Goldilocks Contouring: Enriching One-to-One Medical Apprenticeship with Timely and Diverse Expert Feedback

MATIN YARMAND, The Design Lab, UC San Diego and Computing and Informatics, UNC Charlotte, USA
JOYCE LU, Data Science, UC San Diego, USA
YUNHAO LUO, Computer Science, UC Santa Barbara, USA
MEGAN ORR, Radiation Medicine, UC San Diego Health, USA
MICHAEL SHERER, Radiation Medicine, UC San Diego Health, USA
NADIR WEIBEL, The Design Lab, UC San Diego, USA

The one-to-one apprenticeship in many residency programs, while providing direct supervision from an assigned expert, lacks timely and diverse feedback, especially from other physicians with unique clinical tendencies. This often contributes to subpar training for high-stakes tasks with unstructured solutions, further diminishing education quality and patient safety. This work specifically explores the case of contouring — delineating tumor and healthy tissues — that is often regarded as the weakest link in radiotherapy treatment planning due to pervasive errors that lead to detrimental consequences for patient safety. This paper first offers three design goals aimed to prevent gaming, balance between expert consensus and tendencies, and minimize cognitive load. This work then designs and develops iConTutor, a learning platform that not only provides timely feedback, but also presents crowdsourced expert contours as part of distinct feedback mechanisms. A pre- and post-test study with nine residents showed a 32% increase in contouring accuracy, and the participants benefited from iConTutor's feedback mechanisms in delivering light-touch, skill-based, and awareness-cueing training. The three design goals of this work (implemented as part of feedback strategies in iConTutor) can inform computer-supported learning tools in other healthcare domains that aim to improve apprenticeship training via timely and diverse feedback.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *User studies*; **Systems and tools for interaction design**.

Additional Key Words and Phrases: Contouring Education, Healthcare Apprenticeship, Timely and Diverse Feedback

1 Introduction

Apprenticeship training in many healthcare residency programs [24] is the underlying structure for transferring highly specialized and critical medical skills from expert to novice physicians [49]. At its core, residents learn these skills by observing the clinical practices of an assigned faculty and later re-create their processes [65].

The medical apprenticeship model poses critical challenges given the lack of timely and diverse feedback that encompasses the broader existing tendencies among physicians. Attending faculty take on a dual role of clinician and teacher, and when time is in short supply, patient care takes absolute priority over teaching [56]. As such,

Authors' Contact Information: Matin Yarmand, The Design Lab, UC San Diego, La Jolla and Computing and Informatics, UNC Charlotte, Charlotte, USA, myarmand@ucsd.edu; Joyce Lu, Data Science, UC San Diego, La Jolla, USA, jol072@ucsd.edu; Yunhao Luo, Computer Science, UC Santa Barbara, Santa Barbara, USA, yunhaoluo@ucsb.edu; Megan Orr, Radiation Medicine, UC San Diego Health, La Jolla, USA, m1orr@health.ucsd.edu; Michael Sherer, Radiation Medicine, UC San Diego Health, La Jolla, USA, msherer@health.ucsd.edu; Nadir Weibel, The Design Lab, UC San Diego, La Jolla, USA, weibel@ucsd.edu.

ACM 2637-8051/2026/2-ART
https://doi.org/10.1145/3797889

feedback exchange between residents and faculty can take substantial delays that can degrade the quality of residency training [21]. Besides, the one-to-one model of apprenticeship limits the diversity of feedback for residents, especially in many medical fields where there is not a singular "gold standard" solution even among experts [40]. While this training model provides direct supervision for learning and performing complex medical tasks, facilitating broader feedback can improve the quality and robustness of specialized medical practices.

Specifically, this paper explores contouring education in radiation oncology training; contouring describes meticulous outlining of tumor and healthy tissues, producing a blueprint for high-dose radiation. Contouring is often referred to as *the weakest link* in radiotherapy treatment, as mistakes in contouring are prevalent and detrimental to patient safety. Recent clinical trial studies report that contouring inaccuracies can occur in a staggering 81% of the initial radiation plans [36]. Contouring errors have also been shown to decrease patient survival by 16% [1]. Over- and under-contoured plans lead to excess toxicity to the nearby healthy organs, or insufficient radiation to the tumorous cells which will increase the risk of disease recurrence. Finding just-the-right balance in the dose and placement of radiation (i.e., what we refer to as *goldilocks contouring*) remains the main challenge in radiotherapy treatment.

This work aims to improve one-to-one healthcare apprenticeship by integrating *timely* and *diverse* feedback processes, two overarching principles of this paper. Informed by these two principles, prior work in learning sciences and medicine, as well as in collaboration with two expert physicians throughout the project, we offer three design goals to prevent gaming, balance between consensus guidelines and clinician tendencies, and minimize extrinsic cognitive load. These goals contributed to the design and development of *iConTutor*, a contouring platform with three feedback mechanisms that provide not only an averaged consensus recommendation but also allow users to explore the degree of variability among expert physicians. We evaluated iConTutor in nine-step lab studies with residents (N=9), including pre-/post-tests and retrospective think-alouds. Comparing post-test and pre-test revealed a 32% increase in accuracy, suggesting potential benefits of timely and diverse feedback for patient safety. The participants also valued the three feedback mechanisms for the light-touch, skill-based, and awareness-cueing techniques, and further highlighted the potential of iConTutor in improving contouring skills of novice- and senior-residents. The design goals of this paper have the potential to also transform other healthcare residency programs, contributing to higher quality in healthcare training and ultimately, patient care.

2 Related Work

This section describes the critical role of contouring in radiotherapy treatment, common techniques and resources for contouring education, as well as interactive learning tools that employ timely and diverse feedback.

2.1 Contouring: Critical Step in Radiotherapy Treatment

Radiotherapy is a type of cancer treatment that uses targeted and high-dose radiation to destroy tumorous cells while maintaining the safety of nearby healthy organs. A critical step in this process is the identification and delineation of cancerous and at-risk tissues on volumes of patient cases, sometimes up to hundreds of medical images (e.g., CT scans and MRIs) per case [16]. This complex task, also known as *contouring*, is typically performed by radiation oncologists, a specialized subset of physicians. The primary goal of contouring is to accurately distinguish cancerous tissues from surrounding organs at risks, thereby defining the target regions for radiation therapists to deliver the prescribed dose according to the treatment plan. Contouring interleaves many actions, broadly defined in three categories [66]: *setting up* the layout and images, *delineating* contours via techniques such as outlining, fine-tuning, and smoothing, and *navigating* images on different slices and views. The exact sequence of actions often varies among physicians; for instance, while some might place rough outlines, navigate, and then fine-tune all placed contours, others might choose to finalize contours on each slice before navigating to adjacent images [66].

Contouring is known as the *weakest link* in radiotherapy treatment, due to the complexity of the procedure and the unfortunate wide-scale errors that can lead to detrimental consequences for patient safety. Contouring is a complex procedure that necessitates careful consideration of various anatomical and clinical factors, including treatment context (e.g., clinical symptoms), tumor context (e.g., size and growth direction of tumor), and tumorous areas (e.g., satellite regions)[8]. The difficulty of contouring often leads to wide-scale errors in patient planning; multi-institutional trials have reported that up to 81% of the initial treatment plans contained major errors that required revisions [36]. Accurate contouring in radiation therapy critically depends on the expertise of radiation oncologists, and higher patient volumes often contribute to the development of this expertise. Clinical trials have shown that low-accruing centers (with fewer patients) have both higher protocol violations and worse outcomes, with an 18 percentage point difference in 5-year overall survival compared to high-accruing centers [64]. Given the high stakes involved, contouring errors can lead to detrimental consequences for patient safety, including decreased survival: a major clinical trial for pancreatic cancer revealed that contouring errors led to a 16% decrease of median survival from 1.74 to 1.46 year [1].

