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

Mody, P.P.

5

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

This chapter was adapted from:

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

Abstract

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

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

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

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

5.1 Introduction

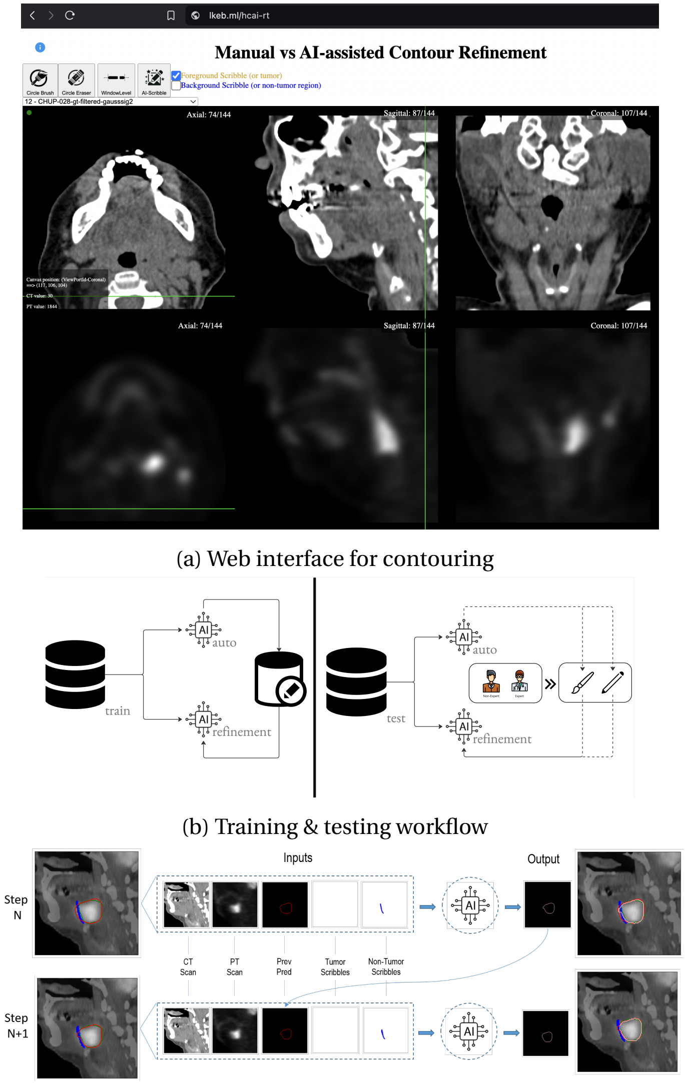
Auto-contouring in radiotherapy has made great progress over the last 5 years with improvements in AI (i.e., deep learning) models and a proliferation of clinically-available commercial tools [179–182]. Widespread use of these AI-based auto-contouring tools can be attributed to the time gains they provide.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

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

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

5.2 Materials and methods

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



(a) Web interface for contouring

(b) Training & testing workflow

(c) Neural network design

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

5.2.1 Dataset

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

5.2.2 Auto-contour and contour-refinement model training

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

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

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

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

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

Finally, we logged the time taken for both refinement tools and compared them. User interaction count, pixels drawn, and slices scrolled were also logged as they directly influenced the total time taken. An interaction was defined as a complete mouse click (press and release). Savings provided by the AI pencil when compared to the manual brush were also shown in percentages as:

