Multi-Metric Clinical Validation of an Auto-Contour Refinement Tool in Head-and-Neck Radiotherapy

Jolijn L. Scharn

TU Delft

Multi-Metric Clinical Validation of an Auto-Contour Refinement Tool in Head-and-Neck Radiotherapy

by

Jolijn L. Scharn

in partial fulfilment of the requirements of
Master of Science
in Biomedical Engineering
Track Neuromusculoskeletal Biomechanics
at the Delft University of Technology,
to be defended publicly on Friday March 27, 2026 at 10:00 AM.

Student number:	5867797	
Project duration:	May 20, 2025 – March 27, 2026	
Thesis committee:	Dr. ir. N. Tumer,	TU Delft, supervisor
	Dr. ir. F.J.W.M. Dankers,	LUMC, supervisor
	Dr. ir. P.P. Mody,	LUMC, supervisor
	Prof. Dr. ir. M. Staring,	LUMC
	Dr. Q. Tao,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

LUMC Leids Universitair Medisch Centrum

Multi-Metric Clinical Validation of an Auto-Contour Refinement Tool in Head-and-Neck Radiotherapy

Jolijn L. Scharn

Background: Interactive segmentation models combine auto-segmentation methods with user interaction to overcome the inconvenience of manually adjusting contours generated by imperfect auto-contouring models. However, these models have not yet been implemented for tumor target volume segmentation in clinical radiotherapy settings. Therefore, this study validates a previously developed auto-contour refinement tool at the LUMC for Head-and-Neck (H&N) radiotherapy, demonstrating its robustness and trustworthiness.
Methods: A user study with six non-expert participants was performed, who iteratively refined a contour-refinement model prediction to align as closely as possible with the corresponding ground truth for six tumor volumes from six patients. The contour-refinement model updated its prior predictions based on user-provided foreground (tumor) and background (non-tumor) scribbles. This enabled Three Dimensional (3D) refinement until a satisfactory result was achieved. User inputs were collected and evaluated using performance metrics such as Dice and Surface Dice to evaluate robustness of the model, along with two newly introduced evaluation metrics proposed in this study to evaluate trustworthiness: local and non-local (Surface) Dice.
Results: Robust behavior is observed, as the model reacts in a highly consistent manner across all users. Only minor differences in model performance (Δ Dice scores of 0.1407 vs. 0.1296) were observed across users when different user inputs were applied. The AI pencil yields a strong initial improvement compared to manual annotations (27.4% vs. 6.4%, Wilcoxon p = 0.047), whereas subsequent iterations show variability. This variability was frequently observed in cases of incorrect user input, distortions caused by dental implants, anatomically complex regions, and during the segmentation of slices at the tumor boundaries. In all other cases the model showed a high trustworthiness, as it follows the users intent during the contouring process.
Conclusion: The incorporation of user feedback into the contour-refinement model results in a rapid improvement in segmentation quality across the entire volume. However, manual refinement by clinicians remains necessary for anatomically complex slices. Overall, this research shows that the model is robust to variations in user input and (apart from the first few iterations) there are no spurious changes in non-local areas. These are important findings when working towards clinical adoption of these interactive contour refinement models.

Keywords: Interactive, Medical, Image, Segmentation, Deep-Learning, Inference, User Prompts, Evaluation Methods, Metrics, Human-AI Interaction, Auto Contouring, Contour-Refinement, 3D, Radiotherapy

I. Introduction and Goal Definition

In modern radiotherapy, accurate tumor target delineation, is the crucial first step in treatment planning [1]. However, manual tumor contouring remains complex, time consuming, and labor intensive, due to variability between observers and

anatomical challenges, particularly in regions such as the head-and-neck region (H&N) [1–4]. This issue is expected to be exacerbated by the increasing incidence of H&N cancer cases [5–7], the projected shortage of physicians in the medical field [8], and the advent of online adaptive radiotherapy, which requires rapid and repeated contouring during treatment [9]. To address these challenges, deep learning-based auto-contouring tools are increasingly used in radiotherapy to delineate structures [10–13]. Currently, in clinical practice, these tools are primarily applied to support treatment planning through the identification and protection of organs at risk [14]. Their application to tumor target volumes in practical scenarios, however, remains limited due to several challenges [15].

First, there is substantial inter-patient variability in tumor volumes, which complicates model generalization. In particular, Convolutional Neural Networks (CNNs), a class of deep learning (DL) models, have limited ability to accurately handle tumor characteristics that were not represented in the training data [16]. Second, there is a need for large, diverse, high-quality training datasets to train these models [17]. However, such datasets are currently not available, and as a result, their applicability remains heavily constrained by the limited task-specific training data [18]. Third, accurately contouring metastatic lymph nodes remains particularly challenging [19], especially in H&N cases, where elective nodal regions often lack clearly imageable tumors [2]. Fourth, traditional automatic segmentation models rely heavily on deterministic predictive behavior that cannot be influenced or adjusted [18]. This is problematic when highly precise segmentations are required, such as in tumor segmentation. Finally, automatic segmentation models operate without considering the needs and context of the user [20], which poses a challenge when clinicians are expected to collaborate with the model. Due to these challenges, auto-contouring models remain imperfect, and clinicians often need to spend considerable time correcting the generated contours. Therefore, despite substantial advances in automatic segmentation, fully automatic approaches often fail to meet the accuracy and robustness requirements necessary for reliable clinical deployment [18, 21, 22].

To overcome the inconvenience of manually adjusting contours generated by auto-contouring models, Mody et al. [23] introduced an auto-contour refinement tool. This tool enables clinicians to refine deep learning-generated auto-contours by making scribbles that provide information about the intended tumor boundaries. This technique, known as interactive segmentation, is a key area of research in medical image analysis that aims to enhance the efficiency of labor-intensive annotations by incorporating human feedback [24]. The AI-pencil, as proposed by Mody et al. [23], enables 2D interactions to directly refine tumor contours in 3D H&N scans. This approach has the potential to significantly accelerate the delineation process.

Due to their potential, interactive segmentation methods have attracted considerable attention in the medical domain and have led to numerous recent studies evaluating 3D medical image segmentation [16, 18, 25–57]. Although these domain-specific adaptations have shown promising progress, many published methods are affected by pitfalls that obscure their effectiveness and hinder appropriate method selection [18]. For example, several studies assess interactive segmentation methods based on a single interaction step [18], whereas in clinical workflows such as radiotherapy, an iterative process is essential. Multiple rounds of tumor target segmentation are needed to gradually refine the contours until satisfactory results are reached. Furthermore, studies often evaluate predictions slice-by-slice or on sub-patches of a 3D volume instead of assessing the full image volume [18]. In addition, there is currently no standardized evaluation framework, leading to inconsistencies across studies [18, 24, 58]. While evaluation is often focused on technical performance, it frequently fails to demonstrate clinical relevance [59]. These missing evaluation frameworks present a challenge in this field, as robust evaluation of these models is a critical step towards clinical adoption.

Previous work by this author [60] has shown a strong reliance on common metrics such as the Dice Similarity Coefficient (Dice) and the Hausdorff Distance (HD) [24, 58, 61], which date back to 1945 [62] and 1914 [63], respectively. This previous work

[60] also demonstrated that almost no new metrics are being introduced, although there are ongoing concerns about the existing metrics. For instance, Dice does not capture boundary accuracy and may therefore be insufficient when boundary quality is the primary focus [58, 64, 65]. Therefore, additional metrics as Surface Dice are needed for the evaluation of the segmentation models at boundary levels. Furthermore, HD is not always suitable, since medical segmentations often contain noise and boundary irregularities, and its high sensitivity to outliers generally discourages direct use in this field [66]. Some research developed advanced segmentation models that demonstrate excellent performance on these metrics [30, 32, 38, 43, 47, 48, 55, 67–74]. However, because researchers in this field rely heavily on traditional evaluation metrics, critical insights into model behavior may be missed. This limits understanding of the internal workings of deep interactive segmentation models, whose decision-making processes remain largely non-transparent to researchers and clinicians [16].

More research on understanding interactive segmentation models could increase the confidence in using these models for clinical diagnosis and ensure the effective application of deep learning-based medical image segmentation in modern practice [16]. Two important aspects for evaluating a model's clinical effectiveness are robustness and trustworthiness. Model robustness reflects the consistency of its responses to different functionally similar corrective user inputs. Quantitative (surface) Dice evaluation only does not capture the robustness of the model and therefore this study additionally focuses on qualitative assessment of interactive segmentation output. Trustworthiness refers to a model's ability to accurately refine the region around the user-provided input without introducing spurious changes elsewhere in the tumor volume. Traditional evaluation metrics may reveal the presence of segmentation errors but do not provide insight into why they occur or where within the tumor volume they are most likely to arise, making their identification difficult. As a result, it is currently difficult to assess the trustworthiness of these models. To close this gap, this study proposes novel evaluation metrics, in addition to the commonly used suite of evaluation metrics such as (surface) Dice. Specifically, these metrics evaluate the model's ability to adhere to user intent in regions both close to the scribble slice and farther away, and are referred to as local and non-local Dice.

By adding these metrics to conventional metrics, this research project at the Leiden University Medical Center (LUMC) focuses on validating the iterative 3D interactive medical image segmentation method proposed by Mody et al. [23]. To achieve this, the same deep learning model is reused on the same patient data, with a different evaluation approach. By these means, this study evaluates whether deep learning-assisted interactive contour refinement is robust and trustworthy and how these terms can be evaluated, which is a valuable and necessary step toward future clinical implementation.

II. Research Questions

This research aims to answer the following question: How can deep learning-based interactive contour refinement models be effectively evaluated for their robustness and trustworthiness?

To explore this for a published model ([23]), this study focuses on these sub-questions:
1. To what extent do segmentation outcomes differ across users during interactive contour refinement?
2. How well does this contour refinement model adhere to the users input in regions local to where the input was provided. Moreover, does it make spurious changes in regions further away?

In this work, robustness is defined as the consistency of the model's response across users when they provide functionally similar corrective input, reflecting the same segmentation intent. Trustworthiness is defined as the model's ability to adhere to user corrections locally, while avoiding unintended spurious changes in non-local regions.

III. Research Methods

A. Dataset Description

A dataset from the HECKTOR2022 challenge was used in this study [75], which includes 524 paired Computed Tomography (CT) and Positron Emission

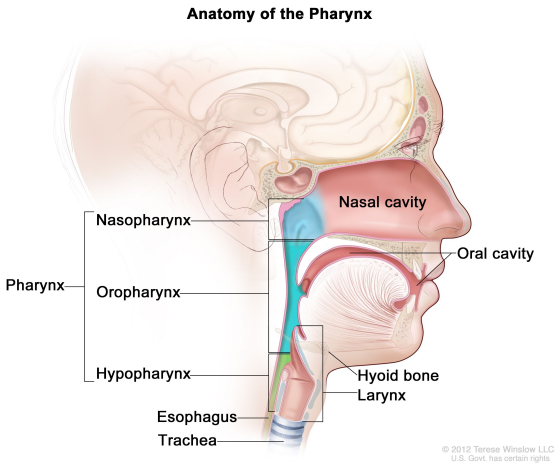


Figure 1. Anatomy of the pharynx. The pharynx is a hollow muscular structure in the neck that connects the nasal cavity to the larynx and esophagus and consists of three regions: the nasopharynx, oropharynx, and hypopharynx [76].

Tomography (PET) scans of the H&N area from seven different medical centers across four different countries. The dataset exclusively includes patients diagnosed with oropharyngeal cancer. This cancer arises in the oropharynx at the back of the throat, encompassing regions such as the base of the tongue and the tonsils. The oropharynx is a region of the pharynx, as shown in Figure 1. The contouring models used in this research were trained and validated by Mody et al. [23]. In that study, data from three countries (Canada, Switzerland, and the United States of America; 452 scan pairs) were used for training and validation, while data from the remaining country (France; 72 scan pairs) were reserved for testing. In the current study, six cases were selected from this test set. The ground truths of these cases are presented in Figure 14. Additional details regarding the dataset are provided in the Appendix Section VIII.A.

B. Auto-Contour and Contour-Refinement Models

This study employs two deep learning models: an auto-contouring model and a contour-refinement model. The auto-contouring model is used to generate an initial tumor segmentation. Subsequently, user-provided input is supplied to the contour-refinement model, which refines the initial segmentation and then iteratively updates its own predictions based on the provided interaction. A visual representation of this process is presented in Figure 2. Both the auto-contouring and contour-refinement models were developed, trained, and validated by Mody et al. [23]. The models employ a standard U-Net architecture (approximately 1.2 million parameters) implemented within the MONAI framework [77] and were trained using a standard cross-entropy loss function. In this study, these pretrained models were used without further modification.

The auto-contouring model generates an initial tumor mask from CT and PET images using ground truth annotations from the HECKTOR dataset. This initial prediction serves as input to the contour-refinement model, which incorporates additional information in the form of user interactions to refine the segmentation. The refinement model operates on multiple input channels (n=5), including the imaging data (CT and PET), the predicted segmentation from the previous iteration, and user-provided foreground (tumor) and background (non-tumor) scribbles. During training of the refinement model (done by Mody et al. [23]), user interaction was simulated by automatically generating two-dimensional scribbles in regions where the initial segmentation deviated from the ground truth. These simulated scribbles guided the refinement model to correct both missing (false negative (FN)) and incorrectly included regions (false positive (FP)), and model outputs were optimized by comparison with the ground truth. Further details on the simulated interaction strategy and the training of the models are described in the original work [23].

C. Graphical User Interface

The experiments have been conducted using a web-based interface, which was originally developed by Mody et al. [23]. In this interface, users can choose between a manual brush and an AI-assisted pencil tool. With the AI-assisted tool, users refine auto-generated tumor contours by placing scribbles (positive or negative) on the image. These scribbles, along with the CT and PET scans and the previous contour prediction, are sent to the backend as model inputs. The deep learning contour-refinement model then generates updated contours, which are returned to the frontend for

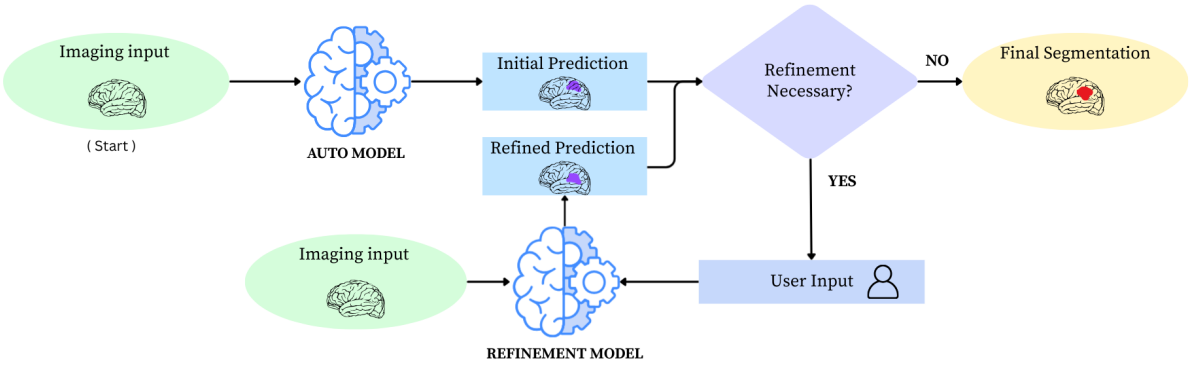


Figure 2. Graphical Flow of Proposed Deep Interactive Image Segmentation. Imaging Input: CT and PET. User input: Foreground and Background Scribbles. Auto-model: Auto-contouring model Section III.B. Refinement Model: Contour-refinement model Section III.B. The auto model had a two-channel input (i.e., CT + PET). The refinement model had a five-channel input (i.e., CT + PET + previous contour + foreground scribble + background scribble).

user review and, if needed, further refinement. Additional details regarding the graphical user interface are provided in the Appendix Section VIII.B.

D. Interactive Contour Refinement Sessions

This study will evaluate the auto-contour refinement tool by performing a user-study by conducting interactive refinement contouring sessions. For these sessions, six non-expert participants were recruited, which were (PhD) students within the Radiotherapy or Radiology department of LUMC. During these sessions, two experiments were conducted.

1. *Experiment 1 - Robustness of Interactive Segmentation to Single-Scribble User Input*

In this experiment, the response of the interactive contour-refinement model is measured to different types of user inputs. Users were asked to refine the segmentation of a single image slice by placing one scribble. Two distinct slices were selected for this experiment: one representing a relatively simple segmentation task with straightforward anatomy, and one representing a more challenging task due to the presence of an air cavity (Figure 3). In both cases, the initial mask was generated by the auto-contour model and the refinement was generated by the contour-refinement model. By presenting the same starting condition to multiple users, i.e. presenting the same slice for user interaction, the robustness to variation in user inputs can be quantified using Dice values, as well as qualitatively analyzed by visual inspection.

The simple slice was selected for this experiment due to its typical straightforward anatomical structure. In this slice, the auto-contouring model produced an incorrect prediction because it oversegmented the tumor region on one side of the tumor. This error is visible in Figure 3 and in the 3D visualization of Patient 4 shown in the Appendix Section VIII.C. Users were asked to refine this oversegmented area by placing a background scribble in the erroneous region. The users were free to determine the precise shape and location of the scribble.

The more challenging slice, shown in Figure 3, was chosen because it contained an air cavity, which was hypothesized to complicate segmentation. The corresponding ground truth in the axial direction consisted of two separated regions, one of which was undersegmented by the auto-contour model. Users were

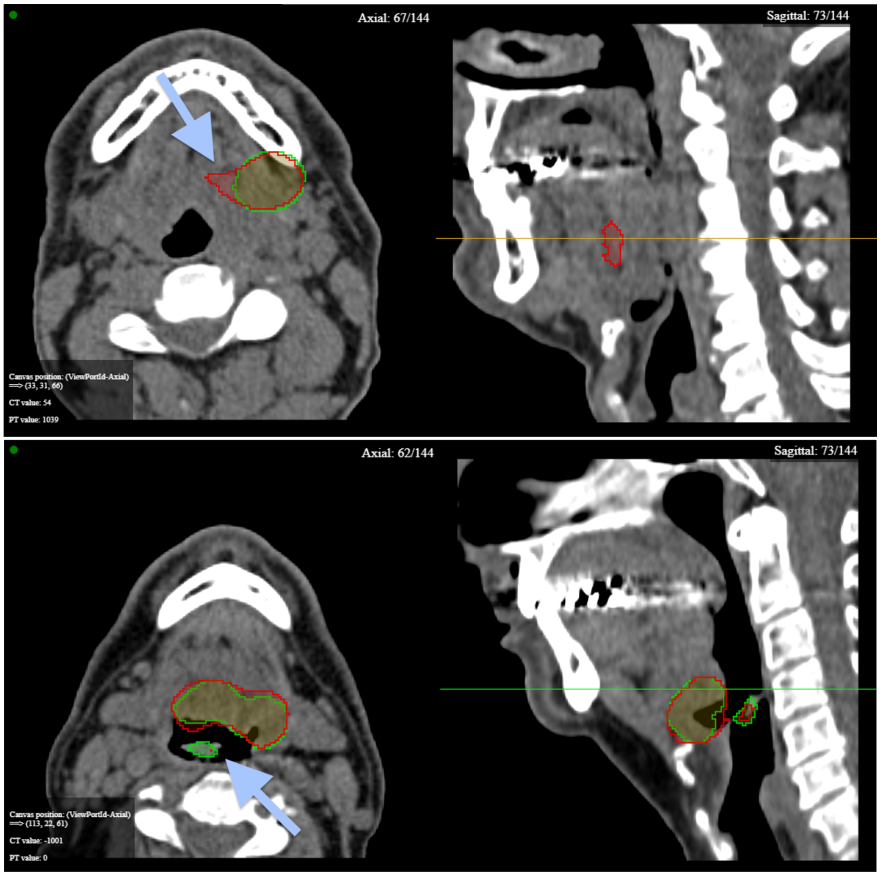


Figure 3. Automatic segmentation results for Experiment 1 in axial (left) and sagittal (right) view. (a) Easy slice with straightforward anatomy (axial slice 67/144, P4). (b) Challenging slice including an air cavity (axial slice 62/144, P3). Green: ground truth. Red: tumor target prediction. Arrows point to the part of the segmentation that needs to be corrected by the user.

instructed to correct this undersegmentation by placing a foreground scribble in the region where tumor tissue was present in the ground truth but missing in the automatic segmentation.

2. Experiment 2 - Interactive Contour Refinement Sessions: AI vs Manual

In this experiment, six users were asked to segment six complete tumor volumes. The objective was to reproduce the ground truth segmentation, which was simultaneously visualized, as accurately as possible using either a manual brush or an AI-assisted pencil. In the manual brush experiments, the contour-refinement model was not used. In contrast, when using the AI pencil, user input was communicated to the contour-refinement model, which iteratively refined the segmentation in three dimensions. Users continued refining the segmentation until they were satisfied with the final result.

Several constraints were applied to standardize the experimental procedure and improve the reliability and comparability between experiments. Scribble refinements were performed with a standard optical mouse, while navigation of slices and changing segmentation tools, e.g., switching between foreground and background scribbles or brushes, was performed using keyboard shortcuts. Users were also restricted to perform all segmentation actions exclusively in the axial plane to facilitate evaluation. Additionally, users were asked to follow a consistent navigation strategy by starting in the middle of the volume (slice 72), proceeding upward through the volume, and then moving downward. This guideline was introduced to improve comparability across users. Nevertheless, users retained

the freedom to navigate through the volume as needed.

Table 1. Overview of patients and segmentation type used per session. AI: An experiment using only the AI pencil. Manual: An experiment using only the manual brush.

Patient	Session 1	Session 2
P1	AI	Manual
P2	Manual	AI
P3	AI	Manual
P4	Manual	AI
P5	AI	Manual
P6	Manual	AI

Patient cases were selected to represent a range of interesting anatomical structures, including relatively straightforward tumors and more challenging cases. The order of the patient cases was determined based on the quality of the initial segmentation produced by the auto-contouring model. The patient with the highest surface Dice score was presented first. The resulting case order was identical for all users and is summarized in Table 1. A one-week interval was introduced between Session 1 and Session 2 to reduce potential memory bias. Before the first session, all users participated in a training session in which they learned how to segment an entire volume. They became familiar with the user interface, the delineation tools (AI pencil and manual brush), and with the contour-refinement model response to scribble inputs.

For each paired experiment (AI and manual), time was normalized using the maximum duration of the two sessions, whether manual or AI. Dice trajectories were interpolated, as users placed scribbles at different time points during the experiments, enabling comparison across users. Median and interquartile ranges (Q1-Q3) were plotted instead of mean and standard deviation to reduce the effects of outliers, given the small number of participants. To quantify early changes in performance, the initial slope of the Dice curve was estimated over the first 5% of the normalized interaction time. This interval was selected to restrict slope estimation to the early phase of the experiment and to ensure local linearity. The slope was computed by linear regression on the median interpolated Dice curve, using the same interpolation scheme as applied for visualization.

E. Performance Metrics

Table 2. Notation used for local and non-local (Surface) Dice and Surface Dice metrics.

Symbol	Description
V	Set of all voxels in the 3D volume
v	A voxel in the 3D volume
z_v	Slice index of voxel v
z_S	Slice index at which the scribble is placed
n	Neighborhood (number of slices cranial and caudal to z_S to form the local region)
G	Ground truth (reference) segmentation mask
P_{before}	Predicted segmentation before refinement
P_{after}	Predicted segmentation after refinement
S_X	Set of surface voxels of segmentation X (morphological boundary of X)
$S_{X,L}$	Surface voxels of X restricted to the local region ($S_X \cap \Omega_{\text{Local}}$)
$S_{X,NL}$	Surface voxels of X restricted to the non-local region ($S_X \cap \Omega_{\text{Non-Local}}$)
$S_{\text{before},NL}$	$S_{P_{\text{before}}} \cap \Omega_{\text{Non-Local}}$
$S_{\text{after},NL}$	$S_{P_{\text{after}}} \cap \Omega_{\text{Non-Local}}$
τ	Surface distance tolerance (2 mm)

1. Volumetric Dice and Surface Dice

To evaluate the model's performance, both conventional and novel metrics will be used. Conventional metrics include overlap-based measures such as the Dice, which quantifies the overall agreement between the predicted and reference contours. However, conventional Dice does not account for spatial context, it only measures total volumetric overlap, regardless of where the differences occur. Therefore, results will also be evaluated using surface Dice, which was first introduced by Nikolov et al. [78]. It assesses the overlap of two surfaces (at a specified

tolerance) instead of the overlap of two volumes. This makes it more sensitive to boundary differences, which are clinically relevant in precise tumor target delineation. The imaging data used in the present study have a resolution of 1 mm (Appendix VIII.A). Therefore, a tolerance of 2mm was selected, as this corresponds to roughly twice the size of a single voxel and accommodates minor variations in boundary segmentation. The formulas for Dice (1) and surface Dice (2) used in this study are given below, where P denotes the predicted segmentation and G the ground truth segmentation. All symbols including their description are presented in Table 2.

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

$$\text{Surface Dice}_\tau(S_P, S_G) = \frac{|S_P^{(\tau)} \cap S_G| + |S_G^{(\tau)} \cap S_P|}{|S_P| + |S_G|} \tag{2}$$

2. Local and Non-Local Dice

Since deep learning-based predictions can result in both local corrections and unintended non-local changes (e.g., spurious activations far from the user-provided input), this study proposes two novel metrics: local Dice and non-local Dice. These metrics provide spatial sensitivity to better evaluate the tool's ability to adhere to user input. Specifically, the local Dice quantifies performance in the immediate proximity of user-provided scribbles, while the non-local Dice captures changes in areas distant from the scribbled regions, identifying unintended effects of the model's prediction. These novel metrics were developed and investigated to answer Research Subquestion 2 (Section II) and provide a more targeted and meaningful evaluation of interactive refinement performance. For these metrics, the masks are divided into a local region and a non-local region. The local region includes the scribble slice and the slices adjacent to it, while the non-local region consists of all remaining slices. In this study, the slices surrounding the scribble slice that define the local region are referred to as the neighborhood (n), where n indicates the number of slices cranial and caudal to the scribble slice (Z_s). A visual representation of the local and non-local Dice regions are shown in Figure 4.

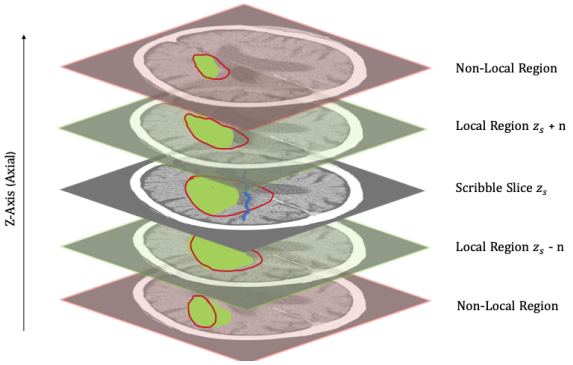


Figure 4. Local and non-local regions for neighborhood example n=1. Green mask: Ground Truth. Red line: Prediction. Blue line: Background Scribble (user input). Local Region: light-green slices + gray scribble slice (Z_s). Non-local Region: light-red slices. n: neighborhood.

