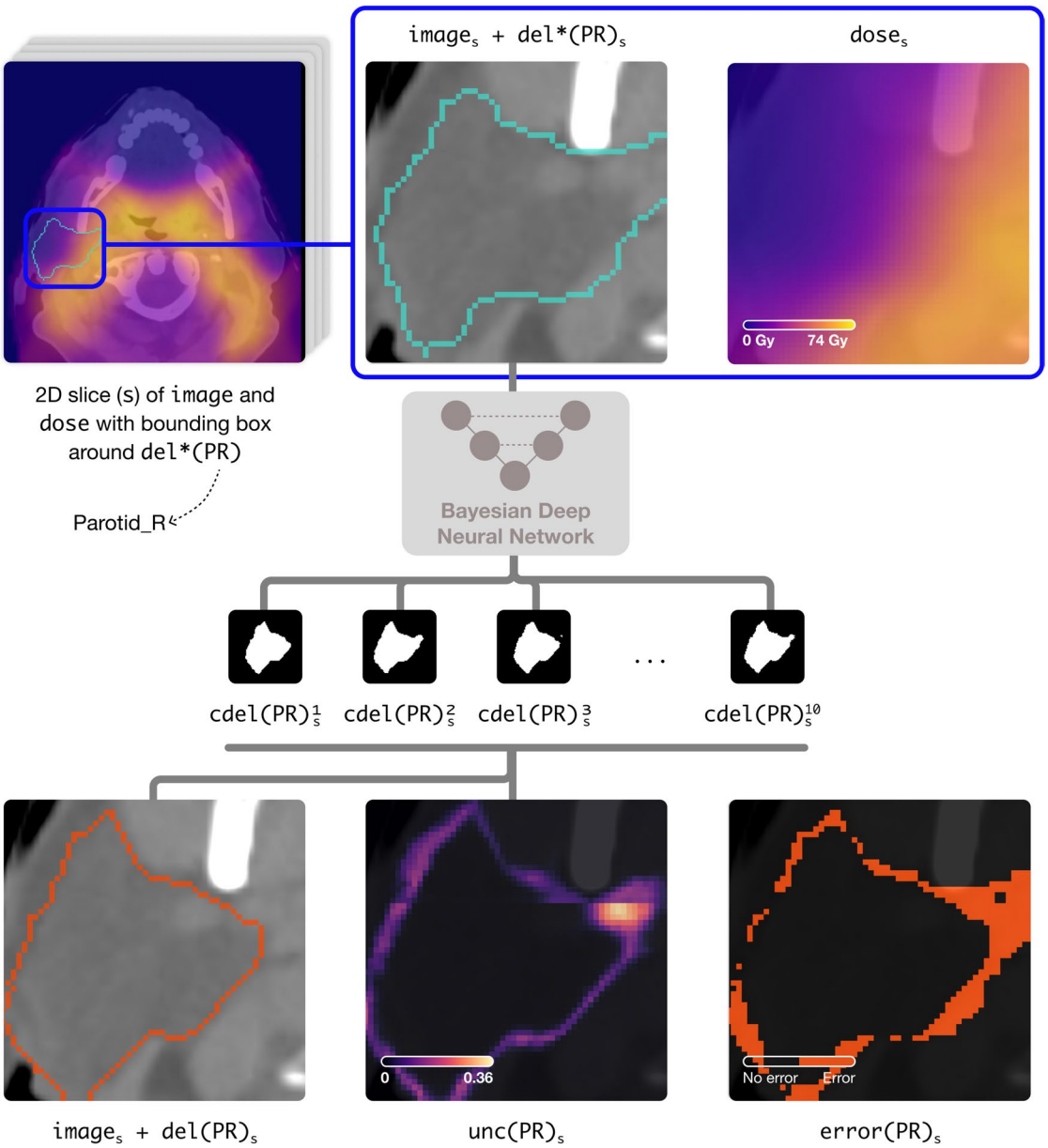

Fig. 2 Example of the information sources used in this paper for one of the patients in the HollandPTC dataset. The top row depicts a slice of the image and dose of the Parotid_R. We used a Bayesian Deep Neural Network to obtain ten delineation candidates based on the image. The bottom row depicts the information sources we derived based on these candidates

one and the rest to zero. Note that the median operation can be thought of as performing a voxel-wise majority vote on the OAR IDs.

From the candidate delineations, we also calculated the AI's uncertainty unc per OAR as the voxel-wise standard deviation of the OAR's candidates:

$$\texttt{unc}(\texttt{OAR}) = \sqrt{\frac{\sum_{i=1}^{10}(\texttt{cdel}(\texttt{OAR})^i - \bar{\mu}(\texttt{OAR}))^2}{9}}, \tag{2}$$

where $\texttt{cdel}(\texttt{OAR})^i$ represents the binary image of the OAR's i^{th} delineation candidate and $\bar{\mu}(\texttt{OAR})$ the mean delineation for a specific OAR.

As the sample unc slice in Fig. 2 illustrates, the computed uncertainty exhibits higher values (brighter spots) in image regions with challenging delineation, such as those lacking inter-tissue contrast. We prefer AI uncertainty over previous hand-engineered feature-based methods because it is readily available from the Bayesian network, requiring less domain-specific knowledge, and is correlated with

delineation errors (Sander et al. 2020; Mody et al. 2022b). Therefore, in our studies we adopt unc as a proxy for delineation errors' location and extent.

The final information source we consider is the delineation error error, calculated as

$$\texttt{error}(\texttt{OAR}) = |\texttt{del}^*(OAR) - \texttt{del}(\texttt{OAR})|, \tag{3}$$

where $|\cdot|$ is the voxel-wise absolute value function. error(OAR) highlights areas where AI predictions and HollandPTC's delineations disagree. Note we do not differentiate between under and over-segmentation errors. Being an error proxy, unc can suffer from false positives and negatives. In the studies, we use error to provide an upper bound to the performance gains, assuming an optimal error detector. Finally, in the user study, we use error as an additional information source to elicit discussion, allowing participants to contrast it with unc.

3.3 Per-slice scores

To enable priority sorting in the DEDS-assisted workflow, for an OAR we compute per-slice scores based on the unc, dose, and error. Computing the priority scores $\boldsymbol{p}(\texttt{OAR})$ of an OAR's slices entails applying an aggregation function to each slice of the OAR and collecting the values in an array:

$$\begin{aligned}\boldsymbol{p}(\texttt{OAR}) = \{&\texttt{agg}(\texttt{vol}(\texttt{OAR})_{s=1})\texttt{agg}(\texttt{vol}(\texttt{OAR})_{s=2}), \dots, \\ &\texttt{agg}(\texttt{vol}(\texttt{OAR})_{s=S})\},\end{aligned} \tag{4}$$

where agg(·) takes as input a set of voxels (in this case those in an axial slice s) and outputs a number. For instance, to obtain the mean uncertainty score, we set $\texttt{vol}(\texttt{OAR}) = \texttt{unc}(\texttt{OAR})$ and agg = mean. We only consider voxels within $\texttt{del}^*(\texttt{OAR})$'s bounding box to avoid assigning scores to unrelated parts of the volume, like slices above and below the OAR. The assumption of correct bounding boxes before QA is not unreasonable, as inspecting and rectifying OARs' bounding boxes is an easy task that could be performed beforehand. In the user study, we considered the minimum (min), maximum (max), mean, and sum aggregation functions to enable discussion. In the simulation study, we focused on the most relevant ones from the user study.

4 User study: workflow comparison

We conducted a two-part user study to investigate clinicians' current (part 1) and DEDS-assisted (part 2) workflows. In the following, we describe the study setup and then present and discuss the main findings, which inform the simulation study in the next section.

4.1 Study setup

4.1.1 Participants

A radiation oncologist (RO) and a radiotherapy technologist (RTT) from Holland Proton Therapy Center (HollandPTC), specialized in the head-and-neck area participated in our study. Both participants have several years of experience and perform delineation tasks routinely. TU Delft's IRB approved this research, and each clinician provided informed consent to be part of the study.