Despite advancements in computer vision and continuous efforts to automate contouring [42, 47], current algorithms face several critical challenges that hinder their adoption in the real-world clinical practice. These deep learning-based algorithms have been shown to yield high variability in segmenting structures, such as just over 70% for esophagus and close to 100% for rigid structures like lungs [47]. The lack of adoption of these algorithms in radiation oncology stems from the lack of "gold-standard" contouring solution, physician accountability in high-stakes medical procedures, as well as the adaptability of these models to real-world patient cases. The most prominent adoption barrier is the inherent subjectivity of contouring, which leads to substantial inter-observer variability [12, 20, 60]. Even among expert physicians, contouring outcomes can differ due to variations in clinical judgment and risk assessment; some may choose broader contours to ensure full coverage, while others may be more conservative and prioritize protecting nearby healthy tissues. Another challenge is the dependence on high-quality training data, which is particularly difficult to obtain in the medical domain due to privacy constraints and the labor-intensive nature of expert annotations [22, 27]. Besides, these models often perform poorly when applied to cases that fall outside the specific patient and anatomical structures that they are trained on [30, 71]. As a result of these limitations, manual contouring by physicians remains the standard in clinical practice, and this paper focuses on improving contouring education within this context. This work aims to lessen contouring errors and contribute to the overall quality of radiotherapy treatment. The design goals introduced in this paper provide a pathway to implement timely and diverse feedback in contouring education.

2.2 Contouring: Training Methods and Tools

Quality contouring education is a cornerstone in improving overall quality of radiotherapy treatment. The current residency programs in radiation oncology rely on an apprenticeship model of training, where residents gain experience and receive feedback from experienced physicians. However, the availability of experienced physicians is often limited due to their dual responsibility in teaching and clinical practice; this ultimately leads to prolonged delays in feedback exchange and reduced hands-on learning opportunities [65]. The rapid advancement of radiotherapy has further underscored the need for improved training in this domain. Thus, there is a growing recognition to incorporate structured radiation oncology education early in medical school curricula [53].

To improve the quality of education in contouring, there has been an increasing shift towards flexible and on-demand digital learning [39]. Multiple platforms have been proposed aiming to enhance access and provide on-demand learning experiences. For example, *eContour* [57] is a browser-based contouring platform that offers a pre-populated atlas of patient cases and contours, and has been shown to improve both contouring accuracy and anatomical knowledge [28]. Yarmand et al. [70] developed a prototype that compares learner contours against experts and provides accuracy (in terms of percentage of overlap) *after* the learner submits their contour

for review. Similarly, Mazur et al. [51] demonstrated that simulation-based training can improve procedural adherence among radiation therapy professionals while fostering the development of new skills and knowledge. Lastly, other works aimed to provide flexible practice opportunities that traditionally rely on standalone desktop and mouse set-up, such as VRContour (that provides a mixed 3D and 2D contouring environment) [17] and iContour (which introduces an interface compatible with everyday touch devices) [66, 69].

While these training strategies and tools improve contouring practice and learning, they generally lack timely and diverse feedback, two critical components of effective learning. Common feedback exchange techniques (e.g., contour hand-offs in the apprenticeship model [65]) often provides feedback *long after* a contouring session where the learner has completed an entire case with sometimes many image slices. Yarmand et al.'s prototype [70] moves toward real-time feedback by providing guidance *right after* each contour submission. However, especially when residents need to contour tens or hundreds of slices [16], early feedback around core anatomical reasoning helps them carry this knowledge across the entire case and avoid wasting time delineating images with flawed understanding [65]. As such, real-time guidance *during* the contouring session has the potential to drastically improve learning. Contouring education (embodied by one-on-one apprenticeship training) also relies on feedback from one expert physician, despite the subjective and unstructured nature of contouring. Understanding broader tendencies in contouring can help residents gain more robust contouring skills. Our work addresses these two overlooked (yet, crucial) factors of quality contouring education via the design and development of iConTutor.

2.3 Timely Feedback: Enhancing Learning Outcomes in Interactive Educational Tools

Many interactive educational tools benefit the learning process via delivery of timely supplementary material and personalized hints. The timing of feedback is particularly critical in these environments, as it directly influences learner engagement and educational outcomes; prior research shows that students' interest and engagement with the learning material lowers as the delay in receiving feedback increases [11]. Another study demonstrated that students who received immediate feedback significantly improved learning outcomes [3].

Given these benefits, many systems (in different settings) explored and designed the use of timely feedback. For instance, CritiqueKit [25] is a mixed-initiative system that provides real-time suggestions to help students give higher-quality peer input; by leveraging a reusable feedback corpus and allowing user corrections, it promotes immediate and actionable feedback. Another example is ADRD (Adjacent Display of Relevant Discussion) [67], which dynamically surfaces and presents contextually relevant forum posts next to instructional videos right at moments of potential confusion. This design enables just-in-time peer feedback without interrupting the learning flow, which leads to increased confusion resolution and engagement compared to traditional forum layouts. Intelligent tutoring systems are also a prominent line of educational technology (e.g., Cognitive Tutor [6] and Auto Tutor [29]) that leverage bespoke-designed decision trees to determine learners' unique cognitive models and provide just-in-time hints.

While these systems studied the value of timely feedback in various forms, there remains a lack of research on immediate feedback mechanisms in the medical training, particularly for contouring tasks that involve a complex, multi-step process. This paper offers design guidelines and implementation techniques that can help foster timely feedback in contouring education.

2.4 Diverse Feedback: Wisdom of the (Expert) Crowds in Learning Tools

Crowdsourcing has been a promising approach in addressing the *gulf of expertise*: the scarcity of skilled experts, in addition to the need for diverse feedback for complex and unstructured tasks, necessitates going beyond the traditional methods of seeking expertise. A systematic review by Jiang et al. [35] found that crowdsourcing has been widely utilized in education to create learning content, provide practical experience, exchange complementary knowledge, and augment abundant feedback. Prior HCI works have introduced systems aimed to

match learners with diverse expertise that, otherwise, would not be available to them. For instance, Atelier [59] connects novice workers with mentors who break down tasks and provide oversight. Besides, Ichinco [34] offers a workflow to crowdsource feedback from experienced programmers and distribute among novice programmers.

In medical education, many works advocated to bridge access to scarce expert physicians, especially for gaining complex and high-stakes medical skills. The apprenticeship model of residency programs is a prominent example. Despite the direct supervision from one expert physician, residents are often left without feedback from a diverse crowd that can reveal crucial tendencies in contouring. Dai et al. [19] conducted a systematic review demonstrating that crowdsourced assessments of surgical skills show strong correlation with expert evaluations, highlighting their potential for providing well-rounded feedback to learners. The effectiveness of crowdsourcing in medical education has been demonstrated in various contexts. For example, Bow et al. [13] developed a crowdsourced learning system that enabled medical students to collaboratively create and edit their study material in the form of flashcards; students who engaged in this system outperformed their predecessors who did not have access to it, while also benefiting from a collaborative learning environment.

In radiation oncology (like many other medical disciplines), the absence of a universally accepted "gold standard" for contouring further complicates training and necessitates the need for tools and approaches that incorporate diverse expert feedback. This paper contributes to this critical gap by developing and testing a learning system that incorporates expert feedback during contouring sessions.

3 System Design and Implementation

We developed iConTutor by first defining key design goals aimed at enhancing learning through timely and diverse expert feedback. Guided by these goals, we designed three feedback mechanisms targeting different contouring needs. Finally, this section describes system implementation details for the three feedback mechanisms.

3.1 Design Goals

The design of this tool followed three goals unique to learning systems, medical apprenticeship, and contouring education, and was informed by the two overarching principles of this work (i.e., diverse and timely feedback).

[DG1] Prevent Gaming Behaviors in Learning Tools — *Gaming the system* commonly occurs in learning environments in which the learners exploit the feedback mechanisms of the system to succeed rather than focus on learning the material [9]. A common behavior (particularly seen in intelligent tutoring systems) is to continuously activate the tutor until the system reveals the correct answer [5], or exhausting all possible options (e.g., in multiple choice questions) until the student selects the correct answer [18]. As such, we aimed to design intentionally vague feedback mechanisms that gently guide residents' awareness to anatomical regions of the images without revealing the final answer. This *awareness-cueing* technique can make it difficult to extract final answers by repeatedly trying out the feedback feature without reflecting on the underlying lessons. In addition, awareness-cueing can especially help learners in fuzzy moments: when learners don't know what they don't know, and hence, are incapable of seeking explicit guidance [67, 68].