The formulas for the local and non-local Dice used in this study are presented below. The local Dice coefficient before refinement quantifies the overlap between the ground truth and the model prediction within the local region surrounding the scribble. The local Dice coefficient after refinement is computed analogously, but uses the refined prediction instead. All variables are defined in Table 2.

$$\Omega_{\text{Local}} = \{v \in V \mid z_S - n \leq z_v \leq z_S + n\} \tag{3}$$

$$\begin{aligned} \text{Dice}_{\text{Local}}^{\text{Before}} &= \text{Dice}\left(G \cap \Omega_{\text{Local}}, P_{\text{before}} \cap \Omega_{\text{Local}}\right) \\ &= \frac{2\,|G \cap P_{\text{before}} \cap \Omega_{\text{Local}}|}{|G \cap \Omega_{\text{Local}}| + |P_{\text{before}} \cap \Omega_{\text{Local}}|} \end{aligned} \tag{4}$$

Similarly, the local Dice after refinement is given by:

$$\begin{aligned} \text{Dice}_{\text{Local}}^{\text{After}} &= \text{Dice}\left(G \cap \Omega_{\text{Local}}, P_{\text{after}} \cap \Omega_{\text{Local}}\right) \\ &= \frac{2\,|G \cap P_{\text{after}} \cap \Omega_{\text{Local}}|}{|G \cap \Omega_{\text{Local}}| + |P_{\text{after}} \cap \Omega_{\text{Local}}|} \end{aligned} \tag{5}$$

The change in local Dice, denoted as Δ_{Local}, is defined as the difference between the post-refinement and pre-refinement local Dice values and reflects the extent to which segmentation accuracy within the local region improves after refinement.

$$\Delta_{\text{Local}} = \text{Dice}_{\text{Local}}^{\text{After}} - \text{Dice}_{\text{Local}}^{\text{Before}} \tag{6}$$

The non-local region consists of all voxels outside the local region, which was indicated by n. The non-local Dice coefficient measures the agreement between the predictions before and after refinement within this non-local region. Ideally, this value remains close to 1, indicating that no unintended changes occur in the non-local area.

$$\Omega_{\text{Non-Local}} = V \setminus \Omega_{\text{Local}} \tag{7}$$

$$\text{Dice}_{\text{Non-Local}} = \frac{2\,|P_{\text{before}} \cap P_{\text{after}} \cap \Omega_{\text{Non-Local}}|}{|P_{\text{before}} \cap \Omega_{\text{Non-Local}}| + |P_{\text{after}} \cap \Omega_{\text{Non-Local}}|} \tag{8}$$

3. Local and Non-local Surface Dice

To also evaluate spatial sensitivity in the local and non-local areas the choice was made to include local and non-local surface Dice as well in this study. The definitions are similar to the local and non-local Dice coefficients. To avoid artificial boundary effects caused by cropping, surface voxels are not extracted from cropped subvolumes. Instead, the surface of each segmentation is first computed on the full volume and subsequently restricted to the local or non-local region, following common practice [79, 80]. This ensures that measured surface differences reflect true segmentation boundaries rather than cropped artificial edges. The formulas for the local surface Dice are presented below. The notation used in the following formulations is summarized in Table 2.

$$\text{SurfaceDice}^{\text{Before}}_{\text{Local},\tau} = \frac{|S^{(\tau)}_{P_{\text{before}},L} \cap S_{G,L}| + |S^{(\tau)}_{G,L} \cap S_{P_{\text{before}},L}|}{|S_{P_{\text{before}},L}| + |S_{G,L}|} \tag{9}$$

Similarly, the local Surface Dice after refinement is given by:

$$\text{SurfaceDice}^{\text{After}}_{\text{Local},\tau} = \frac{|S^{(\tau)}_{P_{\text{after}},L} \cap S_{G,L}| + |S^{(\tau)}_{G,L} \cap S_{P_{\text{after}},L}|}{|S_{P_{\text{after}},L}| + |S_{G,L}|} \tag{10}$$

The change in local surface Dice due to a user-input, denoted as $\Delta_{\text{Local Surface}}$, quantifies the effect of the refinement within the local region and is defined as:

$$\Delta_{\text{Local Surface}} = \text{SurfaceDice}^{\text{After}}_{\text{Local},\tau} - \text{SurfaceDice}^{\text{Before}}_{\text{Local},\tau} \tag{11}$$

The non-local surface Dice measures the agreement between the model predictions before and after refinement in the non-local region ($\Omega_{\text{Non-Local}}$). This metric captures unintended surface changes occurring in the area outside the local area and is defined as:

$$\text{SurfaceDice}_{\text{NL},\tau} = \frac{|S^{(\tau)}_{\text{before,NL}} \cap S_{\text{after,NL}}| + |S^{(\tau)}_{\text{after,NL}} \cap S_{\text{before,NL}}|}{|S_{\text{before,NL}}| + |S_{\text{after,NL}}|} \tag{12}$$

4. Slices without Ground Truth or Prediction

It is possible that the ground truth mask, the prediction mask, or both, are empty for some slices or regions. To ensure consistent handling of such cases across metrics, we apply the same evaluation procedure to volumetric Dice and surface Dice. For Dice, surface Dice, and non-local metrics, a region in which both segmentations contain no voxels is assigned a score of 1.0. This score indicates that no unintended differences occur in the non-local area. For local metrics, cases in which both surfaces are absent are considered undefined and are therefore marked as *NaN*, as the computation would otherwise involve division by zero. These cases are excluded from the analysis in order to measure only changes in overlap between the ground truth and the prediction.

F. Neighborhoods

The variable n specifies the extent of the local region, which in turn defines the size of the non-local region. When a high value of n is chosen, the local region expands while the non-local region contracts. As this metric is introduced for the first time and no exact value for n is defined, the analysis focuses on three different neighborhood sizes: ±1, ±5, and ±10. This choice was made based on the original work [23] and the hypothesis that the interactive contour refinement model works with a gaussian curve effect around the scribble slice, with the highest effect around the scribble slice. This means that, when placing a scribble, the strongest refinement effect is observed in the scribble slice, gradually decreasing toward minus slices and plus slices, until no effect is observed. Because it is not known how large this effect is, the choice was made to focus on these three neighborhoods (±1, ±5, and ±10) to capture a large area around the scribble slice and to evaluate where the local effect is the highest. A visual representation of this Gaussian

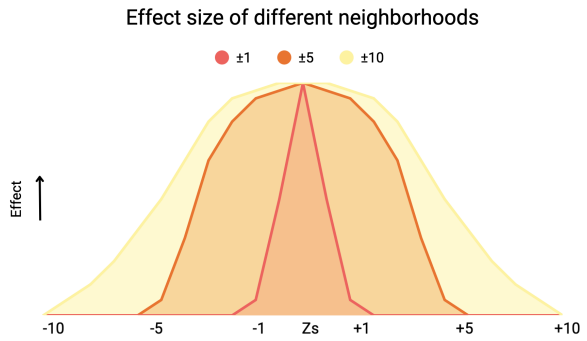


Figure 5. Illustration of the hypothetical effect of user input across the scribble slice (z_S) and its surrounding slices for different neighborhoods (±1, ±5, and ±10).

effect of the user input in the scribble slice is shown in Figure 5.

G. Ethical Approval Statement

This study is based on a publicly available dataset [75], that is fully anonymized.

IV. Results

In this section the outcomes of this study will be presented. This will be done based on the two research sub-questions on robustness (Section IV.A) and trustworthiness (Section IV.B). The results of both Experiment 1 (Section III.D.1) and Experiment 2 (Section III.D.2) are used to investigate these aspects.

A. Robustness

1. Experiment 1

As described in Section III.D.1, users were asked to perform a single refinement on two slices selected based on anatomical complexity: one relatively easy slice and one more challenging slice. These slices, together with the automatic predictions from the auto-contouring model, are shown in Figure 3. After the scribble input of all six users, the contour-refinement model updated the segmentation results, which are shown in Figure 6. Several observations can be made based on these results.

As users were free to choose how to place their scribbles, different interaction strategies emerged. Based on the visual characteristics of the scribbles, three distinct categories of user input were identified (Figure 6):

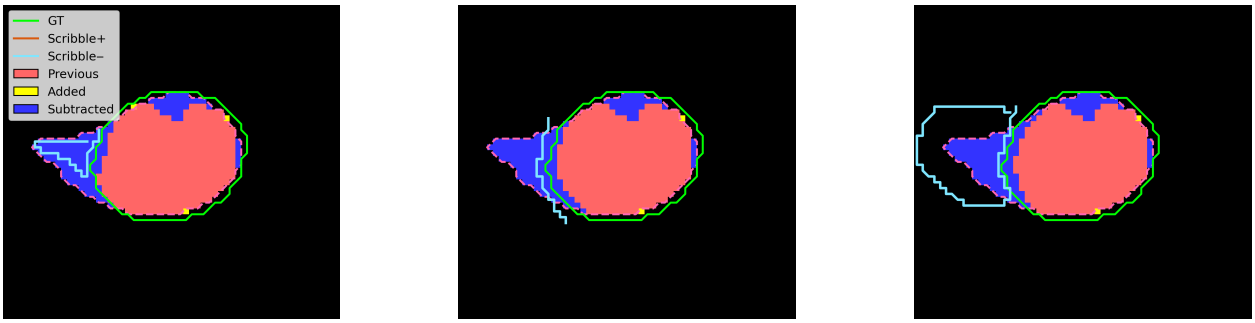
- **Contour-following correction**, where the scribble was placed along the erroneous segmentation boundary.
- **Region-based correction**, where a larger oversegmented area was covered in a single scribble.
- **Localized area-based correction**, where a compact scribble was placed locally at the location corresponding to the ground truth.

For the easy slice (P4), three users applied a contour-following correction strategy (users 2, 4, and 5), two users employed a region-based approach (user 3 and 6), and one user used a localized area-based approach (user 1). For the more challenging slice (P3), all users employed a similar interaction strategy, which can be characterized as a contour-following correction and a localized area-based correction.

Prior to user interaction, the overlap between the automatic segmentation produced by the auto-contouring model and the ground truth was quantified using Dice. Baseline Dice scores were 0.7258 (P4) and 0.8649 (P3). Corresponding surface Dice values with a 2 mm distance threshold were 0.7067 for P4, and 0.8800 for P3. For each user, the Dice values after refinement and the corresponding differences with respect to the Dice values before refinement are reported in Table 3. For the easy slice (P4), an increase in Dice scores is observed for all users (range 0.1296 to 0.1407). The highest change in Dice value (0.1407) was observed for the user who applied a localized area-based scribble, followed by region-based correction (0.1381 and 0.1366), while the lowest Dice values were observed for users who performed contour-following correction (0.1356, 0.1326, and 0.1296). For the difficult slice (P3), a decrease in Dice was observed despite visually correcting the undersegmented region results (range -0.0997 to -0.1015).

As presented in Figure 6, the model reacts in a highly consistent manner across all users, despite differences in user input (foreground versus background) and the resulting increase or decrease in Dice. For P4, the oversegmented region, that needed to be corrected, was correctly removed. However, for two users applying the contour-following correction, a small

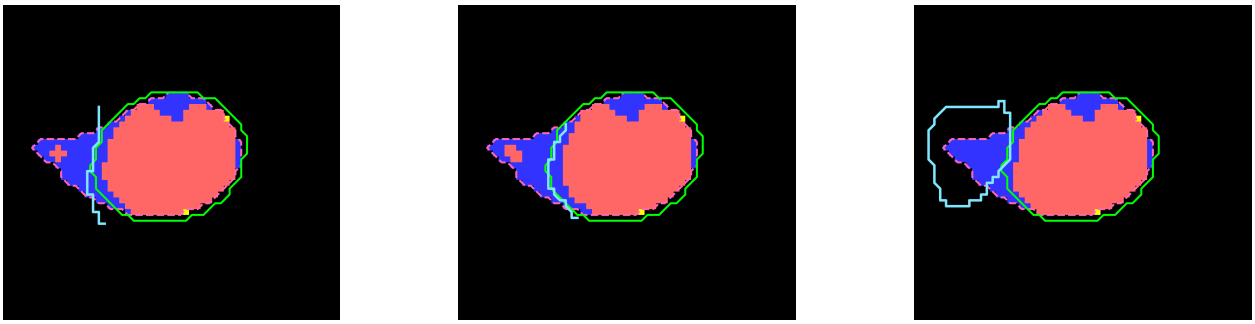
(a) P4 - Easy slice



(a) User 1 (Localized area-based correction)

(b) User 2 (Contour-following correction)

(c) User 3 (Region-based correction)

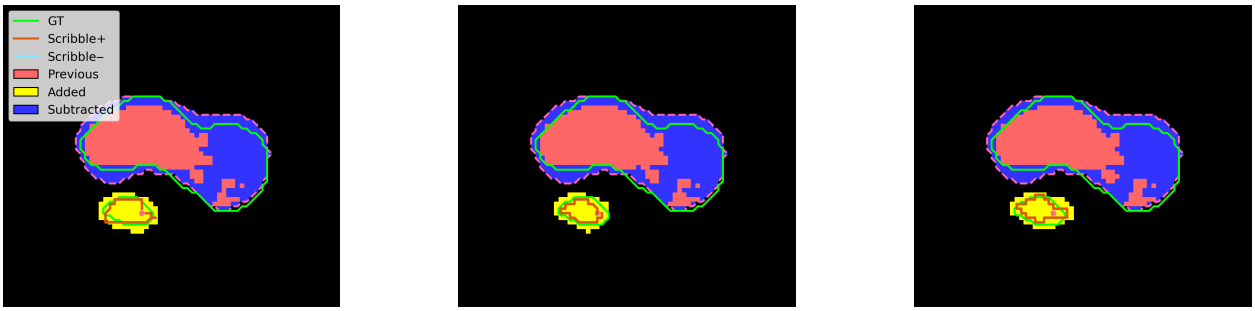


(d) User 4 (Contour-following correction)

(e) User 5 (Contour-following correction)

(f) User 6 (Region-based correction)

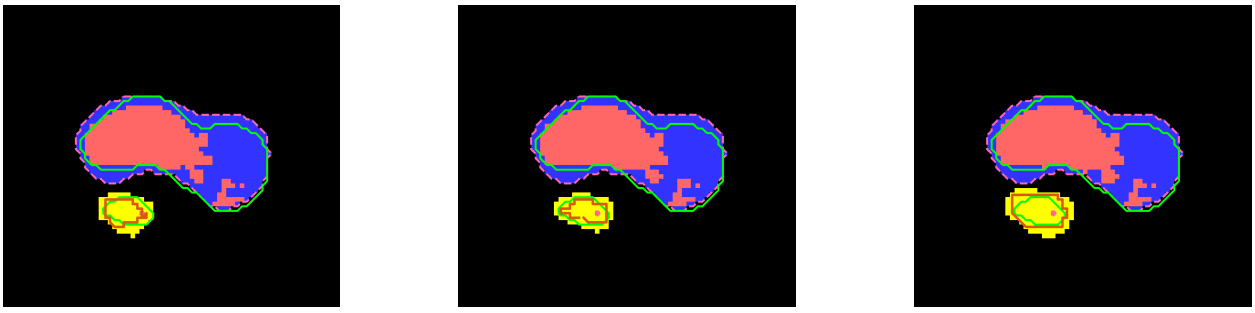
(b) P3 - Difficult slice



(g) User 1

(h) User 2

(i) User 3



(j) User 4

(k) User 5

(l) User 6

Figure 6. Segmentation masks after refinement for P4 (easy slice) and P3 (difficult slice). Green line: Ground Truth, Orange line: Plus Scribble, Light-blue line: Minus Scribble, Pink Mask/dashed line: Previous segmentation mask before refinement, Yellow Mask: Added voxels after refinement, Blue Mask: Subtracted voxels after refinement.

Table 3. Segmentation performance and change after interactive refinement for the easy slice (P4) and difficult slice (P3). Dice Similarity Coefficient (Dice) and surface Dice at 2 mm distance threshold are reported. Positive Δ values indicate improvement in overlap in segmentation masks before and after refinement, negative values indicate deterioration.

User	Performance		Refinement change	
	Dice	sDice	ΔDice	ΔsDice
Easy slice (from P4)				
1	0.8665	0.9152	+0.1407	+0.2085
2	0.8614	0.9098	+0.1356	+0.2031
3	0.8624	0.9144	+0.1366	+0.2077
4	0.8584	0.9029	+0.1326	+0.1962
5	0.8554	0.8947	+0.1296	+0.1880
6	0.8639	0.9136	+0.1381	+0.2069
Difficult slice (from P3)				
1	0.7634	0.7224	−0.1015	−0.1576
2	0.7647	0.7238	−0.1002	−0.1562
3	0.7652	0.7255	−0.0997	−0.1545
4	0.7652	0.7241	−0.0997	−0.1559
5	0.7635	0.7213	−0.1014	−0.1587
6	0.7647	0.7229	−0.1002	−0.1571

portion of the oversegmented region remained. In addition, a small region at the superior part of the tumor was removed for all users, representing an incorrect refinement. Another notable observation is that a few voxels (highlighted in yellow) were consistently added to the segmentation for all users. Thus, the model correctly incorporated the previously missing pixels.

For the foreground scribble results on the P3 slice, several trends can be observed. The user inputs provided to the contour-refinement model were highly similar across all participants, and the resulting model response also showed minimal variability. For all users, placing a foreground scribble in the undersegmented region led to the removal of a previously correctly segmented part of the tumor, while incorrect it was consistent irrespective of the slight variation in the user scribbles. At the same time, the region that was missing in the initial segmentation was successfully added to the prediction. However, for all users, this newly added region was slightly oversegmented when compared to the ground truth for this slice. As a result, the Dice values decreased after refinement (range -0.0997 to -0.1015) (Table 3).

2. Experiment 2

In this experiment, all six users completed one experiment using the AI pencil and one experiment using the manual brush, for all six patients. Figure 8 presents the results from these experiments, including volumetric Dice and surface Dice values. Figure 20 in Appendix Section VIII.E presents the trajectories of each user during their AI and manual experiments. This figure provides a clear idea of how the individual users performed the experiments. After analyzing the experimental results, the decision was made to exclude User 6 from the quantitative analysis of this experiment. Several reasons motivated this decision, which are explained in Appendix VIII.E together with the corresponding graphs.

What can be observed from the graphs from Experiment 2 (Figure 8) is that, at the beginning, Dice values increase (except for the case of P3 where they decrease) more rapidly between iterations when using the AI pencil than when using the manual brush. The AI experiments achieved a significantly larger proportion of its total Dice improvement within the first 5% of the experiment compared to manual annotations (27.4% vs. 6.4%, Wilcoxon p = 0.047), although one patient (P3) exhibited an initial decrease in Dice score. Without this patient, the AI experiments accounted for 47% of the total Dice improvement in the early phase, compared with 6% for manual annotation ($p < 10^{-7}$). The first 5% of the experiment refers to the interval between 0% and 5% of the normalized iteration axis per experiment, where 0% corresponds to the first iteration and 100% to the final iteration of the experiment. All experiments were normalized to this 0-100% scale to allow comparison across experiments with different numbers of iterations. Figure 7 shows the boxplots of these first 5% for all six patients with median early Dice gain and interquartile ranges, highlighting a clear shift toward higher early gains for AI compared to manual annotation. The outliers shown in this

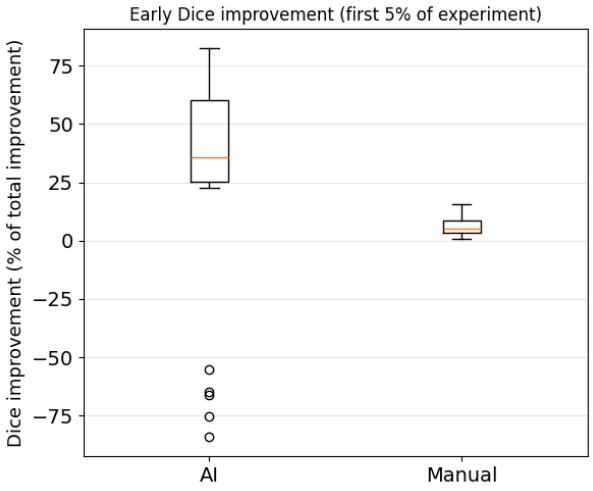


Figure 7. Boxplot showing the percentage of the total Dice score improvement that occurs during the first 5% of the AI (N=30) and manual (N=30) contouring experiments.

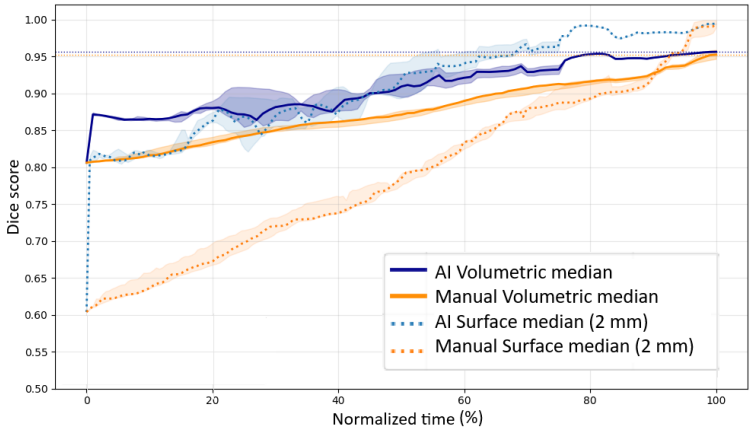
boxplot are all associated with Patient 3. The y-axis represents the Dice improvement during the first 5% of the experiment, expressed as a percentage of the total Dice improvement achieved over the entire experiment.

After this initial increase, the Dice values rise more gradually, with occasional decreases occurring during the later stage (Figure 8). By contrast, the manual brush exhibits a more consistent but slower upward trend over the full course of the experiment. The surface dice follows these trends of the volumetric Dice, but is more accurate for changes of the contour instead of the entire volume. Another noticeable result is that the refinement behavior during these experiments was very similar across users (indicated by the relatively small bands), but differed substantially between patients. For Patient 3, Dice bands ranged from 0.02 to 0.20, whereas for the other patients the bands remained between 0.00 and 0.05. At the end of the experiment, AI-assisted experiments (N = 30) achieved a final mean Dice score of 0.9303, whereas manual experiments (N = 30) reached a higher final mean Dice score of 0.9477.

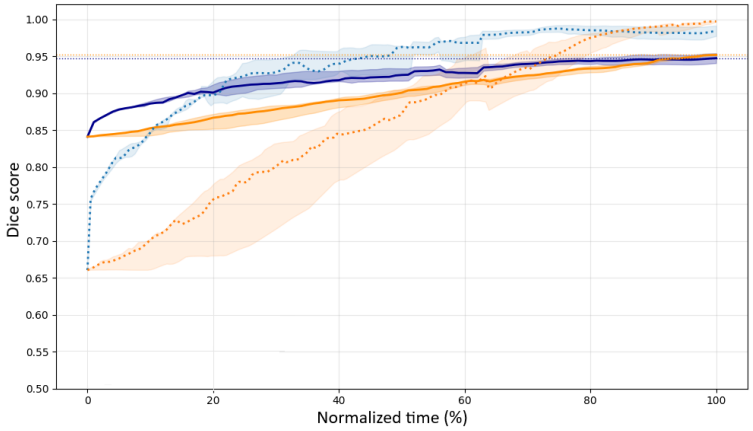
For Patient 3, a drop in Dice scores is observed at the beginning of the experiment (Figure 8). When examining the 3D visualizations of these masks (Figure 9), it can first be observed that the auto-contour model segmented the tumor volume reasonably well. However, an unexpected part, superior to the tumor, was incorrectly segmented. All users tried to remove this part from the segmentation by placing a background scribble in this part of the mask. As shown in Figure 9 this part was partly removed, however, this action also removed a correctly segmented region. This resulted in a decrease in Dice score from 0.8649 to 0.7592 after this first refinement step. Similar decreasing behavior was observed in subsequent iterations and was consistent across all users.

As mentioned earlier, the Dice values for the other patients included in this study show a rapid increase. All auto-contours and their corresponding ground truths are shown in Appendix Section VIII.C. In addition, all refinement processes after the first user interaction are shown in Appendix Section VIII.D. For cases where the initial auto-contour differs substantially from the ground truth, changes in the mask can be observed after the first user interaction.

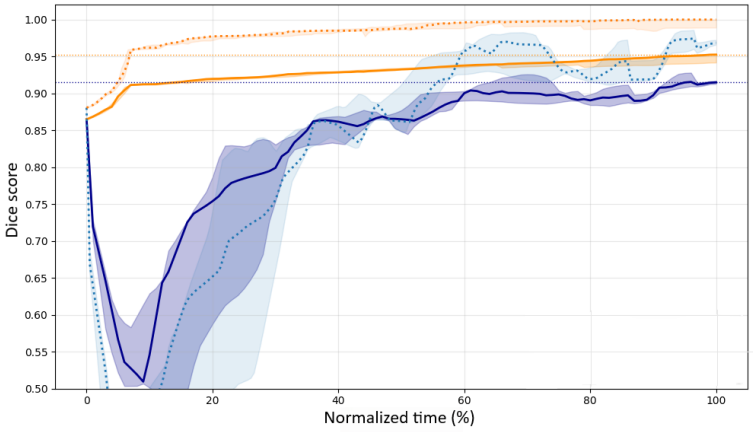
Following the large corrections at the beginning of the refinement process, drops in Dice are more frequently observed in the middle or toward the end of the AI pencil experiment. This occurred mainly in three situations: when users attempted to refine slices at the boundary of the tumor volume, when users segmented slices that included complex anatomical structures like contour curvatures around air cavities (e.g. near the trachea), and when users provided incorrect user input. Figure 10 illustrates these cases. It also shows an example of CT distortion caused by a dental implant in Patient 1, which also negatively affects the accuracy of the segmentation. More examples of erroneousness segmentation masks for each category are presented in Section VIII.F.



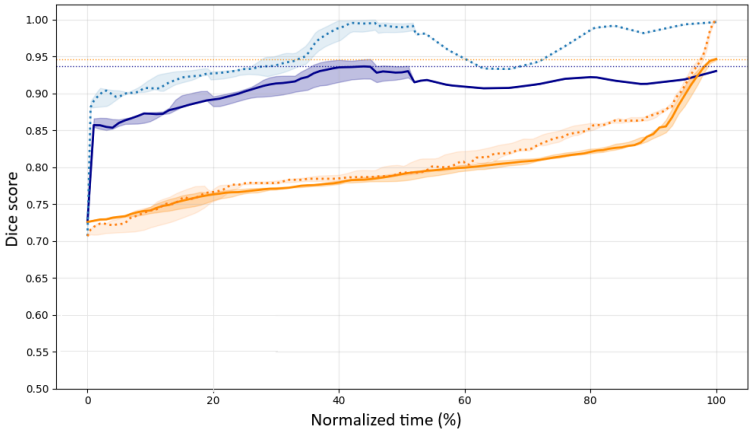
(a) Patient 1



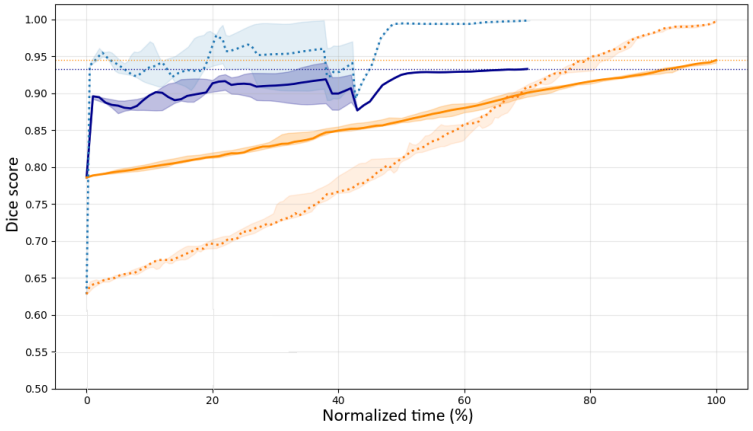
(b) Patient 2



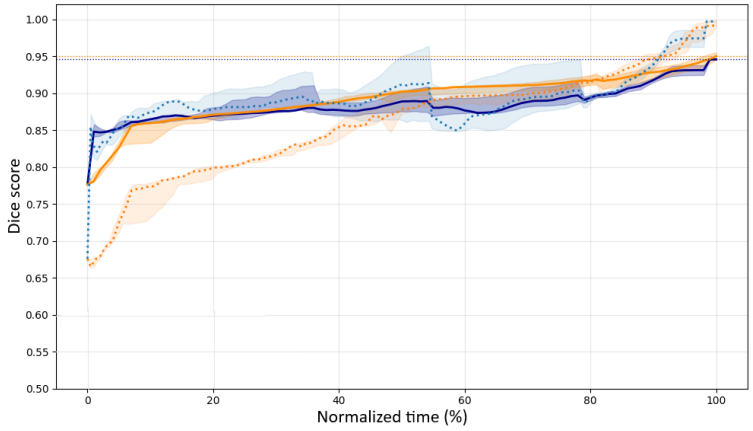
(c) Patient 3



(d) Patient 4



(e) Patient 5



(f) Patient 6

Figure 8. Dice and Surface Dice trajectories (User 1–User 5) plotted against the normalized time axis for all six patients. Median (line) and interquartile range (band) are shown for Users 1 to 5.

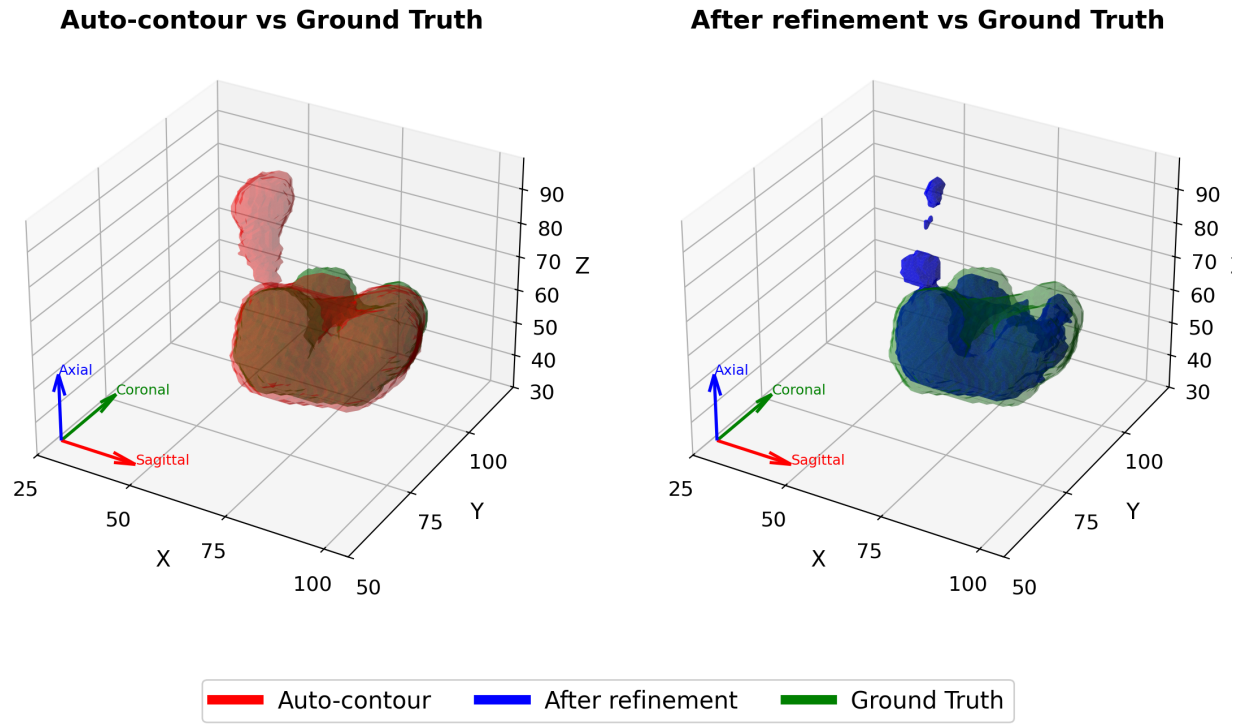


Figure 9. Overlay between ground truth and segmentation masks before (between auto-contour produced by the auto-contouring model and ground truth masks) and after user refinement (between contour prediction produced by the contour-refinement model and ground truth).

