Towards fast human-centered contouring workflows for adaptive external beam radiotherapy

Nicolas F. Chaves-de-Plaza[1], Prerak Mody[2], Klaus Hildebrandt[1], Marius Staring[2], Eleftheria Astreinidou[2], Mischa de Ridder[3], Huib de Ridder[1] and René van Egmond[1]
[1]Delft University of Technology, [2]Leiden University Medical Center, [3]University Medical Center Utrecht, [4]HollandPTC consortium – Erasmus Medical Center, Rotterdam, Holland Proton Therapy Centre, Delft, Leiden University Medical Center (LUMC), Leiden and Delft University of Technology, Delft, The Netherlands Netherlands

Abstract

Delineation of tumors and organs-at-risk permits detecting and correcting changes in the patients' anatomy throughout the treatment, making it a core step of adaptive external beam radiotherapy (EBRT). Although auto-contouring technologies have sped up this process, the time needed to perform the quality assessment (QA) of the generated contours remains a bottleneck, taking clinicians between several minutes up to an hour to complete. The authors of this article conducted several interviews and an observational study at two treatment centers in the Netherlands to identify challenges and opportunities for speeding up the delineation process in the context of adaptive therapies. The present article starts by describing several performance factors uncovered in the study. Then, it discusses two clinician-centered strategies for accelerating the contouring process. First, enable targeted inspection of the generated contours by leveraging AI uncertainty and clinically relevant features such as the proximity of the organs-at-risk to the tumor. Second, minimize the number of interactions needed to edit faulty delineations with redundancy-aware editing tools that provide the user a sense of predictability and control. Tool developers and workflow builders can follow these strategies to increase contouring efficiency without compromising the patient's outcome.

Introduction

External Beam Radiotherapy (EBRT) is the most common form of RT and has become one of humanity's main tools against cancer, together with surgery and systemic treatment. In EBRT, ionizing radiation is directed at the patient's tumor to destroy the malignant cells. Over the last decades, significant technological improvements have been made in treatment planning and delivery, which increased the precision of EBRT. For instance, proton beam therapy (PT) can harness the ability of protons to deposit all their energy (Newhauser & Zhang, 2015; Wilson, 1946). This capability permits PT more precisely shaping the radiation dose to the tumor, minimizing the dose to the surrounding healthy tissue and reducing side effects (Langendijk et al., 2013; Lundkvist et al., 2005; Simone et al., 2011; Thomas & Timmermann, 2020).

In D. de Waard et al., [more] (Eds.) (2022). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

During an EBRT treatment course, which usually lasts several weeks, the patient anatomy can change. When this happens, it is often necessary to update the treatment plan, which entails executing the colored processes in the dose delivery pipeline of Figure 1. The present study focuses on the contouring process, which, despite the introduction of auto-contouring methods, still requires significant intervention of clinicians who must perform an extensive quality assessment of the generated delineations to ensure that they are clinically acceptable (Cardenas et al., 2019; Nikolov et al., 2020; van Dijk et al., 2020).. For instance, in the institutions surveyed in the present work, contouring can take more than an hour. Even when automatically generated initial contours are available.

===== INSERT FIGURE 1 APPROXIMATELY HERE =====

Existing studies seek to make the contouring workflow fit into the reduced time windows of adaptive therapies in two ways. First, bottom-up investigations focus on how clinicians perform the contouring task, providing a deeper understanding of their information needs (Steenbakkers et al., 2006), the optimal workspace conditions (Multi-Institutional Target Delineation in Oncology Group, 2011), and producing design insights for the user interfaces and editing tools to improve contouring performance (Aselmaa et al., 2014, 2017; Ramkumar, 2017; Ramkumar et al., 2016). These studies often explore their research questions in a controlled environment using software developed for this purpose (Steenbakkers et al., 2005), which ignores clinicians' context and limits the scope of their findings to ways of optimizing current software solutions.

On the other hand, articles from the medical domain often take a top-down view of the contouring activity, discussing aspects like the commissioning process of auto-contouring technologies and how to integrate them into clinical practice (Cardenas et al., 2019; Vandewinckele et al., 2020). Even though clinical papers consider the clinical context, their focus on auto-contouring technology distracts them from the clinician's role in the process. The latter can result in clinical workflows like those observed at the surveyed institutions, where even though automation has greatly accelerated contouring, the activity is still time-consuming because clinicians must manually inspect and edit faulty delineations.

The present study aims at bridging the bottom-up and top-down approaches discussed above. Concretely, it is assumed that contouring will remain a human-centered process for the foreseeable future and seeks how to make it suitable for adaptive EBRT workflows. Concretely, this article contributes to the state-of-the-art of clinical contouring workflows in adaptive EBRT in two ways:

1. It reports the results of an observational study in two cancer treatment centers in the Netherlands. The Study of the Contouring Workflow provided a situated account of the current contouring workflows in the context of adaptive EBRT, together with factors that can affect its performance.

towards fast contouring workflows for adaptive proton therapy

2. It discusses acceleration strategies based on the context of adaptive radiotherapy that tool developers and clinicians can leverage to adapt the contouring workflow to time-constrained scenarios.

The Contouring Activity

An exploratory literature review was performed to establish baseline knowledge about the contouring activity and its role in adaptive therapies. The query used for the search (Scopus, PubMed, and Google Scholar) included the keywords: adaptive, adaptation, proton therapy, radiotherapy, contouring, automatic, semi-automatic, workflow, and head-and-neck. The latter term was relevant since the study's participants (next section) were specialists in this region. The search yielded around 50 articles with publishing years ranging between 2008 and 2021.

As Figure 2 depicts, the main inputs of the contouring activity are 3D images (stacks of hundred of 2D images) that describe the patient anatomy. Among these, there is an image to contour, usually a Computerized Tomography (CT), and supporting information such as previous contours of the patient and other image modalities such as Magnetic Resonance Imaging (MR) and Positron Emission Technology CT (PET-CT). Using available information, the contouring task consists of drawing the contours of anatomical structures in the image to contour relevant to the patient's cancer. The two main anatomical groups are the target volumes (TVs), which correspond to areas affected by tumoral cells, and the organs at risk (OARs), which correspond to healthy tissue.

===== INSERT FIGURE 2 APPROXIMATELY HERE =====

As the right panel of Figure 2 indicates, the goal of the contouring activity is to produce contours suitable for creating or updating the patient treatment plan and assessing its quality. Several actors participate in this workflow in the clinic, distributing contouring tasks based on the anatomical structures' groups. In general, radiotherapy technologists (RTTs) start by delineating the OARs. After this, the radiation oncologists (ROs), which are directly responsible for the patient’s outcome, assess the quality of the OARs contours and draw the boundaries of the TVs, the structures with the highest priority. The study that the next section describes was designed based on this understanding of the contouring activity.