4.1.2 Apparatus

The clinicians utilized the DEDS depicted in Fig. 3. We developed the custom DEDS software based on several sessions with two clinicians from Leiden University Medical Center and University Medical Center Utrecht. The design process is detailed in Appendix A. The DEDS incorporates functionality from standard delineation software like the list of OAR to review and a slice-based image viewer that allows inspecting the image volumes with interactions such as navigation, zooming, and panning. This functionality enables traditional error detection workflows. Additionally, as detailed next, the DEDS software implements functionality that permits clinicians to define and execute priority-based workflows.

A more detailed slice-level OAR explorer (slice explorer) allowed participants to inspect OARs' slices and sort them based on a priority score

$$\begin{aligned}\boldsymbol{wp}(\texttt{OAR})_s = &w_1\texttt{agg}_1(\texttt{unc}(\texttt{OAR})_s) + w_2\texttt{agg}_2(\texttt{dose}(\texttt{OAR})_s) \\ &+ w_3\texttt{agg}_3(\texttt{error}(\texttt{OAR})_s),\end{aligned} \tag{5}$$

defined as weighted combination of unc, dose, and error scores. w_i represents weights, normalized to sum to one, and $\texttt{agg}_i$ denotes aggregation functions. Participants selected their preferred aggregation functions and assigned them weights before starting part 2 of the study using the form in the score definition area of the DEDS' GUI in Fig. 3a. We allowed participants to define the priority score to elicit discussion about the relevance of different information sources and aggregation functions.

Although unc can be used as an error proxy, it is not the only option. For instance, the approach of Sander et al. (2020) directly flags errors at the patch level. To facilitate richer discussions, we decided to permit participants to use the error and told them it was computed by an automatic method to prevent overreliance. Participants could overlay the volumes used for the score computation on the image viewer for closer inspection. A panel to the right of the image viewer (contextual information) provided details

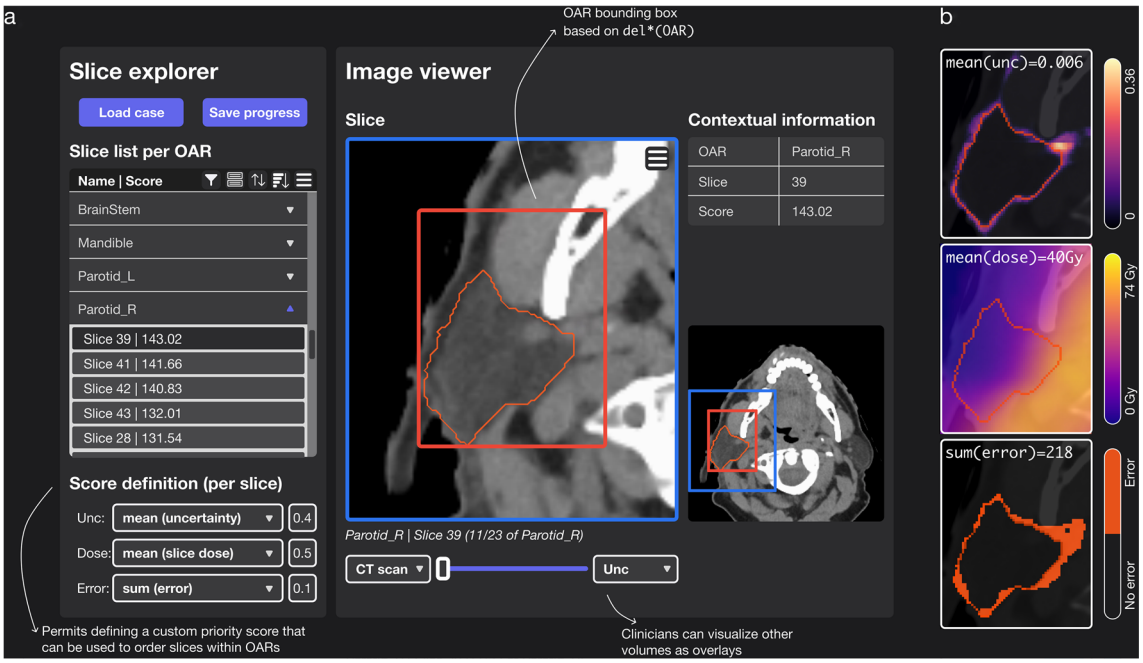

Fig. 3 Custom DEDS software used in the study. **a** Shows the graphical user interface. The main areas are the slice explorer and the image viewer. Using the score definition box, clinicians can define a slice ordering per OAR based on uncertainty, dose, and error information sources. **b** Shows the available information sources for the currently displayed OAR (slice 11 of Parotid_R). It also presents the per-slice value obtained with the user-defined aggregation functions

about the current slice, its score, and its location within the image. Figure 3b presents an example of the different information sources for slice $s = 11$ of OAR=Parotid_R.

4.1.3 Procedure

The RTT and RO participated in a three-stage, 60-min session. In the first stage, we presented the study's goal, introduced the clinicians to the DEDS, explained how to define priority scores based on weights and aggregation functions to sort OARs' slices, and let them interact with the DEDS to gain familiarity. In the second stage, the participants detected delineation errors without (part 1) and with (part 2) DEDS assistance. In part 1, participants performed their usual sequential error-finding workflow, permitting them to gain further familiarity with the tool before introducing assistance. For part 2, participants were instructed to use DEDS guidance by defining a priority score (as defined in Eq. 4) and using it to guide the order in which they visit OARs' slices. In both parts, the participants were instructed to consider OARs' priorities when deciding which to address within a 5-min time window, chosen to induce the need to prioritize delineation errors. Furthermore, OARs were shown in the same order in the graphical user interface, and participants had to complete an OAR before moving on to the next. Finally, the participants were allowed to move back and forth between adjacent slices if needed for sense-making. Because rectifying errors is time-consuming and not within the scope of this study, we asked clinicians to instead indicate per slice if they would edit it via a keyboard shortcut. After finishing each task, we used a 5-min time slot to discuss the clinicians' experience using specific slices they marked as requiring editing, and, in part 1, to define the priority score. In the last 20-min stage, we had a semi-structured discussion about participants' workflows, their choice of information sources for prioritization, and their experiences and challenges for DEDS adoption.

We used a subset (N = 3) of HollandPTC's patients' data (D1, D2 and D3). D1 was used in the familiarization stage. The RO saw data from D2 and D3 in part 1 and part 2. The RTT observed D2 twice. This was unintentional and was not noticed until the data analysis phase. Therefore, we treated these sessions as independent observations, but we acknowledge this duplication as a limitation and have taken it into account when interpreting the results. Table 1 summarizes the structures considered in the user study analysis for D2 and D3. We do not include the mandible because clinicians tend to skip it due to its low clinical significance (Jensen et al. 2020) and the clinicians' high confidence in AI auto-delineations for bony structures. Also note that the parotid glands demand the most effort, with their bounding boxes spanning more slices and containing more voxels per slice than the BrainStem and submandibular glands.

Table 1 Overview of the organs-at-risk (OARs) considered for analysis

OAR	Dataset 2 (D2)			Dataset 3 (D3)		
	Number of slices	Voxels per slice	Volume (mm^3)	Number of slices	Voxels per slice	Volume (mm^3)
BrainStem	25	1666	29,963	25	1872	36,037
Parotid_L	25	2688	35,736	26	4104	36,875
Parotid_R	**26**	**2912**	**36,646**	**24**	**4292**	**39,267**
Submand_L	18	1209	12,498	16	1015	10,410
Submand_R	17	1394	10,970	17	928	9970

The table lists, for each OAR of each dataset, the number of slices and amount of voxels per slice its bounding box spans. It also lists the volume in mm^3 of the OAR's delineation ground truth `del`*. Bold entries indicate the OAR with the largest volume within each dataset

4.1.4 Data analysis

We recorded the screen and the participant's spoken remarks in the sessions. From these, we transcribed clinicians' remarks and timestamped OAR changes, slice changes, and slices marked as "required editing". We recorded slice changes, yielding information about the order in which clinicians inspect the delineations in each condition. These interaction logs allowed us to reconstruct clinicians' workflows.