[DG2] Balance between Consensus and Tendencies — Complex learning tasks are often void of a "gold-standard" solution. These tasks involve many trade-offs, making the final outcome subject to relevant circumstances and even personal preferences. Varying contouring styles stem from the trade-off between applying radiation to tumors and potentially damaging nearby healthy organs: given expert physicians' past experiences and interpretations of the patient case, some might be likely to apply higher dose to tumors, at the risk of damaging nearby healthy tissues, while others might tend to preserve more of nearby organs. The apprenticeship model of training in residency programs lacks diverse feedback and only facilitates feedback exchange with one physician. While consensus guidelines represent the commonalities among multiple experts, there is immense value in learning

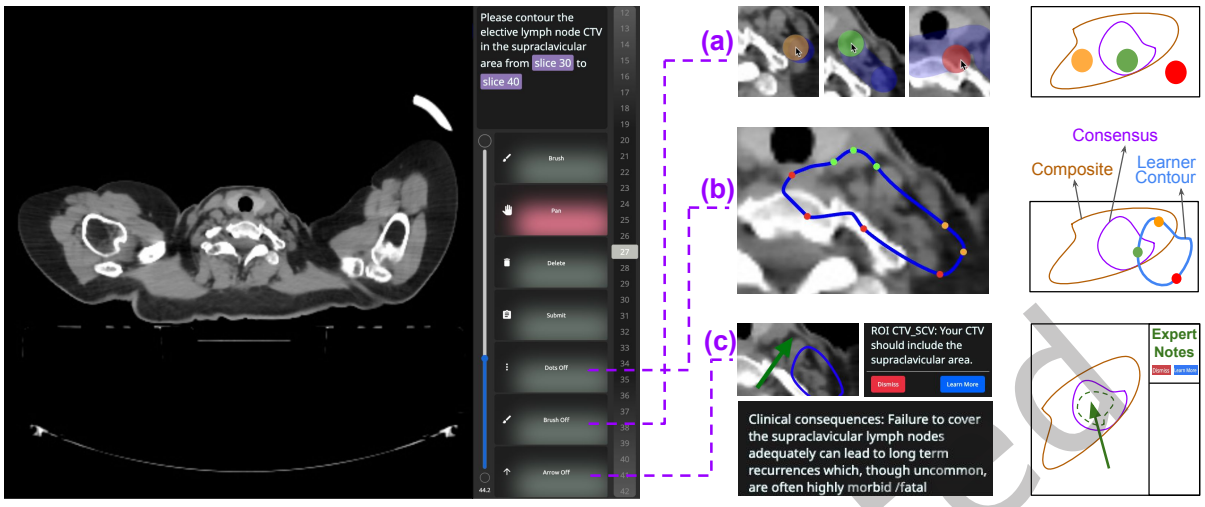


Fig. 1. Interface and feedback mechanisms of *iConTutor*, a learning tool that leverages contours from expert physicians to not only present a consensus solution in real-time, but also show the inter-expert variability that is often hidden from learners due to the one-to-one apprenticeship model of residency programs. This tool offers three types of feedback: (Feedback a) When dragging the brush, if the center of the circle lies within consensus, it turns green. When the brush is outside consensus but within composite (i.e., union of all expert contours), it turns orange. Lastly, a red brush indicates regions outside composite. (Feedback b) A similar color scheme applies when the contour is placed and dots appear on the user contour: for example, orange dots indicate that these dots lie outside consensus and within composite regions. (Feedback c) The tool displays targeted regions to be included and avoided via an arrow and a note which can be expanded for more details.

about granular (and often, differing) tendencies within experts [35]. The feedback mechanisms of this tool aim to distinguish and balance between consensus feedback and the broader expert tendencies.

[DG3] Minimize Extrinsic Cognitive Load in Contouring Education — Contouring demands high cognitive load from physicians who have to simultaneously consider many factors like patient history, areas at risk for harboring tumor, and organs at risk [8]. As such, just-in-time feedback mechanisms should balance between the provided learning resources and the main contouring procedure. This balance between the the mental workload to accomplish the task (*intrinsic*) and the cognitive workload needed to engage with the provided feedback (*extrinsic*) is often discussed in educational psychology literature as part of the Cognitive Load Theory [55]. For a task like contouring that carries a high intrinsic cognitive load (due to the need to account for treatment context, tumor context and tumorous areas [8]), it is essential to minimize extrinsic cognitive load for contouring educational tools. Excessive and obtrusive feedback can not only diminish learning gains, but also prolong contour delineation — that can already take up to 30 hours per patient [7] — and putting patients at risk. Our learning tool aims to facilitate this balance by closely defining the *what* (i.e., digestible content snippets), the *when* (i.e., during idle times), and the *how* (i.e., easily dismissible feedback).

3.2 Design and Development of iConTutor, a Training Tool with Timely and Diverse Feedback

Informed by core contouring actions [66] and the design goals presented in Section 3.1, we developed *iConTutor*, a browser-based contouring application with feedback mechanisms that provide timely and diverse expert feedback. The following three components facilitate the main contouring functionalities (as shown in Figure 1):

Canvas – The largest section of the interface, the canvas, renders selected medical scans from a `DICOM` file stack (i.e., Digital Imaging and Communications in Medicine, the standard format for medical imaging) using `cornerstone API`. The users perform contouring with the brush tool that features a circular marker that follows the user's pointer, implemented via `Paper.js` library. Dragging this brush across the canvas allows users to define enclosed regions. If no existing contour is present, a new structure is outlined. When intersecting with an existing contour, the tool refines its shape: pulling it outward when dragged from inside and pushing it inward when dragged from outside. This switching between outlining and refining streamlines delineation.

Control Buttons - We designed and implemented seven toggle buttons and sliders in the right panel: *Draw*, *Pan*, *Delete*, *Submit*, and toggle buttons for three feedback, including *Dots*, *Brush*, and *Arrow and Notes*. The *Draw* button changes the system to contouring mode, *Pan* button enables panning mode, *Delete* button deletes all contours on the current slice, and *Submit* button submits the contours. The three feedback buttons toggle on or off the corresponding feature. To the left of the buttons, a vertical slider adjusts the brush size. Placing this slider near the canvas accommodates the frequent need for physicians to switch between large brushes for broad outlining and smaller brushes for detailed refinement on each slice.

Navigation – A slice scrolling selector, positioned to the right of the canvas, offers an intuitive way to navigate through image slices. This UI element allows for both precise selection of specific slices and rapid comparison across multiple images. The users can also navigate through slices by using the mouse wheel while the cursor is placed on the scrolling selector.

In addition to core contouring functionalities, iConTutor provides three feedback mechanisms. The feedback content used in this work comes from two sources. First, we worked closely with two experts from UC San Diego Health (one attending faculty and one senior resident) to identify regions of interest and develop medical notes and guidance. The two experts had previously worked together during a medical rotation and collaborated on research projects. Second, we leveraged the crowdsourced expert feedback previously collected as part of the C3RO project (Contouring Collaborative for Consensus in Radiation Oncology) [61]; the C3RO project led to a publicly-available dataset containing segmentations of critical regions for radiotherapy treatment. Specifically, we used the breast cancer case with contours from eight expert physicians, not only for the clinical target volumes, but also the nearby organs at risk. These eight contours contributed to two core areas used for the color schemes in the feedback mechanisms: (1) *consensus* (generated via the `STAPLE` method), a probabilistic algorithm for combining multiple segmentations [62], and (2) *composite*, the union of all eight contours, designed to help learners understand the degree of acceptable variation in contours across multiple experts. The rest of this section describes the three feedback mechanisms and how they align with the design goals in Section 3.1:

(Feedback a) Brush: Changing Brush Color when Outlining Contours — This feedback initiates as early as the learner starts delineating by tracing regions on the medical images (Fig. 1–a). The brush (that is used for contouring) changes colors according to where the center of the brush lies by comparing coordinates of the brush and the underlying contours. If the center is within the consensus contour, the brush becomes a green circle. Moving the brush outside of consensus (but still within the composite) changes the color to orange, and once fully out of the composite, it turns red.

Changing brush color makes a distinction between consensus and tendencies via two different colors [DG2], and aims to minimize cognitive load by solely changing color of an existing element (without adding separate visuals on the canvas) [DG3]. Lastly, given that the center of the circle indicates color, yet the perimeter of the circle shapes contours, this separation can provide sufficient vagueness to prevent gaming the system and encourage learning the underlying anatomy [DG1].