Study of the Contouring Workflow

A study of the contouring workflow was conducted to identify characteristics of adaptive EBRT that affect contouring performance and to identify context-dependent strategies that tool developers can leverage to improve it. The following sections detail the study's design and describe the methodology used for analyzing the resulting data.

Study design

Participants

Two radiation oncologists (RO) and two radiotherapy technologists (RTT) from two cancer treatment centers in the Netherlands specializing in the head-and-neck area joined the study. Table 1 summarizes the participants' information. One of the institutes, the Leiden University Medical Center (LUMC), offers photon-based VMAT treatments. The second, the Holland Proton Therapy Center (HollandPTC), offers proton therapy (PT). Despite the differences in dose delivery technology, both institutions have a similar workflow, performing offline adaptations. The latter means that the patient's treatment plan is updated sparsely during treatment (entails re-executing blue boxes in Figure 1). The Institutional Review Board at the Delft University of Technology approved this research. Each participant provided informed consent to be part of the study.

Procedure

The study had three sessions. The first one, a one-hour-long semistructured interview, permitted establishing rapport with the participants and validated the initial understanding of the EBRT workflow. In the second and third sessions, the participants performed their contouring duties while being recorded. As Table 1 shows, these meetings lasted between one and two hours, depending on the participants' time. In the second session, clinicians performed initial contouring. The third focused on adaptive contouring, where clinicians perform a quality assessment of automatically generated contours. Given the limited clinicians' time, they contoured a subset of anatomical including the tumors and organs close to them that could affect the patient outcome.

Materials

For the observational sessions, clinicians at each center had access to the data of two previously treated head and neck patients. Each patient file included initial treatment planning data such as CT, PET-CT, and MRI scans and daily images such as CBCT and CT, relevant for sessions 2 and 3, respectively. For session 3, starting delineations could have been generated by another clinician or automated methods like deformable or rigid registration and deep learning-based contouring. For inspecting and editing the contours, clinicians used their routine software.

Data Analysis

The recordings of the three sessions were transcribed and analyzed using Thematic Analysis (Braun & Clarke, 2006). The coding process was bottom-up, first labeling patterns in the transcripts and then grouping the resulting fine-grained codes into coarser ones based on their similarity. Table 2 displays the underlying coarser codes, the resulting themes, and sample data excerpts. The screen recordings of sessions 2 and 3 were also relevant as they showcased the way clinicians interact with the user interface during the contouring process. The interactions were mapped onto a timeline like the one that Figure 4 depicts. For the y-axis, the authors drew inspiration from the literature on contouring tasks (Aselmaa et al., 2017) but grouped them into four categories to simplify the coding process and the analysis.

These are direct and indirect manipulation, navigation, and non-contouring interactions.

Initial Contouring

Results

Initial contouring (IC) occurs when executing the plan creation and offline adaptation process in Figure 1 for the first time. At LUMC and HollandPTC, initial contouring (IC) takes two to six hours for head-and-neck (HN) cancers, requiring delineating more than twenty structures. The following paragraphs group the observations about the IC workflow into three characteristics. Finally, this section discusses how these characteristics can affect contouring performance.

Usable Additional Information
At IC, no pre-existing contours of the patients exist, given that this process occurs after they have started treatment. Instead, clinicians use information from multiple image modalities acquired in the simulation process. The main image modality in radiotherapy, CT, usually does not provide enough boundary information when the contrast between adjacent tissues is not enough or when there is noise or artifacts in the image acquisition process. In these cases, clinicians rely on Magnetic Resonance Imaging (MRI) and Positron Emission Technology-CT (PET-CT) scans, acquired for most patients at HollandPTC and LUMC. As Figure 3 shows, MRI helps differentiate soft tissue structures:" MRI makes it easier for us to delineate the parotid glands because you can see them very good at an MRI.". In the case of PET-CT, this modality permits clinicians to locate tumors and estimate their boundaries with higher precision:" We actually scan all of our head and neck patients [with PET-CT] because it makes our delineations that more easy and more accurate, so that is now standard." [P1].

===== INSERT FIGURE 3 APPROXIMATELY HERE =====

In practice, clinicians align additional images to the CT before using them for contouring. This process, known as image registration, can take several minutes per image pair and requires the clinician's intervention to verify the alignment's quality. Registering the images allows clinicians to scroll through them in parallel using the contouring software, enabling direct comparison of the structures in both scans.

Applicable Domain-Specific Knowledge
In some cases, the information in the images is not enough. At IC, this happens when MRI and PET-CT scans are not available and because there are no pre-existing contours of the patients (they just started the treatment). In these cases, clinicians rely on domain-specific knowledge they access in two ways. First, they leverage guidelines (Brouwer et al., 2015) and atlases that describe and indicate what the contours should look like, respectively. Second, they draw on their experience. Experienced clinicians know what areas can be challenging to delineate given the available data. They use this domain-specific anatomical knowledge to direct their attention and estimate contours over unclear image boundaries. An example of this dynamic occurs when the radiation oncologists (ROs) review the delineations

created by the radiotherapy technologists (RTTs): “We [ROs] think that it [delineating the swallowing muscles] is too hard for RTTs, need quite a bit of anatomical knowledge to know where they are exactly. And in this case, this patient doesn't have a very big tumor in the throat, but most of the time patients have quite a big tumor here. And you can't see the swallowing muscles that good. So, then you need to know exactly where they run from to delineate them.” [P1].

Editing Capabilities of Contouring Software

In practice, at IC, clinicians create the contours from scratch. As the timeline on the top section of Figure 4 depicts, this entails starting with an empty delineation and gradually building the contours through a series of interactions. At the surveyed institutions, clinicians favored a semi-automatic workflow, which consisted of two phases. First, they generated initial contours using the between-slice interpolation tool, which only requires contouring a subset of the slices that the structure spans. After auto-completing the rest of the delineations (indirect editing around eighty seconds in), they proceeded to correct potential inaccuracies manually with the brush tool. As the timeline shows, the generation of contours takes more time than the refinement, and clinicians spend most of the time directly editing the delineations with the brush.

===== INSERT FIGURE 4 APPROXIMATELY HERE =====

Discussion

Clinicians use contours produced at IC to create the patient’s treatment plan. Therefore, they seek maximal accuracy, often at the expense of longer task durations. The three characteristics of the IC context described before affect contouring time in several ways. First, extra image modalities reduce the task difficulty, which can result in reduced dwelling times to determine where the contour should go. Nevertheless, additional images need to be registered to the main one, a time-consuming process that could offset the performance benefits gains that the process offers. Second, domain-specific knowledge can reduce the extent of the contouring task by letting clinicians direct their attention to where it is needed. Yet, following the accuracy directive, they still must go through the whole volume to ensure no inaccuracy remains. Finally, the semi-automatic between slice interpolation tool spares clinicians from needing to edit several slices but still requires significant manual effort to initialize the method.