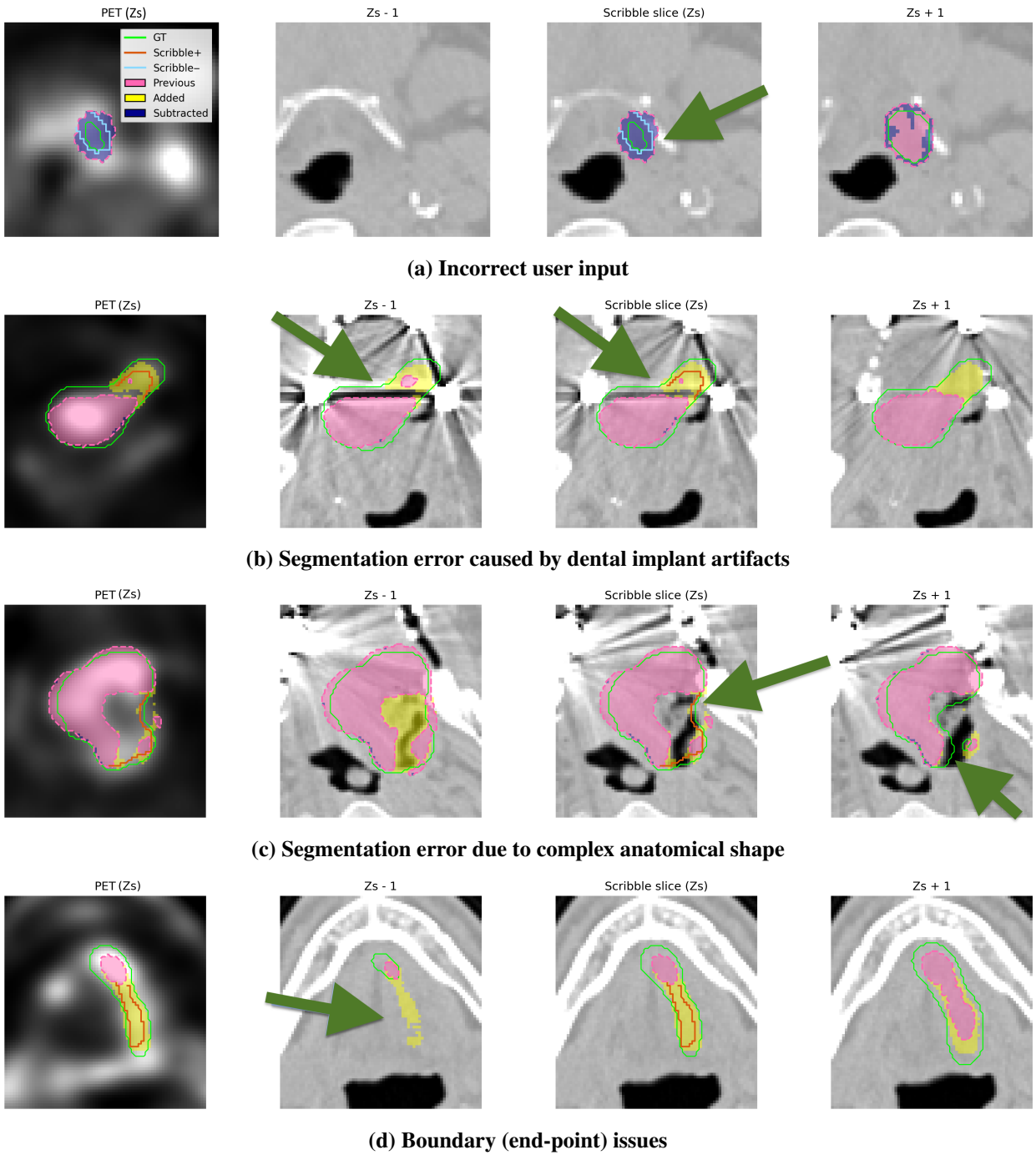


Figure 10. Overview of common issues in segmentation masks after user refinements. PET slice corresponding to the scribble slice is shown in the left figure, remaining figures are CT, with the central CT slice corresponding to the scribble slice. Green line: Ground truth. Orange line: User refinement (foreground scribble). Light blue line: User refinement (background scribble). Dashed pink line: Contour previous segmentation mask. Pink mask: Previous segmentation. Yellow mask: Added region after refinement. Blue mask: Removed region after refinement.

B. Trustworthiness

1. Experiment 1

Table 4 shows the Δ local Dice, non-local Dice, and surface Dice values obtained after the single refinement step in Experiment 1 (III.D.1). Mean values are included and were computed across all six users. For the easy slice (Patient 4), local improvement is observed for all users in the ±1 neighborhood, with the exception of user 5, who exhibits a minor decrease in Dice (-0.0075) in this area. All Δ local surface Dices increased for Patient 4. Larger improvements, however, are evident for the ±5 and ±10 neighborhoods in both Dice and Surface Dice metrics. Between neighborhoods ±5 and ±10, only minor differences (0.000 – 0.0041 (Dice) and (0.0024 – 0.0159 (surface Dice)) in metric values are observed. The strongest changes occur within neighborhoods located approximately two to three slices away from the scribble location, extending up to five slices. This behavior is shown in Figure 24 in Appendix Section VIII.G, where tumor regions above the scribble slice are correctly added, while regions below the tumor are correctly removed. These results show that the best local segmentation performance for this case is achieved for a neighborhood size of ±5 slices. The non-local Dice values, indicating differences outside the local interaction region, range between 0.7489 and 0.6072. Furthermore, the non-local Dice decreases as the neighborhood size increases. All surface Dice values follow the same behavior as the volumetric Dice values.

For Patient 3, a consistent decrease in local Dice and surface Dice values is observed across all users and neighborhood sizes, as shown in Table 4. The mean Δ local Dice shows larger decreases for the ±1 and ±5 neighborhoods (−0.1788 and −0.1807, respectively) compared to the ±10 neighborhood (−0.1490). The Δ Local surface Dice showed the biggest change in ±10 slices around the scribble. As shown in Figure 25 in Appendix Section VIII.G, a large portion of the correct prediction mask is removed in all five slices above the scribble slice. Similarly, parts of the mask are incorrectly deleted in the five slices below the scribble slice. The decrease in local Dice values is similar for all users, in contrast to Patient 4, where greater variation between users was observed. For example, for the ±1 neighborhood, inter-user variation in Δ local Dice was larger for Patient 4 (range: -0.0075 to 0.0305) than for Patient 3 (range: -0.1899 to -0.1719). In contrast to the Δ local Dice, the non-local Dice in Patient 4 remains relatively stable across different neighborhood sizes. However, the magnitude of the values indicates that changes occur in the non-local region, since a non-local Dice value of 1.0 represents perfect overlap in the non-local area. The region outside the ±10 neighborhood shows the lowest overlap before and after refinement.

2. Experiment 2

Local Dice

Figure 11 illustrates the distribution of all Δ local Dice values (with a neighborhood of ±5) for Experiment 2. The ±5 neighborhood was chosen because it became evident from Experiment 1 that the most changes occur within this area. The figure includes an aggregate boxplot of all data points, alongside zoomed boxplots that provide detailed insight into the per-patient variations. A Wilcoxon signed-rank test across all paired observations (n = 30) revealed a significant difference between AI and Manual measurements ($W = 27.0$, $p = 2.35 \times 10^{-6}$). Median values were 0.0050 for AI and 0.0025 for Manual, with a median paired difference of 0.0036. The paired rank-biserial effect size was 0.884, indicating a large non-parametric effect. Per-patient analyses showed effects in the same direction, however did not reach significance due to the limited statistical power (n = 5 per patient).

Both the AI-assisted and manual strategies predominantly produce positive values. So, overall, the model more frequently refined the segmentations in the local region following user input, thereby improving segmentation quality. In general, the experiments using the AI pencil show the highest values in Δ local Dice across all experiments as well as for each individual patient. In contrast, the manual experiments exhibit a more stable median and overall behavior. Compared to the AI experiments, the manual experiments required a substantially higher number of iterations to segment the entire volume. Appendix Section VIII.H presents additional Δ local Dice analyses, including plots per iteration for a single user and corresponding histograms.

Table 4. Dice and surface Dice results (Δ local and non-local) in a neighborhood of ±1, ±5, and ±10 slices around the scribble slice, reported per user for the easy slice (P4) and difficult slice (P3). Non-local values represent the overlap between ground truth and the refined prediction outside the defined local neighborhood. A ±1 neighborhood corresponds to a local region of three slices, with the non-local region comprising all remaining slices.

User	Dice						Surface Dice					
	Δ Local			Non-local			Δ Local			Non-local		
	±1	±5	±10	±1	±5	±10	±1	±5	±10	±1	±5	±10
Easy slice – P4												
1	0,0305	0,0404	0,0368	0,7483	0,7077	0,6069	0,0866	0,1151	0,1022	0,7709	0,7685	0,7214
2	0,0176	0,0302	0,0289	0,7490	0,7097	0,6074	0,0852	0,1087	0,0928	0,7717	0,7721	0,7721
3	0,0103	0,0282	0,0302	0,7447	0,7066	0,6067	0,0846	0,1133	0,1001	0,7655	0,7678	0,7205
4	0,0043	0,0220	0,0239	0,7504	0,7103	0,6075	0,0574	0,0903	0,0805	0,7786	0,7741	0,7228
5	-0,0075	0,0149	0,0190	0,7531	0,7117	0,6077	0,0094	0,0686	0,0662	0,7809	0,7759	0,7235
6	0,0167	0,0326	0,0326	0,7477	0,7074	0,6069	0,0852	0,1135	0,0992	0,7687	0,7692	0,7218
Mean	**0.0200**	**0.0281**	**0.0286**	**0.7489**	**0.7089**	**0.6072**	**0.0681**	**0.1016**	**0.0902**	**0.7727**	**0.7713**	**0.7304**
Difficult slice – P3												
1	-0,1814	-0,1816	-0,1500	0,6998	0,7002	0,6330	-0,2190	-0,1992	-0,2466	0,5886	0,5589	0,5642
2	-0,1719	-0,1803	-0,1487	0,6995	0,7000	0,6319	-0,2142	-0,1988	-0,2460	0,5846	0,5535	0,5578
3	-0,1735	-0,1798	-0,1484	0,7011	0,7013	0,6336	-0,2113	-0,1966	-0,2446	0,5905	0,5596	0,5628
4	-0,1812	-0,1822	-0,1491	0,7016	0,7026	0,6356	-0,2116	-0,2001	-0,2468	0,5902	0,5604	0,5648
5	-0,1746	-0,1800	-0,1492	0,6990	0,6990	0,6307	-0,2205	-0,1988	-0,2463	0,5854	0,5543	0,5571
6	-0,1899	-0,1801	-0,1486	0,7061	0,7057	0,6395	-0,2157	-0,1951	-0,2435	0,6016	0,5707	0,5695
Mean	**-0,1788**	**-0,1807**	**-0,1490**	**0,7012**	**0,7015**	**0,6341**	**-0,2154**	**-0,1981**	**-0,2456**	**0,5902**	**0,5596**	**0,5627**

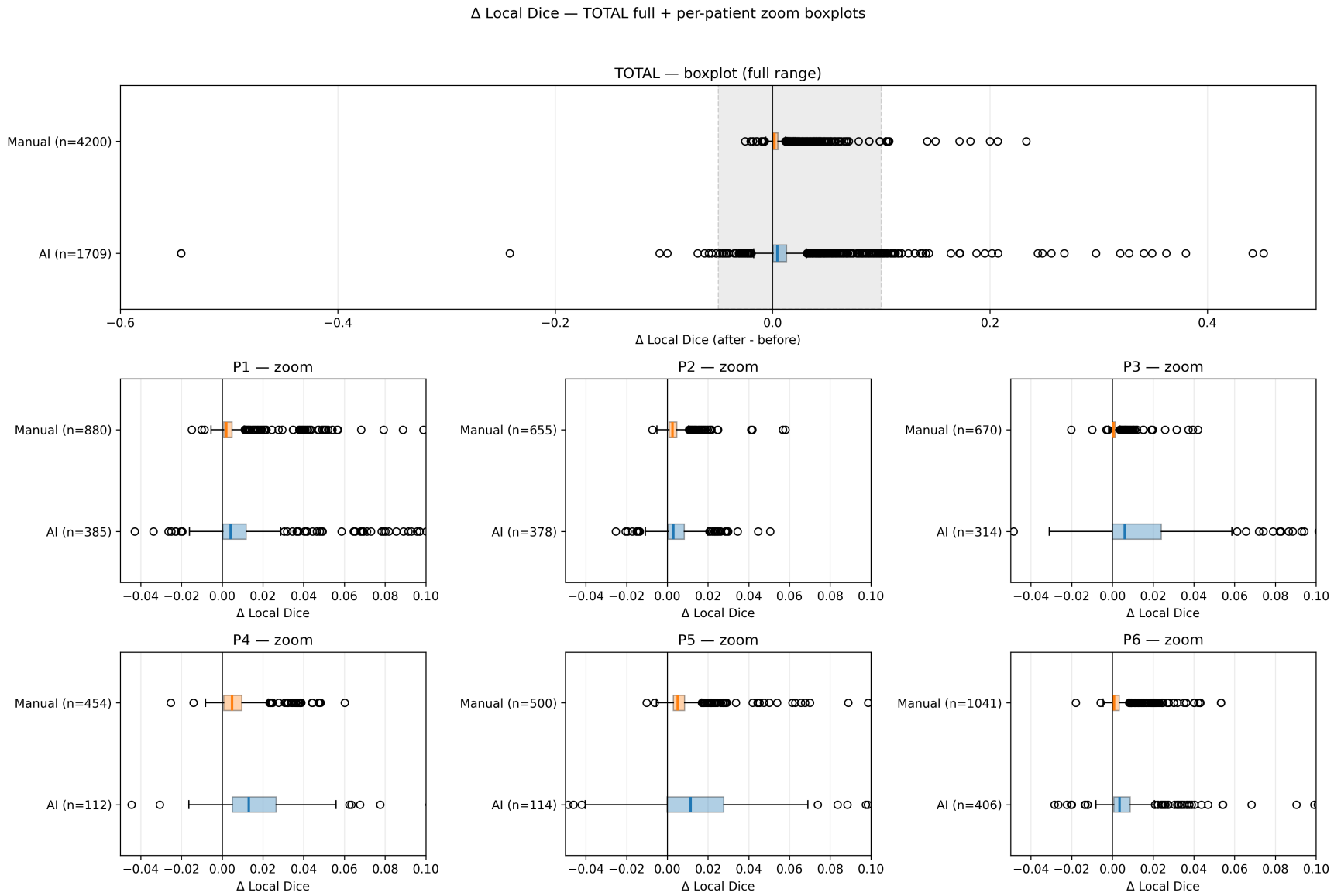


Figure 11. Boxplots of Experiment 2: Presenting the results of all experiments across users and patients, accompanied by zoomed-in boxplots displaying per-patient results for all users.

Non-Local Dice

Figure 12 presents non-local Dice results for one user and one patient plotted together with the volumetric Dice during experiment 2. Results are shown for one AI experiment and for a manual experiment on Patient 4. For the AI experiment a strong increase in non-local Dice is observed at the beginning, reflecting a positive change in Dice between the masks before and after refinement. This effect occurs only at the beginning. Subsequently, the non-local Dice quickly reaches 1.00 and remains around this value during the remainder of the experiment. In this stage a much more stable effect is observed compared to the Δ local Dice. However, for certain iterations, small changes (maximum of 0.02) are observed in the non-local Dice. In these cases, the non-local Dice first drops, but is subsequently restored in later refinement steps. Comparing the results of the experiments using the AI pencil and the manual brush in Figure 12, it is clearly observed that, in the manual experiment, the non-local Dice remains exactly 1.00 for the entire duration of the experiment since there is no non-local effect for manual delineations. The volumetric Dice shows a consistent upward trend, with a steeper increase toward the end of the experiment. More non-local Dice results are shown in the Appendix Section VIII.I. These graphs show that the same effect is observed for the other patients.

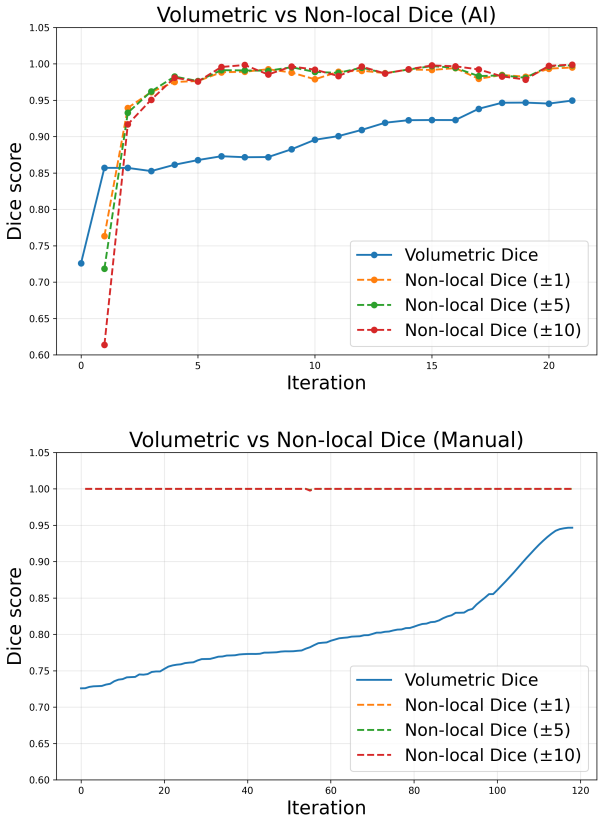
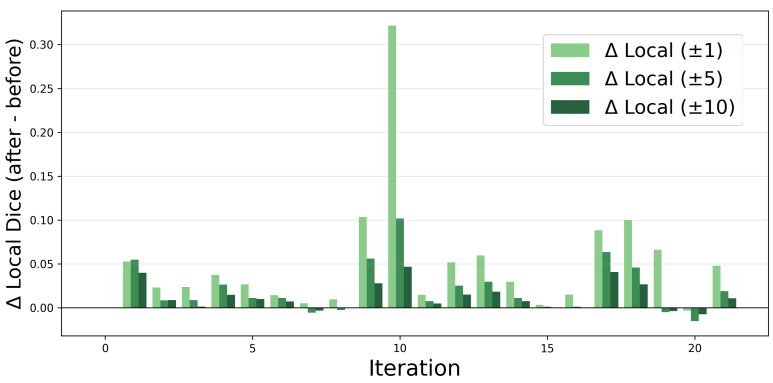


Figure 12. Non-local Dice results (for the three different neighborhoods ±1, ±5, and ±10) plotted against volumetric Dice results for the AI experiment on Patient 4 performed by one user.

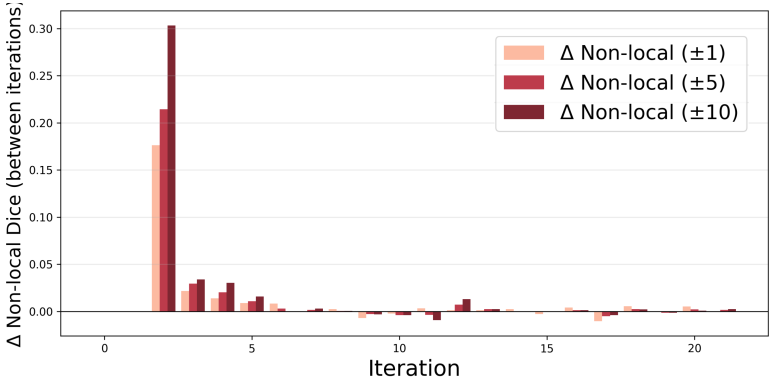
Neighborhoods

This study evaluates the effect of user input across different neighborhoods. Neighborhoods are defined as the number of slices encompassing the local area, as illustrated in Figure 4. Figure 13 presents results from one user during the experiment using the AI pencil for local and non-local Δ (surface) Dice values for three neighborhoods: ±1, ±5, and ±10. Several patterns are visible in these graphs. Δ local Dice is present for the majority of iterations in all three neighborhoods, showing that in most iterations the model refines an area around the scribble slice. The highest Δ local Dice values per iteration are observed within the ±1 neighborhood (Figure 13). Local Dice results for different neighborhoods for the remaining five patients are presented in Appendix Section VIII.J. These figures indicate that the same effects are observed across all six patients.

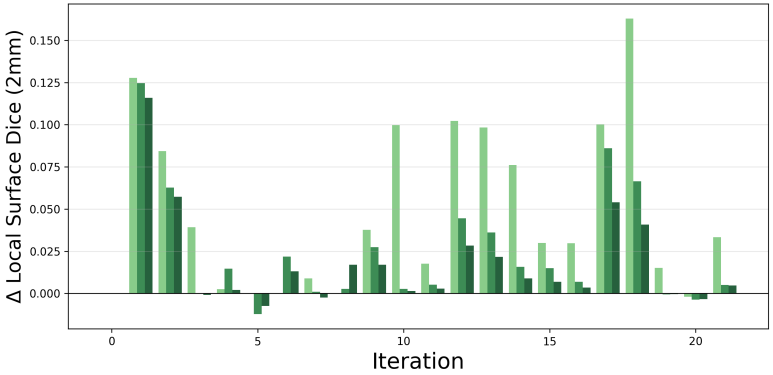
For the Δ (surface) non-local Dice (Figure 13), indicating the change between non-local results between iterations, a different trend compared to the local Dice is observed. High values appear at the beginning of the experiment, while only minor values occur during the middle and later stages. For Patient 4 (Figure 13), the highest Δ non-local Dice values are observed outside the ±10 neighborhood. Across the other patients (Appendix Section VIII.J), mixed patterns appear: in some cases (Patients 1, 3, and 6) the area outside the ±1 neighborhood shows the largest change, whereas in others (Patients 2, 4, and 5) the area outside the ±10 neighborhood exhibits the highest values. Nevertheless, for all patients, high positive Δ non-local Dice values occur at the beginning, with only minor values thereafter. For Patient 4, improvements were higher for the volumetric non-local Dice, whereas for the other patients the non-local surface Dice showed higher values.



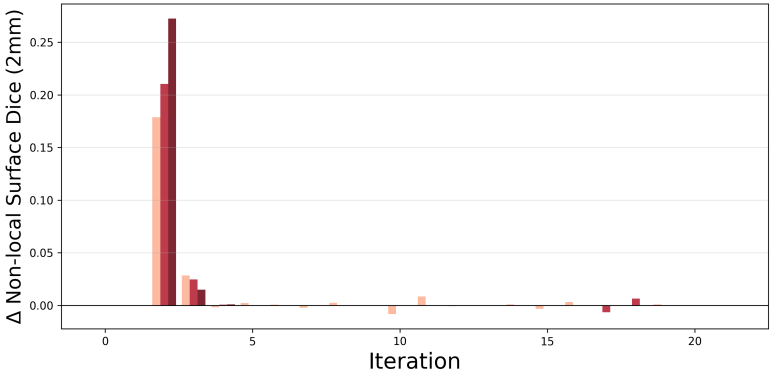
(a) Δ Local Dice results



(b) Δ Non-local Dice results



(c) Δ Local surface Dice results



(d) Δ Non-local surface Dice results

Figure 13. Δ Local and non-local Dice and surface Dice results across neighborhoods (±1, ±5, and ±10) for one AI experiment of Patient 4. (a) Δ Local Dice, (b) Δ Non-local Dice, (c) Δ Local Surface Dice, and (d) Δ Non-local surface Dice. Iteration 0 represents the auto-contour. Δ non-local values begin at iteration 2, because non-local Dice measures the change between consecutive predictions. The first non-local value, at iteration 1, corresponds to the Dice between the auto-contour and the prediction following the first refinement in the non-local area. The non-local value at iteration 2 represents the overlap between the prediction before and after refinement 2. The Δ non-local Dice represents the change between these non-local Dice values.

V. Discussion

In H&N radiotherapy, it is important to speed up the tumor segmentation process. This is a time consuming process that can be significantly accelerated by introducing auto-contouring models. However, fully automatic contouring models for tumor target segmentation in the H&N region often fail to meet the stringent requirements for reliable clinical deployment [18, 22, 81]. As a result, medical experts need to manually adjust the auto-contours generated by these models, which offsets the gains from auto-contouring. Therefore, this study focused on validating previous work [23] that refines these auto-contours in 3D by introducing an interactive contour refinement tool based on a DL model. The performance of this model is evaluated based on its robustness and trustworthiness, i.e., whether it generates contours where the user expects them, both of which are important for future clinical applications. The results show that the model follows the users' intent for both aspects, indicating promising potential for future application.

A. Robustness

1. Scribble Types

Scribbles are considered a user-friendly prompt type because they allow users to freely draw annotations [81, 82]. Experiment 1 demonstrated that users employed different approaches when placing scribbles (Figure 6). Within this study, a custom taxonomy was defined, consisting of contour-following correction, region-based correction, and localized area-based correction (Section IV.A.1). This study's observation of varying scribble approaches between users, and sometimes within a single user, aligns with previous work reporting wide variability in scribble types among users [83]. The model showed small performance differences for the various scribble types, with the highest values for the localized area-based scribble (0.1407 vs. 0.1296), indicating that the placement of the scribble potentially influences the model behavior. Previous work has shown that interactive scribble-based methods outperform traditional methods due to their enhanced accuracy and flexibility, thereby reducing user workload [82]. However, the results of this study indicate that it is also important to consider which types of scribbles are possible and which perform best. To address these differences in scribble approaches between users, Wong et al. [84] demonstrated that training the model using different scribble types (simulated line, centerline, and contour scribbles) improves model performance. The findings of the present study confirm the importance of this approach for proper model performance.

2. Similarities in model behavior

The differences observed in this study in the model's response across users were very minor, both quantitatively and qualitatively. Regardless of the scribble type (foreground or background), the scribble form (contour-following, region-based, or localized area-based), or the individual user, the model refined or reduced the same region of the tumor. Together, these findings indicate that the model exhibits robust and consistent behavior. This consistent behavior is clearly visible in the results of Experiment 1 (Figure 6), where the model's performance was evaluated in two different slices. For example, the small isolated pixels added in the segmentation of Patient 4 (Figure 6), indicate high consistency across users. Across all users, the model refined or reduced nearly the same regions of the tumor volume. Another observation is that user input in the relatively easy slice (Patient 4) resulted in positive refinement, whereas user input in the more challenging slice (Patient 3) led to a decline in the overall Dice and surface Dice values for all users. This indicates that the refinement model is highly sensitive to the anatomical structure of the tumor and its surroundings, while being less sensitive to differences in user input. Thus, refinement behavior is more strongly influenced by patient-specific anatomy than by inter-user variability.

Although different scribble behavior was observed for Patient 4, the model still showed highly consistent performance. When these results are analyzed over the complete duration of tumor segmentation in Experiment 2 (Figure 8), this similarity between users is also clearly visible. For all patients, with the exception for Patient 3, refinement steps using the AI pencil contribute to a rapid increase in Dice, most notably in the early phases of the experiments.

In later stages of the experiments, the results show more fluctuation between users. This indicates that incorporating user input leads to a significant improvement of the auto-contour, while further improvement in tumor delineation slows down for later iterations. However, the overall effect indicates improved efficiency, as users require significantly fewer iterations to complete the task compared with using the manual brush.

3. Errors in segmentation performance

Fluctuations in Dice values were observed particularly in the following situations: incorrect user input, distortions by dental implants, anatomically complex regions, and segmentation of slices at the tumor boundaries. Figure 10 illustrated these slices with erroneous segmentation results observed in this study.