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

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

5.2.4 Web-based tool

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

5.2.5 User Cohort

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

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

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

5.3 Results

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

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

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

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

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

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

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

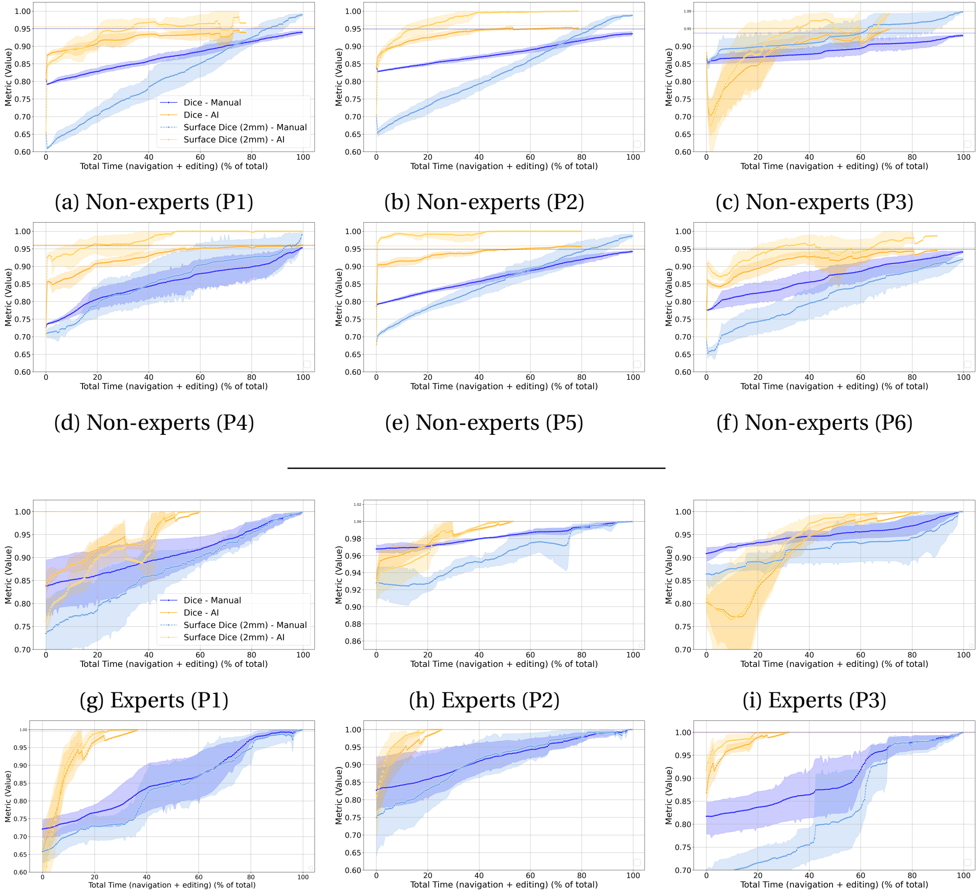
	P1	P2	P3	P4	P5	P6
E1	99 (77%)	19 (55%)	41 (67%)	117 (88%)	117 (95%)	190 (97%)
E2	119 (85%)	40 (78%)	66 (80%)	76 (80%)	69 (86%)	143 (88%)
E3	91 (81%)	18 (72%)	33 (66%)	78 (90%)	57 (90%)	87 (93%)
E4	72 (74%)	38 (84%)	43 (62%)	79 (90%)	26 (92%)	104 (95%)
Range (min, max)	[72,119] ([74%,85%])	[19,40] ([55%,84%])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

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

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

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

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



(a) Non-experts (P1)

(b) Non-experts (P2)

(c) Non-experts (P3)

(d) Non-experts (P4)

(e) Non-experts (P5)

(f) Non-experts (P6)

(g) Experts (P1)

(h) Experts (P2)

(i) Experts (P3)

(j) Experts (P4)

(k) Experts (P5)

(l) Experts (P6)

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

Additionally, Figure 5.3 shows a histogram plot of pixels drawn during contour refinement which is [63% – 93%] and [81% – 99%] less for the AI pencil than the manual brush for non-expert and expert sessions, respectively (Section 5.7.4). Figure 5.4 shows visual examples of scribbles and the contours they produce. AI pencil scribbles can be used to deal with both false positives (Figure 5.4a, 5.4c, 5.4d, 5.4f, 5.4g, 5.4h, 5.4i) and false negatives (Figure 5.4b, 5.4e, 5.4j). While most scribbles were shorter in length others were lengthier and explicit with their feedback (Figure 5.4f, 5.4i). Regardless of the style of the scribble, a sparse scribble in 2D (on slice *s*) propagates its changes in 3D. This is seen when one observes the updated contour on slices *s-1* and *s+1*. While some interactions align the refined contour with the reference contour with a single scribble (Figure 5.4g, 5.4h), others still require more interaction (Figure 5.4i, 5.4j)

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

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

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

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

Table 5.7: Interobserver Variability (IoV) across patients.

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

5.4 Discussion

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

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

5.4.1 Time for auto-contour refinement

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

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

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

5.4.2 Contour Consistency

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

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

5.4.3 Tooling

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

5.4.4 Future Work

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

5.5 Conclusion

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

5.6 Acknowledgement

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

5.7 Appendix

5.7.1 Dataset

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

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

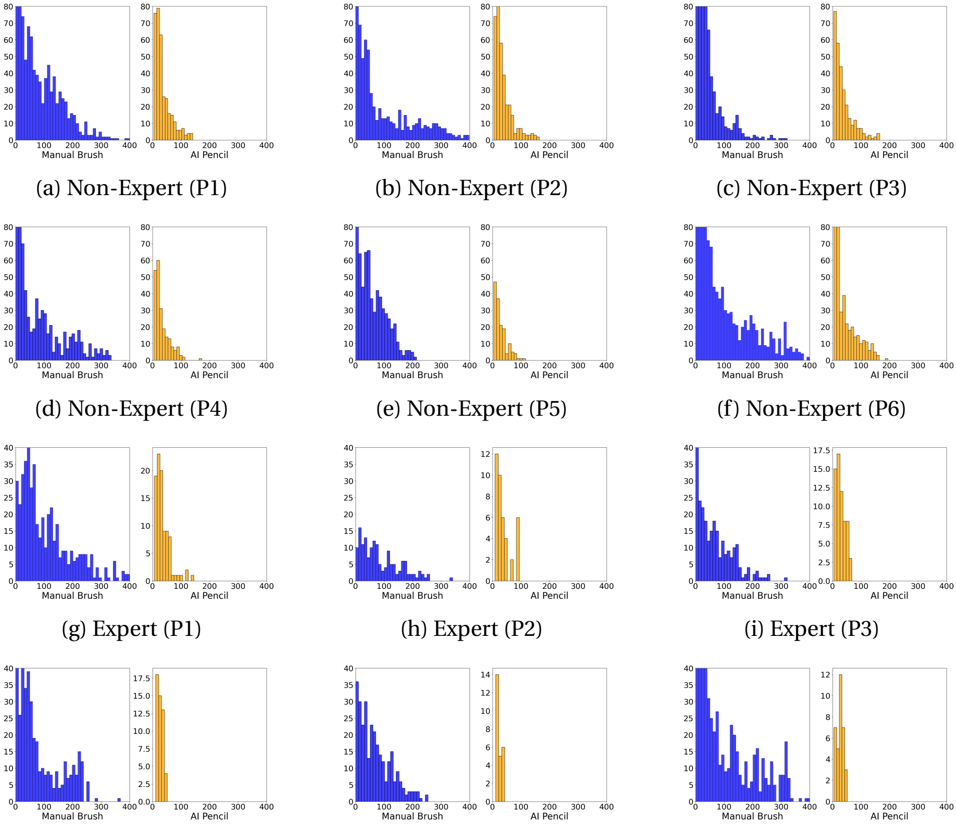
5.7.2 Auto-contouring and contour-refinement models