Adaptive Contouring

Results

LUMC and HollandPTC implement an offline-adaptive dose delivery pipeline, which entails updating the treatment plan several times during treatment by repeating the plan creation and offline adaptation process between fractions. Adaptive contouring (AC) occurs in this setting and differs from initial contouring (IC) in that the time is more critical and the resources scarcer. At the surveyed institutions, AC takes one to two hours for head and neck cancer patients. Like the previous section, the following paragraphs detail the AC context and discuss how it affects the process' performance.

Usable Additional Information

In contrast with IC, at AC, no extra images of the patient are acquired. Therefore, clinicians have access to the image to contour, a CT at LUMC and HollandPTC, the images acquired for IC, and the approved IC contours. In practice, clinicians only use the latter and do so in two ways. First, because IC contours document all the clinical decisions made for the current patient, they use them as a patient-specific atlas to resolve complex contouring tasks. Regarding having an atlas for contouring, P4 mentioned that" it's always nice to have it [the atlas] like a verification. Because the brainstem isn't that difficult, but like if you have the swallowing muscles or something, that's really something. If you have the atlas side by side, it really can come in handy." [P4] Second, clinicians use approved IC contours to create an initial segmentation. For this, they align, or register, the IC and AC images and then "propagate" the contours from the former to the latter.

Applicable Domain-Specific Knowledge

In addition to general anatomical knowledge, at AC, clinicians use knowledge about dosimetry and the patient tumor to structure and guide the contouring process. On the one hand, it can help them direct their attention to critical areas. On the other, it lets them modulate the contouring based on the structure's relevance to the patient's treatment plan. For instance, P2 mentioned that while some contours require maximal attention and precision:" with this type of organs, so all the nervical organs, as in optical nerves and brain stem and spinal cord, when it's critical, so when the PTV is nearby, then it's very important that we draw this very precise." Others accept rougher contours as they will not significantly impact the patient's outcome:" this submandibular gland, it gets too much dose, so it won't work. After irradiation, this one is gone. So, at that point, we can decide to delineate, but it isn't, it's OK if it isn't quite perfect."

Editing Capabilities of Contouring Software

As mentioned before, clinicians do not start delineating from scratch at AC. Instead, they generate a starting point by propagating the contours from the initial scan to the current one. Therefore, the goal at AC is to perform a quality assessment (QA) of these delineations. The timeline in the bottom section of Figure 4 exemplifies the series of interactions that clinicians usually perform during the QA process. In the timeline, it is possible to see how starting from partial delineations, they reach the final ones after a series of relatively long direct editing interactions interleaved with brief navigation operations ones. Between slice interpolation, the tool clinicians use for contouring from scratch, does not work for contour refinement. Therefore, for extensive errors across multiple slices like the one that Figure 5 depicts, clinicians face two options. Either manually fix the contour on every slide or delete the delineation and re-do it from scratch using between-slice interpolation.

===== INSERT FIGURE 5 APPROXIMATELY HERE =====

Discussion

While clinicians use IC contours for creating the treatment plan, they use AC contours to update it. For this reason, at this stage, their primary concern seemed to be to faithfully translate IC contours to the current patient anatomy. The identified contextual characteristics affect AC performance in several ways. First, having information about the role that each structure plays in the patient's treatment helps direct clinicians' attention to delineations that can affect the patient outcome. A potential pitfall of the current prioritization approach is that it is purely heuristic and based on clinicians’ experience instead of available information such as the planned dose. Second, by using IC-approved contours, clinicians can reduce the time for analyzing and editing complex or large regions by propagating them via registration. Nevertheless, same as with other image modalities at IC, the time it takes to perform the registration might offset the time gains. Finally, although contouring is overall faster at AC due to the contours being pre-generated, there is no tool to efficiently perform QA, requiring clinicians to invest significant manual effort.

Discussion

The Study of the Contouring Workflow provided an understanding of several characteristics that affect contouring performance in different contexts. This section takes these observations as input and lays down several ways of accelerating the adaptive contouring activity, which is increasingly time-pressured due to clinics implementing more responsive adaptative workflows. The discussion differentiates between the inspection, navigation, and editing tasks, which account for most of the delineation time. Figure 6 summarizes the study's findings and the resulting context-dependent acceleration strategies.

===== INSERT FIGURE 6 APPROXIMATELY HERE =====

Inspection and Navigation

In adaptive contouring, clinicians focus on inaccurate delineations that affect the patient's treatment. For example, they prioritize inspection of tumors contours because an error could result in overexposure of surrounding organs to radiation or, worse, in underexposure of the cancerous tissue (Aliotta et al., 2019). This observation suggests that it is possible to use patient-specific treatment-level information to define the contouring priority of anatomical structures. For instance, those areas that, when creating the treatment plan, showed to affect patient outcome, with metrics such as the Normal Tissue Complication Probability (NTCP) models (Brouwer et al., 2014), should be prioritized. Figure 7 presents an example of prioritization based on the local characteristics of the dose distribution. As can be observed, while a potential inaccuracy in the tumor delineation has a high priority, errors in the parotid glands are less urgent due to their lower impact on the patient’s treatment.

Before prioritizing errors, they need to be detected and highlighted to the user. Several methods have been proposed in the literature for this purpose. They vary in the information and the mechanism used to perform the search. As for the former, it is possible to compute shape (Heimann & Meinzer, 2009; Hermann & Klein, 2015) and image or appearance-related (Gao et al., 2010) characteristics of the contours. For instance, the surface area or the intensity histogram, respectively. Another possible indicator of the contours' quality is their uncertainty or variability, which can come from historical patient data (Chu et al., 2013), the auto-contouring algorithm (LaBonte et al., 2020; Mody et al., 2021), or directly the image-to-contour (Top et al., 2011). After gathering all these sources of information, available techniques surface potential errors in two ways. On the one hand by letting a classifier automatically find data-based rules for separation of the inaccurate from the accurate regions (Altman et al., 2015; Chen et al., 2015; Hui et al., 2018; Kalpathy-Cramer et al., 2014; McIntosh et al., 2013; Rhee et al., 2019; Sandfort et al., 2021). On the other, they delegate the search task to the users, presenting them with the traditional two-dimensional image and contour slices together with informative overlays such as uncertainty iso-lines (Al-Taie et al., 2014; Prassni et al., 2010) and contour boxplots (Whitaker et al., 2013). These two-dimensional visualizations have been augmented by adding three-dimensional views (Lundström et al., 2007; Raidou et al., 2016) and letting the user interact with the data by filtering and sorting mechanisms (Furmanová et al., 2021; Saad et al., 2010).