A representative example of the model's difficulty in contouring the tumor volume is observed in Patient 3, where the tumor was located near an air cavity at the height of the pharynx. The difficulty encountered by the model for these slices aligns with earlier research. Mansoor et al. [85] found that the presence of air cavities can lead to inaccurate boundary identification. One possible explanation for the increased segmentation difficulty of tumors near air cavities is their complex anatomical shape and the high differences at these places between patients, whereas many DL models are trained on more commonly occurring spherical shapes. Song et al. [86] reported that, in paranasal sinus segmentation, the lowest model performance was observed for cases with the highest anatomical complexity. Due to their complex geometry and interleaved structures, boundary ambiguity can arise, leading to a reduced performance of DL models such as CNNs [81, 86].

Another possible reason for these erroneous predictions is that the model also incorporates Positron Emission Tomography (PET) imaging data as input. The hypothesis is that the model used in this study relies heavily on this imaging data due to its strong ability to distinguish regions of tumor activity. However, PET has relatively poor spatial resolution, which is disadvantageous because its distribution is sampled on a voxel grid. Therefore, most voxels include different types of tissues, as the contours of the voxels do not match the actual contours of the tracer distribution [87]. As a result, the PET signal may spill over the tumor boundaries, including into adjacent air cavity regions.

Near the cranial and caudal endpoints of the tumor volume, the model also showed segmentation difficulties. At these locations, the tumor ends, such that one slice contains tumor tissue while the adjacent slice no longer does. As the model performs refinements in 3D, it considers not only the current slice but also its surrounding slices. When the tumor volume abruptly ends, the model may have difficulty identifying the exact boundaries. This results in prediction errors in these slices, as shown in Figure 10 and Figure 23 in Appendix Section VIII.F.

B. Trustworthiness

1. Local Dice

Considering all iterations in Experiment 2, Δ Local Dice is non-zero for most iterations, indicating that the model refines the volume within the neighborhood of the scribble slice. The majority of these values are positive, indicating that incorporating user input often improves Dice locally (i.e. agreement with the ground truth). A statistically significant improvement in Δ Local Dice was observed when using the AI pencil compared to the manual brush. Consequently, users required fewer iterations to segment the same volume and to achieve a similar final Dice score. However, the use of the AI pencil was also associated with negative Δ Local Dice values. This is consistent with observations made throughout the study, indicating that although the AI pencil generally outperforms the manual brush, it also introduces fluctuations in Dice. Variations in magnitude during the refinement process are expected and could have multiple explanations, as the local Dice depends on the region of the tumor being refined, the quality of the existing auto-contour in that area, and the extent of prior user refinement.

The results of Experiment 1 (Table 4) suggest that, for the easy slice in Patient 4, the positive outcome is mainly driven by refinement effects in slices located outside the ± 1 neighborhood and within the ± 5 and ± 10 neighborhoods. In contrast, for the difficult slice

(Patient 3), the largest Δ Local changes occur within 5 slices around the scribble slice (neighborhood ±5). However, the Δ Local surface Dice showed the biggest negative change in ±10 around the scribble. These values contradict with each other. This discrepancy is likely due to the shape of the segmentation masks in these regions and the differences between the metrics in penalizing changes between the prediction masks and the ground truth. Surface Dice is more sensitive to boundary roughness than standard volumetric Dice. Upon closer examination of the auto-contour mask and the subsequent mask after user refinement (Figure 9), it can be observed that large errors occur in the local area around scribble slice 62, where the scribble was placed. These errors also extend over a wider area, affecting nearly the entire ±10 neighborhood.

The decline in both local and volumetric Dice for Patient 3 may have multiple causes. One possible reason is that the initial auto-contouring model produced a highly inaccurate segmentation (Figure 9). The figure shows that the user attempted to remove the false positive portion of the tumor. However, this input also removed parts of the correctly segmented tumor volume. Since the auto-contour was used as input for the contour-refinement model, this incorrect information may have led the model to make an inaccurate estimate after the initial user correction.

Another possible explanation is difficulty associated with segmentation in this specific direction. Interestingly, in the case of Patient 3, this decrease in volumetric Dice was observed within the local area, while the non-local Dice values even increased after this refinement. A closer inspection of Figure 9, together with the fact that all users started in slice 62, shows that the negative changes occur far from the scribble location in the sagittal direction, although within the same axial slice. In this axial direction, the tumor segmentation is divided into two separate parts due to the tumor's shape around the cavity. This is clearly visible in the shape of the tumor's ground truth in Figure 14. The model responded incorrectly, possibly because it had difficulty contouring two separate regions within a single slice. A similar effect is observed in other slices where the segmentation was fragmented, for example in Figure (c) of Figure 22 where the tumor mask encompassed non-tumor regions. This suggests that when focusing only on the axial direction, as in this study, the model performs poorly in cases where the tumor mask is divided within a single slice. However, allowing users to refine the segmentation in other directions (coronal and sagittal), where features such as cavities may not complicate the segmentation in this particular case, could potentially improve the results.

2. Non-Local Dice

The results of Experiment 1 show that both patients exhibited large non-local differences between the initial auto-contour and the contour after refinement. For the easy slice, the non-local Dice values were 0.7727, 0.7713, and 0.7304, whereas for the difficult slice they were 0.5902, 0.5596, and 0.5627 (Table 4). A non-local Dice value of 1.000 corresponds to perfect overlap in the non-local area between iterations. Thus, for the easy slice, approximately one-quarter of the non-local volume between iterations did not align, whereas for the difficult slice, nearly half of the non-local volume was misaligned.

The results of Experiment 2 (Figure 8) show a large volumetric Dice refinement at the beginning of the segmentation process. When evaluating local and non-local Dice, it becomes apparent that this initial change is present in the total volume of the tumor. This illustrates the benefit of using both metrics to gain insight into which parts of the tumor are modified by user input. This provides a spatial description of where the model refines contours and therefore brings us one step closer to understanding its behavior, which is a necessary step toward clinical adoption. When examining the behavior of the non-local Dice over the entire experiment in Figure 30 and Figure 29, it is observed that non-local effects occur almost only during the initial user inputs. Afterward, the impact on the non-local Dice decreases, meaning that later user inputs almost exclusively refine the local area around the scribble.

This heterogeneity likely arises from two interacting factors: the spatial distribution of inaccuracies in the automatic prediction across tumor regions, and the manner in which the model responds to user-provided

scribbles. Through training, the model acquired prior expectations regarding the segmentation shape. During the first interactive steps of inference, the model partially relies on the users input and partially on it's own understanding of the tumor shapes. This leads to changes in non-local regions, even though the user did not provide specific information about those regions. As the interaction steps proceed and the user inputs accumulate, the model's decisions rely more on the user input. Sometimes this creates tension when a user provides input, as this could contradict with the internal-memory (i.e. neural weights). This behavior was observed in Patient 3. Initially, the model did not adhere to the user input, and the contour worsened after each iteration. However, if the user input is strong or repeated, the model gradually learned more about the specific tumor shape and the prediction will increasingly follow the user corrections rather than the model's original prior. Multiple studies [25, 81, 88] have focused on this aspect, as additional user input through the experiment influences and strengthens the model's predictions by altering the network's activation patterns. For all other patients evaluated in this study the model immediately adapted to the refinement, resulting in positive changes in the non-local area. Therefore, in the majority of cases, the model possesses an accurate prior representation of the expected tumor appearance.

3. Neighborhoods

As it was unclear how far the refinement effect of a scribble extends to surrounding slices, this study also evaluated Δ local and non-local Dice and surface Dice for different neighborhood definitions (±1, ±5, and ±10). The largest changes in local Dice between predictions are most frequently observed in the ±1 neighborhood, indicating that user input primarily refines the region closest to the scribble. In some iterations, higher Δ local Dice values are observed in the larger neighborhoods (±5 and ±10). It remains uncertain whether these changes are caused by the scribble itself or by other inputs of the contour-refinement model, such as CT, PET, and the auto-contour.

For the non-local Dice, mixed results are observed. In some patients the area outside the ±1 neighborhood shows the largest changes, whereas in others the area outside the ±10 neighborhood shows higher values. Notably, these larger non-local changes mainly occur at the beginning of the experiment, suggesting that they may depend strongly on the shape of the initial auto-contour. In later iterations the non-local effects diminish, indicating that these changes are likely not directly caused by the user scribble.

These observations raise the question of which neighborhood size adequately captures the local effects of user refinement throughout the experiment. Therefore, it is advisable to evaluate multiple neighborhood sizes during the refinement process, as the spatial meaning of a given neighborhood may change as the model and user interaction evolve.

C. Limitations and Future Recommendations

In this study, users were limited to refinement in axial slices. In certain cases, refinement could have been performed more efficiently in the sagittal or coronal planes. It could be valuable to examine whether the model predictions converge faster to the ground truth when the user is not restricted to the axial plane, but can place scribbles in either of the cardinal planes. Furthermore, model outcomes should be assessed in these directions using local and non-local Dice to investigate how refinement behaves in different directions and whether this behavior is similar across directions.

Difficulties occurred frequently when users attempted to segment tumor regions surrounding an air cavity. This was common because the HECKTOR dataset [75] contains only oropharyngeal cases, which are located near the upper airway as shown in Figure 1 [76]. The difficulty the model exhibits when contouring more complex tumors may pose a problem, as such cases are expected to occur frequently in clinical practice. Therefore, this model should be refined if it is used in other high-complex cancer sites, but can already be employed in anatomical sites containing less complex structures. These sites are not tested yet and could be a valuable addition for possible implementation in other cancer sites.

In this work, all users completed a training session

prior to participating in the experiments. However, the case of user 6 raises the question of whether this training was sufficient to fully understand the model's behavior and learn to use it. In general, variations in scribble placement resulted in minor differences in outcomes, and users learned under which conditions the model followed their intent and when it did not. The results of this study show that users occasionally made mistakes when providing scribble input. This is undesirable, as all scribbles influence the model's interpretation of the tumor volume. Although the model showed high robustness to variations in user input, it still followed the user's intent when an incorrect scribble was placed, resulting in erroneous segmentations (Figure 21). Therefore, it is important that clinicians are properly trained and supported in understanding how the model responds to user input. Improved training and familiarity with the model's behavior could increase clinicians' confidence in using these systems and help ensure the effective integration of deep learning-based medical image segmentation in clinical practice [16].

Another important consideration is the number of participants. Although this study identified a statistically significant improvement, it lacks sufficient power to draw the same conclusions at the level of individual patients. Including more participants would enable for analysis of a larger dataset. In such a scenario, a greater number of scribbles and patterns of user behavior could be examined. Future work including a larger number of users and scribble types could enable a more comprehensive analysis of scribble taxonomies and their influence on model outputs.

The participants in this study were non-experts, meaning that they did not know how to segment tumors. Therefore the outline of the ground truth was visible during the total experiments. Their assignment was to align the proposed prediction created by the contour-refinement model as much as possible with the presented ground truth. However, the ground truth data used in this study originate from the open HECKTOR dataset [75], and it can be questioned whether this ground truth is always correct or whether the HECKTOR dataset may contain inaccuracies. This raises a fundamental question of which serves as the more reliable reference: the model's refined output or the defined HECKTOR ground truth. Consequently, the conclusions drawn from the non-expert use of the tool and the model's response may not fully reflect expert behavior or real-world clinical scenarios. Future studies should therefore involve expert users instead of non-experts. In such settings, the final refined segmentation produced by the experts could be considered the reference ground truth for calculating Dice scores, thereby better reflecting the clinical correctness of the segmentation.

The models used in this study were also trained on the HECKTOR dataset [75]. When involving clinical experts, it should also be considered to include training and testing on real hospital data, ideally from the hospital where the experts work. High-quality data are essential for accurate tumor segmentation because lower contrast hinders users in identifying tumor boundaries [81]. Inaccurate delineation of tumor boundaries can adversely affect patients, since precise target segmentation is essential to prevent tumor underdosage or overdosage of adjacent organs at risk. This level of quality was not consistently achieved in the HECKTOR dataset [75]. This is a common problem, because existing public medical segmentation datasets currently do not fully meet high-quality standards, limiting their suitability for comprehensively supporting and evaluating interactive models [89]. Therefore, future work should focus on increasing the availability of high-quality clinical data and on training and evaluating the model on such data, as conclusions drawn from suboptimal datasets may differ from real-world clinical outcomes.

Complex anatomy, decreased performance near tumor edges, and surrounding structures such as cavities increase the difficulty of model segmentation. A potential solution to this problem is to use both the AI pencil and the manual brush within one experiment. The GUI used in this research already includes both interaction methods, however users were restricted to only use one during the experiments. By introducing this dual workflow, the strengths of both approaches can be combined: rapid 3D refinement over the entire volume using the AI model, complemented by

precise manual correction in regions where the model fails to perform adequately. Diaz-Pinto et al. [90] presented a comparable approach by integrating three methods: fully automatic segmentation, interactive segmentation with foreground and background clicks, and refinement of the segmentation using clicks limited to regions requiring correction. This resulted in increased workflow flexibility, which is an important factor for incorporating interactive segmentation into clinical practice.

AI models, such as the one validated in this study, can be difficult to interpret, as their internal workings are not yet fully understood. This phenomenon is often referred to as "black-box" behavior. Studies like the present one, which examine human-AI interaction, are necessary to build confidence in the use of AI models for clinical tasks. Further research in this area is needed to enable the safe adoption of AI in clinical workflows. This study also introduced two novel metrics, local and non-local (surface) Dice, which provide additional insights into how the model follows the user's intent and into its stability throughout the contouring process. These contributions represent an important step toward understanding the behavior of contour-refinement models. Expanding the evaluation of such models beyond conventional Dice metrics could lead to a better understanding of their performance and reliability in clinical practice.

VI. Conclusion

This study focused on the robustness and trustworthiness of an auto-contour refinement tool in H&N Radiotherapy. It validated that the contour-refinement model integrated in the refinement tool responds in a highly consistent manner across different users, thereby indicating a high degree of robustness. Small changes in performance are observed however, for different kinds of scribbles. This underscores the importance of adequately training clinicians before using this tool. The results demonstrate a high level of trustworthiness throughout the entire tumor segmentation process in almost all cases. The model is able to adhere to user corrections locally while avoiding unintended spurious changes farther away from the user input. At the same time, it evaluates the full 3D volume at every iteration and attempts to refine the entire volume, with the largest refinements occurring at the beginning by correcting the initial auto-prediction. However, when the anatomy of the tumor becomes more complex, often because of the location of the tumor for example near air cavities, the model fails to follow the users intent. The same effect is observed in the edges of the tumor. The results of this study indicate that incorporating user feedback into the contour-refinement model leads to a rapid improvement in segmentation quality over the full volume, which underscores the potential impact of these models to enhance clinical workflows. After this initial improvement in segmentation quality, the model increasingly relies on the user's input to perform local refinements. Nevertheless, clinicians are still required to manually refine small differences in challenging slices. This highlights the necessity of research into the model's internal priors and how these interact with user corrections. It is crucial to gain a deeper understanding of the internal workings of the AI model and to further investigate human-AI interaction in order to build greater trust in such systems. This research shows that the model is robust to variations in user input and (apart from the first few iterations) there are no spurious changes in non-local areas. These are important findings when working towards clinical adoption of these interactive contour refinement models.

VII. Acknowledgements

The author would like to thank Frank Dankers, Prerak Mody, and Nazli Tümer for the thesis supervision, insightful discussions and valuable feedback on this study. The author gratefully acknowledges all participants for their contribution to this study.

References

[1] Rasch CRN, Duppen JC, Steenbakkers RJ, Baseman D, Eng TY, Fuller CD, et al.. Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices; 2011.

[2] Harari PM, Song S, Tomé WA. Emphasizing Conformal Avoidance Versus Target Definition for IMRT Planning in Head-and-Neck Cancer. International Journal of Radiation Oncology Biology Physics. 2010;77:950-8.

[3] Oreiller V, Andrearczyk V, Jreige M, Boughdad S,

Elhalawani H, Castelli J, et al.. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Elsevier B.V.; 2022.

[4] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiotherapy and Oncology. 2015 10;117:83-90.

[5] Purkayastha M, McMahon AD, Gibson J, Conway DI. Trends of oral cavity, oropharyngeal and laryngeal cancer incidence in Scotland (1975-2012) – A socioeconomic perspective. Oral Oncology. 2016 10;61:70-5.

[6] Smith CDL, McMahon AD, Purkayastha M, Creaney G, Clements K, Inman GJ, et al. Head and neck cancer incidence is rising but the sociodemographic profile is unchanging: a population epidemiological study (2001-2020). BJC Reports. 2024 9;2.

[7] Gormley M, Creaney G, Schache A, Ingarfield K, Conway DI. Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. British Dental Journal. 2022 11;233:780-6.

[8] Zhu H, Chua MLK, Chitapanarux I, Kaidar-Person O, Mwaba C, Alghamdi M, et al. Global radiotherapy demands and corresponding radiotherapy-professional workforce requirements in 2022 and predicted to 2050: a population-based study. The Lancet Global Health. 2024 12;12:e1945-53.

[9] Albertini F, McWilliam A, Winey B. Editorial: Advances in online and real-time adaptive radiotherapy. Institute of Physics; 2025.

[10] Zhao X, Pan H, Bai W, Li B, Wang H, Zhang M, et al. Interactive segmentation of medical images using deep learning. Physics in Medicine and Biology. 2024 2;69.

[11] Zhang Y, Liu H, Hu Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. 2021 7. Available from: http://arxiv.org/abs/2102.08005.

[12] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 9351. Springer Verlag; 2015. p. 234-41.

[13] Doolan PJ, Charalambous S, Roussakis Y, Leczynski A, Peratikou M, Benjamin M, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. Frontiers in Oncology. 2023 8;13. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10436522/.

[14] Ye X, Guo D, Ge J, Yan S, Xin Y, Song Y, et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nature Communications. 2022 12;13.

[15] Sakinis T, Milletari F, Roth H, Korfiatis P, Kostandy P, Philbrick K, et al. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:190308205. 2019.

[16] Gao Y, Jiang Y, Peng Y, Yuan F, Zhang X, Wang J. Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. Tomography. 2025 4;11(5):52. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC12115501/.

[17] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks;. Available from: http://code.google.com/p/cuda-convnet/.

[18] Ulrich C, Wald T, Tempus E, Rokuss M, Jaeger P, Maier-Hein K. RadioActive: 3D Radiological Interactive Segmentation Benchmark. 2025.

[19] Reinders FCJ, Savenije MHF, de Ridder M, Maspero M, Doornaert PAH, Terhaard CHJ, et al. Automatic segmentation for magnetic resonance imaging guided individual elective lymph node irradiation in head and neck cancer patients. Physics and Imaging in Radiation Oncology. 2024 10;32.

[20] Gao Y, Jiang Y, Peng Y, Yuan F, Zhang X, Wang J. Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. Multidisciplinary Digital Publishing Institute (MDPI); 2025.

[21] Wang G, Zuluaga MA, Li W, Pratt R, Patel PA, Aertsen M, et al. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. IEEE Trans Pattern Anal Mach Intell. 2019;41:1559-72. Available from: https://ieeexplore.ieee.org/ielx7/34/8730438/08370732.pdf?tp=&arnumber=8370732&isnumber=8730438&ref=.

[22] Sharma N, Ray AK, Shukla KK, Sharma S, Pradhan S, Srivastva A, et al.. Automated medical image segmentation techniques; 2010.

[23] Mody P, de Plaza NC, Gooding M, de Jong M, de Ridder M, den Hans N, et al. Manual Brush vs AI Pencil: Evaluating tools for auto-contour refinement of head-and-neck tumors on PET/CT scans 1 Authors Manual Brush vs AI Pencil: Evaluating tools for auto-contour refinement of head-and-neck tumors on PET/CT scans; 2026. Available from: https://ssrn.com/abstract=6108793.

[24] Marinov Z, Jager PF, Egger J, Kleesiek J, Stiefelhagen R. Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.

2024.

[25] Asad M, Williams H, Mandal I, Ather S, Deprest J, D'hooge J, et al. Adaptive Multi-scale Online Likelihood Network for AI-Assisted Interactive Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2023;14221 LNCS:564-74. Available from: https://link-springer-com.tudelft.idm.oclc.org/chapter/10.1007/978-3-031-43895-0_53.

[26] Hallitschke VJ, Schlumberger T, Kataliakos P, Marinov Z, Kim M, Heiliger L, et al. Multimodal Interactive Lung Lesion Segmentation: A Framework for Annotating PET/CT Images based on Physiological and Anatomical Cues. 2023 1. Available from: http://arxiv.org/abs/2301.09914http://dx.doi.org/10.1109/ISBI53787.2023.10230334.

[27] Shen CY, Li WH, Xu QS, Hu B, Jin B, Cai HB, et al. Interactive medical image segmentation with self-adaptive confidence calibration. Frontiers of Information Technology & Electronic Engineering. 2023;24:1332-48. Available from: https://link.springer.com/content/pdf/10.1631/FITEE.2200299.pdf.

[28] Tian F, Tian Z, Chen Z, Zhang D, Du S. Surface-GCN: Learning interaction experience for organ segmentation in 3D medical images. Medical Physics. 2023;50:5030-44.

[29] Tian M, Chen X, Gao Y. A dynamic interactive learning framework for automated 3D medical image segmentation. arXiv preprint arXiv:231206072. 2023.

[30] Zhou T, Li L, Bredell G, Li J, Unkelbach J, Konukoglu E. Volumetric memory network for interactive medical image segmentation. Med Image Anal. 2023;83:102599. Available from: https://www.zora.uzh.ch/entities/publication/dccc1143-3636-4d0a-bc86-bd7f0a8d06fd.

[31] Zhuang M, Chen Z, Wang H, Tang H, He J, Qin B, et al. Efficient contour-based annotation by iterative deep learning for organ segmentation from volumetric medical images. Int J Comput Assist Radiol Surg. 2023;18:379-94. Available from: https://link.springer.com/content/pdf/10.1007/s11548-022-02730-z.pdf.

[32] Li H, Oguz B, Arenas G, Yao X, Wang J, Pouch A, et al. Interactive Segmentation Model for Placenta Segmentation from 3D Ultrasound images. 2024. Available from: https://github.com/MedICL-VU/PRISM-placenta.

[33] Shen B, Chang L, Chen S, Guo S, Liu H, Chang L. Lightweight Method for Interactive 3D Medical Image Segmentation with Multi-Round Result Fusion. 2024. Available from: https://github.com/goodtime-123/LIM-Net.

[34] Borkar K, Reen AS, Jawahar CV, Arora C. No Prompting Frozen Foundation Models: Interactive Medical Volume Segmentation using Continual Test Time Adaptation of Compact Models. Association for Computing Machinery; 2024. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85216725853&doi=10.1145%2F3702250.3702275&partnerID=40&md5=c9bf33ee5821d0960b8c2eb239a5f533.

[35] Chen H, Chen Z, Zhao J, Li H, Li J, Liu Y, et al. MSI-UNet: A Flexible UNet-Based Multi-Scale Interactive Framework for 3D Gastric Tumor Segmentation on CT Scans. In: Proceedings - International Symposium on Biomedical Imaging. IEEE Computer Society; 2024. .

[36] Hadlich M, Marinov Z, Kim M, Nasca E, Kleesiek J, Stiefelhagen R. Sliding Window Fastedit: A Framework for Lesion Annotation in Whole-Body Pet Images. IEEE Computer Society; 2024. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203324162&doi=10.1109%2FISBI56570.2024.10635459&partnerID=40&md5=8a4056aac37f0170d8f78d021d094d7ehttps://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10635459&ref=.

[37] Le QA, Pham XL, van Walsum T, Dao VH, Le TL, Franklin D, et al. Precise ablation zone segmentation on CT images after liver cancer ablation using semi-automatic CNN-based segmentation. Med Phys. 2024;51:8882-99.

[38] Li H, Liu H, Hu D, Wang J, Oguz I. Prism: A promptable and robust interactive segmentation model with visual prompts. Springer; 2024.

[39] Mikhailov I, Chauveau B, Bourdel N, Bartoli A. Sharing is Caring: Concurrent Interactive Segmentation and Model Training using a Joint Model. Institute of Electrical and Electronics Engineers Inc.; 2023. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85182945737&doi=10.1109%2FICCVW60793.2023.00257&partnerID=40&md5=5d920f0027fc188ec32134ca31b0a89ehttps://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10350959&ref=.

[40] Shen C, Li W, Shi Y, Wang X. Interactive 3d medical image segmentation with sam 2. arXiv preprint arXiv:240802635. 2024.

[41] Sydorskyi V. Interactive Decision Support System

for Lung Cancer Segmentation. System Research and Information Technologies. 2024;2024:68-81. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204067796&doi=10.20535%2FSRIT.2308-8893.2024.2.05&partnerID=40&md5=1d324dbffe5376bc7ccb1d3553555fc2http://journal.iasa.kpi.ua/article/download/290923/301102.

[42] Du Y, Bai F, Huang T, Zhao B. SegVol: Universal and Interactive Volumetric Medical Image Segmentation. 2025. Available from: https://github.com/BAAI-DCAI/SegVol.

[43] Wong HE, Gonzalez JJ, Databricks O, Guttag J, Dalca AV. MultiverSeg: Scalable Interactive Segmentation of Biomedical Imaging Datasets with In-Context Guidance MultiverSeg User Interactions (optional) In-Context Inputs. 2025. Available from: https://multiverseg.csail.mit.edu.

[44] Gu Y, Lei W, Chen H, Zhang X, Zhang S. Interactive Segmentation and Report Generation for CT Images. 2025.

[45] Orbes-Arteaga M, Lucena O, Ourselin S, Cardoso MJ. MAIS: Memory-Attention for Interactive Segmentation. 2025.

[46] Archit A, Freckmann L, Nair S, Khalid N, Hilt P, Rajashekar V, et al. Segment Anything for Microscopy. Nat Methods. 2025;22:579-91. Available from: https://www.nature.com/articles/s41592-024-02580-4.pdf.

[47] Li H, Oguz B, Arenas G, Yao X, Wang J, Pouch A, et al. PRISM Lite: A lightweight model for interactive 3D placenta segmentation in ultrasound. Proc SPIE Int Soc Opt Eng. 2025;13406.

[48] Li Y, Zhang Q, Zhou H, An Y, Li J, Li X, et al. Deep leaning-based interactive segmentation of three-dimensional blood vessel images. Biomedical Signal Processing and Control. 2025;104:107507.

[49] Orsmaa L, Saukkoriipi M, Kangas J, Rasouli N, Järnstedt J, Mehtonen H, et al. Interactive AI annotation of medical images in a virtual reality environment. Int J Comput Assist Radiol Surg. 2025. Available from: https://link.springer.com/content/pdf/10.1007/s11548-025-03497-9.pdf.

[50] Putz F, Beirami S, Schmidt MA, May MS, Grigo J, Weissmann T, et al. The Segment Anything foundation model achieves favorable brain tumor auto-segmentation accuracy in MRI to support radiotherapy treatment planning. Strahlenther Onkol. 2025;201:255-65. Available from: https://link.springer.com/content/pdf/10.1007/s00066-024-02313-8.pdf.

[51] Shen Y, Ding H, Shao X, Unberath M. Performance and nonadversarial robustness of the segment anything model 2 in surgical video segmentation. SPIE-Intl Soc Optical Eng; 2025. p. 13.

[52] Shen Y, Dreizin D, Inigo B, Unberath M. ProtoSAM-3D: Interactive semantic segmentation in volumetric medical imaging via a Segment Anything Model and mask-level prototypes. Comput Med Imaging Graph. 2025;121:102501. Available from: https://www.sciencedirect.com/science/article/pii/S0895611125000102?via%3Dihub.

[53] Shen Y, Shao X, Romillo BI, Dreizin D, Unberath M. FastSAM-3DSlicer: A 3D-Slicer Extension for 3D Volumetric Segment Anything Model with Uncertainty Quantification. Found Models Gen Med AI (2024). 2025;15184:1-9. Available from: https://link.springer.com/content/pdf/10.1007/978-3-031-73471-7_1.pdf.

[54] Shi WT, He JJ, Shen YQ. SIT-SAM: A semantic-integration transformer that adapts the Segment Anything Model to zero-shot medical image semantic segmentation. Biomedical Signal Processing and Control. 2025;110. Available from: https://www.sciencedirect.com/science/article/pii/S174680942500597X?via%3Dihub.

[55] Sun YK, Zhang SJ, Li JS, Han Q, Qin YH. CAISeg: A Clustering-Aided Interactive Network for Lesion Segmentation in 3D Medical Imaging. IEEE Journal of Biomedical and Health Informatics. 2025;29:371-82. Available from: https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?tp=&arnumber=10693449&ref=.

[56] Wei Z, Ren J, Eriksen JG, Jensen K, Mortensen HR, Korreman SS, et al. An Interactive Deep-Learning Workflow for Head and Neck Gross Tumour Volume Segmentation. Ssrn. 2025;2. Available from: https://www.ssrn.com/index.cfm/en/https://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=empp&DO=10.2139%2fssrn.5219763http://catalogue.leidenuniv.nl/openurl/UBL/UBL_services_page?sid=OVID:Embase+Preprint&issn=1556-5068&isbn=&volume=&issue=&spage=&date=2025&pid=%3Cauthor%3EWei+Z.%3C%2author%3Ehttps://papers.ssrn.com/sol3/papers.cfm?abstract_id=5219763.