4.2 Part 1: Non-assisted workflow

The RTT and RO conducted the error-finding task as in clinical practice. Figure 4 shows the sequence of slices followed by the RTT and RO for the BrainStem (a) and Parotid_L (b). Figure 4a.1 and b.1 display the clinicians' and optimal slice change sequences using the per-slice sum of errors as the priority score. The y-axis is trimmed to slices within the bounding box of `del`*(OAR) and sorts the slices based on their 3D position within the image volume. Despite opposing starting directions, both clinicians share similar navigation behavior, following a sequential approach (unlike the optimal sequence's "jumpy" behavior), with the RTT moving from bottom to top and the RO mostly in reverse. They frequently revisited adjacent slices to verify multi-slice errors, particularly in the slice range [14, 19] of the BrainStem.

To compare the slice sequences of different workflows, we calculated the number of slice change interactions required to review slices suggested by a DEDS. A subset S of an OAR's slices consists of the $|S|$ slices exceeding the threshold. We evaluated the interactions needed for slice subsets of increasing size as the threshold decreased, including clinician workflows with redundant interactions removed and hypothetical scenarios: an optimal sequence ordered by decreasing erroneous voxels per slice, a worst-case sequence reversing the optimal, and five random permutations of the optimal sequence, with the mean and 95% confidence interval.

Figure 4a.2 and b.2 show slice change interactions as a function of suggested slice subset size for clinicians' workflows and hypothetical scenarios. The optimal workflow forms a diagonal line with a unit slope, indicating slice changes match the subset size. The worst-case scenario appears as a horizontal line since the highest error slice is reviewed last. Random samples lie between the optimal and worst-case scenarios, approaching the latter as the subset size grows, reflecting higher chances of critical slices appearing later. Clinicians' workflows generally deviate from the optimal path and often exceed the worst-case due to redundant interactions. Removing redundancy improves the RO's performance, aligning closer to or surpassing random workflows but still falling short of the optimal. The RTT's workflows remain near the worst-case, often missing critical slices early. The RO's workflows are faster than the RTT's, indicating shorter per-slice analysis times.

Table 2 compares the performance of different workflows for inspected OARs. Performance is quantified by the area under the curve relative to the optimal sequence, normalized per OAR. Scores closer to zero indicate near-optimal performance, while scores closer to one approach the worst-case scenario. Values above one reflect redundant interactions. Removing redundant visits (RTT' and RO') significantly improves scores. Trimmed RO workflows (RO') perform best, outperforming RTT and random sequences, but still deviate from the optimal, especially for the BrainStem and parotid glands, suggesting DEDS guidance could further reduce interactions and save time.

4.3 Part 2: DEDS-assisted workflows

In part 2, the RTT and RO were offered and instructed to use DEDS assistance to find slices that required attention. They started by defining a priority metric as a weighted combination of `unc`, `dose`, and `error` to sort the slices in priority order. Table 3 shows the combinations of information sources clinicians defined for different OARs. Both expressed reservations about the redundancy of uncertainty

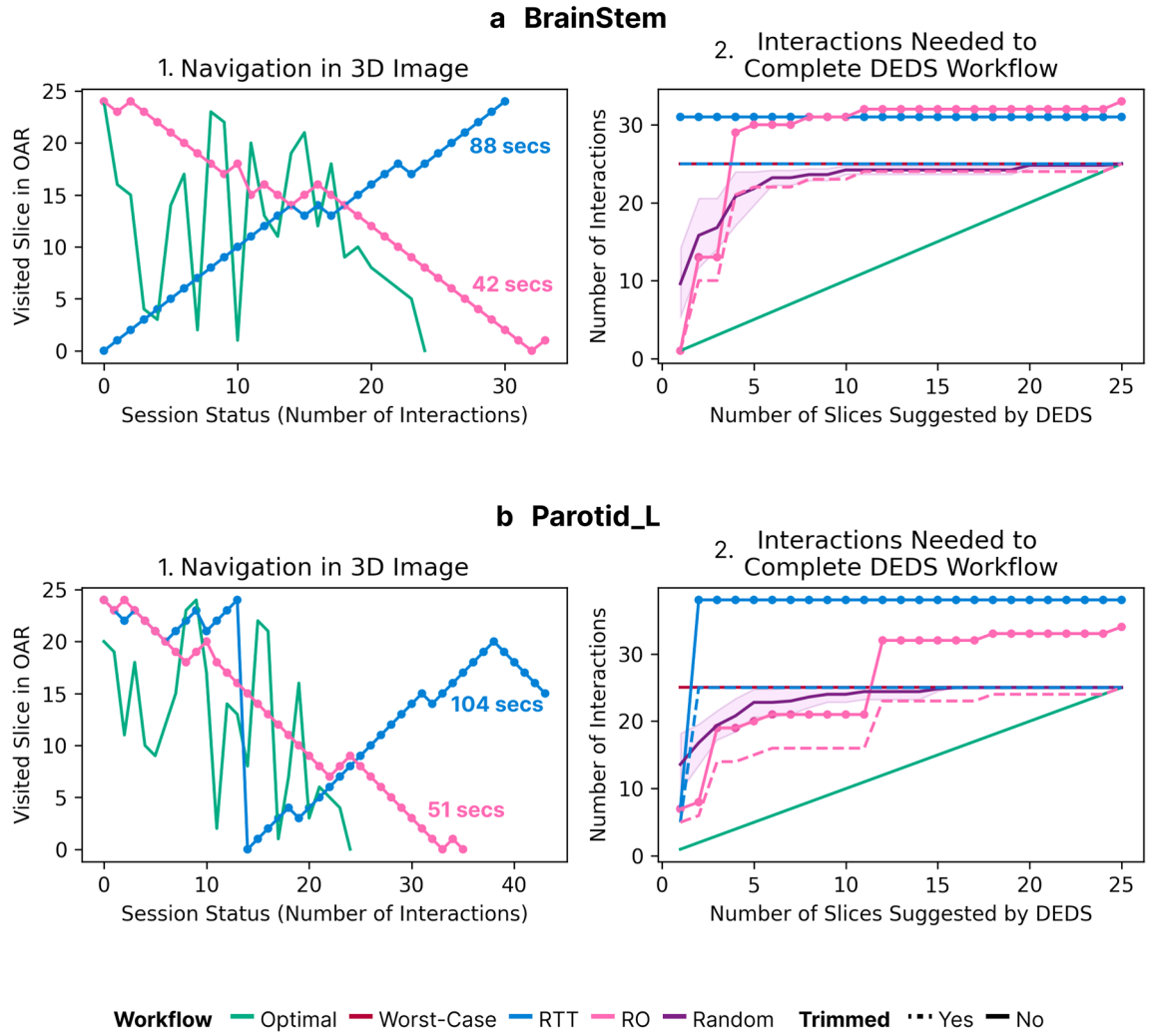

Fig. 4 Unassisted workflows for BrainStem (**a**) and Parotid_L (**b**) for the RTT and RO. **a.1** and **b.1** Depict slice changes as the session progresses, and (**a.2**) and (**b.2**) show the interactions needed to complete a DEDS-suggested workflow, encompassing subsets of OAR's slices of increasing cardinality corresponding to decreasing threshold values for the prioritization scores. We compare the observed workflows with versions in which redundant interactions have been trimmed and with several hypothetical scenarios. The purple shaded area corresponds to the 95% confidence interval of the random scenario

Table 2 Performance of various error detection workflows

OAR	RTT	RTT'	RO	RO'	Random
BrainStem	1.50	1.00	1.32	**0.71**	0.81
Parotid_L	1.98	0.93	1.10	**0.52**	0.86
Parotid_R	–	–	1.11	**0.69**	0.84
Submand_L	–	–	0.30	**0.18**	0.80
Submand_R	–	–	0.21	**0.21**	0.75

For a given workflow, its score corresponds to the difference between the areas under the workflow's and the optimal workflow's curves. The scores are normalized per OAR to provide comparable scores. The optimal and worst-case sequences have scores of zero and one, respectively. Clinicians' workflows with redundant slice visits removed are indicated by the apostrophe. Bold values highlight the smallest difference per OAR

and error and their reliability in time-sensitive scenarios. This might be why clinicians emphasized dose-based risk measures, assigning lower weights to `unc` and `error`. Information sources, aggregation functions, and weights remained generally consistent across OARs. The sole exception was the aggregation function for dose-based slice scores for the parotid glands, where the RO adjusted it to the mean.