(Feedback b) Dots: Color-coded Dots Placed on Learner Contours — The second type of feedback (Fig. 1–b) activates immediately after the color-coded brush feedback (with every release of the mouse), in which dots appear on the learner contour to point out clues about the regions underneath. Instead of directly validating the placed contour, the dots inform whether the regions that they fall onto are within consensus (green), outside consensus but within composite (orange), or outside composite (red). To calculate the frequency and placement of the dots, iConTutor first calculates the number of distinct segments within the contour, multiply this number by 0.7 while upper-bounding by 10 (defined via trial and error with expert physicians), and randomly select segments that encapsulate the dots. In this method, the areas that are refined more and have more segments are likely to contain more dots; meaning, learners are likely to receive more feedback on more complex and clinically significant areas. The dots should appear at a controlled frequency that while guiding the learners towards important regions of contouring, avoid revealing too much information.

Limiting the density of the dots crudely informs the correctness of the placed contour and aims to lessen the likelihood of gaming the platform [DG1]. Also, the small size of the added dots mitigate crowding the interface, and hence, the cognitive load of contouring [DG3]. Lastly, the color-scheme (same as the dots) distinguishes and balances between contouring guidelines and tendencies [DG2].

(Feedback c) Arrows and Notes: Targeted Avoid and Include Regions — This feedback aims to emphasize nearby Regions of Interest (ROIs) that according to consensus guidelines should always be included (i.e., unique areas subject to tumor recurrence) or fully avoided, such as lungs. Overall, we included five avoidance regions and three inclusion areas, and for each region, we developed short and long explanations of why the region should be avoided or included. When learner under-contours an include region or over-contours an avoid region (thresholds defined and tested with medical collaborators), the tool displays an arrow pointing to the region. The arrows are rendered by `HTML Canvas 2D` API, and the placement of the arrow is determined by the center of the targeted ROI via extracting coordinate points of the contours and calculating the center points. All points are calibrated to the current zoom scale and pan location to ensure the correct placement of ROI relative to the medical image. Simultaneously, the short description appears in-situ of the medical images and provides a *learn more* option to show the longer description, if the learner chooses to engage with the feedback in more depth. Given that learners' contours can conflict with multiple critical regions at once, this feedback mechanism orders conflicts based on severity (determined by overlap percentage) and only provides the arrow and note for the most severe region each time feedback is triggered. The two participating experts collaboratively curated the content for both the short and long descriptions and defined severity levels for each ROI.

This feedback mechanism satisfies [DG1] by showing an arrow pointing to ROIs instead of displaying the exact region boundaries. Also, the combination of arrow and notes prioritizes learning core topics based on consensus guidelines, more so than the existing tendencies [DG2]. Lastly, besides showing only the most severe conflict at every point, providing content in-situ of the main workspace is a common technique used to minimize the cognitive load of learning tasks [DG3].

4 Methods

We conducted nine, 65-minute user sessions with radiation oncology residents (out of a cohort of ten residents at a large teaching hospital) who went through nine study phases: starting from establishing a contouring baseline, the residents solved real-world contouring tasks via iConTutor (with automatic and manual feedback), and lastly participated in retrospective think-alouds. We performed a mix of quantitative and qualitative analysis methods. The Institutional Review Boards (IRB) approved the study protocols.

4.1 Study Design

The participants engaged in a number of real-world contouring tasks. We define a contouring task as a unique combination of a patient case (e.g., breast) and a specific volume of the CT scan (i.e., a distinct range of image slices). Given this definition, we curated different tasks for this study. We used the same task (the same target structure and range on identical DICOMs) for both pre- and post-contouring steps in order to ensure fair comparison and accurate capture of learning gains; it is common for pre-/post- test studies to facilitate identical assessment instruments in order to measure exact learning gains, such as in STEM [48] and medical education [4]. We designed different tasks for the two feedback steps to facilitate broader anatomical learning and avoid direct task-specific learning effects. The residents participated in 65-minute individual sessions; we allocated the times for each step (including contouring durations) in close consultation with the two expert physicians who had extensively worked with trainees at this level. The sessions consisted of the following nine steps:

(1) [2 mins] Pre-survey — Residents first completed a brief survey covering their demographic information (i.e., age and residency year), prior experience with breast cancer cases (the anatomical focus of this study), and training strategies (i.e., educational resources used).

(2) [7 mins] Familiarization with interface and case — We then introduced the participants to the iConTutor interface and system features. The first author guided the residents through the interface layout and core contouring functionalities, including the canvas, tool controls, slice navigation, and brush behavior. Participants were also shown how to adjust brush size, delete contours, zoom, pan, and move across slices. At this stage, we aimed to mainly familiarize the participants with the core contouring features and the patient case, without introducing any of the feedback mechanisms. For the last 5 minutes of this phase, participants self-explored the breast case and interface. To further structure this session and cue the attention of participants to important components of contouring in the later steps, we provided the following prompt: *"please use the next 5 minutes to contour the lung anywhere in the range of slices 20 to 100. Please use this time to play around with the interface and get familiar with the anatomy of this case, which is a postmastectomy breast and regional nodal radiation that will be treated with IMRT based planning."* This exploratory task intended to ensure that participants gained sufficient familiarity with the tools and anatomy prior to the main study activities. It also aimed to minimize carry-over bias in later phases by reducing the likelihood that performance differences were caused by interface unfamiliarity rather than feedback interventions.

(3) [3 mins] Pre-contouring — Participants completed a 3-minute baseline contouring task (Task A). This task aimed to establish baseline expertise, and served as a reference for evaluating performance changes in later parts of the study. The prompt was: *"please contour CTV Chestwall in slices 65-70 in the next, at most, 3 minutes. The earlier you finish, the earlier we can move on to the next part of the study."* This phrasing aimed to balance accuracy and time spent, two critical (and often, conflicting) success factors in contouring [69]. We disabled all feedback mechanisms during this step.

(4) [5 mins] Familiarization with feedback — Following exposure to the core contouring functionalities, this step introduced the three contouring mechanisms embedded in the system: (1) changing brush color when outlining contours, (2) color-coded dots placed on learner contours, (3) and arrows and notes for under or over coverage of target regions. The first author explained the purpose and behavior of each feedback type using a demonstration without the underlying medical images (i.e., the right-most column in Figure 1). This stage aimed to familiarize participants with interpreting and interacting with the provided feedback in subsequent sessions.

(5) [10 mins] iConTutor with automatic feedback — The participants completed a new 10-minute contouring task (Task B). All three feedback types were activated automatically throughout the session, in order to expose the participants to all three feedback types. The automatic activation is determined by the overlap percentage

between user's contour and the target regions with set thresholds (defined and tested with medical collaborators): for instance, if the learner contour covers more than 90% of lungs (a region that is meant to be avoided), the system gauges severe enough error to initiate feedback. The prompt given was: *"please contour CTV chest wall in slices 70-80 in the next 10 minutes."* Participants were given a verbal reminder half way in the session to help them pace their work. We recorded the screen to later conduct a retrospective think-aloud. This phase facilitated real-time and system-driven feedback of iConTutor, and also aimed to provide hands-on experience with the three feedback mechanisms before the next, manual session.

(6) [10 mins] iConTutor with manual feedback — Participants then completed another 10-minute contouring task (Task C). Unlike the last phase, participants had full control over enabling or disabling any of the three feedback mechanisms in the session. The prompt in this step was: *"please contour the elective lymph node CTV in the supraclavicular area in slices 30-40 in the next 10 minutes."* Participants were also given a verbal reminder 5 minutes before the end of the session. Similar to the automatic phase, we recorded the screen for further analysis. This phase assessed how residents chose to engage with feedback features when given full control over their use.

(7) [3 mins] Post-contouring — To assess potential performance change and learning gains after the feedback sessions, participants repeated the same task as in the pre-contouring phase (Task A, including the same set of images) for 3 minutes under identical conditions and instructions. This allowed for a within-subject comparison of contouring performance before and after interacting with the iConTutor system and the incorporated feedback.

(8) [5 mins] Post-survey — The participants completed a post-session survey assessing the *perceived usability* of the tool (via the ten questions of System Usability Scale [10]), and eight questions for *perceived learnability* [43]. The learnability questions measured different aspects of learning such as motivation and real-world application.