[57] Zhou Y, Ye M, Ye H, Zeng S, Shu X, Pan Y, et al. Semi-automated segmentation of breast tumor on automatic breast ultrasound image using a large-scale model with customized modules. Sci Rep. 2025;15:17329. Available from: https://www.nature.com/articles/s41598-025-97098-w.pdf.

[58] Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE, et al. Understanding

metric-related pitfalls in image analysis validation. Nature Methods. 2024 2;21:182-94.

[59] Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al.. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Elsevier Ireland Ltd; 2021.

[60] Scharn JL. A Systematic Review of Evaluation Metrics in Deep Interactive Medical Image Segmentation with Inference-Time User Prompts [Masterthesis Literature Study]. Delft University of Technology; 2026. In preparation for publication.

[61] Chau M, Vu H, Debnath T, Rahman MG. A scoping review of automatic and semi-automatic MRI segmentation in human brain imaging. Radiography. 2025 3;31:102878. Available from: https://www.sciencedirect.com/science/article/pii/S1078817425000197?via=ihub.

[62] Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945;26(3):297-302.

[63] Hausdorff F. Grundzüge der Mengenlehre. Leipzig: Veit & Comp.; 1914.

[64] Mohammadi M, Mollazade K, Behroozi-Khazaei N. Under- and over-segmentation: New metrics for image segmentation accuracy measurement. Array. 2025 12;28.

[65] Rainio O, Klén R. Modified Dice Coefficients for Evaluation of Tumor Segmentation from PET Images: A Proof-of-Concept Study. Journal of Imaging Informatics in Medicine. 2025.

[66] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Medical Imaging. 2015 8;15.

[67] Li Z, Zhao K, Wang Y, Wang S. Adaptive interactive segmentation for multimodal medical imaging via selection engine. arXiv preprint arXiv:241119447. 2024.

[68] Okel SE, Viviers CGA, Ramaekers M, Hellström TAE, Tasios N, Mavroeidis D, et al. Advancing Abdominal Organ and PDAC Segmentation Accuracy with Task-Specific Interactive Models. vol. 14313. Springer Science and Business Media Deutschland GmbH; 2024. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177225101&doi=10.1007%2F978-3-031-47076-9_6&partnerID=40&md5=c3f4ebd0a2b139004f06a9c186cbb2bchttps://link.springer.com/content/pdf/10.1007/978-3-031-47076-9_6.pdf.

[69] Tang Y, Li Y, Zou H, Zhang X. Interactive Segmentation for Medical Images Using Spatial Modeling Mamba. Information. 2024;15:633.

[70] Zhang Y, Chen J, Ma XX, Wang G, Bhatti UA, Huang MX. Interactive medical image annotation using improved Attention U-net with compound geodesic distance. Expert Systems with Applications. 2024;237.

[71] Roddan A, Czempiel T, Xu C, Elson DS, Giannarou S. SAMSA: Segment Anything Model Enhanced with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation. 2025.

[72] Roddan A, Czempiel T, Xu C, Elson DS, Giannarou S. SAMSA 2.0: Prompting Segment Anything with Spectral Angles for Hyperspectral Interactive Medical Image Segmentation. 2025.

[73] Shi ZD, Chen T, Zhou Q, Zou H, Xiao X. IMedSeg: Towards efficient interactive medical segmentation. Neurocomputing. 2025;624. Available from: https://www.sciencedirect.com/science/article/pii/S0925231225000918?via%3Dihub.

[74] Xu W, Liang Z, Anthony H, Ibrahim Y, Cohen F, Yang G, et al. Continuous Online Adaptation Driven by User Interaction for Medical Image Segmentation. arXiv preprint arXiv:250306717. 2025.

[75] Andrearczyk V, Oreiller V, Abobakr M, Akhavanallaf A, Balermpas P, Boughdad S, et al. Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 13626 LNCS. Springer Science and Business Media Deutschland GmbH; 2023. p. 1-30.

[76] Centers for Disease Control and Prevention (CDC). HPV and Oropharyngeal Cancer; 2024. Updated Sep. 17, 2024. Accessed Jan. 22, 2026. https://www.cdc.gov/cancer/hpv/oropharyngeal-cancer.html.

[77] Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: An open-source framework for deep learning in healthcare. 2022 11. Available from: http://arxiv.org/abs/2211.02701.

[78] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res. 2021 Jul;23(7):e26151. Available from: https://www.jmir.org/2021/7/e26151.

[79] MONAI Consortium. monai.metrics.surface_dice - MONAI Documentation; 2024. https://monai-dev.readthedocs.io/en/latest/_modules/monai/metrics/surface_dice.html.

[80] Podobnik G, Vrtovec T. Understanding implementation pitfalls of distance-based metrics for image

segmentation. 2025 7. Available from: http://arxiv.org/abs/2410.02630.

[81] Wang G, Zuluaga MA, Li W, Pratt R, Patel PA, Aertsen M, et al. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation.

[82] Wang Z, Wang J, Budd C, MacCormac O, Shapey J, Vercauteren T. Scribble-Based Interactive Segmentation of Medical Hyperspectral Images. arXiv preprint arXiv:240802708. 2024.

[83] Jiang B, Ren T, Bei J. Automatic Scribble Simulation for Interactive Image Segmentation Evaluation. In: Tian Q, Sebe N, Qi GJ, Huet B, Hong R, Liu X, editors. MultiMedia Modeling. Cham: Springer International Publishing; 2016. p. 596-608.

[84] Wong HE, Rakic M, Guttag J, Dalca AV. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. 2023 12. Available from: http://arxiv.org/abs/2312.07381.

[85] Mansoor A, Bagci U, Foster B, Xu Z, Papadakis GZ, Folio LR, et al. Segmentation and image analysis of abnormal lungs at CT: Current approaches, challenges, and future trends. Radiographics. 2015 7;35:1056-76.

[86] Song D, Yang S, Han JY, Kim KG, Kim ST, Yi WJ. Comparison of segmentation performance of cnns, vision transformers, and hybrid networks for paranasal sinuses with sinusitis on CT images. Scientific Reports. 2025 12;15.

[87] Soret M, Bacharach SL, Buvat I. Partial-Volume effect in PET tumor imaging. Journal of Nuclear Medicine. 2007 5;48(6):932-45. Available from: https://jnm.snmjournals.org/content/48/6/932?utm_source=copilot.com.

[88] Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, et al. Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning. IEEE Trans Med Imaging. 2018;37:1562-73. Available from: https://ieeexplore.ieee.org/ielx7/42/8400497/08270673.pdf?tp=&arnumber=8270673&isnumber=8400497&ref=.

[89] Cheng J, Fu B, Ye J, Wang G, Li T, Wang H, et al. Interactive Medical Image Segmentation: A Benchmark Dataset and Baseline; 2025. Available from: https://www.cancerimagingarchive.net.

[90] Diaz-Pinto A, Mehta P, Alle S, Asad M, Brown R, Nath V, et al. DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images. 2023 5. Available from: http://arxiv.org/abs/2305.10655http://dx.doi.org/10.1007/978-3-031-17027-0_2.

VIII. Appendix

A. Dataset

All imaging data originate from the HECKTOR challenge dataset [75], which contains multi-center PET/CT scans of H&N cancer patients. The contributing centers and their corresponding number of scans and splits are shown in Table 5. In the original study by Mody et al. [23], data from the HMR center were retained as in-distribution validation data, while data from the French center (CHUP) were reserved exclusively for out-of-distribution testing. In this study patients from the CHUP test-set were also chosen for the experiments.

Prior to model training, the data underwent preprocessing as described in the study by Mody et al. [23]. This included resampling the PET/CT scans to an isotropic voxel spacing of 1 mm using B-spline interpolation, while segmentation masks were resampled using nearest-neighbor interpolation. All scans used for training and testing were cropped to a fixed size of $144 \times 144 \times 144$ voxels centered around the primary H&N tumor. Intensity normalization was performed using Hounsfield Unit (HU) windowing of $[-250, 250]$ for CT images and standardized uptake value (SUV) windowing of $[0, 25]$ for PET images. Finally, all scans were z-normalized, following standard practice in deep learning-based medical image analysis [23].

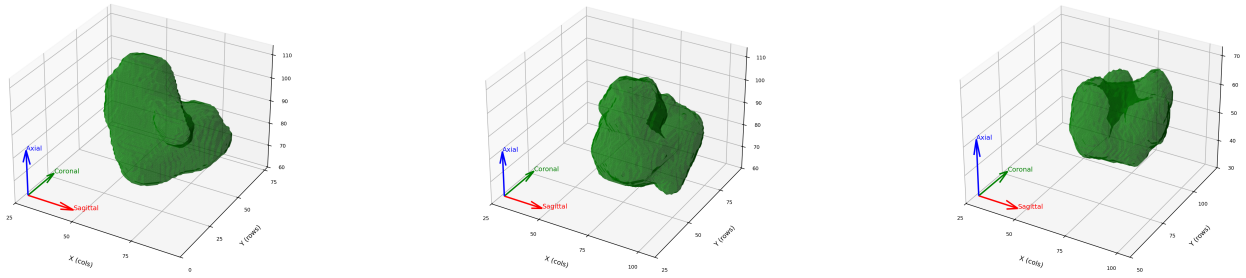
Table 5. Overview of hospital centers used in this study, including country, abbreviation, and number of cases per task [75].

Country	Center	Abbr.	Split	Cases
Canada	Centre Hospitalier de l'Université de Montréal	CHUM	Train	56
	Centre Hospitalier Universitaire de Sherbrooke	CHUS	Train	72
	Hôpital Général Juif, Montréal	HGJ	Train	55
	Hôpital Maisonneuve-Rosemont, Montréal	HMR	Train & Validation	18
USA	MD Anderson Cancer Center	MDA	Train	198
Switzerland	Centre Hospitalier Universitaire Vaudois	CHUV	Train	53
France	Centre Hospitalier Universitaire Poitiers	CHUP	Test	72
Total				**524**

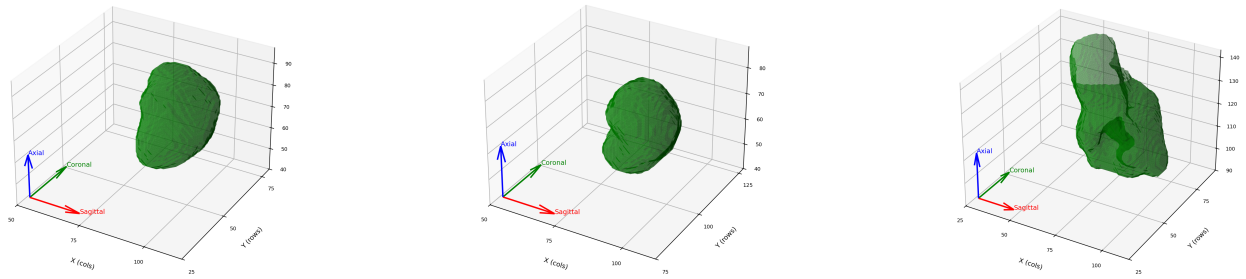
The ground truth segmentations of the six patients included in this study are presented in Figure 14.

B. Graphical User Interface

A screenshot of the graphical user interface (GUI) used in this study is shown in Figure 15. The interface allows users to create scribbles using either an AI-assisted scribble tool or a manual brush. Controls for switching between foreground and background scribbles, as well as for selecting the brush or eraser, are located in the upper-left panel of the interface. The GUI supports panning, zooming, and scrolling within both the CT and PET images. Three orthogonal views are displayed (Axial, Sagittal, and Coronal), enabling users to interactively inspect and adjust the segmentation in all three spatial dimensions. In addition to mouse-based interaction, the interface provides several keyboard shortcuts, including options to change between foreground and background brushes, adjust brush size, and show or hide the segmentation contours.



(a) (Patient 1) (b) (Patient 2) (c) (Patient 3)



(d) (Patient 4) (e) (Patient 5) (f) (Patient 6)

Figure 14. 3D Ground Truth Segmentation masks for all six patients included in this study. Blue arrow: axial direction. Green arrow: coronal direction. Red arrow: sagittal direction.

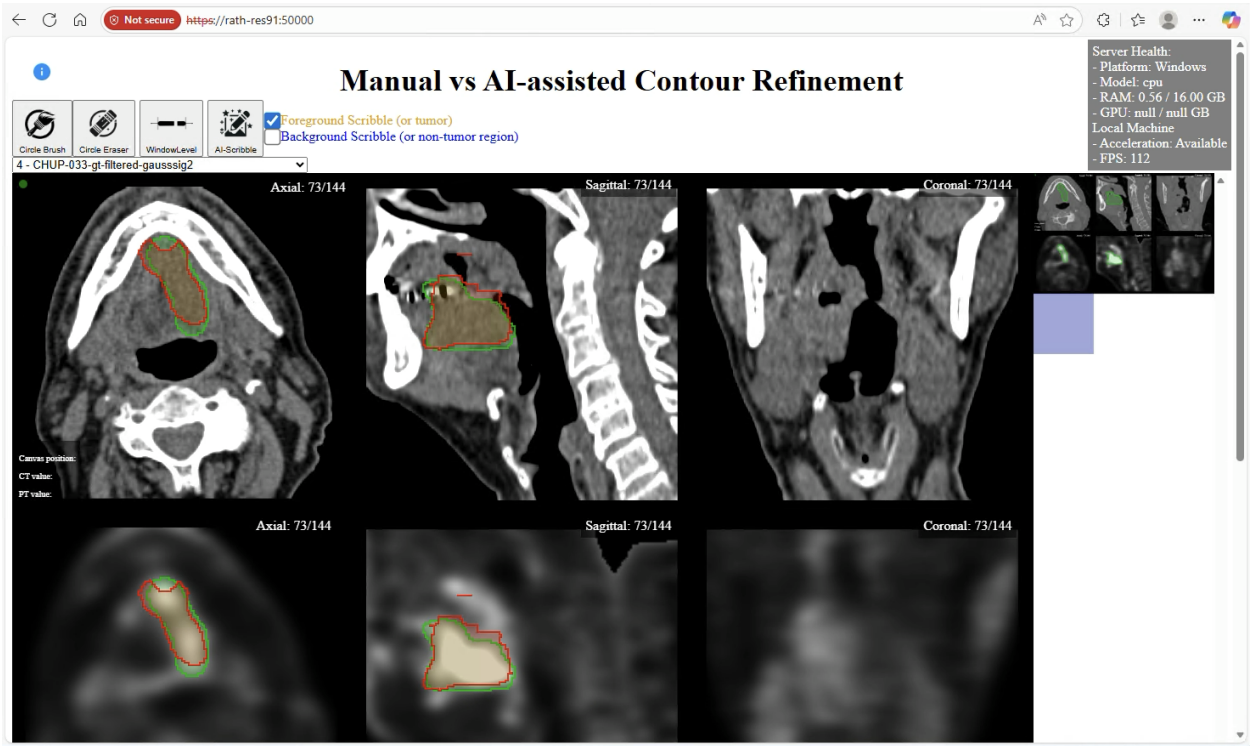
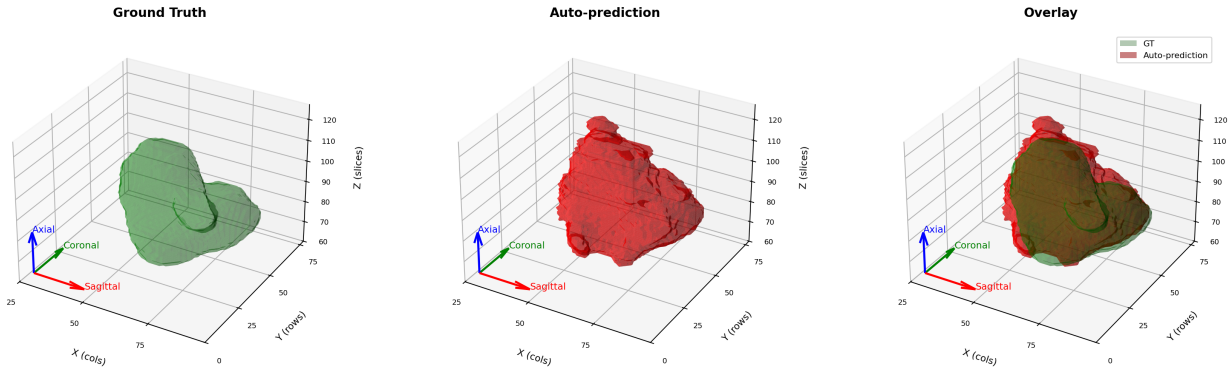


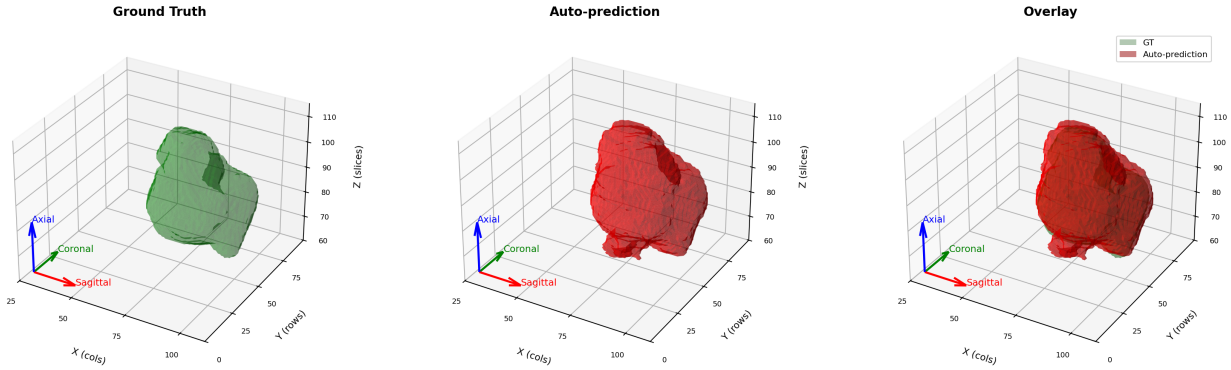
Figure 15. Screenshot of the Graphical User Interface. Green mask: ground truth. Red mask: Prediction.