The RTT and RO found following the priority order to be cumbersome and fatiguing, echoing the RO's view that "jumping between slices is not logical" and disrupts the 3D perception. Figure 5 illustrates this sentiment in the Parotid_R's workflow data. The RO (a) struggled with the initial sorting order provided by DEDS, leading to a reverse inspection (following ascending rather than descending priority score order), which led to a mirrored

Table 3 Settings the RTT and RO used to define the priority score for sorting the slices of the different OARs in part 2 of the user study

Source	BrainStem				Parotid_L				Parotid_R			
	RTT		RO		RTT		RO		RTT		RO	
	agg	*w*	agg	*w*	agg	*w*	agg	*w*	agg	*w*	agg	*w*
`unc`	Mean	0.50	Mean	0.25	Mean	0.50	Mean	0.25	Mean	0.50	Mean	0.25
`dose`	Max	0.50	Max	0.65	Max	0.50	Max	0.65	Max	0.50	Mean	0.65
`error`	Sum	0	Sum	0.10	Sum	0	Sum	0.10	Sum	0	Sum	0.10

agg denotes the aggregation functions and *w* the weights clinicians applied per information source and OAR

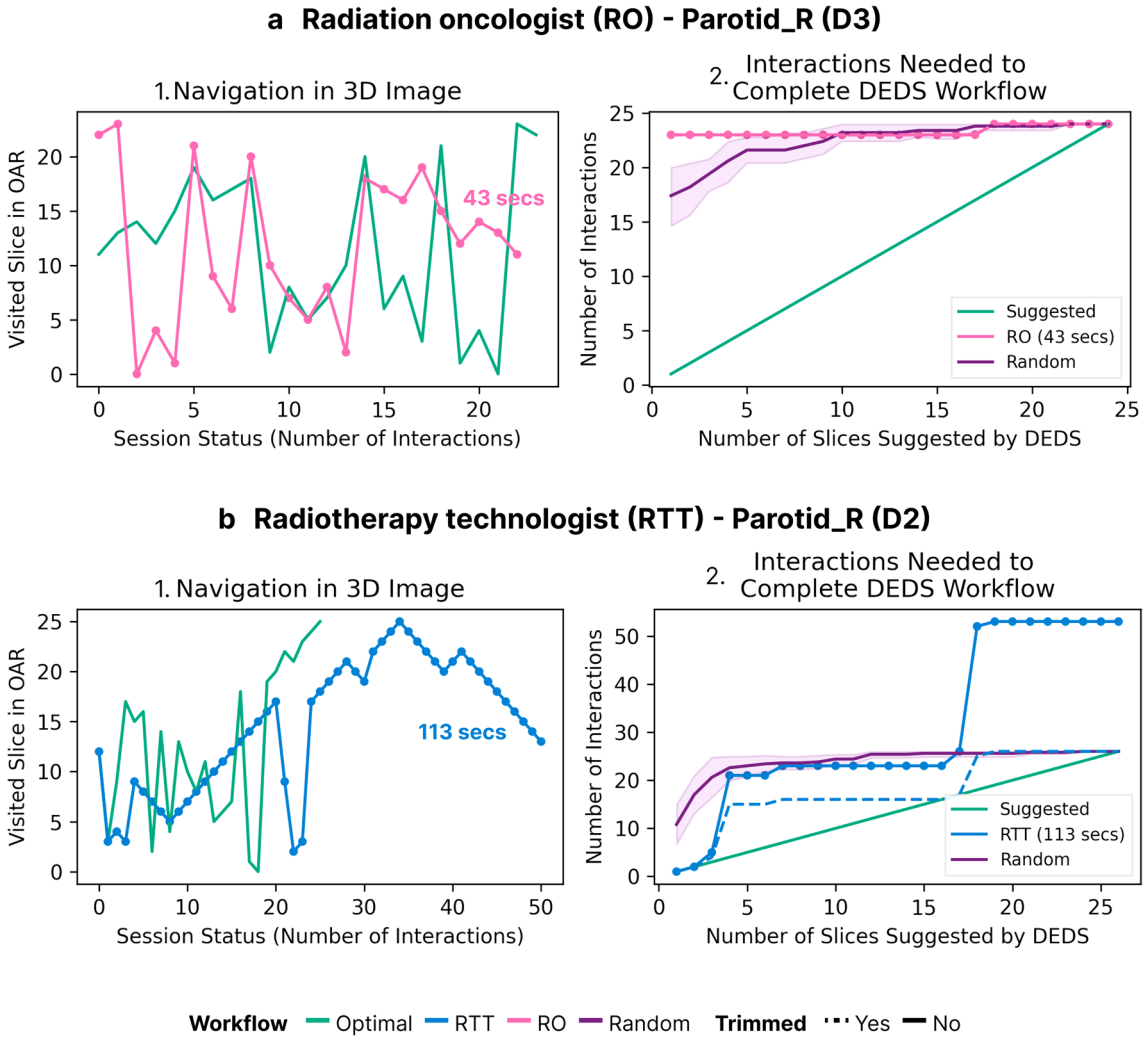

Fig. 5 Assisted workflows of the RO (**a**) and RTT (**b**) for Parotid_R. **a.1** and **b.1** depict slice changes as the session progresses, and **a.2** and **b.2** show the interactions needed to complete a DEDS-suggested workflow, encompassing subsets of OAR's slices of increasing cardinality corresponding to decreasing threshold values for the prioritization scores. We compare the observed workflows with versions in which redundant interactions have been trimmed and with several hypothetical scenarios. The purple shaded area corresponds to the 95% confidence interval of the random scenario

slice sequence as shown in (a.1). The RTT (b) intermittently followed the DEDS suggestions but often reverted to traditional navigation, as depicted in (b.1). Figure 5a.2 and b.2 show that deviations from the suggested sequence led to suboptimal performance. A similar pattern is evident in the BrainStem and parotid glands, as presented in Table 4. The trimmed RTT workflows (RTT') tend to perform better, as the RTT intermittently followed DEDS

Table 4 Performance of various error detection workflows

OAR	RTT	RTT'	RO	RO'	Random
BrainStem	0.92	0.92	0.95	0.95	**0.42**
Parotid_L	0.57	**0.39**	1.08	1.00	0.40
Parotid_R	1.39	**0.34**	0.94	0.94	0.42

For a given workflow, its score corresponds to the difference between the areas under the workflow's and the optimal workflow's curves. The scores are normalized per OAR to provide comparable scores. The optimal and worst-case sequences have scores of zero and one, respectively. Clinicians' workflows with redundant slice visits removed are indicated by the apostrophe. Bold values highlight the smallest difference per OAR

pointers, avoiding unnecessary slice visits, especially for the parotid glands.

4.4 Discussion

Part 1 investigated clinicians' error detection workflows. Both the RO and RTT followed a sequential strategy, inspecting adjacent slices. They favored such workflow because it helps them to orientate spatially, leveraging their mental representations of the OARs. Nevertheless, the comparison of clinicians' workflows with other scenarios revealed that redundant and suboptimal sequences decrease their performance. Part 2 focused on investigating clinicians' use of DEDS systems. The RTT and RO had problems accepting this approach, complaining about fatigue, losing their spatial orientation, and, in the case of the RTT, repeatedly falling back to the sequential workflow. These issues need to be solved in the future since the workflow comparison again convincingly demonstrates that DEDS can reduce the number of needed interactions, which can also impact overall spent time.

Concerning the three information sources considered, both clinicians expressed their doubts regarding the intelligibility and trustworthiness of the uncertainty and error information sources. The dose was less problematic as an information source, likely due to participants' experience in adaptive radiotherapy where heuristics like stimating the delineation error's proximity to the tumor are employed. They mentioned that the maximum dose could provide a guiding signal because false positives and negatives are problematic in slices with a max dose higher than the OAR-specific limit. We leverage this observation in the next section to develop a computational model of the DEDS workflow.