(9) [20 mins] Retrospective think-aloud — Lastly, the participants engaged in a retrospective think-aloud [58] while viewing recordings of their own contouring sessions (both automatic and manual). The purpose of having retrospective (as opposed to concurrent) think-aloud sessions was to reduce the additional mental load of verbalizing for the participants and enhance their natural performance [31]. They had full control of playback (pause, rewind, and fast-forward) to reflect on their decisions, points of confusion, and use of feedback. We ended the session inquiring about overall perceptions of iConTutor, and to what extent the residents would prefer to engage with this tool in their residency programs.

4.2 Participants

We recruited participants through the mailing list of the Radiation Medicine department at UC San Diego Health, one of the largest research and teaching hospitals in the United States. Nine residents participated in the study (out of a cohort of 10 residents). Participant demographics, collected via a pre-survey, showed an age range of 28 to 44 years and included seven male and two female residents. Training levels ranged from residency year 2 (PGY2) to PGY5, with a median level of PGY3, reflecting diversity in both age and clinical experience. At this institution, residents are required to complete 350 contouring cases across the 5 years of residency. All participants had completed at least one clinical rotation in breast cancer, with two participants reporting multiple rotations; each rotation often includes 30 cases. In addition, all participants reported prior use of eContour [57], a widely-adopted, browser-based contouring atlas that provides pre-populated patient cases and contours to support anatomical learning. All participants, however, had never seen iConTutor prior to the study, including any of the feedback features.

4.3 Data Analysis

The analysis of the participant sessions followed a mix of quantitative and qualitative techniques.

Statistical Analysis — To assess differences in contouring accuracy between the pre-contouring and post-contouring phases, we employed the Wilcoxon signed-rank test [63]: a non-parametric test for comparing two related samples. We applied this test to participants' average Dice Similarity Coefficient (DSC) per slice to determine whether post-contouring performance significantly improved accuracy over the baseline. DSC is a common measure in radiation oncology [23] to gauge the similarity between two structures with a score from 0 (no overlap) to 1 (perfect overlap): the two structures in our analysis are the participant's contour, and the consensus contour of the target volume comprising eight expert faculty from the C3RO dataset [61].

Log-based Analysis — During each contouring session, we collected detailed interaction logs capturing every user action. Each data point in the log contained the following information:

- type of action (e.g., outlining, refining, and navigating),
- start and end time of each action,
- slice number that the action was initiated in,
- resulting DSC score with the target volume, and
- percentage overlaps between the user's contour and each target region.

All logs were stored on a backend server and later used to analyze user behavior and tool usage patterns, as well as changes in contouring performance.

We later generated visualizations that represent these actions over the course of a session. Time-series plots were generated to track DSC progress across the session duration, with overlaid annotations indicating feedback periods (during the automatic phase) and specific feedback mechanisms used (in the manual phase). To assess overall improvement, we used bar plots to compare average DSC values before and after feedback intervention. These visualizations support both individual-level analysis and group-level comparisons across conditions.

Multi-modal Thematic Analysis — We examined the think-aloud sessions via a reflexive thematic analysis [14, 15], including the text-based transcript and video-based screen recordings. The first author — an experienced HCI researcher with extensive research in healthcare (and especially, radiation oncology) — initially built familiarity with the data by actively reading (and re-reading) the transcripts that were mapped to the video recordings, and later conducted data-driven coding (i.e., open coding) on content related to the feedback mechanisms and overall contouring education. Lastly, pairs of authors iteratively searched and reviewed the codes to create sub-themes, and later themes.

5 Results

This section describes the results of a mixed-method analysis of the contouring sessions.

5.1 iConTutor improved contouring accuracy and yielded high usability and learnability scores

Figure 2 shows that the participants significantly improved contouring performance when comparing post- and pre-test results, as measured by Dice Similarity Coefficient (DSC). The DSC scores increased from an average of 0.6371 (pre-test) to 0.8408 (post-test), a significant 32% improvement in contouring accuracy (p<.0001). Appendix C shows the DSC scores of each participant during the pre- and post-tests, in which they generally stayed consistent throughout the session.

The residents also rated iConTutor highly in terms of usability and learnability. The overall System Usability Score (SUS) for iConTutor was 77.78 which falls well in the "acceptable" range (>70) [10]. As demonstrated by Figure 3, the participants highly valued the learning benefits of iConTutor. Notably, all participants assigned *agree* or *strongly agree* when applying the learning benefits gained from iConTutor for real-context scenarios, a critical factor for bridging between educational and clinical contexts.

5.2 Residents benefited from timely and diverse feedback, yet raised need to balance load

The participants valued the timeliness and diversity of feedback provided by iConTutor: especially in contouring education that revolves around one-to-one apprenticeship, residents typically have to wait for an extended amount of time before they get feedback from their assigned faculty. Many residents expressed the instrumental benefit of timely feedback, in contrast to what they typically get as part of residency programs. The participants especially appreciated the role of iConTutor when learning new disease sites, as mistakes can be pointed out early before spending a long time contouring while having an incorrect understanding of the anatomy:

> I definitely think for my first prostate cases, this tool would be very helpful, because we don't have any formal, like, this is how you do this. We just draw something and then eventually our attending will look at it and say, this is what you did wrong. Whereas this tool could tell me at the time. (P2)

Some participants further indicated that unlocking the full potential of iConTutor requires more exposure and getting used to the idea of expecting and receiving feedback during contouring sessions, as explained by P8:

> I'm not used to having real time feedback as I'm contouring. Usually, you spend an hour contouring something and then you get the feedback afterwards that it was terrible and you wasted your time. So it's nice to get it in real time. Although that'll take some getting used to. (P8)

P8 shared this point after having missed an arrow and note about the scar tissue because he was too focused on the contouring task; while this participant highlighted the importance of including this region, he originally missed the feedback given that *"he was too busy chasing [his] edges out of the lung"* (P8).

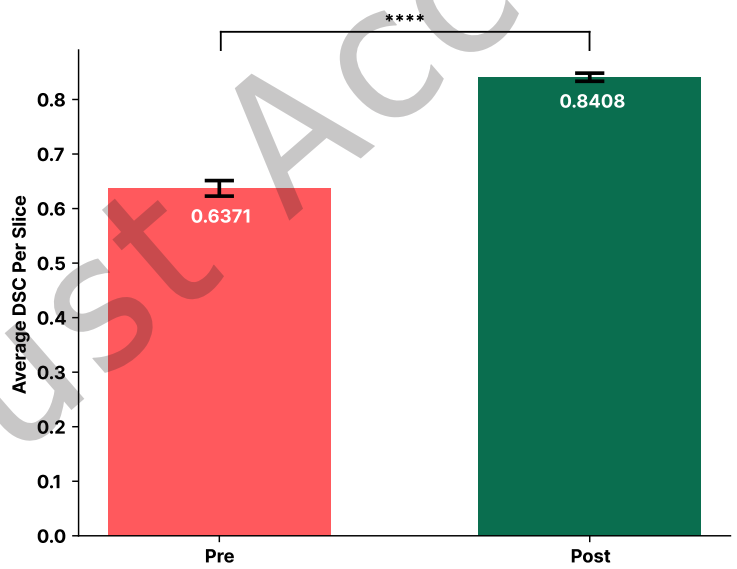


Fig. 2. Average contouring accuracy per slice (measured by DSC) for pre and post tests. The red bar represents the pre-test DSC, while the green bar corresponds to the post-test DSC. Error bars indicate the standard error of the mean. Post-test produced higher average DSC per slice (0.8408) compared to pre-test (0.6371) with a 32% increase ($p<.0001$).

The unfamiliarity with real-time feedback (and residents' varying adaptability skills and contouring expertise) likely contributed to mixed opinions about the mental load introduced by iConTutor. While some participants found getting all three feedback *"too over-stimulating"* (P2) — and as such, opted to only turn on dots in the manual phase — others felt that *"it's not overwhelming to have all three on"* (P7).