C. 3D Masks Auto-Contour

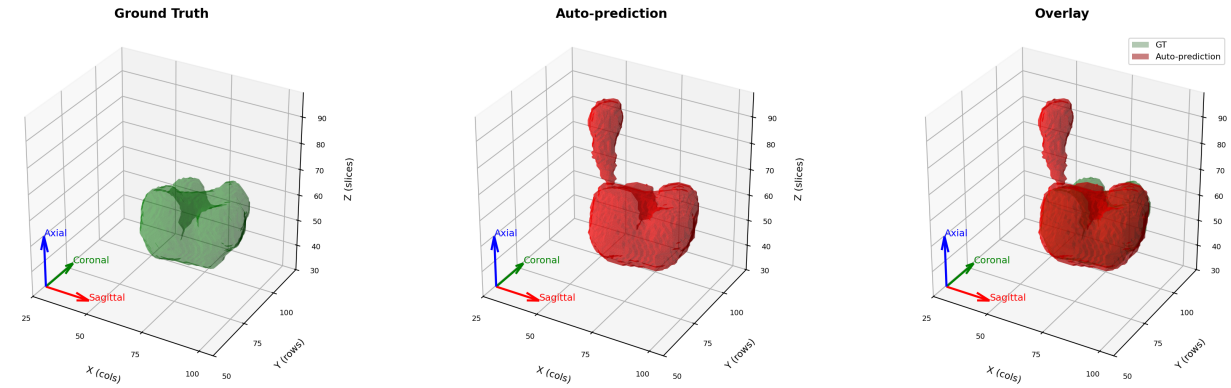
GT vs Before - Axial | P1 - User 2 - AI



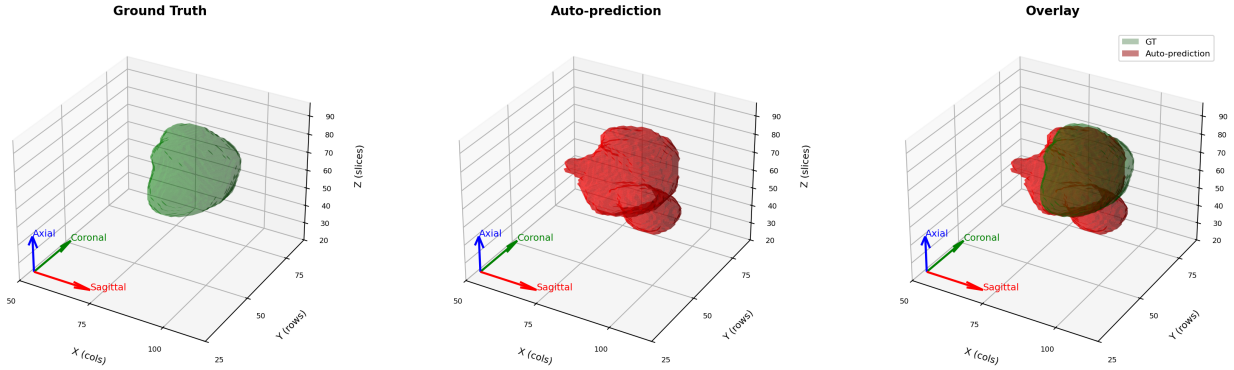
GT vs Before - Axial | P2 - User 2 - AI



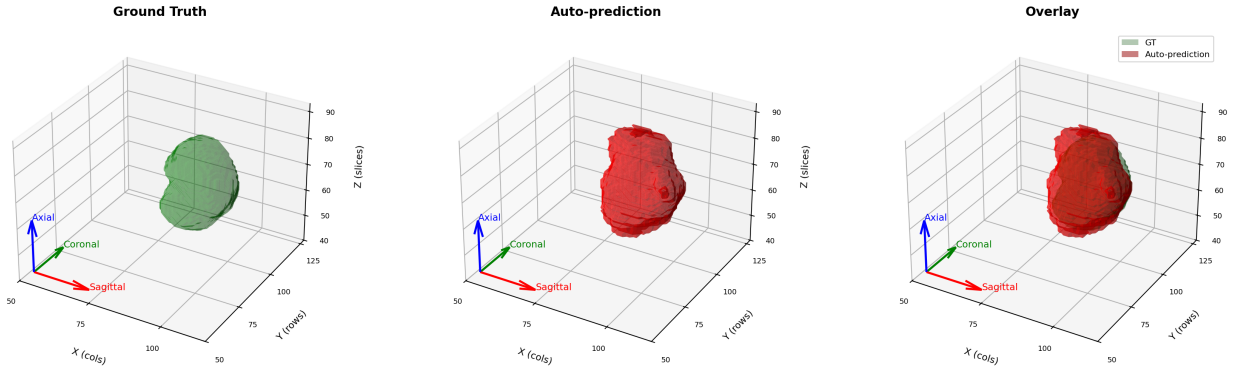
GT vs Before - Axial | P3 - User 2 - AI



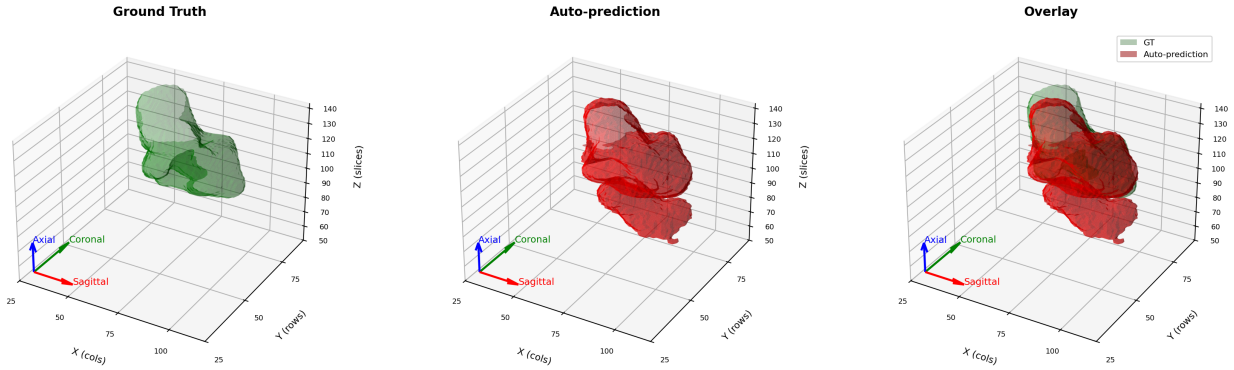
GT vs Before - Axial | P4 - User 2 - AI



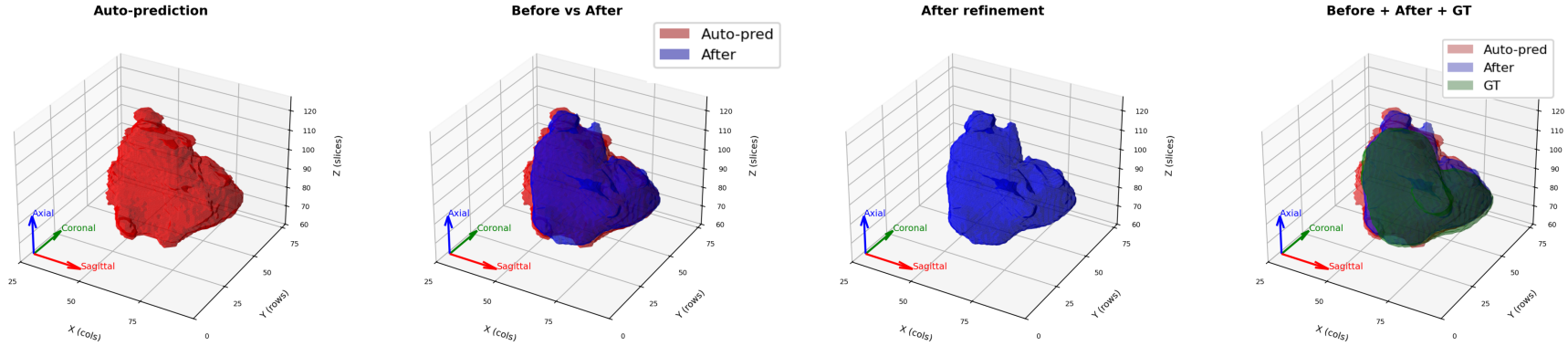
GT vs Before - Axial | P5 - User 2 - AI



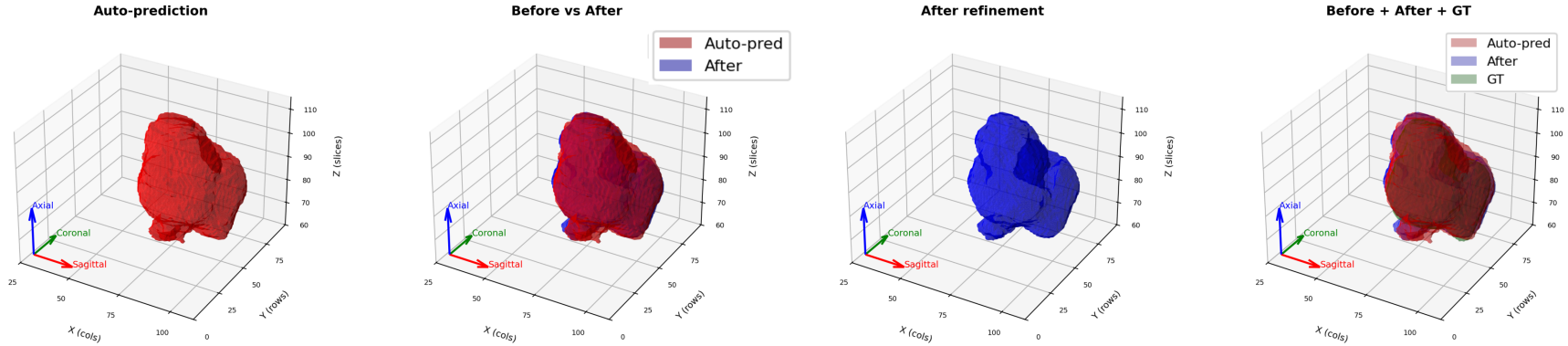
GT vs Before - Axial | P6 - User 2 - AI



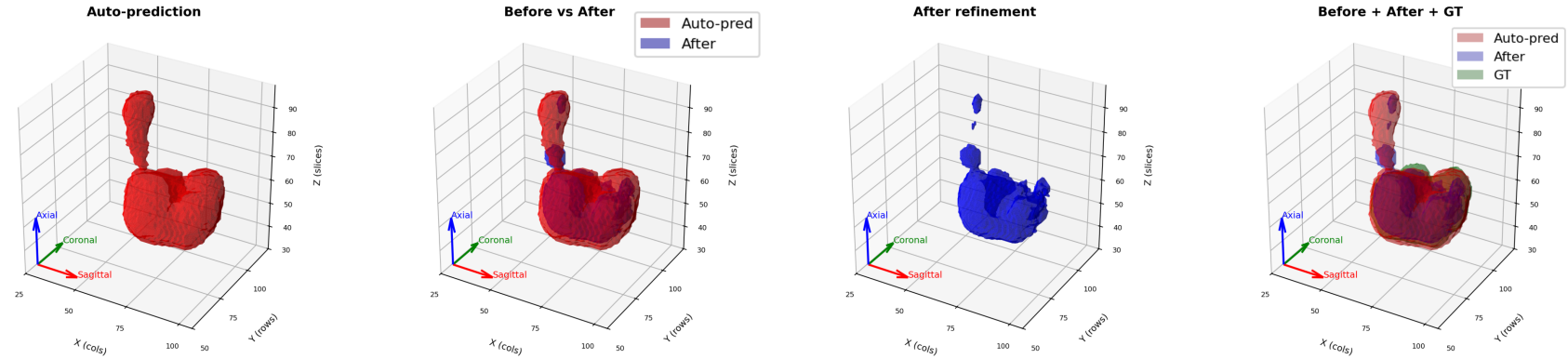
D. 3D Masks After First Refinement



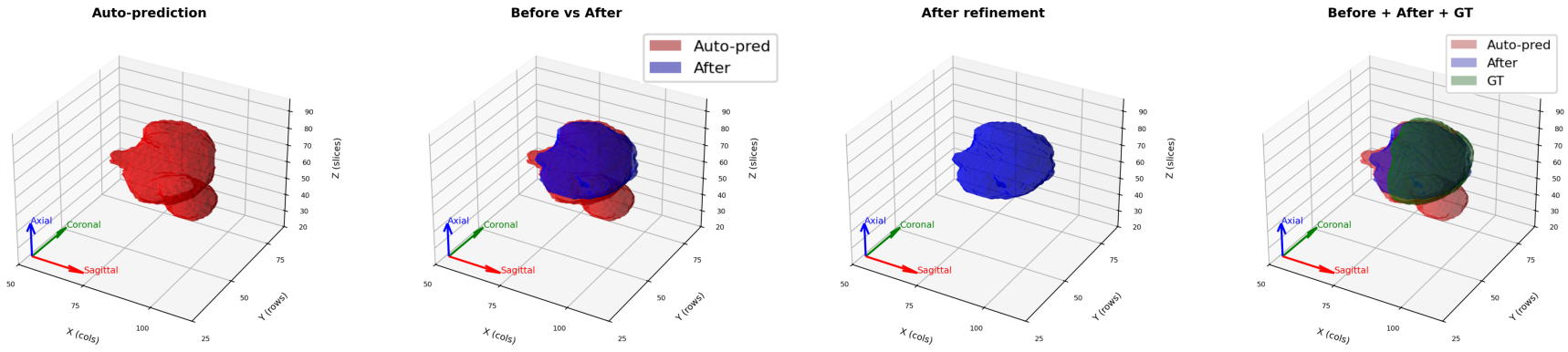
(a) Patient 1



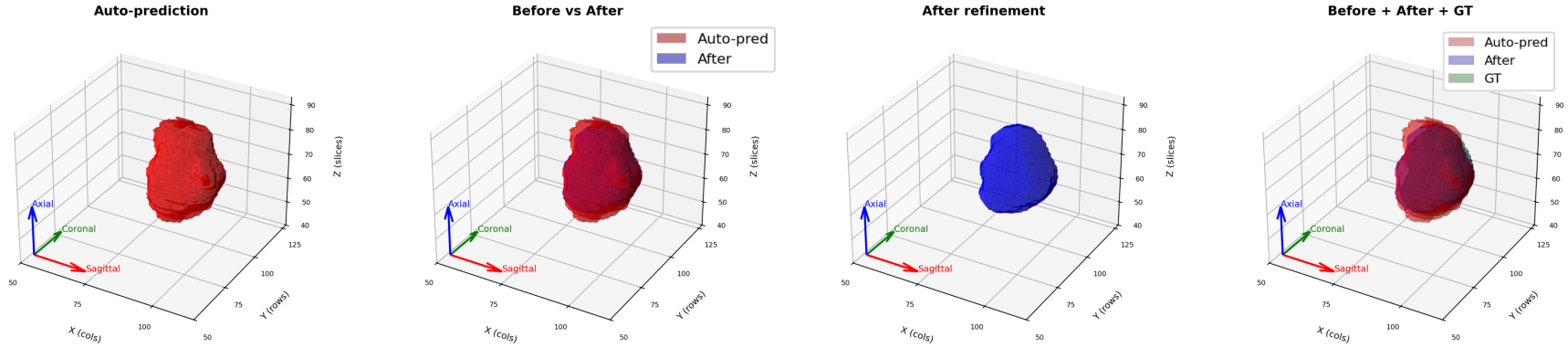
(b) Patient 2



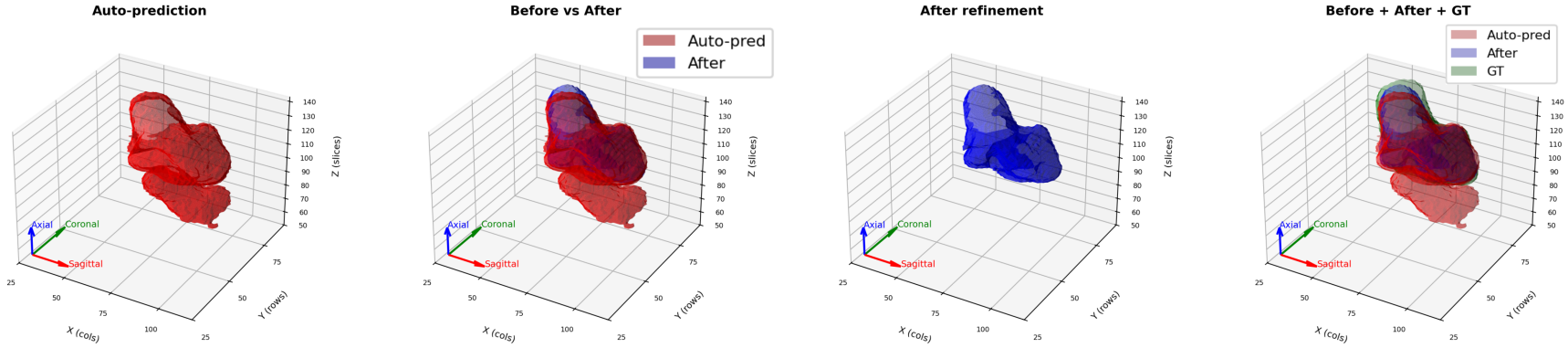
(c) Patient 3



(a) Patient 4



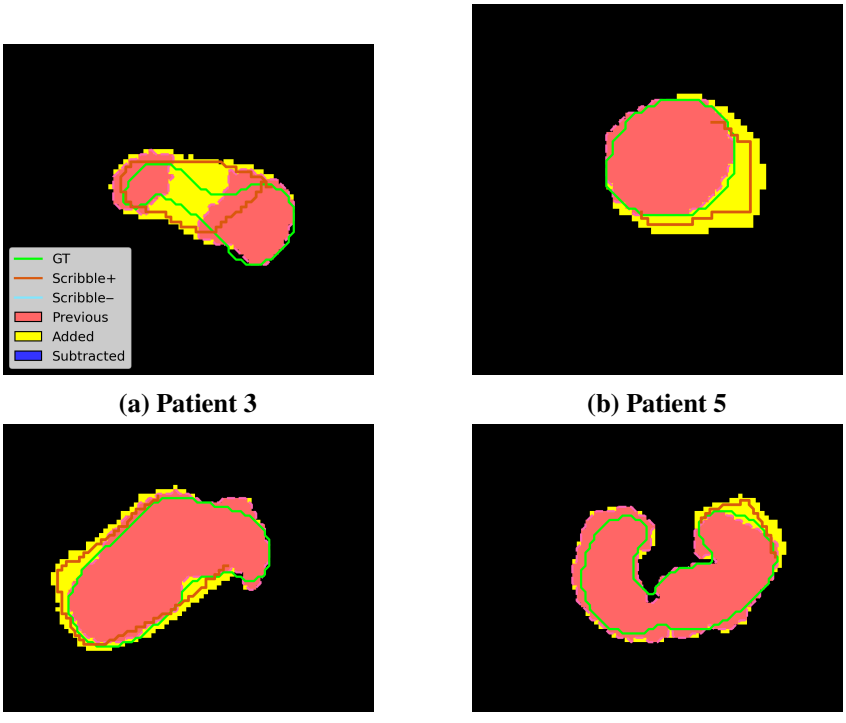
(b) Patient 5



(c) Patient 6

E. User 6 Exclusion

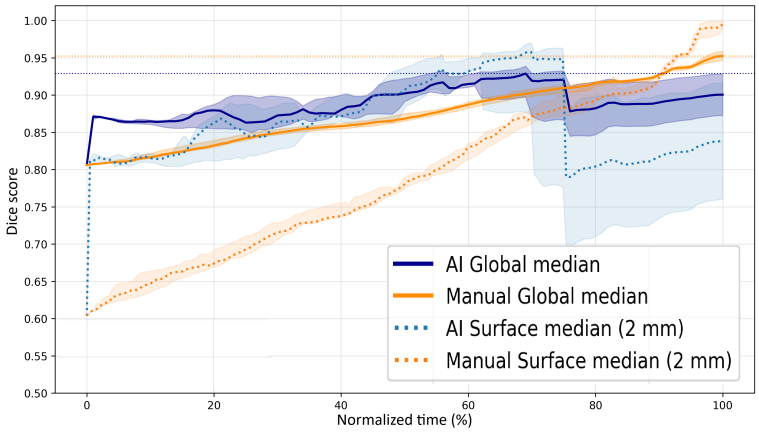
After analyzing the results of Experiment 2 (see III.D.2), the decision was made to exclude user 6 from this experiment. This choice was motivated by the observation that her results consistently deviated from those of the other participants, indicating that the user was a systematic outlier. This effect is shown in Figure 20 where her trajectory differs significantly from the other users. Several possible reasons may explain this behavior. First, the participant did not adhere fully to the instructions provided. This may be attributed to the fact that, afterwards the participant was not a typical non-expert, as the user had more prior knowledge than expected beforehand. to segment the entire tumor rather than only the intended regions (i.e., delineating the Planning Target Volume (PTV) instead of the Gross Tumor Volume (GTV)). This effect is shown in Figure 18, where segmentations frequently extend beyond the indicated boundaries of the ground truth. For example, in Figure 18b, the segmentation produced by the model is almost perfect, except that a small region on the right side of the tumor was not included. Because the user knew that, in daily radiotherapy practice, it is important to target the entire tumor during treatment, the user added this missing region. However, this addition extended far beyond the actual tumor boundaries, resulting in a large incorrectly segmented area. Another example is shown in Figure 18d, where the tumor was over-segmented in multiple regions. Nevertheless, user 6 considered it more important to add a small under-segmented area rather than correcting the other over-segmented parts of the segmentation In addition, user 6 did not used the available shortcuts consistently, whereas all other users did. Furthermore, the graphic user interface crashed once during the segmentation of patient 2, due to an excessive number of open tabs on the computer of the user. This resulted in a lower DSC score for patient 2. Global and Surface Dice results for Experiment 2 including user 6 are presented in Figure 19.



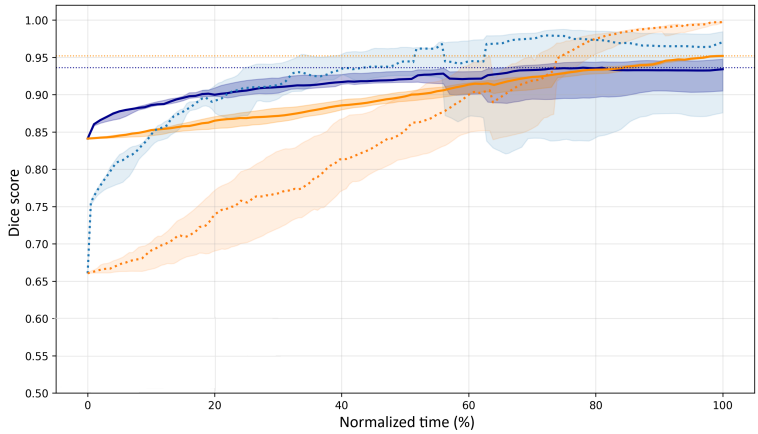
(a) Patient 3 **(b) Patient 5**

(c) Patient 6 **(d) Patient 6**

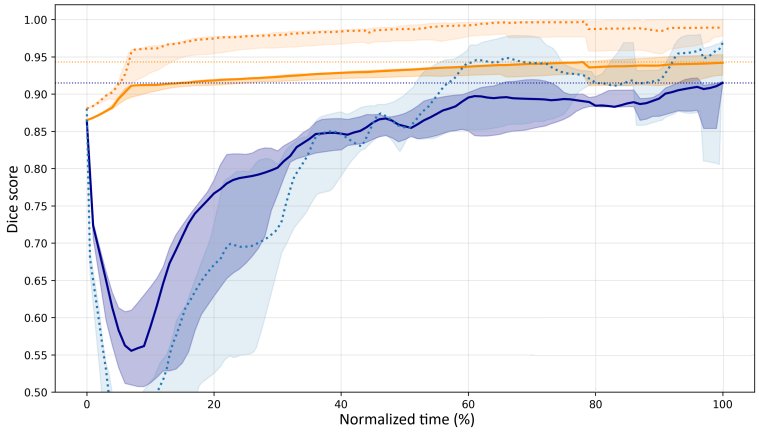
Figure 18. Overview of segmentation behavior of User 6



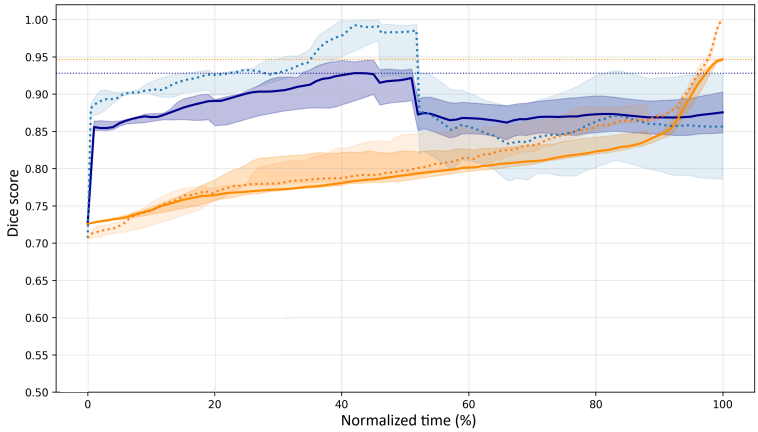
(a) Patient 1



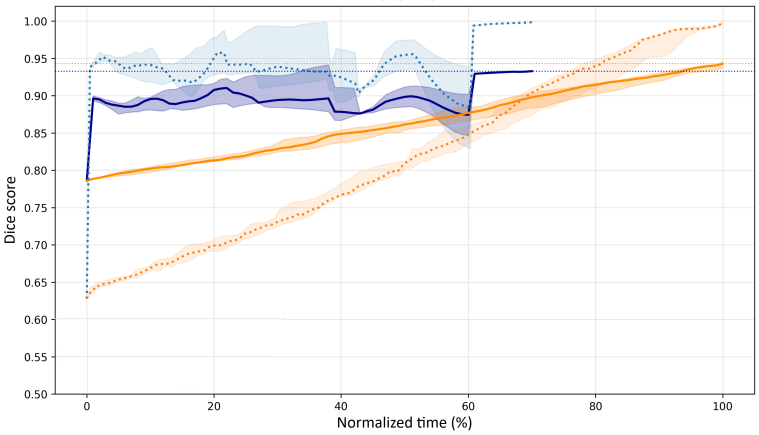
(b) Patient 2



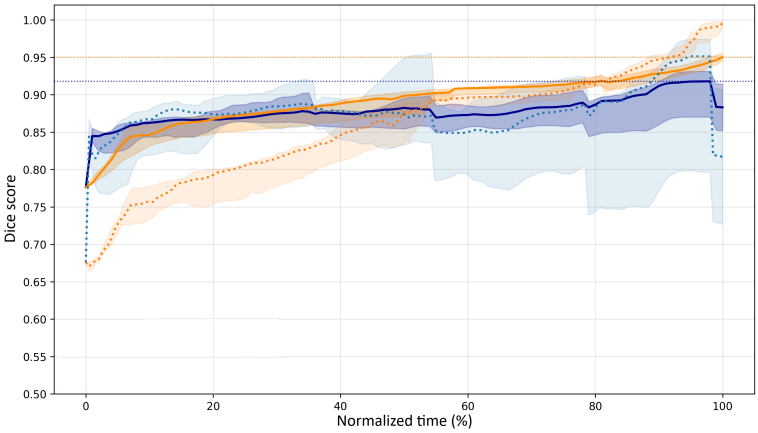
(c) Patient 3



(d) Patient 4

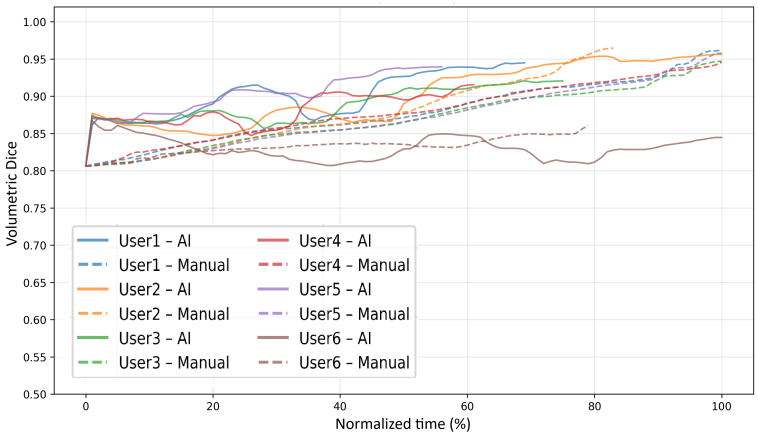


(e) Patient 5

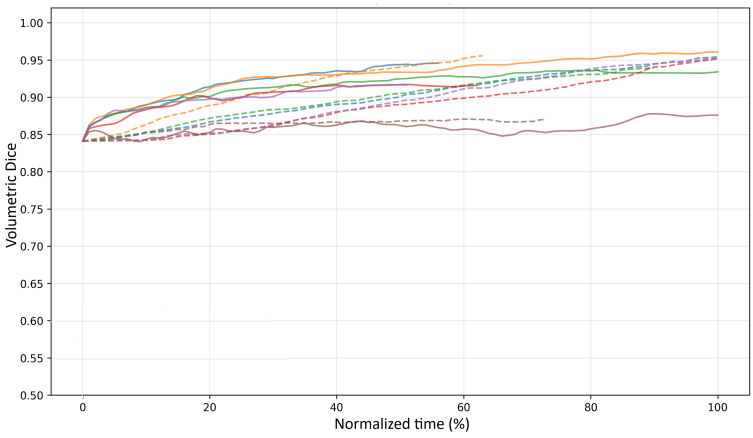


(f) Patient 6

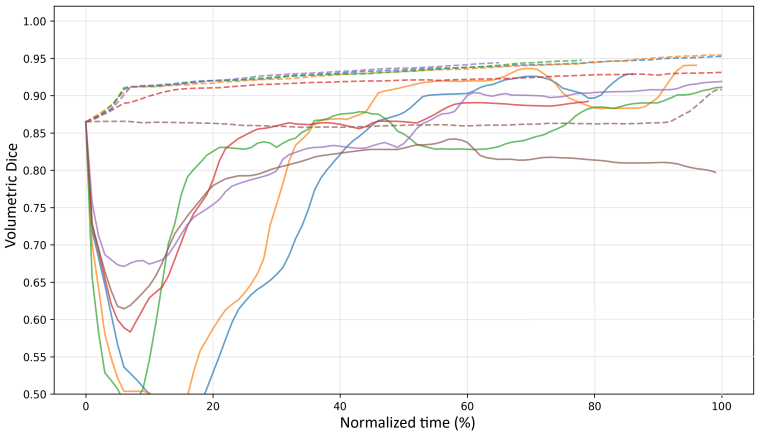
Figure 19. Dice and Surface Dice trajectories including user 6. Results are plotted against the normalized time axis for all six patients.



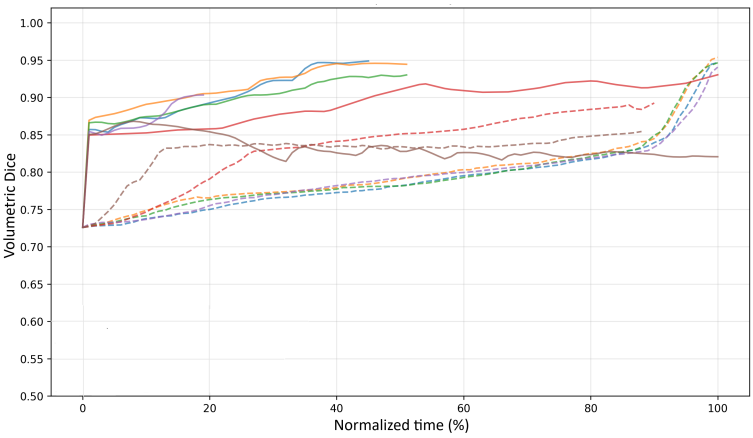
(a) Patient 1



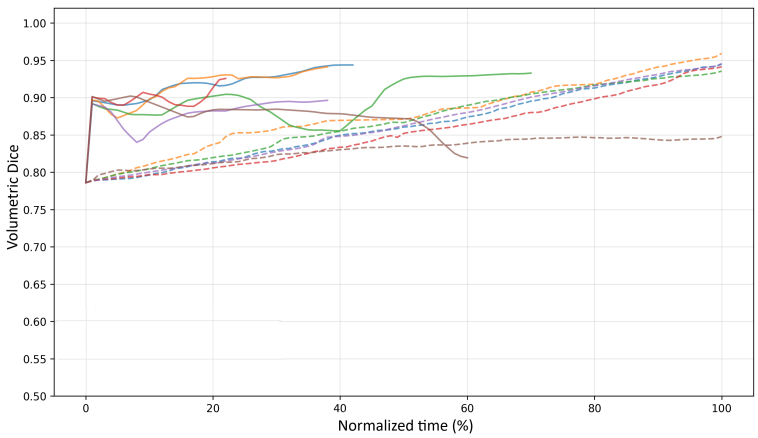
(b) Patient 2



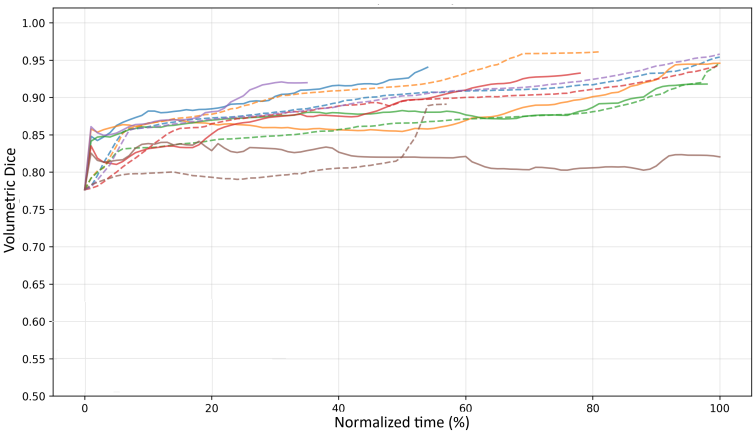
(c) Patient 3



(d) Patient 4



(e) Patient 5



(f) Patient 6

Figure 20. Dice trajectories for all users, including user 6, across all patients plotted against normalized time.

F. User input on PET and CT images

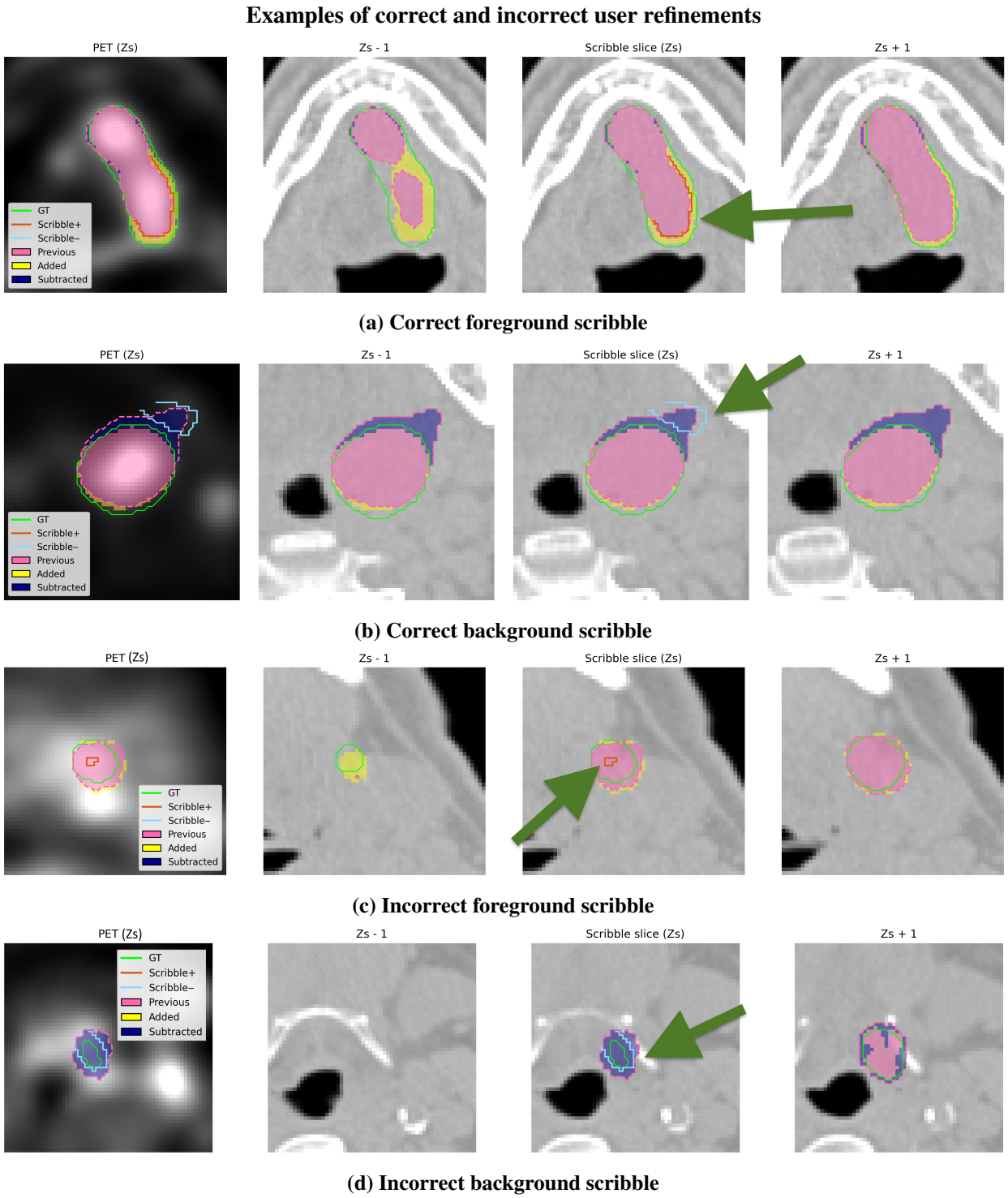
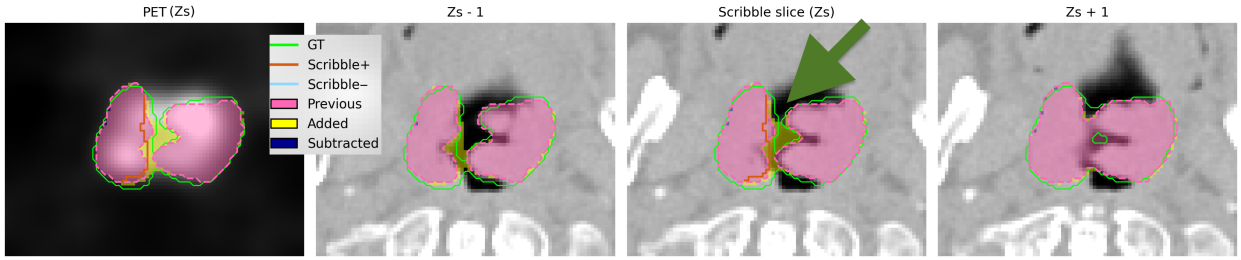
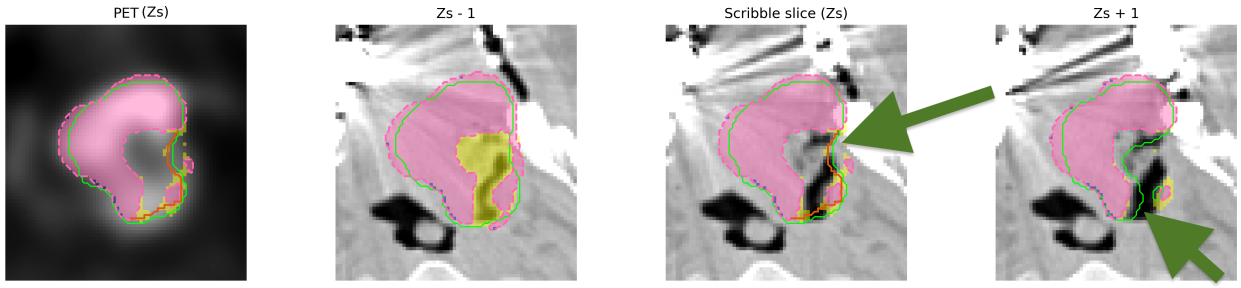


Figure 21. Examples of correct and incorrect user refinements using foreground and background scribbles. In each example, the PET image corresponding to the scribble slice is shown on the left, which was also visible to the user in the GUI. The remaining images are CT slices, with the central CT slice corresponding to the scribble slice. Arrows point to the scribble placed by the user. Green line: Ground truth. Orange line: User refinement (foreground scribble). Light blue line: User refinement (background scribble). Dashed pink line: Contour previous segmentation mask. Pink mask: Previous segmentation. Yellow mask: Added region after refinement. Blue mask: Removed region after refinement.