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

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

5.7.3 Web interface

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



(a) Non-Expert (P1)

(b) Non-Expert (P2)

(c) Non-Expert (P3)

(d) Non-Expert (P4)

(e) Non-Expert (P5)

(f) Non-Expert (P6)

(g) Expert (P1)

(h) Expert (P2)

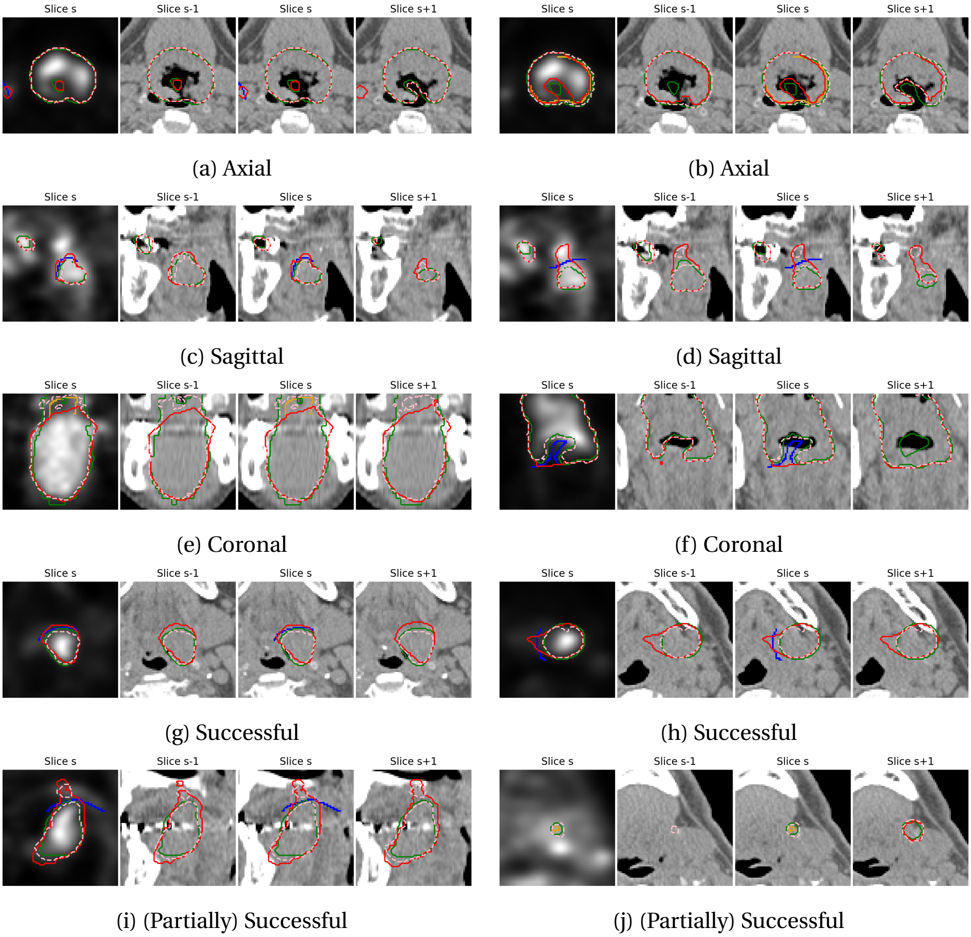
(i) Expert (P3)

(j) Expert (P4)

(k) Expert (P5)

(l) Expert (P6)

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



(a) Axial

(b) Axial

(c) Sagittal

(d) Sagittal

(e) Coronal

(f) Coronal

(g) Successful

(h) Successful

(i) (Partially) Successful

(j) (Partially) Successful

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

5.7.4 Additional results on user effort

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

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

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

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

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

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

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

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

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

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

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