The main problem of existing techniques for guiding the user search for contouring errors is their reliability. On the one hand, they often surface areas that are not inaccurate, which could result in the user unnecessarily editing the contours, the most expensive operation (false positives). On the other, they can fail to spot inaccuracies that could affect the patient's treatment (false negatives). Invariably, the latter low sensitivity would erode the user's trust in the system, which might explain why these systems have not been widely deployed in clinical practice.

===== INSERT FIGURE 7 APPROXIMATELY HERE =====

Editing

Currently, clinicians use mostly manual tools when fixing an inaccuracy. For errors that occupy a large portion of the volume, like the example in Figure 5, this often means that the user will perform similar edits across slices. Existing semi-automatic interactive contouring techniques mitigate this issue by extrapolating rough feedback provided by the user. Their general workflow consists of two steps. First, the user provides a rough indication of the change to be made or the area to update via coarse inputs such as scribbles, points, or a bounding box. Based on this input, the algorithm proceeds to update the segmentation. Traditionally Markov Random Field-based algorithms were used (Kato & Zerubia, 2012; Rother et al., 2004). Recently, deep learning based implementations have appeared that offer more sophisticated suggestions based on the user’s input (Dai et al., 2015; Lin et al., 2016; Maninis et al., 2018).

The adoption of these semi-automatic interactive editing tools in the clinic remains challenging. Based on discussions with clinicians, the reason for their resistance to these interactive editing tools seems to be that they perceive scribbles as a blunt tool for communicating to the algorithm what they want. Therefore, more research is needed to determine which type of input mechanism the clinicians prefer and how the algorithm should respond. For instance, do they prefer coarse inputs like scribbles? Or would they be more comfortable with high precision inputs such as selecting a contour from an ensemble of candidates (Ferstl et al., 2016)? With editing being the most time-consuming QA operation, obtaining a synergy between humans and AI is paramount.

Limitations and Future Work

A limitation of this work is the reduced number of treatment centers and clinicians surveyed in the study, which might have led to weighting heavily on custom institutional practices and personal preferences. In the future, a questionnaire like the one reported in (Bertholet et al., 2020) could be prepared to validate the conclusions with a larger pool of participants. Another limitation is the qualitative nature of the timelines used to illustrate the dynamics between the clinicians and the contouring software. In the future, we plan to use keystroke logging software to include more fine-grained actions and more accurate timings. The latter would be especially valuable for comparing different segmentation tools.

In terms of future work, we will translate the findings of this study into a practical human-centered contouring protocol that clinicians can adapt to their institution-specific adaptive EBRT capabilities and constraints. In addition to the clinician-level considerations that the present article considered, such protocol will also account for team dynamics, which also emerged as a performance factor in the surveyed institutions.

Conclusion

This study characterized the contouring workflows in EBRT. An observational study at two treatment centers in the Netherlands revealed several context-dependent characteristics that influence delineation performance. Based on these, strategies for accelerating the inspection, navigation, and editing tasks were discussed. By applying these when developing and commissioning tools, tool builders and clinicians can decrease the delineation time, which will increase the suitability of this process for time-critical therapies like online-adaptive EBRT.

Acknowledgement

The authors of this work are grateful for the assistance and collaboration of the personnel at both Holland Proton Therapy Center and Leiden University Medical Center. The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022), and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate.

References

Aliotta, E., Nourzadeh, H., & Siebers, J. (2019). Quantifying the dosimetric impact of organ-at-risk delineation variability in head and neck radiation therapy in the context of patient setup uncertainty. *Physics in Medicine & Biology*, *64*(13), 135020. https://doi.org/10.1088/1361-6560/ab205c

Al-Taie, A., Hahn, H. K., & Linsen, L. (2014). Uncertainty estimation and visualization in probabilistic segmentation. *Computers & Graphics*, *39*, 48–59. https://doi.org/10.1016/j.cag.2013.10.012

Altman, M. B., Kavanaugh, J. A., Wooten, H. O., Green, O. L., DeWees, T. A., Gay, H., Thorstad, W. L., Li, H., & Mutic, S. (2015). A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Physics in Medicine and Biology*, *60*(13), 5199–5209. https://doi.org/10.1088/0031-9155/60/13/5199

Aselmaa, A., Goossens, R., Rowland, B., Laprie, A., Song, W., & Freudenthal, A. (2014, July 19). Medical Factors of Brain Tumor Delineation in Radiotherapy for Software Design. *Proceedings of the 5th International Conference on Applied HumanFactors and Ergonomics*.

Aselmaa, A., van Herk, M., Laprie, A., Nestle, U., Götz, I., Wiedenmann, N., Schimek-Jasch, T., Picaud, F., Syrykh, C., Cagetti, L. V., Jolnerovski, M., Song, Y., & Goossens, R. H. M. (2017). Using a contextualized sensemaking model for interaction design: A case study of tumor contouring. *Journal of Biomedical Informatics*, *65*, 145–158. https://doi.org/10.1016/j.jbi.2016.12.001

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brouwer, C. L., Steenbakkers, R. J. H. M., Gort, E., Kamphuis, M. E., van der Laan, H. P., van't Veld, A. A., Sijtsema, N. M., & Langendijk, J. A. (2014). Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. *Radiotherapy and Oncology*, *111*(1), 148–152. https://doi.org/10.1016/j.radonc.2014.01.019

Cardenas, C. E., Yang, J., Anderson, B. M., Court, L. E., & Brock, K. B. (2019). Advances in Auto-Segmentation. *Seminars in Radiation Oncology*, *29*(3), 185–197. https://doi.org/10.1016/j.semradonc.2019.02.001

Chen, H.-C., Tan, J., Dolly, S., Kavanaugh, J., Anastasio, M. A., Low, D. A., Li, H. H., Altman, M., Gay, H., Thorstad, W. L., Mutic, S., & Li, H. (2015). Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy. *Medical Physics*, *42*(2), 1048–1059. https://doi.org/10.1118/1.4906197

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Nimura, Y., Rueckert, D., & Mori, K. (2013). Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *16*(Pt 2), 165–172. https://doi.org/10.1007/978-3-642-40763-5_21

Dai, J., He, K., & Sun, J. (2015). *BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation*. 1635–1643. https://openaccess.thecvf.com/content_iccv_2015/html/Dai_BoxSup_Exploiting_Bounding_ICCV_2015_paper.html towards fast contouring workflows for adaptive proton therapy

Ferstl, F., Kanzler, M., Rautenhaus, M., & Westermann, R. (2016). Visual Analysis of Spatial Variability and Global Correlations in Ensembles of Iso-Contours. *Computer Graphics Forum*, *35*(3), 221–230. https://doi.org/10.1111/cgf.12898