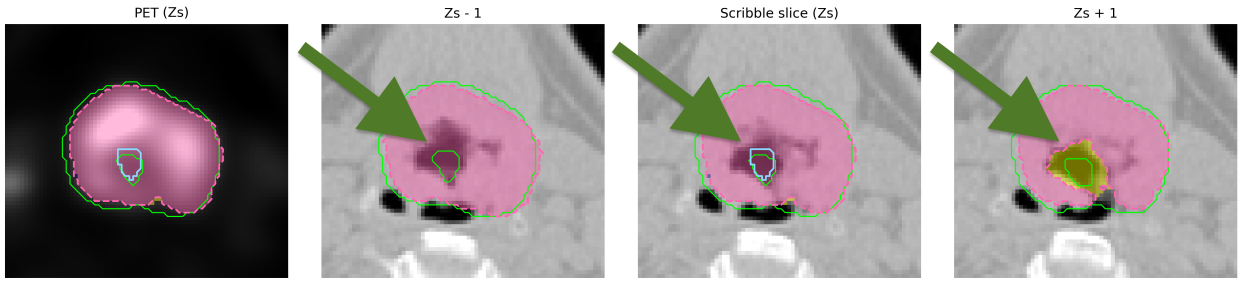
Segmentation errors due to complex anatomy



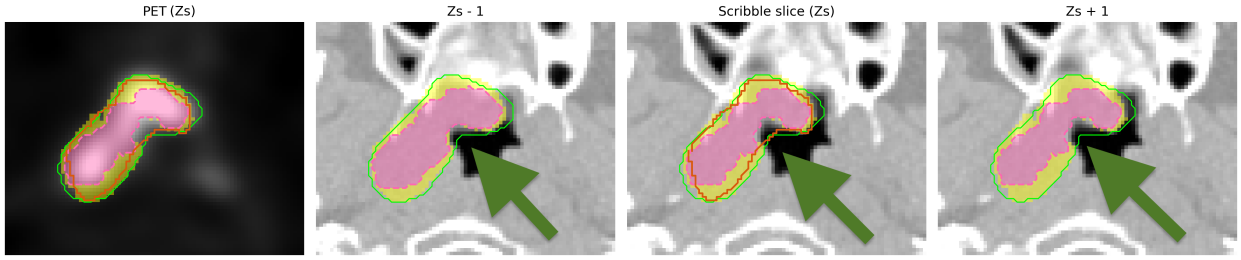
(a) Patient 6 (foreground scribble)



(b) Patient 2 (foreground scribble)



(c) Patient 3 (background scribble)



(d) Patient 6 (foreground scribble)

Figure 22. Errors due to complex anatomical structures like air cavities in the trachea region. Arrows point to the over- or under-segmented areas due to complex anatomical structures. The left image shows the PET scribble slice; the remaining images are CT slices, with the central slice corresponding to the scribble. Arrows point to the wrongly segmented areas. Green line: Ground truth. Orange line: User refinement (foreground scribble). Light blue line: User refinement (background scribble). Dashed pink line: contour previous segmentation mask. Pink mask: Previous segmentation. Yellow mask: Added part after refinement. Blue mask: Removed part after refinement.

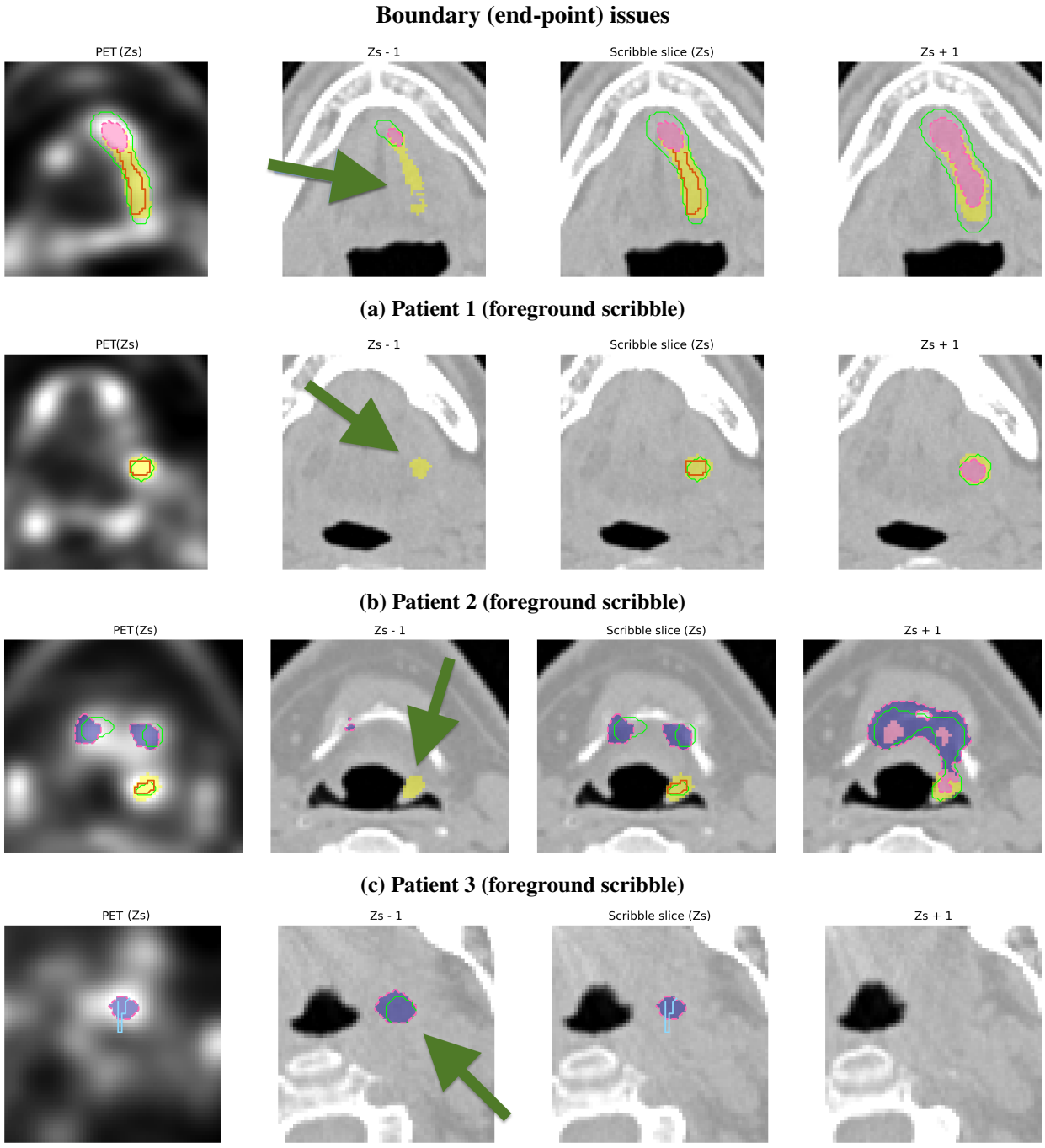


Figure 23. Boundary issues near the cranial and caudal endpoints of the segmentation mask. Arrows point to the over- or under-segmented areas due to boundary issues. The left image shows the PET scribble slice; the remaining images are CT slices, with the central slice corresponding to the scribble. Arrows point to the areas of false positive or false negative areas after user refinement. Green line: Ground truth. Orange line: User refinement (foreground scribble). Light blue line: User refinement (background scribble). Dashed pink line: contour previous segmentation mask. Pink mask: Previous segmentation. Yellow mask: Added part after refinement. Blue mask: Removed part after refinement.

G. Experiment 1 - RQ2 (Trustworthiness)

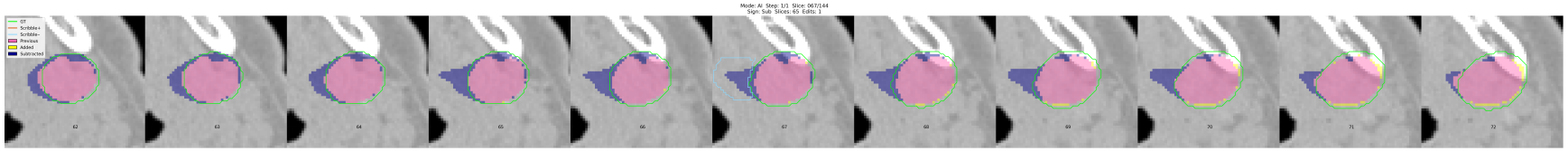


Figure 24. Experiment 1 (RQ2): Refinement results for the easy slice (P4) for Users 3. The central image corresponds to the slice on which the scribble was placed (slice 67/144). Slices to the right show the ±5 neighboring slices above this location (slices 68-72), while slices to the left show the ±5 neighboring slices below it (slices 62-66). Green: Ground Truth. Light Blue: scribble. Pink: prediction before. Yellow: added after refinement. Blue: removed after refinement.

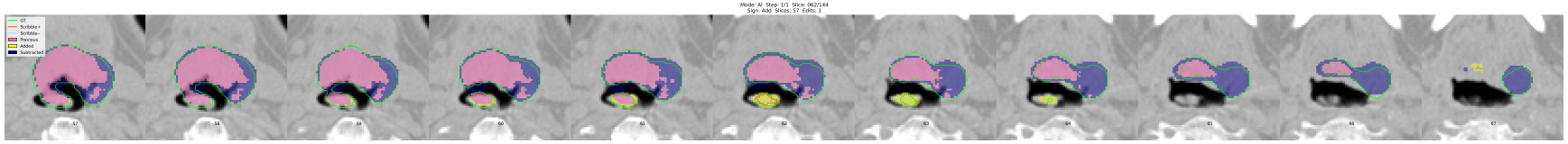
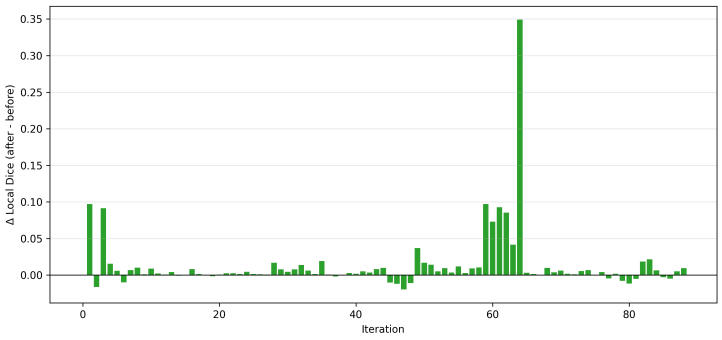
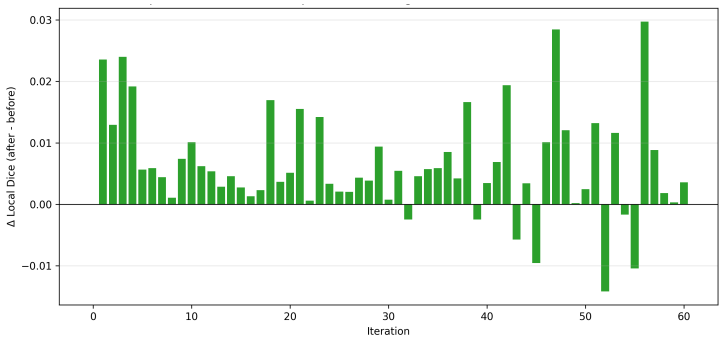


Figure 25. Experiment 1 (RQ2): Refinement results for the difficult slice (P3) for Users 3. The central image corresponds to the slice on which the scribble was placed (slice 62/144). Slices to the right show the ±5 neighboring slices above this location (slices 63-67), while slices to the left show the ±5 neighboring slices below it (slices 58-62). Green: Ground Truth. Red: scribble. Pink: prediction before. Yellow: added after refinement. Blue: removed after refinement.

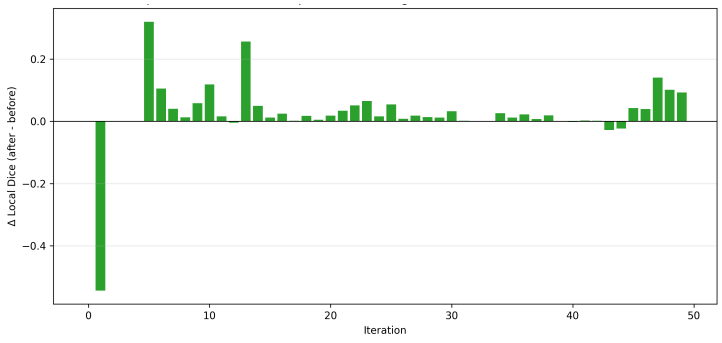
H. Δ Local Dice Graphs



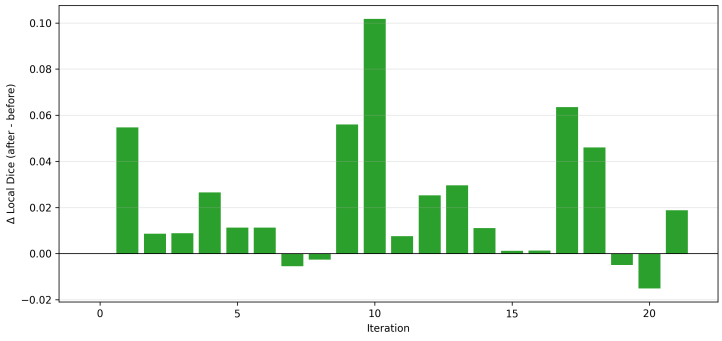
(a) Patient 1



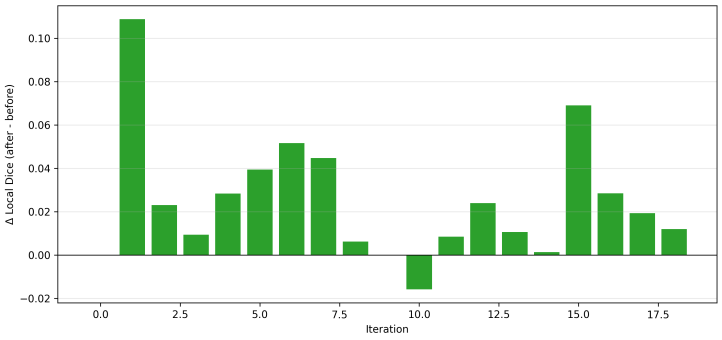
(b) Patient 2



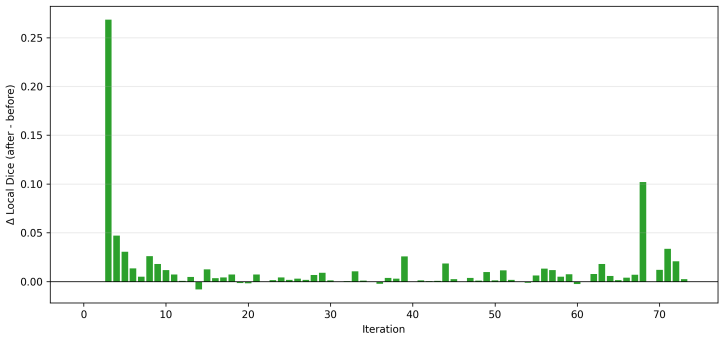
(c) Patient 3



(d) Patient 4

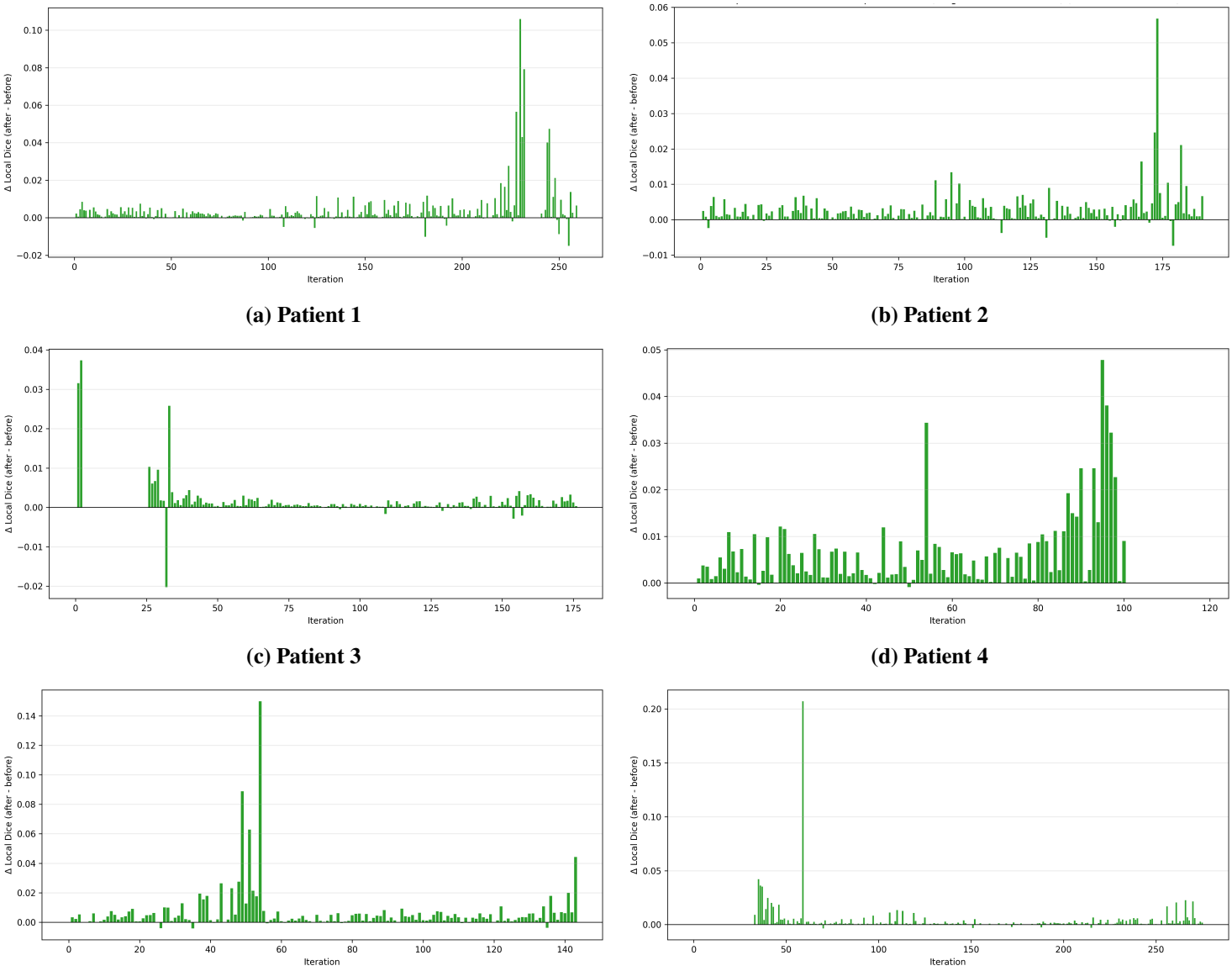


(e) Patient 5



(f) Patient 6

Figure 26. Δ Local Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (±5 Neighborhood)



(a) Patient 1

(b) Patient 2

(c) Patient 3

(d) Patient 4

(e) Patient 5

(f) Patient 6

Figure 27. Δ Local Dice outcomes (using manual brush) for all six patients following Experiment 2, performed by User 1. (±5 Neighborhood)

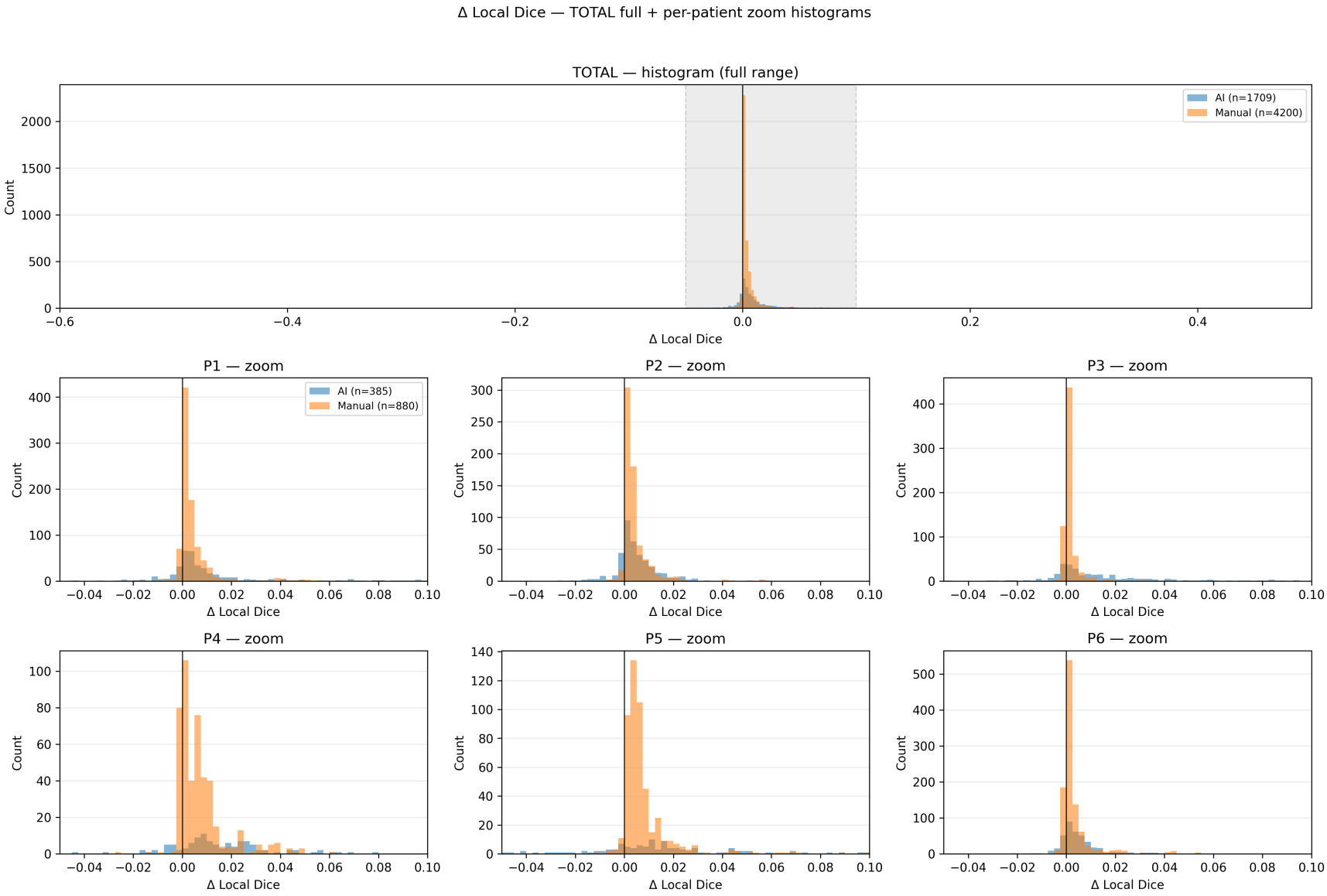
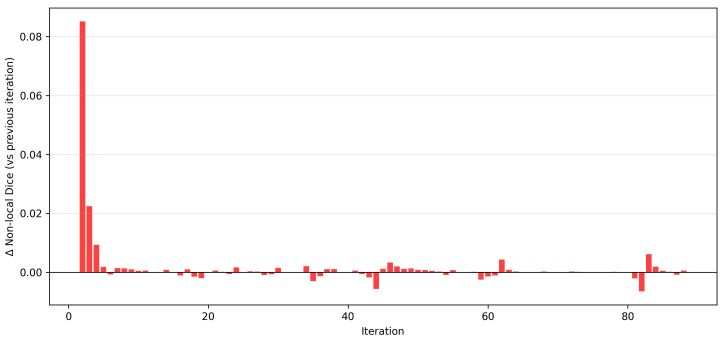
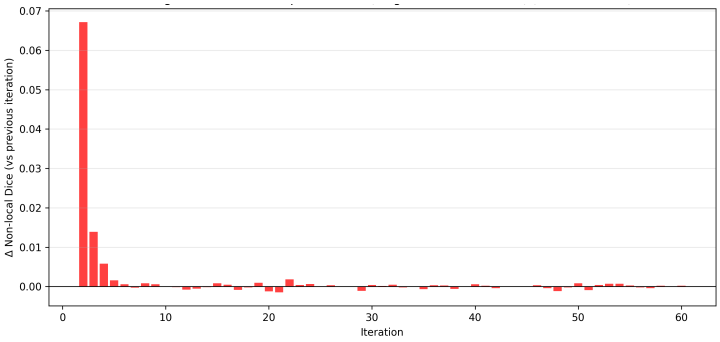


Figure 28. Histograms showing the full range and zoomed views per patient for Δ local Dice: AI (blue) and manual (orange).