The main limitation of the user study is the very small sample size. To test the insights from the user study on a larger dataset, we performed a quantitative evaluation of the DEDS-assisted QA workflow using a simulation approach. To this end, we introduce a computational model of the complete QA workflow, including analysis and editing, which we use to investigate the viability of DEDS workflows. Specifically, we analyze the impact of varying per-slice analysis times on overall QA performance for the complete HollandPTC dataset.

5 Simulation study: assessing DEDS-induced time gains

5.1 Simulation setup

To examine the potential time savings achievable with DEDS, we compare DEDS workflows with the current unassisted clinical workflow. Figure 6 depicts a computational model of the quality assessment process (QA). In our simulation, we consider three variations of this process that arise when using different slice sequences.

In the first variation (baseline), the simulated clinician begins either at the cranial or caudal slice with an equal probability ($Pr = 0.5$) and progresses towards the opposite end (next slice step), analyzing all slices. In the second (error) and third (dose) variations, the clinician visits the slices in order of their decreasing error extent and max dose, respectively. In these DEDS-assisted workflows, the simulated clinician evaluates a slice only if it has an error (error threshold equals zero) or its max dose exceeds a preset limit l(OAR), respectively. l(OAR) is an OAR-specific limit based on constraints proposed by Jensen et al. (2020).

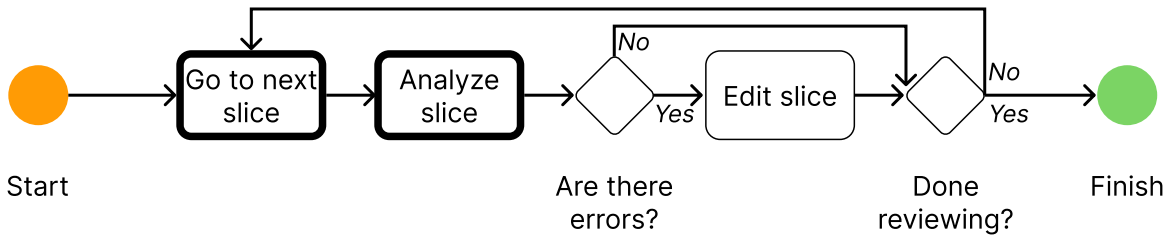

Fig. 6 Scheme of the delineation quality assessment (QA) process for an OAR. The analyze slice and edit slice rectangles have an associated time cost. The workflow variations we implement differ in the implementation of the go to next slice and analyze slice steps, which have a thicker border

In the error variations, we use delineation error instead of AI uncertainty because AI uncertainty serves as a proxy for delineation errors. By using the actual error, we simulate a best-case scenario where AI uncertainty perfectly identifies delineation errors.

For this study, the same OARs and bounding boxes per OAR as described in Sect. 3 were used. We preprocessed the `error` following the protocol proposed by Sander et al. (2020) to remove tolerated errors. This filtering process excludes slices with errors that can be attributed to interobserver variation. An OAR's erroneous voxel is considered a tolerated error if it is within 2 pixels from the border of $\texttt{del}^*$ (OAR), not part of a region of erroneous voxels of at least ten voxels in size, and not outside the top and bottom delineation limits. The slice metric we use for the error workflow is the sum of the non-tolerated erroneous voxels.

We use the dose as a proxy of the clinical significance of potential delineation errors for the patient's treatment. We selected the maximum as the aggregation function for the per-slice dose metric. Jensen et al. (2020) consider the mean dose, but we opted for the max based on the results of the user study. $\max(\texttt{dose}(\text{OAR})_s)$ is a more stringent constraint, representing a worst-case scenario for dosimetric deviations caused by erroneously delineated voxels in slice s. The max of the dose per slice indicates a lower risk in areas where the dose is consistently lower than the OAR's dosimetric constraint. The first three columns of Table 5 display the OARs, their max-dose constraints, and average slice numbers across patients for the baseline.

We simulate clinician behavior, relying on existing literature to estimate time costs for different steps. Based on Aselmaa et al. (2017), we model the time for analyzing a slice s in the baseline condition as $t_a(s) \sim \mathcal{N}(4.2,\ 3.2)$ seconds. For the error and dose conditions, we model the analysis time as $t_a^{\epsilon}(s) \sim \mathcal{N}(4.2 + \epsilon,\ 3.2)$ seconds. Here, ϵ represents the additional time required for analyzing DEDS suggestions, which are often not contiguous, resulting in jumps between non-sequential slices. In the simulation, we consider $\epsilon \in \{0, 4\}$ seconds, which allows us to assess the magnitude of the effect introduced by increasing analysis times. Finally, we assume a two-dimensional brush of size $bs = 10$ pixels for editing and model the time for editing a group of bs pixels as $t_{epix} \sim \mathcal{N}(1,\ 0.1)$ seconds. The time for editing a faulty slice is computed as $t_{ed}(s) = (t_{epix} \cdot \sum_{vox} \texttt{error}_s)/bs$. Note that the editing time modeling may vary depending on the editing tools used. In this case, we assume manual pixel brushing for simplicity. The total time per workflow execution is calculated as

$$T_{tot} = T_a + T_{ed} = \sum_{s \in S} t_a(s) + t_{ed}(s), \tag{6}$$

Table 5 Results of the simulation study conducted on a retrospective cohort of N = 42 patients

OAR	l_{OAR}	Number of slices			Attended errors (%)			Total elapsed time (sec)				
	(Gy)	*Baseline*	*Error*	*Dose*	*Baseline*	*Error*	*Dose*	*Baseline*	*Error*($\epsilon = 0$)	*Error*($\epsilon = 4$)	*Dose*($\epsilon = 0$)	*Dose*($\epsilon = 4$)
BrainStem	54	23 ± 3	18 ± 5	7 ± 6	100 ± 0	100 ± 0	20 ± 21	182 ± 94	157 ± 100	225 ± 111	48 ± 47	76 ± 70
Mandible	72	35 ± 3	18 ± 10	18 ± 11	100 ± 0	100 ± 0	42 ± 38	277 ± 405	203 ± 421	270 ± 437	111 ± 86	180 ± 124
Parotid_L	26	27 ± 3	21 ± 5	22 ± 8	100 ± 0	100 ± 0	79 ± 30	175 ± 44	150 ± 53	231 ± 69	144 ± 63	231 ± 90
Parotid_R	26	26 ± 3	22 ± 5	22 ± 8	100 ± 0	100 ± 0	83 ± 29	190 ± 55	169 ± 62	253 ± 77	160 ± 71	244 ± 96
Submand_L	35	14 ± 2	10 ± 3	13 ± 4	100 ± 0	100 ± 0	93 ± 26	111 ± 39	93 ± 47	134 ± 57	102 ± 48	153 ± 59
Submand_R	35	14 ± 2	10 ± 3	13 ± 4	100 ± 0	100 ± 0	92 ± 25	101 ± 30	82 ± 34	121 ± 44	94 ± 39	146 ± 53
Total	–	140 ± 9	98 ± 19	96 ± 30	100 ± 0	100 ± 0	68 ± 19	1036 ± 445	855 ± 474	1233 ± 516	659 ± 227	1031 ± 333

The table lists the organs-at-risk (OARs) considered in the study and their dosimetric limits in Grays (Gy). For each workflow variation, it provides the average and standard deviation of the number of slices reviewed by the simulated clinicians and the percentage of errors addressed. For the total time taken to complete the QA process, results are further detailed by scenario within each workflow. Decimal places are omitted for clarity

where S is the set of slices to review and T_a and T_{ed} represent the total analysis and editing time, respectively. To assess workflow quality, we calculated the percentage of attended errors for each workflow by dividing the sum of errors in the visited slices by the total amount of errors within the OAR's volume.

We conducted one hundred workflow runs for each combination of patient, OAR, and experimental condition (workflow variation).[1] In the results, we aggregate numerical quantities like slice numbers and times across the workflow runs within each OAR of each patient to obtain a statistical overview of the differences between conditions.