The residents further appreciated having access to diverse contouring styles, beyond what they would normally get from their assigned faculty. As mentioned in Sec. 2.1, contouring plans are not one-size-fits-all solutions; expert physicians' differing strategies can range from applying mild radiation to preserve healthy tissues, to more aggressively targeting the tumorous areas. For instance, P8 expressed interest in learning about contouring tendencies that may prioritize protecting the airways even with the tradeoff of delivering less radiation:

> Depending on what faculty you work with, these volumes can be different. People like to do different things or stay further away from the airways. So I'm surely interested to see what this tool would recommend. (P8)

Despite the potential learning benefits of displaying tendencies, some participants revealed that the additional cognitive load can hinder contouring, especially for junior residents who can be new to contouring. To manage the load for early residents, P9 suggested de-coupling the features for consenus and composite contours:

> [for someone in the first week of their rotation] I think probably it just helps to have the consensus as your gold standard and then maybe you have a feature for turning on the union of contours. You get a sense of what is acceptable variation, but I think if you gave that from the outset, it might be information overload. (P9)

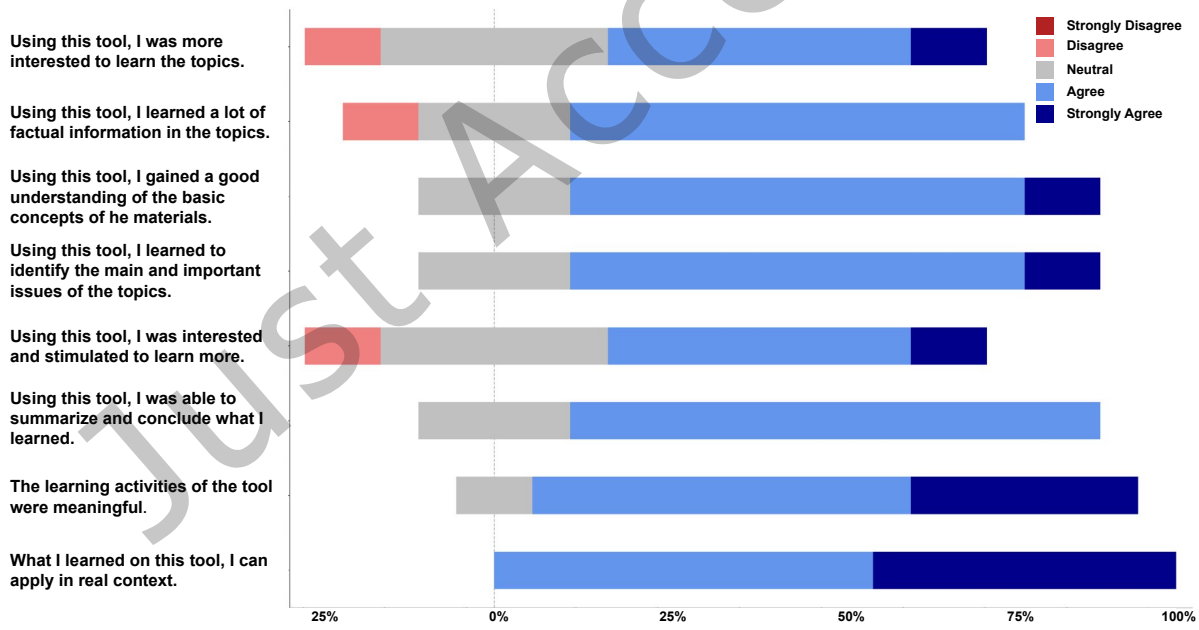
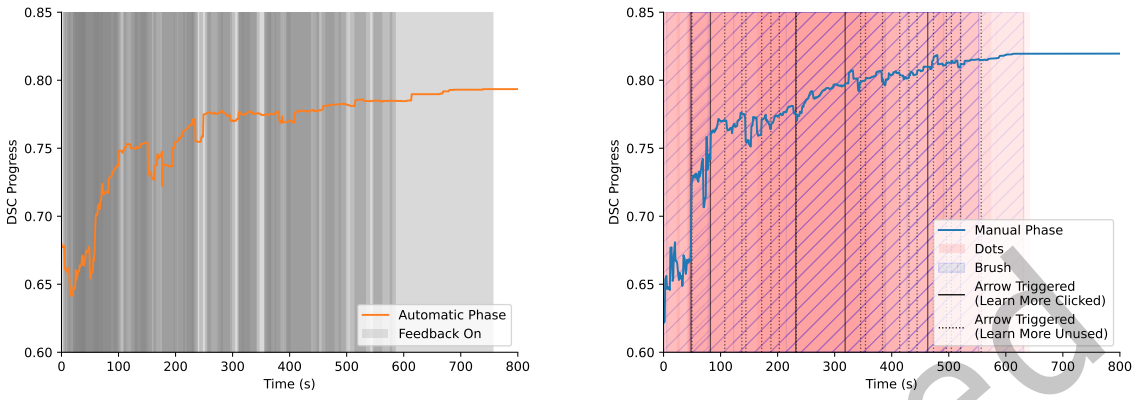


Fig. 3. Eight Likert-style statements regarding the learnability of iConTutor. Most participants highly rated the learning benefits of this tool. Notably, all participants answered "agree" or "strongly agree" to the last statement about applying learning gains in real-world settings, a critical factor for bridging between educational and clinical contexts.



(a) Automatic Phase (b) Manual Phase

Fig. 4. The change in contouring accuracy (measured by averaged DSC scores across all participants) for (a) iConTutor with automatic feedback and (b) iConTutor with manual feedback. (a) illustrates the DSC progress over time during the automatic phase. The background shading denotes periods where feedback mechanisms were active; darker regions represent more participants having received feedback. The consistent upward trend in DSC suggests continuous progress over time, with notable improvements during early parts of the session. (b) illustrates the evolution of the DSC progress over time during the manual phase. The shaded red region in the background denotes periods during which the dots feature was turned on, while the blue diagonal hatching highlights intervals when the brush feature was turned on. Solid vertical lines represent moments when an arrow was triggered and the participant clicked to "learn more" about the region, whereas dotted vertical lines represent arrow triggers where the "learn more" feature was not used. Across all participants in the manual phase, the arrow was triggered 24 times, with participants clicking "learn more" on 5 instances.

5.3 Three feedback mechanisms facilitated three unique learning benefits

Figure 4 demonstrates the improvement in contouring accuracy when iConTutor triggered the three feedback mechanisms, automatically (a) and manually (b). Both line graphs denote a persistent upward trend starting from approximately 65% accuracy tapering to around 80% by the end of the session, with faster rates in the beginning. As evident in the cumulative automatic phase in Figure 4(a), as well as the individual graphs in Appendix A, the trigger of feedback (i.e., dark regions) generally led to an increase in accuracy given that it prompted the learners to fix the raised issues (e.g., remove the red dots that appear on the contour). During the manual phase, Figure 4(b) showcases the durations in which the residents kept dots and brush feedback toggled on, as well as the instances that triggered the arrow and notes feedback, and when residents sought to view the longer descriptions via the *learn more* button. Appendix B depicts feedback usage in the manual phase, per participant.

Arrows and Notes — The participants indicated important learning opportunities from the arrows and notes, and further highlighted that novice residents might benefit the most from this feedback mechanism. Most residents already knew about the regions represented by the arrows, yet the arrows helped them to pay additional attention to these critical landmarks, as pointed out by P7:

> After I read [the note], I think I spent a little more time in that area, especially when I did it the second time, to be careful about covering that. So I think that was helpful because the focus of my mind shifted more to that, than necessarily being perfect elsewhere with my contours. (P7)

P7 proceeded to be *"more generous up into the scar"* (P7) after the arrow recommended more coverage of the scar region. Many residents further acknowledged that these arrows can provide substantially more benefits to junior residents who might not have mastered understanding these core regions, compared with mid- and senior-level residents. For instance, P3, a third-year resident who already had prior experience contouring breast cases, paid limited attention to the arrows:

> [The arrow] was pointing to the scars. I knew what it was saying, you know, cover that. So that's what I assumed. But if I was coming at it without any background, I think I would have taken more time and maybe looked around some more. (P3)

P4 compared arrows and notes to a watchalong session, a method of feedback exchange used for training junior residents in a hands-on format: P4 deemed arrows and notes similar to how *"[their] attending points out this or that structure, and that's an important reference point for this reason or that reason"* (P4) and ultimately *"would have liked more of it to point out useful anatomy or avoidance structures"* (P4).

Dots — Dots was the most frequently enabled feedback mechanism, as the participants used it to fine-tune boundaries informed by consensus (green dots) and tendencies (orange dots). Much of the residents' preference over the dots stemmed from its ease-of-use and intuitiveness to refine region boundaries, as evident by P9: *"I like the dots a lot. I feel like as a radiation oncologist, I'm always thinking about the boundary, and so I think the dots are just very intuitive"* (P9).