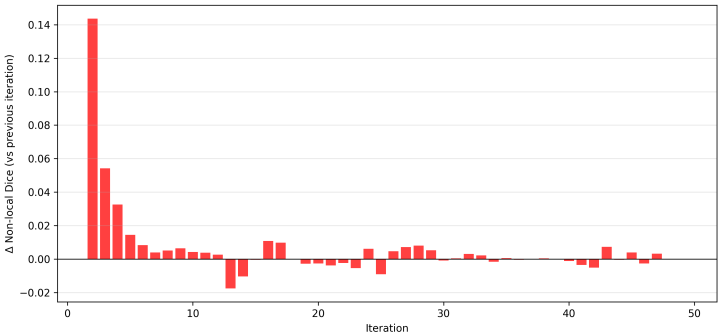
I. Δ Non-Local Dice Graphs



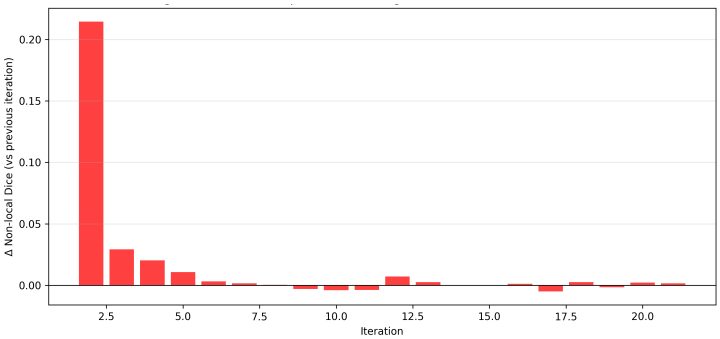
(a) Patient 1



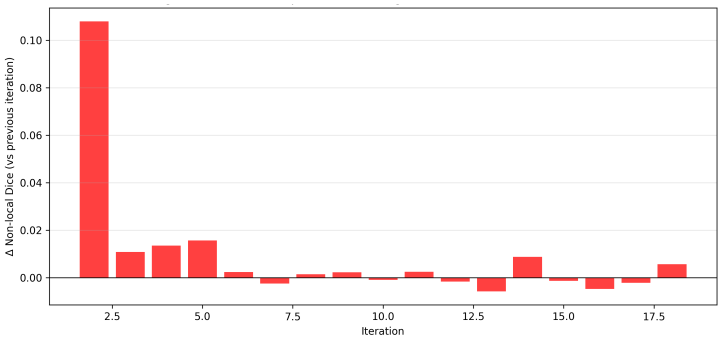
(b) Patient 2



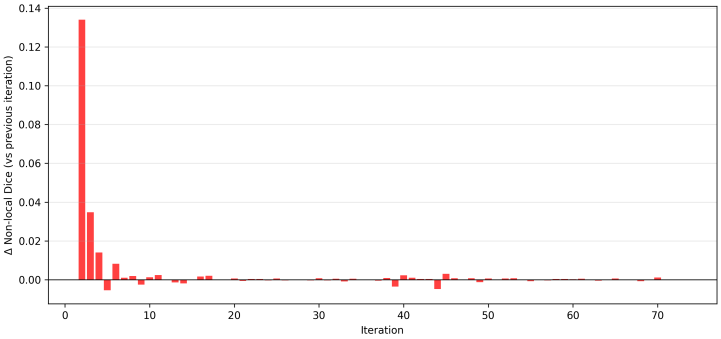
(c) Patient 3



(d) Patient 4

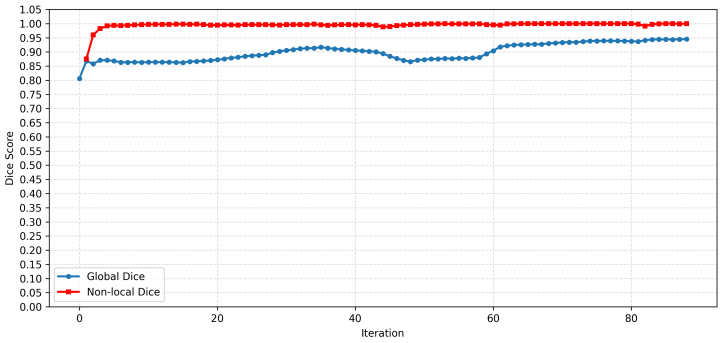


(e) Patient 5

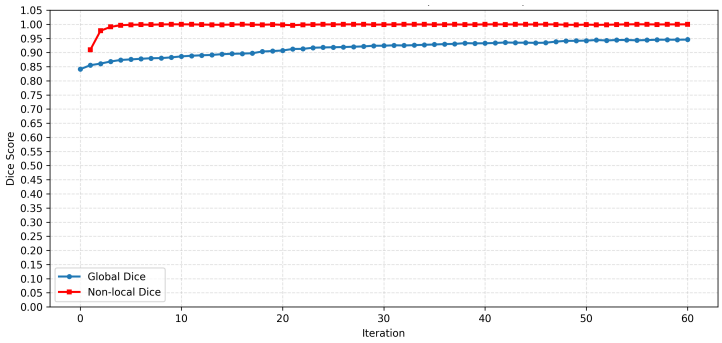


(f) Patient 6

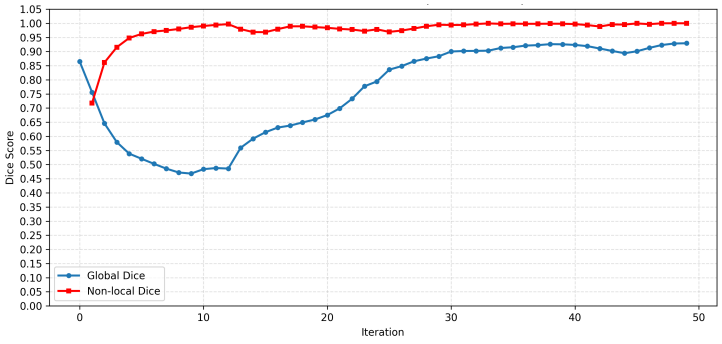
Figure 29. Non-Local Dice changes between iterations (using AI pencil) for all six patients following Experiment 2, performed by User 1. (±5 **Neighborhood)**



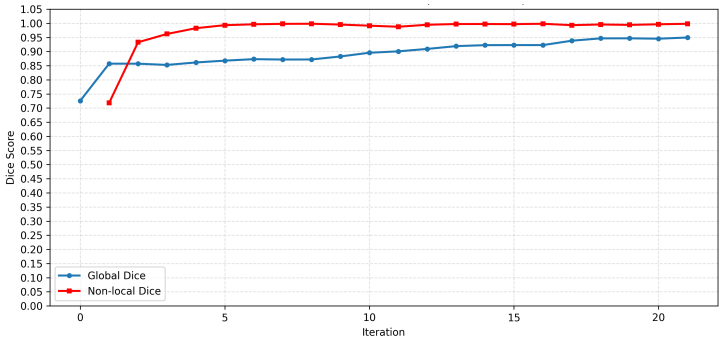
(a) Patient 1



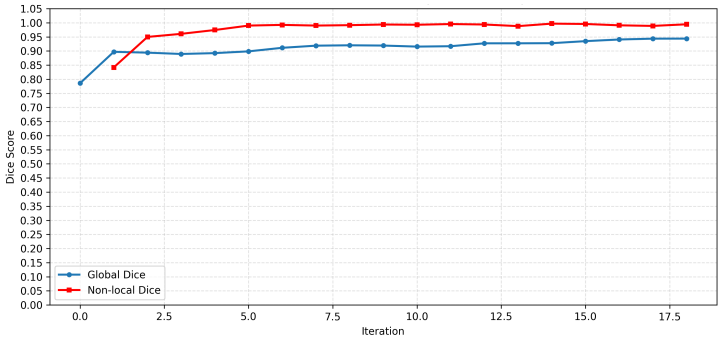
(b) Patient 2



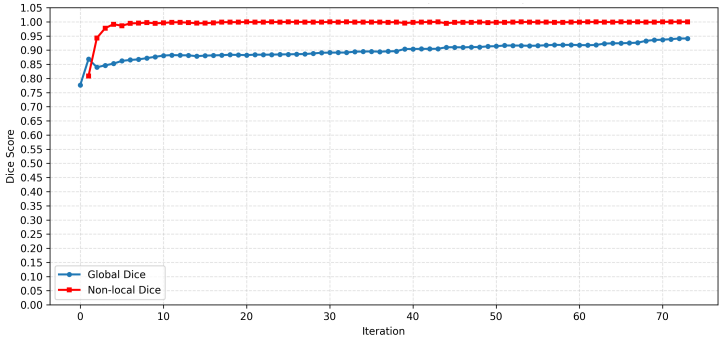
(c) Patient 3



(d) Patient 4

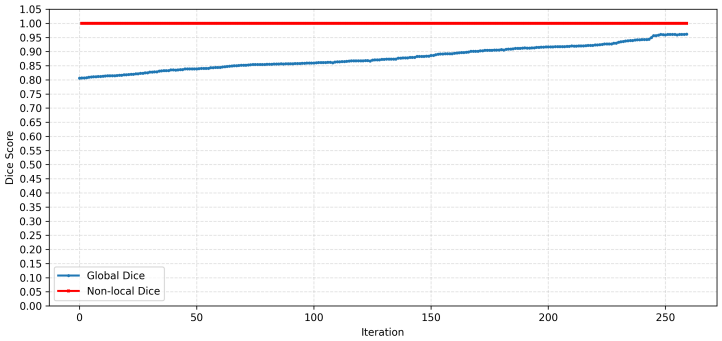


(e) Patient 5

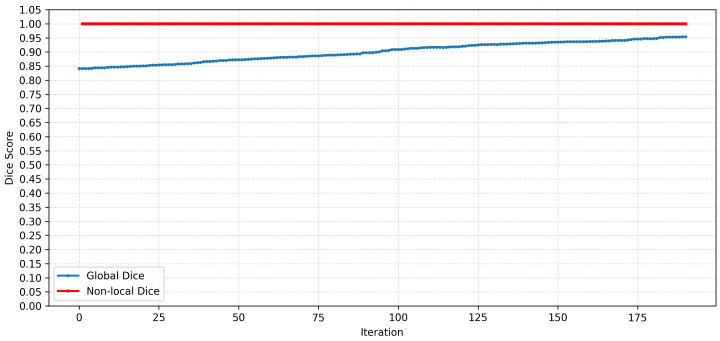


(f) Patient 6

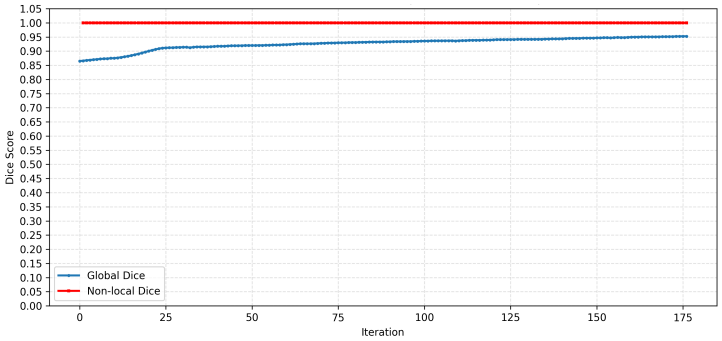
Figure 30. Global and Non-Local Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (±5 **Neighborhood)**



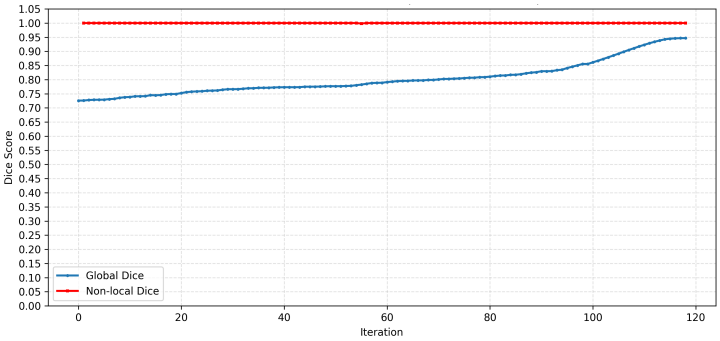
(a) Patient 1



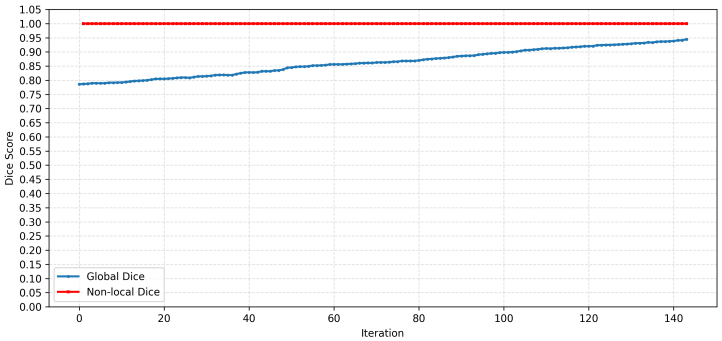
(b) Patient 2



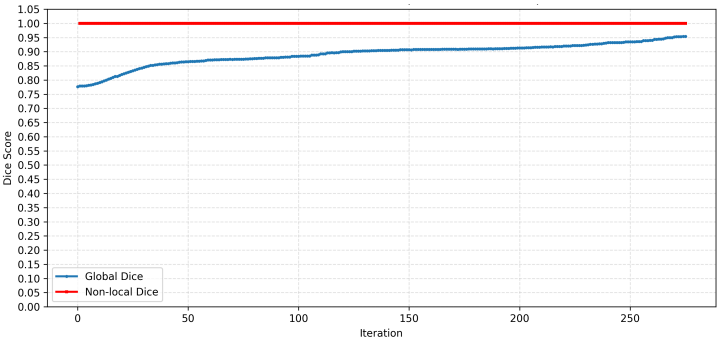
(c) Patient 3



(d) Patient 4



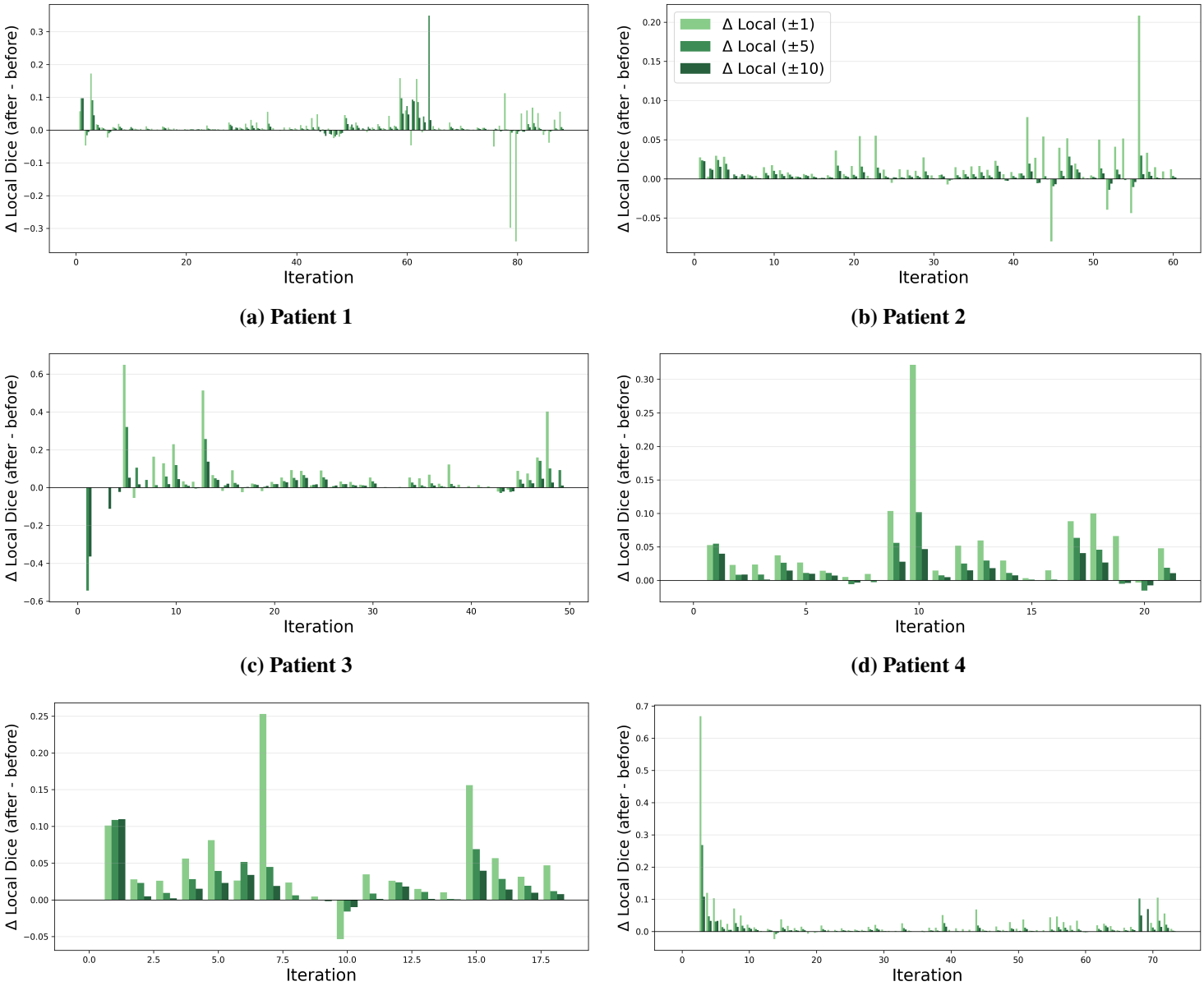
(e) Patient 5



(f) Patient 6

Figure 31. Global and Non-Local Dice outcomes (using manual brush) for all six patients following Experiment 2, performed by User 1. (±5 Neighborhood)

J. Local and Non-local Dice and Surface Dice results for different neighborhoods



(a) Patient 1

(b) Patient 2

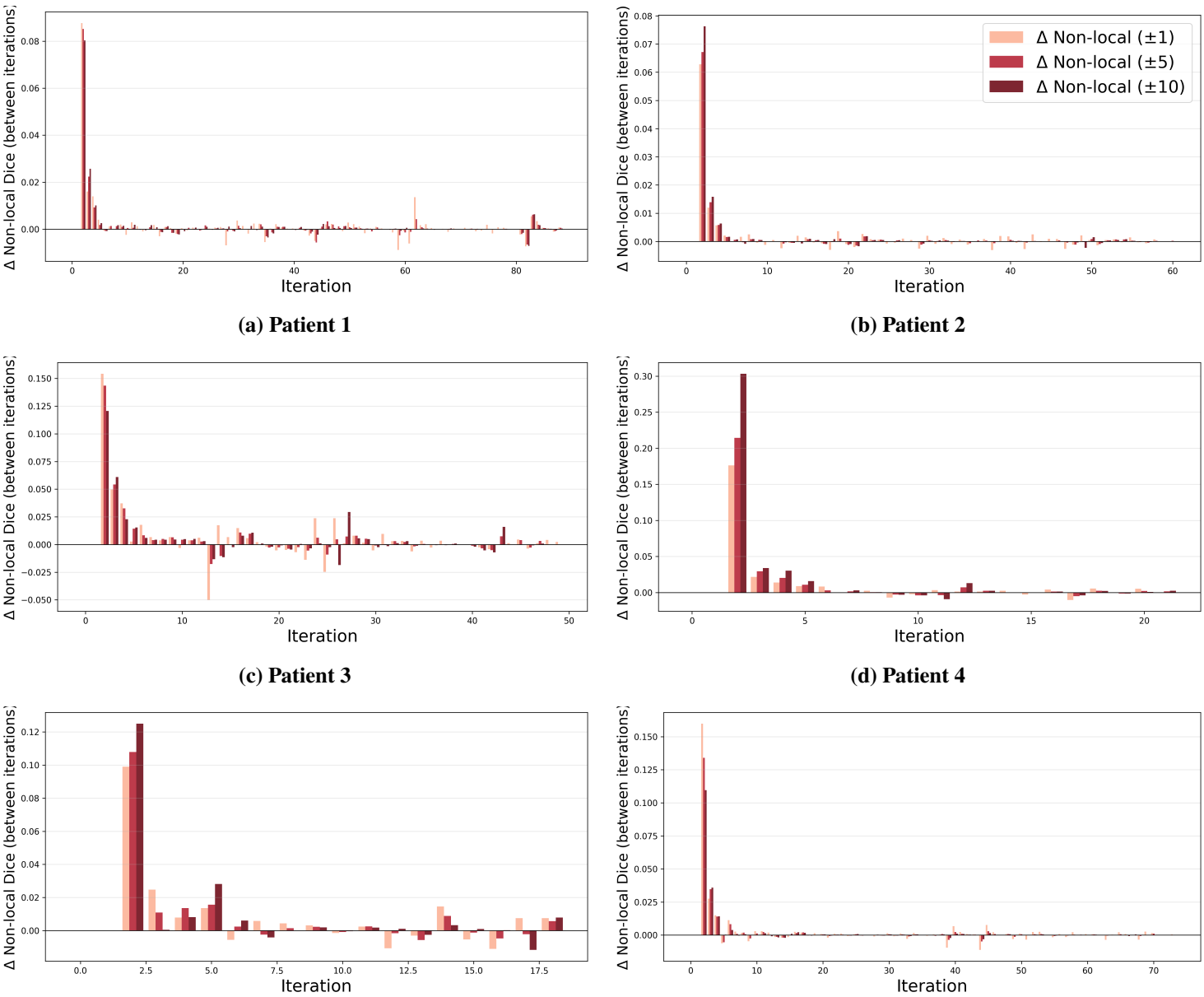
(c) Patient 3

(d) Patient 4

(e) Patient 5

(f) Patient 6

Figure 32. Δ Local Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (Neighborhoods ±1 ±5 ±10)



(a) Patient 1

(b) Patient 2

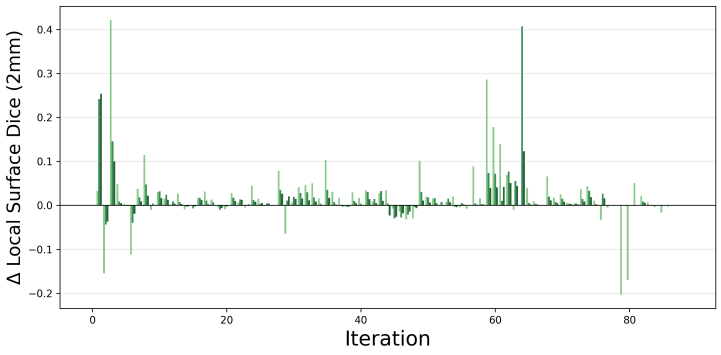
(c) Patient 3

(d) Patient 4

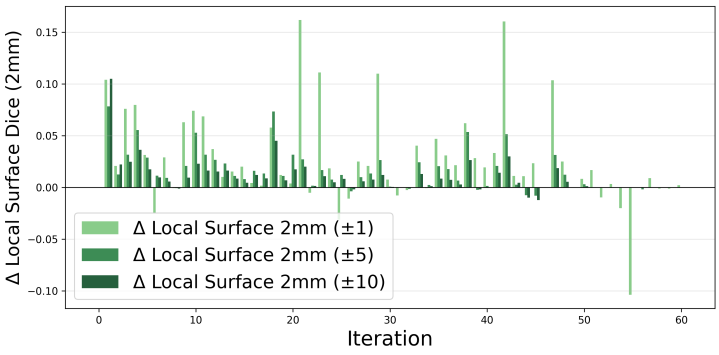
(e) Patient 5

(f) Patient 6

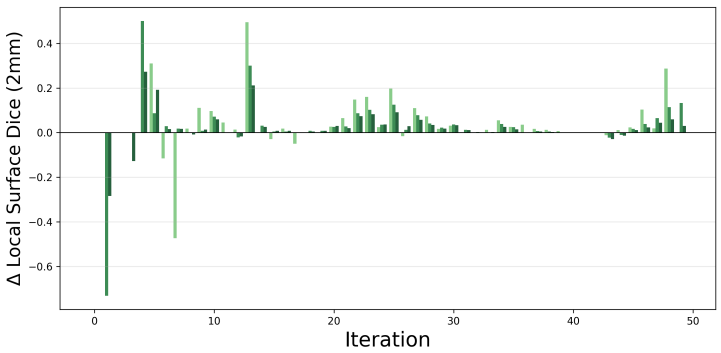
Figure 33. Δ Non-Local Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (Neighborhoods ±1 ±5 ±10)



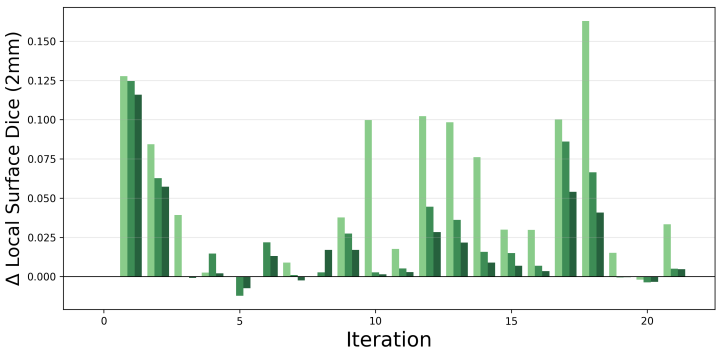
(a) Patient 1



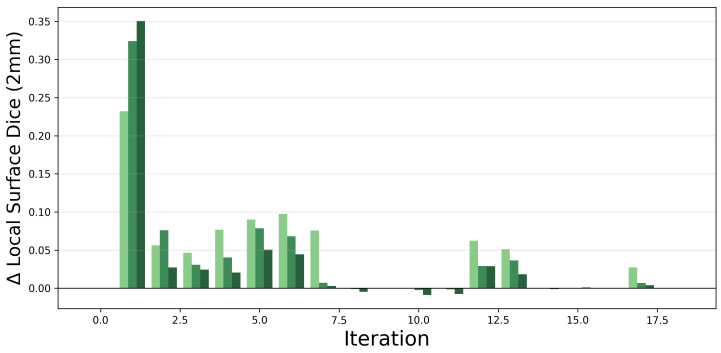
(b) Patient 2



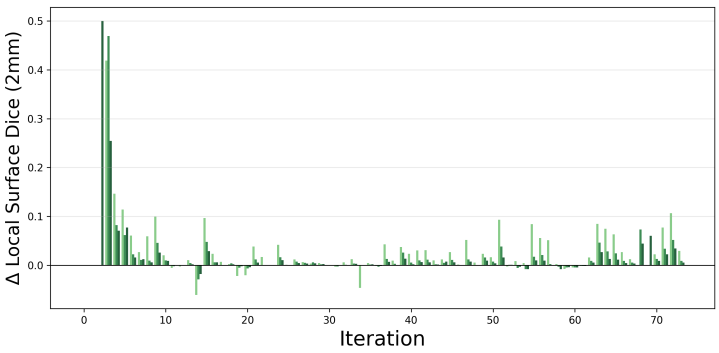
(c) Patient 3



(d) Patient 4

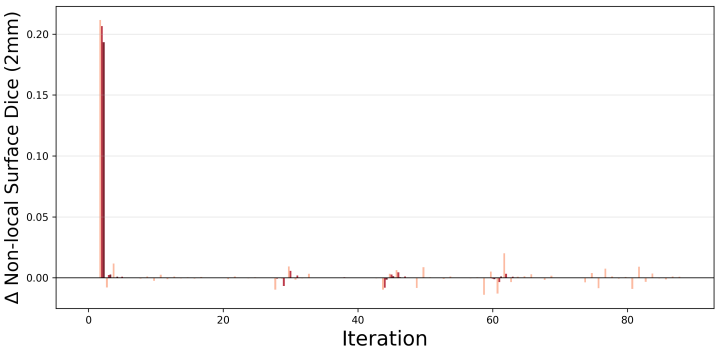


(e) Patient 5

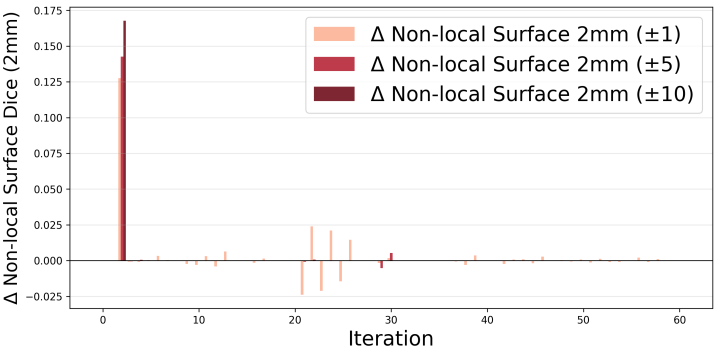


(f) Patient 6

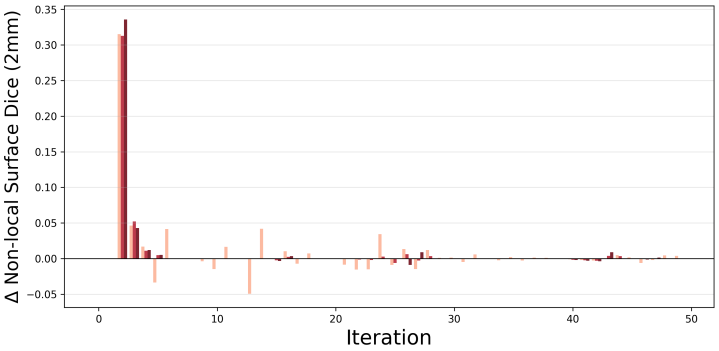
Figure 34. Δ Local Surface Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (Neighborhoods ±1 ±5 ±10)



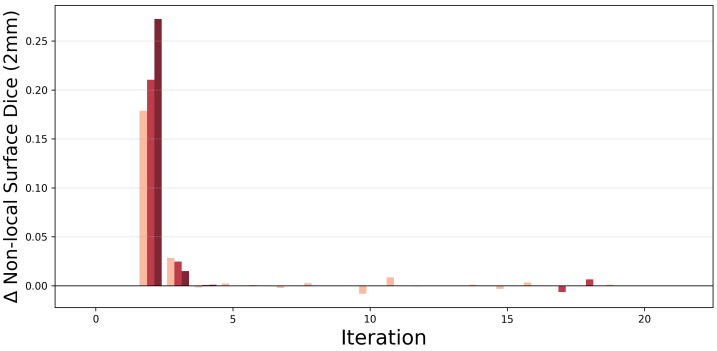
(a) Patient 1



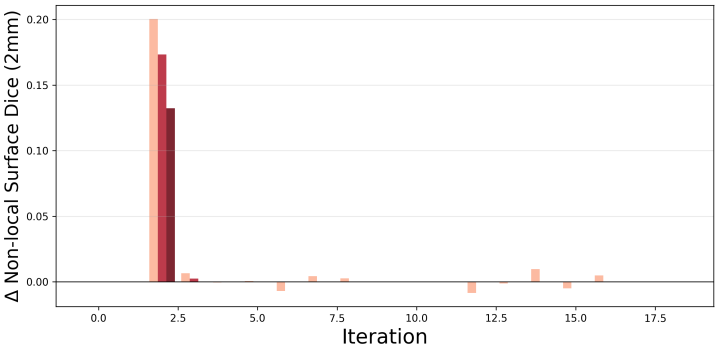
(b) Patient 2



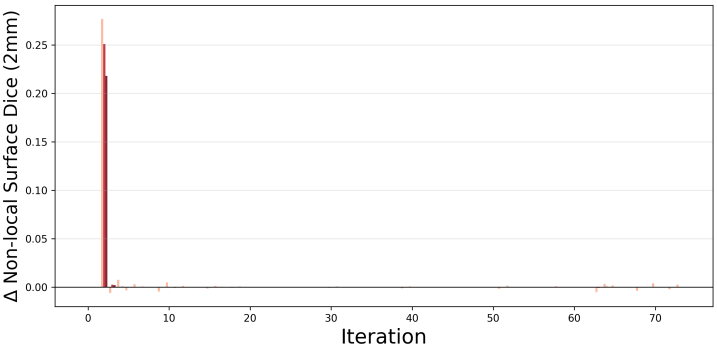
(c) Patient 3



(d) Patient 4



(e) Patient 5



(f) Patient 6

Figure 35. Δ Non-Local Surface Dice outcomes (using AI pencil) for all six patients following Experiment 2, performed by User 1. (Neighborhoods ±1 ±5 ±10)