5.2 Results and discussion

Table 5 aggregates slice numbers, percentages of attended errors, and total elapsed QA times across patients. The last row of the table indicates that, on average, the baseline workflow takes longer than dose-based workflows and the optimistic error-based one. In the baseline workflow, which takes 1034s, the simulated clinician spends an average of 7.4 s per slice. In the error and dose workflows, the time per slice is 8.72 and 6.86 s for the optimistic scenario ($\epsilon = 0$) and 12.58 and 10.73 s for the pessimistic one ($\epsilon = 4$). Even if the time per slice is higher in the DEDS workflows, the total elapsed time generally turns out lower because clinicians do not need to check all slices. Regardless of the scenario, we observe a two-second difference in per-slice times between the dose-based and error-based workflows. These differences translate to total time savings of around two hundred seconds for both scenarios. However, these time gains come at the cost of quality. The table shows that while the baseline and error-based workflows addressed all errors, the dose-based ones only attended to 69% of them. A similar speed/quality tradeoff is expected if a higher threshold is used in the error-based workflows to limit the subset of slices for review. Focusing on individual OARs, we observe similar trends. Noteworthy are the BrainStem and the Mandible for which dose-based DEDS workflows obtain significant speedups. The dose-based workflows had the lowest percentage of addressed errors for the Mandible and BrainStem, indicating that many slices were skipped because they did not exceed the dosimetric constraints. This prioritization strategy, along with the larger size of these structures, accounts for the observed time savings. Skipping more slices, especially those with significant errors, reduces analysis and editing times but compromises delineation quality (Chaves-de-Plaza et al. 2022).

Focusing on the difference between scenarios, it is possible to observe how increasing the difficulty of the slice analysis task, and consequently, the time it takes leads to longer T_{tot}. Although the pessimistic dose scenario is competitive with the baseline, the error one significantly exceeds it. At the OAR level, we note that larger structures like the BrainStem and the Mandible, although closer to the baseline, still outperform it in most cases. This shows that, even with increased analysis times, DEDS can be particularly time-saving when used to review large anatomical structures, at the expense of confusing clinicians as seen in the user study.

To understand the contributions of the analysis (T_a) and editing (T_{ed}) times to the total QA time, in Fig. 7 we visualize the total analysis (a) and editing (b) times per OAR per patient averaged across simulation runs. Each column of gray horizontal lines within an OAR's area corresponds to a simulated condition, denoted by the color of the diamond on the column. Each line corresponds to the average time per patient and the diamond presents the average across patients. In general, we observe that in the optimistic scenarios, the analysis times are consistently below the baseline. In the pessimistic scenario, DEDS analysis times are less favorable but stay close to the baseline for larger structures like the BrainStem and the Mandible, a similar trend to the one we observed for T_{tot} before. Except for the BrainStem, the dose-driven workflow consistently requires more time than the error-driven one for $\epsilon = 0$ and $\epsilon = 4$. This indicates that the max(`dose`) criteria designate more slices as high-risk compared to error-free slices.

Concerning editing times, the figure indicates that the baseline and the error-based DEDS workflows perform similarly because, without a priority metric or error tolerance, the simulated clinician has to amend all the delineation errors in the error-based workflows. In contrast, the dose-based DEDS workflows are faster because they focus solely on slices with a high max dose, which are not necessarily the ones with the errors that take the longest to edit. In line with the results in Table 5, the improved performance of dose-based workflows is notable for the BrainStem and the Mandible, which are the largest structures and, therefore, tend to have more extensive erroneous regions. Finally, note that the times between scenarios do not change because we assumed the editing mechanism remains the same and is unaffected by the slice sequence.

In summary, the results of the simulation study suggest that DEDS workflows can reduce QA times. As the results for the dose-based workflows show, more significant time gains can be achieved by using more stringent thresholds to select the subset of slices to review at the cost of decreased delineation quality. This reduction in quality might be acceptable if it can be established that the bypassed errors are not clinically relevant. Our findings show diminishing DEDS advantages over the baseline workflow for smaller

[1] The simulation and analysis codes are available at https://graphics.tudelft.nl/study-deds.

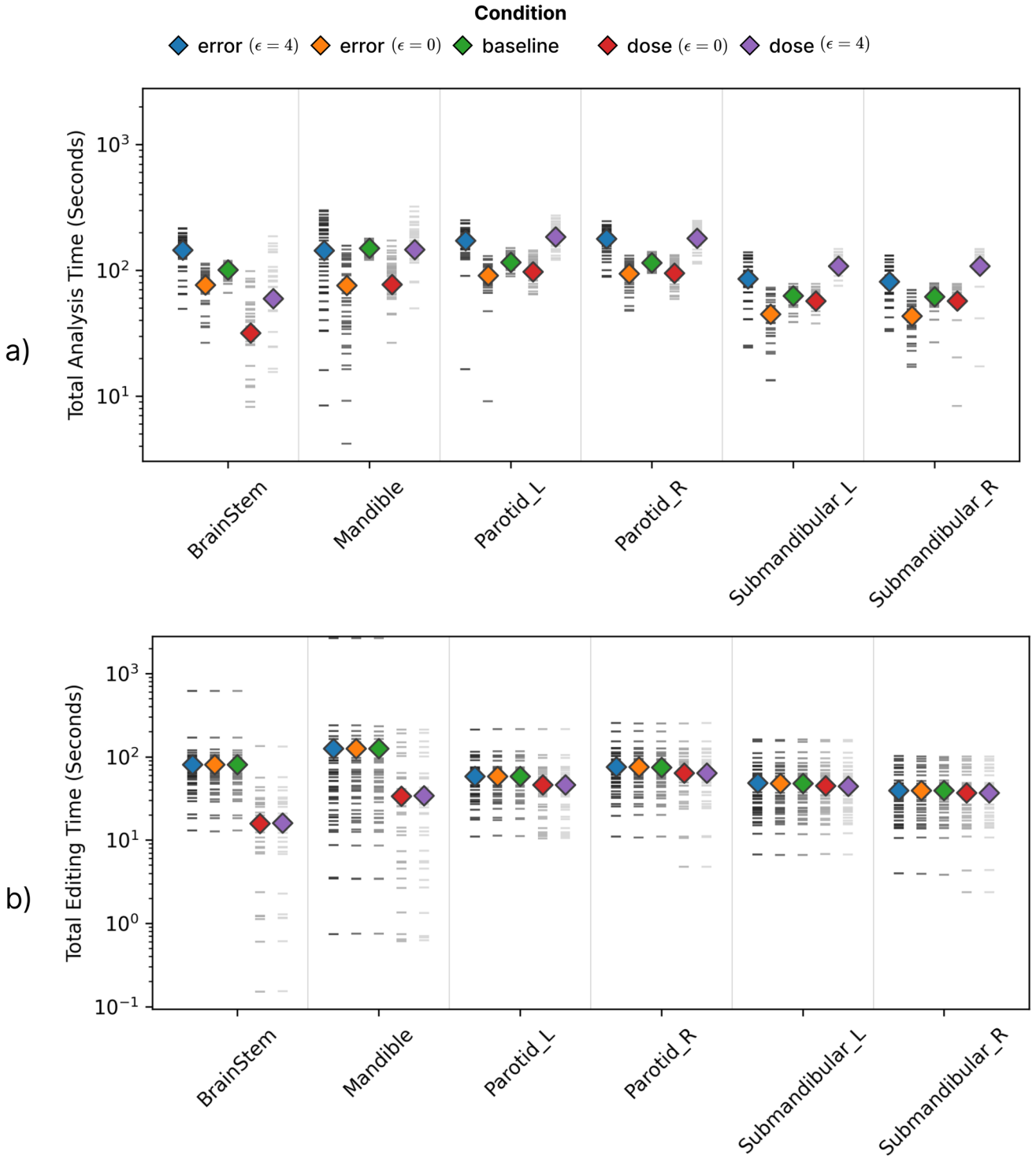

Fig. 7 Mean total analysis (**a**) and editing (**b**) times per OAR per patient in the cohort for the five simulated conditions. Each column within an OAR's area corresponds to the condition indicated by the color of the diamond. Gray horizontal lines within each column correspond to the patient's times, averaged across simulation runs. The colored diamond indicates the mean time per condition. The y-axis uses a logarithmic scale to enhance comparability and reduce empty space in the plot. Note that the y-axes of the two subplots have different ranges

structures and when $\epsilon > 0$. Therefore, it is essential to reduce analysis time to justify the practical use of DEDS.