While the participants always removed the red dots, they refined contours differently given green and orange dots. Their strategies often stemmed from their perception of the existing variation in expert contours. Some residents followed a more conservative approach and strictly aimed to only allow green dots: *"in my mind, green is the right answer"* (P7). Others, however, fully treated orange dots as part of the correct answer: *"I'm OK with orange because contouring is very attending dependent and so some attendings cover more than other attendings. So for me, if any attending would have drawn that, then that's appropriate in my eyes"* (P2). Lastly, some residents were cautious about incorporating orange dots as part of their contour, indicating that their decision relied on the anatomy of the region:

> To me, orange is acceptable in this case where I have clear boundaries, like the lung interface with the chest wall [...] I probably wouldn't change my contour based on the orange in these situations if I have a clear line or a clear boundary. (P5)

Brush — The brush feature especially helped residents decide on large regions and determine the general direction of shrinking or expanding contours. While the participants primarily used dots to refine their contour boundaries on a granular level, sweeps of the brush tool informed excluding or including large structures. P9, for instance, shaved off large portions of his contour with three sweeps of the brush that indicated red throughout: *"I was tracking, is my circle turning orange at all? And then I realized, wow, they really don't want me to cover this, because classically this is the entire blood vessel"* (P9). The participants also used the brush tool to determine how to expand their contours (i.e., what areas to include). Often, the residents started with small contours and gradually expanded to cover the entire tumor regions; when unsure about the direction of expansion, the changing colors of brush tool was a convenient and intuitive method: *"with the brush, it can lead me towards that direction of how far out I should go"* (P1). P1 particularly reflected on the instance shown in Figure 5, in which while the initial contour displayed only green dots, the brush helped discovering an entire area that was originally overlooked by the resident. P7 expressed a similar strategy, and further indicated that he would *"keep going bigger and bigger until [he] saw it change orange"* (P7).

5.4 Mixed Opinions and Strategies on Gaming the System

Despite iConTutor's explicit aim to minimize gaming behaviors, some participants acknowledged contouring to the system and not the learning goals. The design of iConTutor emphasized gentle prompting to important areas of coverage, as opposed to distinctly revealing the final contour like many existing resources (e.g., eContour). Some residents, however, devised workarounds to circumvent the intentional vagueness included as part of the feedback mechanisms. For instance, P3 found out that by repeatedly clicking inside of a contour, the dots get re-generated; while a limited number of dots lean more towards gentle prompting, P3 managed to repeat this process (i.e., clicked for more than 10 times in a row) until he got a clear understanding of the entire contour. P1, a second-year resident, expressed self-awareness of their gaming behaviors, yet highlighted the importance of learning the anatomical foundations underlying the feedback mechanisms:

> At certain points, I felt like I was contouring to the feedback, vs the anatomy itself. So I think, you know, I wouldn't even say a drawback [of the tool], but some things that people should consider is understanding why the green dots are there rather than contouring to the green dots. (P1)

Most residents did not employ techniques to game the feedback mechanisms of iConTutor, and some even went further to restrict their access to feedback during the manual feedback phase. Some participants temporarily disabled all feedback mechanisms to mitigate gaming tendencies and decided to do a few slices before turning the feedback on *"to not let it influence [them] off the bat"* (P8). Similarly, P4 decided to go through all 10 slices without feedback, before going back and evaluating his contours:

> I was like it's been a very long time since I've drawn superclav nodes by breast standards as opposed to head/neck standards. So I was like, I'm going to take a shot at drawing this on my own and then seeing how it lines up. (P4)

5.5 Process-oriented tool for junior, and anatomy-reminder for senior residents

The participants noted unique benefits of iConTutor not only for novice residents, but also for experienced residents and even attending faculty. Reflecting on their first clinical rotations, many residents expressed benefiting from this tool when learning to contour a new case, as highlighted by P1, a second-year resident: *"if it's a new disease site that I have never contoured before, I would definitely use [this tool] for sure"* (P1). Some residents went as far as suggesting firm curricular changes to include iConTutor as a core part of the program: *"I would require it to be mandatorily used as a PGY2, or any first rotation that you do"* (P2). Many learners attributed the benefit of iConTutor to the ability to perform contouring and receive direct feedback; in existing solutions like eContour, *"you only get the final results and you don't get the logic of it [...] and there is huge benefit actually contouring*

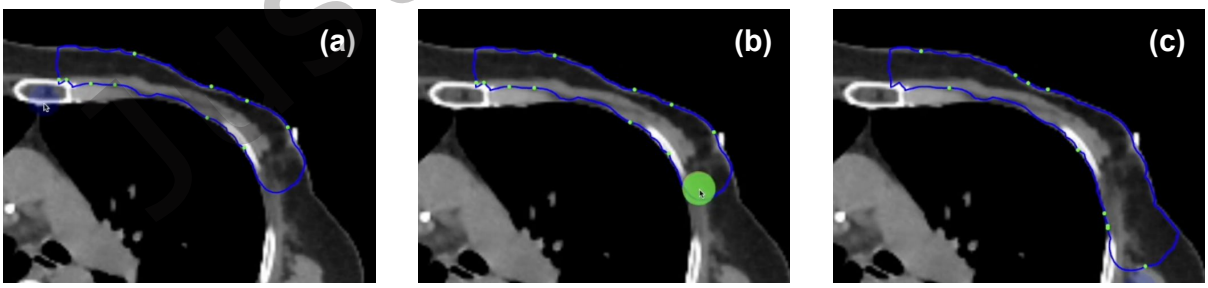


Fig. 5. Three snapshots of P1's contouring session. (a) The initial placed contour that despite showing green dots, severely under-covers the target tumor. (b) The changing brush color gently cued the attention of the learner to the lateral parts of the anatomy as the resident incrementally enlarged the contour. (c) the final contour as the resident learns about the importance of including lateral chest wall, a critical area for treating this disease site.

yourself" (P8). The more senior residents also valued the feedback mechanisms of iConTutor, not necessarily for learning new lessons, but more so pointing residents to their blind spots. P9, a fifth-year resident, refined significant portions of his contours based on the red and orange dots that appeared. He later reflected that these mechanisms reminded him of granular anatomical boundaries that he might have forgotten over the years:

> To me, contouring is like muscle memory, exactly how to go along the pelvic bone. So, for [the feedback] to pop up like, oh, there's red and orange, I was like, oh, shoot, have I been doing this wrong? Like this whole time for my breast cases, I've been doing it this way. So I was like, oh shoot, I should go back and look at something. [...] I see this as a reminding tool. These are the things that we should be thinking about, and making sure we're keeping in mind. (P9)

Thinking ahead of future faculty careers, some residents especially valued iConTutor and the incorporated feedback mechanisms on varying contouring tendencies: *"because I'm preparing to be an attending, I need to know what I want to do myself, and then compare it to what everyone else would [in their contours]"* (P9).

6 Discussion and Future Work

This section first discusses the potential of iConTutor to fundamentally improve radiation oncology education (Sec. 6.1). Sec. 6.2 then presents and interprets unique patterns observed during the contouring sessions and offers design implications. Sec. 6.3 highlights how the three design goals can provide a pathway to introduce timely and diverse feedback in broader medical education. The last part (Sec. 6.4) offers limitations and future work.

6.1 Potential to Transform Contouring Education at Scale

The user study with residents revealed important benefits with providing timely and diverse feedback (two overarching principles behind iConTutor), in contrast with one-to-one apprenticeship models that facilitate direct supervision with only one faculty and often with significant delays in feedback exchange. The participants not only achieved a 32% increase in contouring accuracy, but also rated iConTutor highly in usability and learnability, to the extent of expressing desire to use iConTutor in different stages of residency.

iConTutor has the potential to impact contouring education at scale, via building on the three mechanisms that were developed in close collaboration with expert physicians. Computer-supported tools have the potential to scale novel educational techniques to broad range of physicians and mitigate the *quality gap* of contouring: prior research shows that rural medical institutions (with fewer volume of patients) provide substandard treatment compared to the counterpart urban providers with higher patient volume [2, 45]. To scale iConTutor, many patient cases, disease sites, and consensus contours can be imported via digital atlases (e.g., eContour) given that iConTutor builds on top of standard DICOM files (Digital Imaging and Communications in Medicine). To incorporate expert tendencies, a *snowball effect* strategy can be employed, in which the earlier contours (from both experts and residents) are collected and stored in the server and transformed into visualizations that can be provided to later users. The contours from experts can present differing tendencies, while resident contours can provide peer learning benefits, such as learning common mistakes and new perspectives [41].