Furmanová, K., Muren, L. P., Casares-Magaz, O., Moiseenko, V., Einck, J. P., Pilskog, S., & Raidou, R. G. (2021). PREVIS: Predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics*, *97*, 126–138. https://doi.org/10.1016/j.cag.2021.04.010

Gao, X., Su, Y., Li, X., & Tao, D. (2010). A Review of Active Appearance Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(2), 145–158. https://doi.org/10.1109/TSMCC.2009.2035631

Heimann, T., & Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, *13*(4), 543–563. https://doi.org/10.1016/j.media.2009.05.004

Hermann, M., & Klein, R. (2015). A visual analytics perspective on shape analysis: State of the art and future prospects. *Computers & Graphics*, *53*, 63–71. https://doi.org/10.1016/j.cag.2015.08.008

Hui, C. B., Nourzadeh, H., Watkins, W. T., Trifiletti, D. M., Alonso, C. E., Dutta, S. W., & Siebers, J. V. (2018). Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. *Medical Physics*, *45*(5), 2089–2096. https://doi.org/10.1002/mp.12835

Kalpathy-Cramer, J., Awan, M., Bedrick, S., Rasch, C. R. N., Rosenthal, D. I., & Fuller, C. D. (2014). Development of a Software for Quantitative

Evaluation Radiotherapy Target and Organ-at-Risk Segmentation Comparison. *Journal of Digital Imaging*, *27*(1), 108–119. https://doi.org/10.1007/s10278-013-9633-4

Kato, Z., & Zerubia, J. (2012). Markov Random Fields in Image Segmentation. *Foundations and Trends® in Signal Processing*, *5*(1–2), 1–155. https://doi.org/10.1561/2000000035

LaBonte, T., Martinez, C., & Roberts, S. A. (2020). We Know Where We Don't Know: 3D Bayesian CNNs for Credible Geometric Uncertainty. *ArXiv:1910.10793 [Cs, Eess]*. http://arxiv.org/abs/1910.10793

Langendijk, J. A., Lambin, P., De Ruysscher, D., Widder, J., Bos, M., & Verheij, M. (2013). Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiotherapy and Oncology*, *107*(3), 267–273. https://doi.org/10.1016/j.radonc.2013.05.007

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). *ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation*. 3159–3167. https://openaccess.thecvf.com/content_cvpr_2016/html/Lin_ScribbleSup_Scribble-Supervised_Convolutional_CVPR_2016_paper.html

Lundkvist, J., Ekman, M., Ericsson, S. R., Jönsson, B., & Glimelius, B. (2005). Proton therapy of cancer: Potential clinical advantages and cost-effectiveness. *Acta Oncologica*, *44*(8), 850–861. https://doi.org/10.1080/02841860500341157

Lundström, C., Ljung, P., Persson, A., & Ynnerman, A. (2007). Uncertainty Visualization in Medical Volume Rendering Using Probabilistic

Animation. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1648–1655. https://doi.org/10.1109/TVCG.2007.70518

Maninis, K.-K., Caelles, S., Pont-Tuset, J., & Van Gool, L. (2018). *Deep Extreme Cut: From Extreme Points to Object Segmentation*. 616–625. https://openaccess.thecvf.com/content_cvpr_2018/html/Maninis_Deep_Extreme_Cut_CVPR_2018_paper.html

McIntosh, C., Svistoun, I., & Purdie, T. G. (2013). Groupwise Conditional Random Forests for Automatic Shape Classification and Contour Quality Assessment in Radiotherapy Planning. *IEEE Transactions on Medical Imaging*, *32*(6), 1043–1057. https://doi.org/10.1109/TMI.2013.2251421

Mody, P., Chaves-de-Plaza, N., Hildebrandt, K., van Egmond, R., de Ridder, H., & Staring, M. (2021). Comparing Bayesian Models for Organ Contouring in Head and Neck Radiotherapy. *ArXiv:2111.01134 [Cs, Eess]*. http://arxiv.org/abs/2111.01134

Multi-Institutional Target Delineation in Oncology Group. (2011). Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices. *Journal of Digital Imaging*, *24*(5), 794–803. https://doi.org/10.1007/s10278-010-9341-2

Newhauser, W. D., & Zhang, R. (2015). The physics of proton therapy. *Physics in Medicine and Biology*, *60*(8), R155–R209. https://doi.org/10.1088/0031-9155/60/8/R155

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S. A., Consortium, D. R., Montgomery, H., … Ronneberger, O. (2020). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *ArXiv:1809.04430 [Physics, Stat]*. http://arxiv.org/abs/1809.04430

Prassni, J., Ropinski, T., & Hinrichs, K. (2010). Uncertainty-Aware Guided Volume Segmentation. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1358–1365. https://doi.org/10.1109/TVCG.2010.208

Raidou, R. G., Marcelis, F. J. J., Breeuwer, M., Gröller, E., Vilanova, A., & Wetering, H. M. M. van de. (2016). *Visual Analytics for the Exploration and Assessment of Segmentation Errors*. The Eurographics Association. https://doi.org/10.2312/vcbm.20161287

Ramkumar, A. (2017). *HCI in interactive segmentation: Human-computer interaction in interactive segmentation of CT images for radiotherapy*. https://repository.tudelft.nl/islandora/object/uuid%3A0f0259f1-0c33-442f-b851-86a846e736fc

Ramkumar, A., Dolz, J., Kirisli, H. A., Adebahr, S., Schimek-Jasch, T., Nestle, U., Massoptier, L., Varga, E., Stappers, P. J., Niessen, W. J., & Song, Y. (2016). User Interaction in Semi-Automatic Segmentation of Organs at Risk: A Case Study in Radiotherapy. *Journal of Digital Imaging*, *29*(2), 264–277. https://doi.org/10.1007/s10278-015-9839-8

Rhee, D. J., Cardenas, C. E., Elhalawani, H., McCarroll, R., Zhang, L., Yang, J., Garden, A. S., Peterson, C. B., Beadle, B. M., & Court, L. E. (2019). Automatic detection of contouring errors using convolutional neural networks. *Medical Physics*, *46*(11), 5086–5097. https://doi.org/10.1002/mp.13814

Rother, C., Kolmogorov, V., & Blake, A. (2004). “GrabCut”: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, *23*(3), 309–314. https://doi.org/10.1145/1015706.1015720

Saad, A., Möller, T., & Hamarneh, G. (2010). ProbExplorer: Uncertainty-guided Exploration and Editing of Probabilistic Medical Image Segmentation. *Computer Graphics Forum*, *29*(3), 1113–1122. https://doi.org/10.1111/j.1467-8659.2009.01691.x