6 Discussion

In this paper, we evaluated the clinical suitability of delineation error detection systems (DEDS). In particular, can DEDS speed up the Quality Assessment process without losing quality? To this end, we co-designed a DEDS with two experienced head and neck radiation oncologists from Utrecht University Medical Center and Leiden University Medical Center. The system was then used by two clinicians from HollandPTC to perform the assisted and unassisted DEDS workflows based on slice-wise statistics of the uncertainty, dose, and error. Based on insights from the user study, we addressed the question of whether DEDS can contribute to speeding up the clinical QA workflow using a simulation approach. A contribution of this work is a computational model of the QA process, which we used to simulate and compare several workflows. Researchers can use and extend this model to benchmark novel and existing DEDS proposals.

In the user study, we identified two key challenges to DEDS adoption. First, the information sources require refinement. Clinicians appreciated using dose information for its clarity, as it helped filter out clinically insignificant slices, but found the uncertainty and error metrics confusing, unnecessary, and potentially unreliable. This issue might be addressed by allowing more time for familiarization, introducing clearer indicators of uncertainty, and enhancing system-user compatibility in clinical settings (Bansal et al. 2019; McCrindle 2021; Bansal et al. 2019). Second, DEDS workflows often require navigating between non-contiguous slices, which clinicians found cumbersome and fatiguing. This navigation mode led clinicians to revert to conventional, sequential slice inspection, increasing the number of interactions. The challenge of maintaining a mental frame when jumping between slices could explain this behavior (Aselmaa et al. 2017). Providing less intrusive guidance or better tools to update clinicians' mental models could alleviate these issues (Musleh et al. 2023).

The simulation study showed that DEDS can improve QA times over the current baseline, especially for large anatomical structures where only a subset of slices is relevant according to a predefined metric. Nevertheless, considering smaller subsets of potentially non-adjacent slices poses two challenges. First, analysis times increase because clinicians cannot inspect slices sequentially. A mitigation strategy could be to offer clinicians chunks of contiguous slices to allow more effective sense-making. Second, and perhaps more critical for the adoption of DEDS-based workflows, it should be possible to be certain that bypassed errors are not clinically relevant-a non-trivial challenge that requires improving AI uncertainty estimates and developing clinically relevant metrics (Roberfroid et al. 2024). For instance, DEDS could leverage clinical measurements or heuristics like distance to target volumes as a priority metric when the error or dose are unavailable. The proposed framework can directly accommodate new metrics by defining a per-slice aggregation and a weight, allowing for combination with other metrics if needed.

Finally, there are several future work avenues. First, the present study applies to OARs, but other high-priority structures like target volumes and elective lymph nodes could also be considered. Target volumes likely face challenges to adoption because clinicians are less willing to forego reviewing all slices due to the high risk they represent to the patient. For example, missing errors in target volumes could directly impact treatment outcomes, making clinicians cautious about skipping slices. Lymph node fields are more promising because of their large extent (which makes them cumbersome to delineate), high priority, and relative stability across the population, facilitating the recent development of auto-delineation technologies (Cardenas et al. 2021). Second, the user and simulation studies could be extended to include other auto-delineation AIs and anatomical regions, which might have different error modes. Finally, the computational model of the QA process can be enriched, such as by using skewed distributions for modeling reaction times, which can be more appropriate but need substantial empirical data to estimate their parameters (Wolfe et al. 2010).

7 Conclusion

This study evaluated delineation error detection systems (DEDS) for improving the Quality Assessment (QA) process in clinical settings. A user study identified two main challenges that must be addressed to increase DEDS' adoption. First, clinicians preferred dose-based prioritization for error detection, finding it more intuitive than other metrics like uncertainty and error, which were seen as confusing and less reliable. Second, the non-sequential navigation required by DEDS disrupted clinicians' natural workflow, making it harder to make sense of the DEDS' suggestions. A computational model was introduced to benchmark different DEDS workflows. Simulations showed that DEDS could significantly reduce QA times, particularly for large structures, but this speed-up comes at the cost of delineation quality. Therefore, improving the accuracy of error proxies, such as AI uncertainty estimates, and developing metrics to assess the clinical significance of errors are crucial. Researchers can use and extend the computational model to further evaluate and refine DEDS.

Appendix A ADEDS development

In this section, we outline the development of our Delineation Error Detection System (DEDS) used in the workflow comparison user study (Sect. 5). We engaged in a co-development process with RO1 (RO from Utrecht UMC) and RO2 (RO from Leiden University Medical Center), involving multiple sessions where they used the tool for error detection and participated in structured discussions regarding tool usability and information source suitability. Our analysis involved logging clinicians' interactions and transcribing discussions, with relevant excerpts provided below.

A.1 Clinical delineation software

Figure 3's top panel displays a standard open-source delineation software's graphical user interface (GUI), consisting of two primary sections: the slice explorer (light blue rectangle) listing anatomical structures for delineation and the slice viewer (orange rectangle) for navigating 3D images via scrolling or navigation keys, supporting zooming and panning, and enabling pixel editing using tools like brushes or polygon pens. Our custom implementation, based on this GUI, was developed to support the slice-based error detection task. While we initially considered using existing delineation software, their closed source code or complexity hindered our envisioned extensions. Therefore, we re-implemented essential functionalities, excluding editing features, and instead used key presses to indicate editing intentions, as described in the subsequent section on extending the prototype.

A.2 Error detection and prioritization via per-slice scores

The bottom panel of Fig. 3 shows the GUI of the DEDS prototype. Similar to delineation software it has a slice explorer and viewer. Nevertheless, we extended the slice explorer with two features that permit slice-driven error detection. First, the list offers a higher slice-level granularity level. Traditional software only allows browsing a list of OARS. The DEDS slice explorer permits drilling down the OAR into the slices that it spans. Furthermore, it permits sorting each OAR's slices based on user-defined scores as defined in Sect. 3.3. The bottom left area of the slice explorer in Fig. 3 shows the score definition widget.

A.3 Clinicians' feedback

The DEDS prototype underwent significant changes based on feedback from RO1 and RO2, including the addition of contextual information and image overlay features, customization of color maps, and simplification of score displays. Clinicians' feedback influenced workflow improvements, such as grouping slices by structure in the slice explorer for a less overwhelming experience. Initial impressions of `unc` and `error` were mixed, with clinicians finding them limited and potentially misleading, leading to reduced trust in the system. To address this, explanations were provided during the workflow comparison study. In contrast, clinicians reacted positively to `dose` information, suggesting predefined settings per organ, with an emphasis on maximum dose and gradient magnitude (`grad_dose`) as valuable additions to the information sources. These enhancements aimed to enhance DEDS usability and effectiveness.

Acknowledgements 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.

Data availability The data that support the findings of this study are available from the authors but restrictions apply to the availability of these data. Interaction data from the user study are available from the corresponding author upon reasonable request. Patient data used in the user and simulation studies were provided by Holland Proton Therapy Center (HollandPTC) in the Netherlands for the current study, and so are not publicly available. These data are, however, available from the authors upon reasonable request and with permission from the Research Office at HollandPTC.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article. 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.