Sandfort, V., Yan, K., Graffy, P. M., Pickhardt, P. J., & Summers, R. M. (2021). Use of Variational Autoencoders with Unsupervised Learning to Detect Incorrect Organ Segmentations at CT. *Radiology: Artificial Intelligence*, *3*(4), e200218. https://doi.org/10.1148/ryai.2021200218

Simone, C. B., Ly, D., Dan, T. D., Ondos, J., Ning, H., Belard, A., O’Connell, J., Miller, R. W., & Simone, N. L. (2011). Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer. *Radiotherapy and Oncology*, *101*(3), 376–382. https://doi.org/10.1016/j.radonc.2011.05.028

Steenbakkers, R. J. H. M., Duppen, J. C., Fitton, I., Deurloo, K. E. I., Zijp, L. J., Comans, E. F. I., Uitterhoeve, A. L. J., Rodrigus, P. T. R., Kramer, G. W. P., Bussink, J., De Jaeger, K., Belderbos, J. S. A., Nowak, P. J. C. M., van Herk, M., & Rasch, C. R. N. (2006). Reduction of observer variation using matched CT-PET for lung cancer delineation: A three-dimensional analysis. *International Journal of Radiation Oncology*Biology*Physics*, *64*(2), 435–448. https://doi.org/10.1016/j.ijrobp.2005.06.034

Steenbakkers, R. J. H. M., Duppen, J. C., Fitton, I., Deurloo, K. E. I., Zijp, L., Uitterhoeve, A. L. J., Rodrigus, P. T. R., Kramer, G. W. P., Bussink, J., Jaeger, K. D., Belderbos, J. S. A., Hart, A. A. M., Nowak, P. J. C. M., Herk, M. van, & Rasch, C. R. N. (2005). Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: A 'Big Brother' evaluation. *Radiotherapy and Oncology*, *77*(2), 182–190. https://doi.org/10.1016/j.radonc.2005.09.017

Thomas, H., & Timmermann, B. (2020). Paediatric proton therapy. *The British Journal of Radiology*, *93*(1107), 20190601. https://doi.org/10.1259/bjr.20190601

Top, A., Hamarneh, G., & Abugharbieh, R. (2011). Active Learning for Interactive 3D Image Segmentation. In G. Fichtinger, A. Martel, & T. Peters (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011* (pp. 603–610). Springer. https://doi.org/10.1007/978-3-642-23626-6_74

van Dijk, L. V., Van den Bosch, L., Aljabar, P., Peressutti, D., Both, S., J.H.M. Steenbakkers, Roel., Langendijk, J. A., Gooding, M. J., & Brouwer, C. L. (2020). Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiotherapy and Oncology*, *142*, 115–123. https://doi.org/10.1016/j.radonc.2019.09.022

Vandewinckele, L., Claessens, M., Dinkla, A., Brouwer, C., Crijns, W., Verellen, D., & Elmpt, W. van. (2020). Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, *153*, 55–66. https://doi.org/10.1016/j.radonc.2020.09.008

Whitaker, R. T., Mirzargar, M., & Kirby, R. M. (2013). Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2713–2722. https://doi.org/10.1109/TVCG.2013.143

Wilson, R. R. (1946). Radiological Use of Fast Protons. *Radiology*, *47*(5), 487–491. https://doi.org/10.1148/47.5.487

Table 1. Participants of the qualitative sessions. There were two radiation oncologists (RO) and two radiotherapy technologists (RTT) from two institutions in the Netherlands. Due to their tight schedules, not all of them could participate in all the sessions.

ID	*Institution*	*Role*	*Session*	*Time (hours)*
P1	LUMC	RO	1, 2, 3	5
P2	LUMC	RTT	2, 3	2
P3	HollandPTC	RO	1, 2	3
P4	HollandPTC	RTT	1, 2, 3	5

Table 2. The first column presents the themes that emerged during the Thematic Analysis of the transcripts of the semi-structured interviews and observational sessions of the Study of the Contouring Workflow. The second column presents the coarser codes obtained after several grouping iterations finer ones. Lastly, the third column displays, for each theme, a representative example from the transcribed data.

Theme	*Codes*	*Example*
Adaptive contouring context	Clinical workflow, standardization, physical and clinical artifacts, training, institution specific considerations, EBRT technology	"Now it takes one day to do the whole plan. So, we have to make a new calculation and it has to go into the the LINAC so it has to get another check." [P2]
Structure priority and effect of innacuracies on patient's treatment	Anatomical knowledge, downstream effects, characteristics of different anatomical structures, clinical priorities, tumor-related considerations	"I guess if it's an inner region where for instance the cheek region here. Those are minor [edits], but if we see this region where you have the parotid gland. There it could influence dose to the OARs quite significantly. So there. Then I would say it's a major [edit]." [P1]
Dealing with uncertain regions in the image-to-contour	Anatomical knowledge, image modalities, papers and guidelines, information required for certainty	"With the nasopharyngeal cancers, then I will take an MRI and then I will draw on the MRI. So, then I know exactly where the brainstem is." [P4]
Editing capabilities of contouring software	Characteristics of contouring software, experience with the tools, use of automation	"It seems to me that it's a model based one [automatically generated contour] because the model based one always has trouble here at the head of the mandible at the joint." [P3]
Distribution of labor and clinicians experience	Experience with the contouring task, collaboration, task distribution, protocols	"When an RTT does it [a contour]? Sometimes it's very nice and when a not so experienced RTT does it it's not a very good delineation and then it costs me either a lot of time to adjust every slice or I just start again and that's most of the time." [P3]

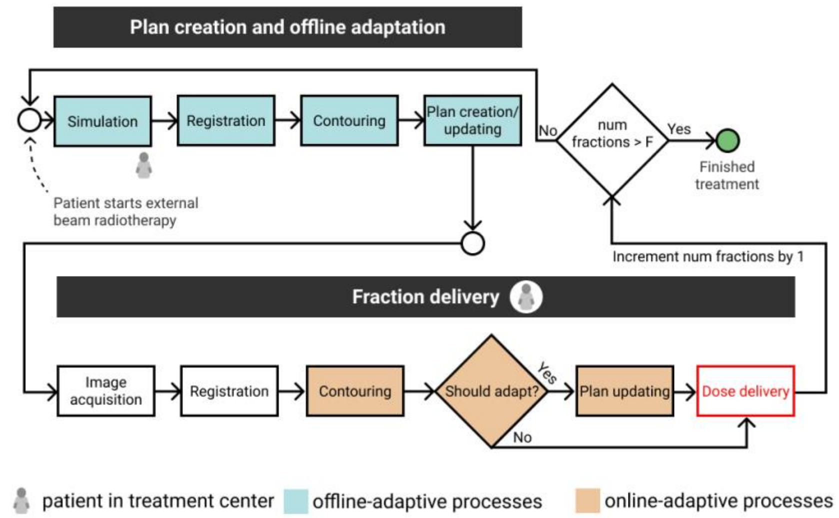

Figure 1. Schematic of an external beam radiotherapy (EBRT) dose delivery pipeline. Each box corresponds to one process, and the diamonds to decisions in the workflow. The goal is to deliver the prescribed dose to the patient in F fractions spread over several days (red box). Adaptive strategies help mitigate dose deviations due to changes in the patient's anatomy during the treatment. Adaptation can be online within a fraction (orange boxes) or offline between fractions (blue boxes).

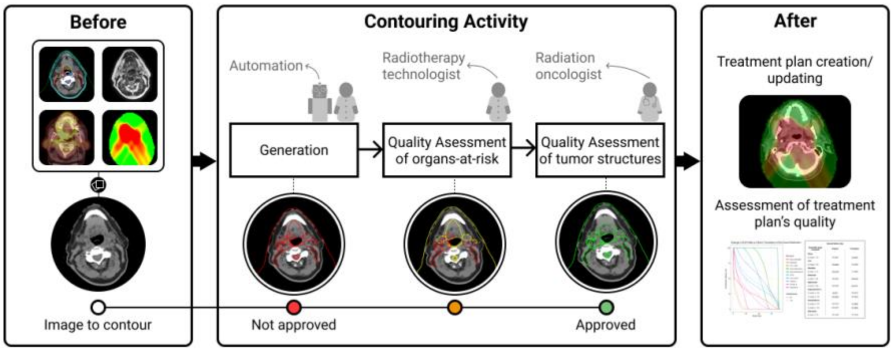

Figure 2. Components of the contouring activity. The inputs (left) are the image to contour and, optionally, other three-dimensional datasets like MRI and PET-CT scans and dose distribution volumes. The contouring activity has two main processes that several actors perform: generation of contours and quality assessment of the delineations. Subsequent processes use the resulting contours for creating/updating the patient's treatment plan and assessing its quality.

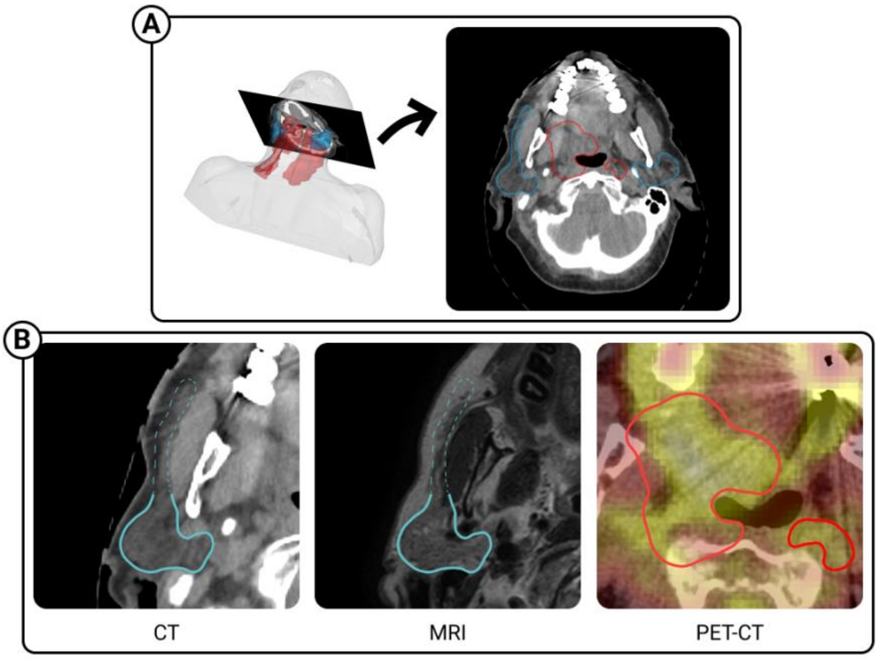

Figure 3. Available information available at contouring. The central input is the image to contour which, as panel A depicts, is a three-dimensional image made out of several 2D slices. Other three-dimensional images available at the surveyed centers are magnetic resonance imaging (MRI) and positron imaging technology CT (PET-CT) scans. As panel B shows, MRI helps differentiate soft tissue, and PET-CT aids in detecting and delineating tumors.

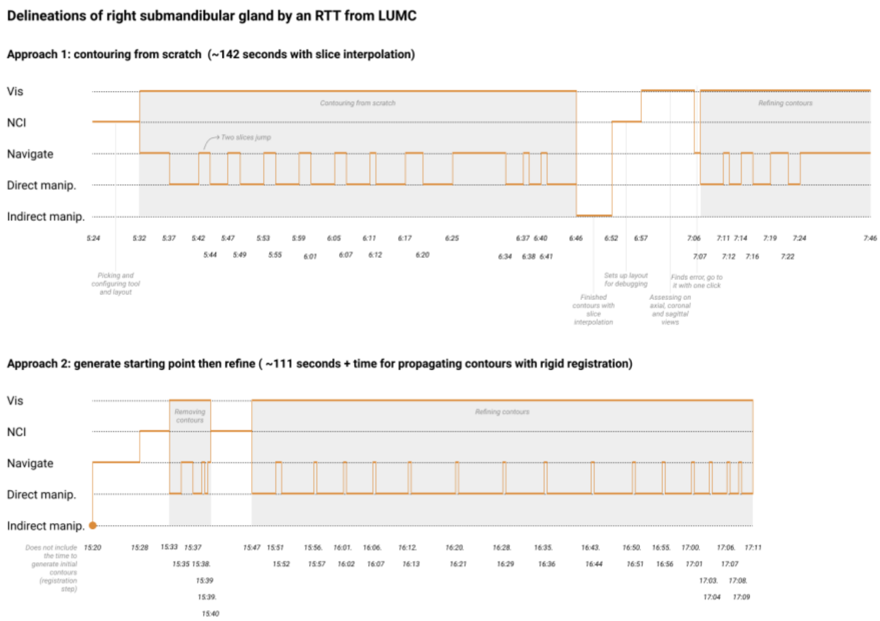

Figure 4. Interaction timelines for initial and adaptive contouring. In both cases, P2, a radiotherapy technologist from LUMC, delineated the right submandibular gland of a head and neck cancer patient. The x-axis encodes time, and the y-axis differentiates the principal interaction categories while delineating. Non-contouring interactions (NCI) correspond to changes in the interface that do not affect the contours, like changing the layout or visualization parameters. Navigation refers to changing the view of the image to contour. Finally, direct and indirect manipulations entail altering the delineations in the 2D slice or through a button in the menu, respectively. Note how initial contouring starts from scratch (empty circle) while adaptive contouring starts with pre-generated delineations (partially filled circle).