References

Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A (2020) Online daily adaptive proton therapy. Br J Radiol 93(1107):20190594

Altman MB, Kavanaugh JA, Wooten HO, Green OL, DeWees TA, Gay H, Thorstad WL, Li H, Mutic S (2015) A framework for automated contour quality assurance in radiation therapy including adaptive techniques. Phys Med Biol 60(13):5199. https://doi.org/10.1088/0031-9155/60/13/5199

Aselmaa A, van Herk M, Laprie A, Nestle U, Götz I, Wiedenmann N, Schimek-Jasch T, Picaud F, Syrykh C, Cagetti LV, Jolnerovski M, Song Y, Goossens RH (2017) Using a contextualized sense making model for interaction design: a case study of tumor contouring. J Biomed Inform 65:145–158

Aselmaa A, van Herk M, Song Y, Goossens RHM, Laprie A (2017) The influence of automation on tumor contouring. Cogn Technol Work 19(4):795–808

Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E (2019) Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. Proc AAAI Conf Artif Intell 33(01):2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

Bansal G, Nushi B, Kamar E, Lasecki W, Weld D, Horvitz E (2019) Beyond accuracy: The role of mental models in human-AI team performance. In: HCOMP 2019. AAAI

Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB (2019) Advances in auto-segmentation. Semin Radiat Oncol 29(3):185–197. https://doi.org/10.1016/j.semradonc.2019.02.001

Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Rhee DJ, McCarroll RE, Netherton TJ, Gay SS, Zhang L, Court LE (2021) Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. Int J Radiat Oncol Biol Phys 109(3):801–812

Castadot P, Lee JA, Geets X, Grégoire V (2010) Adaptive radiotherapy of head and neck cancer. Semin Radiat Oncol 20(2):84–93. https://doi.org/10.1016/j.semradonc.2009.11.002

Chaves de-Plaza NF, Mody PP, Hildebrandt K, Staring M, Astreinidou E, de Ridder M, de Ridder H, van Egmond van René (2022) 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

Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, Hu W (2021) The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiat Oncol 16(1):113

Hui CB, Nourzadeh H, Watkins WT, Trifiletti DM, Alonso CE, Dutta SW, Siebers JV (2018) Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. Med Phys 5(5):2089–2096. https://doi.org/10.1002/mp.12835. https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12835

Jensen K, Friborg J, Hansen CR, Samsøe E, Johansen J, Andersen M, Smulders B, Andersen E, Nielsen MS, Eriksen JG, Petersen JBB, Elstrøm UV, Holm AI, Farhadi M, Morthorst MH, Skyt PS, Overgaard J, Grau C (2020) The danish head and neck cancer group (dahanca) 2020 radiotherapy guidelines. Radiother Oncol 151:149–151. https://doi.org/10.1016/j.radonc.2020.07.037. (**2023/03/08**)

Kalpathy-Cramer J, Awan M, Bedrick S, Rasch CRN, Rosenthal DI, Fuller CD (2014) Development of a software for quantitative evaluation radiotherapy target and organ-at-risk segmentation comparison. J Digit Imaging 27(1):108–119

Maruccio FC, Eppinga W, Laves MH, Navarro RF, Salvi M, Molinari F, Papaconstadopoulos P (2024) Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. Phys Med Biol 69(3):035007

Mazur LM, Mosaly PR, Hoyle LM, Jones EL, Marks LB (2013) Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. Pract Radiat Oncol 3(4):e171–e177

Mazur LM, Mosaly PR, Hoyle LM, Jones EL, Chera BS, Marks LB (2014) Relating physician's workload with errors during radiation therapy planning. Pract Radiat Oncol 4(2):71–75

McCrindle B, Zukotynski K, Doyle TE, Noseworthy MD (2021) A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation. Radiology 3(6):e210031. https://doi.org/10.1148/ryai.2021210031

Mody PP, Chaves-de Plaza N, Hildebrandt K, van Egmond R, de Ridder H, Staring M (2022) Comparing Bayesian models for organ contouring in head and neck radiotherapy. In: Colliot O and Išgum I (eds) Medical imaging 2022: image processing, Vol. 12032. International Society for Optics and Photonics: SPIE, pp 120320F

Mody P, de-Plaza NF Chaves, Hildebrandt K, Staring M (2022) 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, Singapore, September 18, 2022, Proceedings. Springer-Verlag, Berlin, Heidelberg, pp 70–79

Mövik L, Bäck A, Pettersson N (2023) Impact of delineation errors on the estimated organ at risk dose and of dose errors on the normal tissue complication probability model. Med Phys 50(3):1879–1892

Musleh M, Muren LP, Toussaint L, Vestergaard A, Gröller E, Raidou RG (2023) Uncertainty guidance in proton therapy planning visualization. Comput Graphics 111:166–179

Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, Patel Y, Meyer C, Askham H, Romera-Paredes B, Kelly C, Karthikesalingam A, Chu C, Carnell D, Boon C, D'Souza D, Moinuddin SA, Garie B, McQuinlan Y, Ireland S, Hampton K, Fuller K, Montgomery H, Rees G, Suleyman M, Back T, Hughes CO, Ledsam JR, Ronneberger O (2021) Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. J Med Internet Res 23(7):e26151. https://doi.org/10.2196/26151

Ramkumar A, Dolz J, Kirisli HA, Adebahr S, Schimek-Jasch T, Nestle U, Massoptier L, Varga E, Stappers PJ, Niessen WJ, Song Y (2016) User interaction in semi-automatic segmentation of organs at risk: a case study in radiotherapy. J Digit Imaging 29(2):264–277

Ramkumar A, Stappers PJ, Niessen WJ, Adebahr S, Schimek-Jasch T, Nestle U, Song Y (2017) Using goms and nasa-tlx to evaluate human-computer interaction process in interactive segmentation. Int J Hum-Comput Interact 33(2):123–134

Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, Albrecht T, Gass T, Langguth C, Lüthi M, Jung F, Knapp O, Wesarg S, Mannion-Haworth R, Bowes M, Ashman A, Guillard G, Brett A, Vincent G, Orbes-Arteaga M, Cárdenas-Peña D, Castellanos-Dominguez G, Aghdasi N, Li Y, Berens A, Moe K, Hannaford B, Schubert R, Fritscher KD (2017) Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. Med Phys 44(5):2020–2036. https://doi.org/10.1002/mp.12197

Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, Garden AS, Peterson CB, Beadle BM, Court LE (2019) Automatic detection of contouring errors using convolutional neural networks. Med Phys 46(11):5086–5097. https://doi.org/10.1002/mp.13814

Roberfroid B, Lee JA, Geets X, Sterpin E, Barragán-Montero AM (2024) Dive-art: a tool to guide clinicians towards dosimetrically

informed volume editions of automatically segmented volumes in adaptive radiation therapy. Radiother Oncol 192:110108
Sander J, de Vos BD, Išgum I (2020) Automatic segmentation with detection of local segmentation failures in cardiac MRI. Sci Rep 10(1):21769. https://doi.org/10.1038/s41598-020-77733-4
Sandfort V, Yan K, Graffy PM, Pickhardt PJ, Summers RM (2021) Use of variational autoencoders with unsupervised learning to detect incorrect organ segmentations at ct. Radiology 3(4):e200218. https://doi.org/10.1148/ryai.2021200218
Sonke JJ, Aznar M, Rasch C (2019) Adaptive radiotherapy for anatomical changes. Semin Radiat Oncol 29(3):245–257. https://doi.org/10.1016/j.semradonc.2019.02.007
Steenbakkers RJ, Duppen JC, Fitton I, Deurloo KE, Zijp L, Uitterhoeve AL, Rodrigus PT, Kramer GW, Bussink J, Jaeger KD, Belderbos JS, Hart AA, Nowak PJ, van Herk M, Rasch CR (2005) Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'big brother' evaluation. Radiother Oncol 77(2):182–190
Steenbakkers RJ, Duppen JC, Fitton I, Deurloo KE, Zijp LJ, Comans EF, Uitterhoeve AL, Rodrigus PT, Kramer GW, Bussink J, De Jaeger K, Belderbos JS, Nowak PJ, van Herk M, Rasch CR (2006) Reduction of observer variation using matched ct-pet for lung cancer delineation: a three-dimensional analysis. Int J Radiat Oncol Biol Phys 64(2):435–448
van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF (2019) Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiat Oncol Biol Phys 104(3):677–684
Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, van Elmpt W (2020) Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. Radiother Oncol 153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008
Wolfe JM, Palmer EM, Horowitz TS (2010) Reaction time distributions constrain models of visual search. Vision Res 50(14):1304–1311
Zhou T, Li L, Bredell G, Li J, Unkelbach J, Konukoglu E (2023) Volumetric memory network for interactive medical image segmentation. Med Image Anal 83:102599. https://doi.org/10.1016/j.media.2022.102599