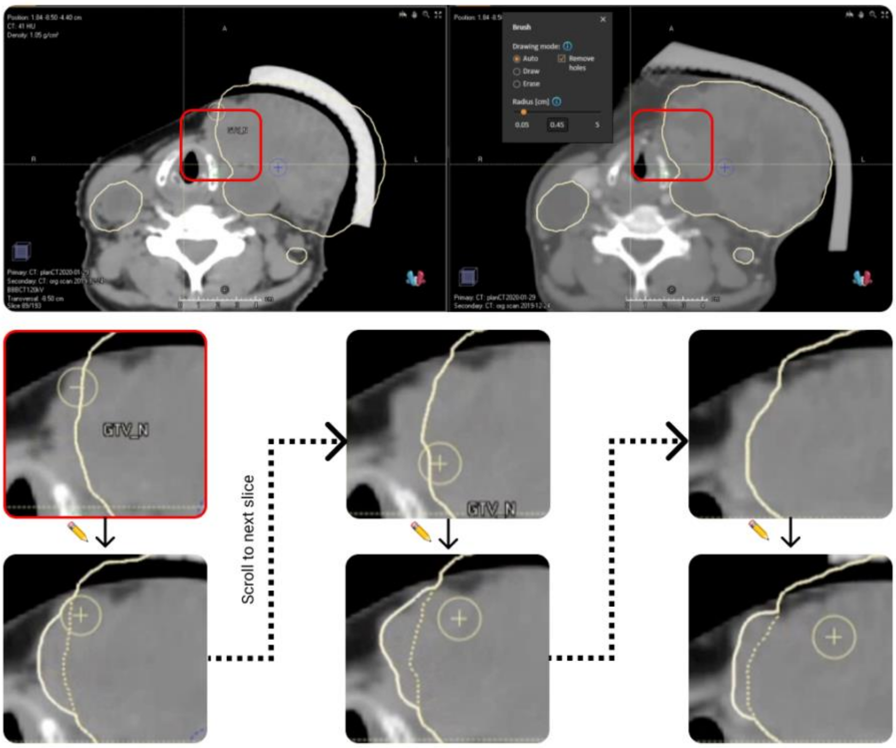

Figure 5. Editing faulty delineations often entails redundant interactions. The top image presents an inaccurate contour of a tumor-related structure (left) generated via rigid registration-based propagation from the planning CT (right) by P2 at LUMC. After the propagation, the internal side of the delineation fails to include the whole structure, which causes an error that spans several slices. The images below present the sequence of steps that P1 followed to amend the inaccuracy.

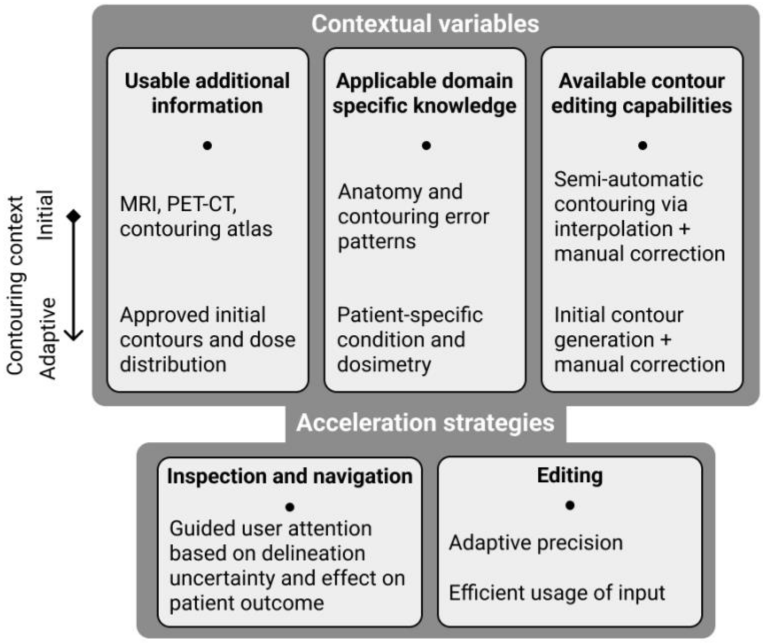

Figure 6. Schematic of the approach that the present study followed. First, it characterized the different contexts in which contouring operates in terms of three items that affect its performance. Using these as input, it discusses context-based strategies for accelerating the inspection, navigation, and editing tasks.

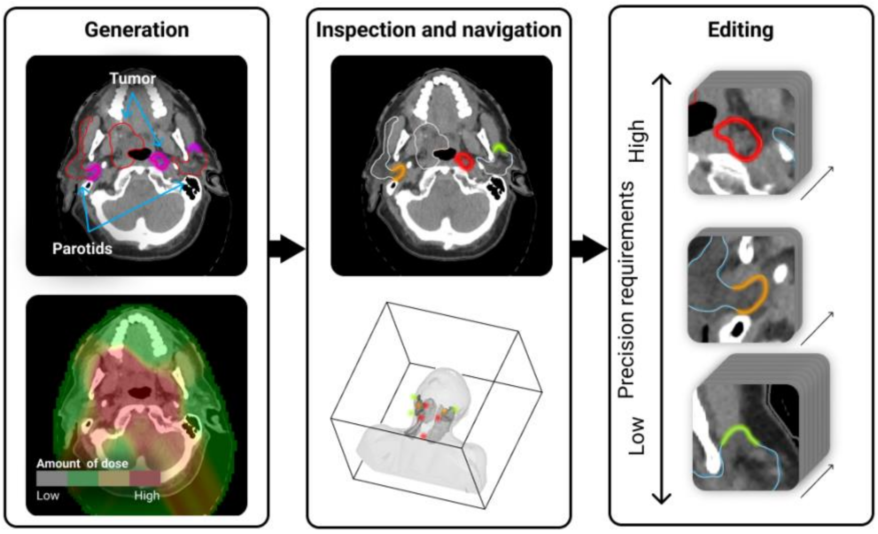

Figure 7. Components for accelerating the inspection, navigation, and editing tasks. The first step (leftmost column) is to generate the contours and gather extra information like delineation variability and the dose distribution (DD). Based on these sources, potential

errors can be flagged and categorized depending on their effect on the patient outcome. In the example, an error in the tumor's delineations was flagged as high priority (red) because it can significantly change the treatment plan. As for the parotid glands, the orange inaccuracy is located in a region where the DD varies more quickly than in the case of the green one. Therefore, subsequent processes (like treatment plan updating) that rely on the orange contours could be more sensitive to changes in these contours.