6.2 Unique Patterns and Design Implications

This section offers interpretations and design implications informed by unique patterns observed when the participants engaged with iConTutor.

Early Sharp Increase in Accuracy — The contouring sessions (as measured by quantitative logged events) showed a sharp early increase in accuracy, followed by more gradual improvement across both automatic and manual feedback conditions (Fig. 4). This pattern mirrors many interactive learning tools where performance rises quickly at first and then plateaus [50]. Qualitative evidence indicates that early gains in this case were driven by learning the core anatomical regions, many of which that the residents were initially unaware of. As

described in Section 5.3, the brush tool helped residents know of these anatomical regions that they had initially overlooked. Without first understanding these regions, residents cannot progress to the more challenging task of granularly refining boundaries and defining margins, and importantly, they may not even know to seek guidance from their attending faculty. Education literature highlights this phenomenon as when learners *don't know what they don't know* [54]. This illustrates an advantage over purely post-contouring feedback that requires learners to submit contours and initiate receiving support (e.g., [70]), as learners who lack this awareness may not know when, or what, to ask for guidance.

Interpreting Feedback Colors — This study defined and used three color categories to convey feedback; although the color meanings were thoroughly explained in each session, participants may still have been influenced by common color associations. For example, because green is widely interpreted as "correct" or "good" [33], some residents may have unconsciously viewed green regions as the only acceptable answer, even though orange indicated areas endorsed by a subset of experts. Sec. 5.3 presents a variety of perceptions around green and orange, with some participants only leaving green dots on their contour. To reduce potential bias introduced by distinct color categories and to avoid unintentional nudging, future versions of the iConTutor system can adopt a continuous color spectrum from green to red. Such a scheme could be further informed by factors like the number of physicians who agree on a region and their respective levels of expertise; greater consensus or higher expertise would shift the color closer to green.

Tailoring Textual Feedback — Compared to the dots and brush feedback, Sec. 5.3 presents that the arrow and notes feature appeared to be especially helpful for junior residents who were less familiar with the anatomical regions being highlighted. In addition, only a small proportion of participants engaged with the *learn more* button to access the extended descriptions that were curated by the participating experts in this work. Future versions of iConTutor could leverage Large Language Models (LLMs) to tailor these explanations to individual needs and backgrounds. For instance, these models can transform the original descriptions into digestible yet detailed explanations for junior residents, and summarize the texts for experienced physicians who need to contour cases with many slices, and as such, can benefit from more concise guidance. Aligned with LLM-supported personalization techniques in STEM education [72], the notes can also boost unique characteristics of the provided feedback (e.g., specificity and actionability) which can improve engagement and learning.

6.3 Reflecting on the Design Goals for Broader Impact

Evaluating the three Design Goals (DGs in Sec 3.1) can provide a pathway to transform medical apprenticeships and ease the teaching load on the faculty who often have to prioritize clinical duties [65]. This section describes the potential of the DGs to curate awareness-cueing (DG 1), skill-based (DG 2), and light-touch (DG 3) feedback.

Awareness-Cueing Feedback via DG1 — Following DG 1, the feedback mechanisms of iConTutor avoided displaying the final contours, and rather leaned into cueing the awareness of learners to areas of importance. This design goal aimed to prevent gaming behaviors, when learners exploit feedback to maximize score without learning the material [9]. As presented in Section 5.4, some participants extended the idea of preventing gaming behaviors by deliberately limiting their access to feedback, so they could first test themselves. This suggests potential system designs for iterative feedback applied to only a small subset of slices, or systems that reveal feedback after a batch of contours is completed. Such approaches would allow learners to periodically review feedback, integrate what they learn, and then continue contouring with improved understanding.

Besides the original intention of preventing gaming behaviors, awareness-cueing can especially benefit adult learners who engage in unstructured and complex tasks. According to the theory of adult learning or *andragogy* [38], adults — given their self-concept of being responsible for own lives — are highly motivated and often, prefer self-directed learning methods; medical residents (who are typically between late-20s and early-40s)

can especially benefit from maintaining control over their learning (while being guided to best practices), as opposed to being told about the exact methods and results to perform a learning task. As highlighted in Sec. 5.5, the residents also valued iConTutor as it focused on the *process* of contouring, rather than the *product* (e.g., eContour). Guiding learners through the process of a complex task via awareness-cueing techniques, while perhaps a lot more involved than simply showing final contours, has the benefit of engaging learners in *productive struggle*: a curated state of demand from learners (often with medium-level difficulty [52]) that leads to growth and mastery of learning concepts [37].

Skill-Based Feedback via DG2 — DG2 aimed to balance between consensus guidelines and the differing tendencies among expert physicians. Residency programs often employ one-to-one apprenticeship models that provide direct supervision over high-stakes clinical tasks. Yet, these models fall short of familiarizing residents with expert variations. As presented in Sec 5.2, expertise level can determine balancing between consensus guidelines and expert tendencies: the participating residents pointed out that junior residents might need to focus more on solely the consensus contours, before getting exposed to variation in expert contours that might be more beneficial for experienced residents. In practice, we envision iConTutor to distinctly highlight the assigned faculty's contour in addition to the consensus and tendencies of other expert physicians. Due to the existing power dynamics in residency programs [44], residents often express satisfying the exact requirements of their faculty [65]. As such, the faculty's styles can be highlighted, especially in the earlier years of residency.

Feedback design in contouring training system should enable experienced physicians to negotiate conflicts, especially when residents might find their contouring tendencies differ from the provided feedback by the system. In this study, the participants' tendencies ranged from including all of orange regions in their contours, to trimming them down to include only the green areas (Sec 5.3). In these cases, the design of dots and brush of iConTutor helped residents negotiate conflicts by taking in the information and deciding to pursue the strategy they deemed more appropriate. Given the subjective and unstructured nature of contouring, feedback interfaces should incorporate ways to negotiate potential conflicts between the learner's tendencies and system feedback.

Light-Touch Feedback via DG3 — The third DG promoted designing light-touch feedback mechanisms, such as short and dismissible text snippets adjacent to the main workspace, as well as changing colors of minimal user interface elements. This strategy aimed to lessen the extrinsic cognitive load, as defined in educational psychology [55]; while *intrinsic* cognitive load stems from the inherent complexity of the learning task itself, *extrinsic* cognitive load focuses on the presentation of the learning task and feedback through instructional design. As evident by Sec. 5.3, the residents found the mechanisms minimally intrusive which later allowed them to conveniently accept or dismiss feedback according to their expertise. The flexibility and non-intrusiveness of iConTutor's feedback mechanisms can especially benefit residents who pursue unique (and at times, differing) contouring strategies [66]: for instance, physicians who prefer to place outlines on multiple images before refining individual slices, are aware of their imperfect initial contours. As such, when feedback is gently provided (without majorly interfering with the contouring workflow), these residents can continue contouring and only engage with the feedback at appropriate times. There may also be risks in making feedback too minimally intrusive. As noted in Section 5.2, some participants missed the text snippets adjacent to the main workspace, likely because the contouring task was highly demanding. Future work can explore how to better balance extrinsic cognitive load by considering resident expertise, feedback importance, and case urgency.

The design goals altogether shaped the foundation of the three feedback mechanisms in iConTutor (light-touch, skill-based, and awareness-cueing feedback) which can improve contouring outcomes and lead to higher patient safety. These design goals can further shape learning tools in other residency domains, especially ones involving imaging-based tasks like pathology (e.g., identifying tumor cells [26]) and dentistry (e.g., describing cavities [46]).

6.4 Limitations and Future Work

Despite the novel insights offered in this work (for both contouring education and broader residency training) some limitations exist:

(1) All nine participants, as well as the two participating experts, attended the same teaching hospital, and hence, likely have developed similar training strategies. While focusing on the same institution enabled us to leverage the network of our medical collaborators to engage physicians in in-depth user studies, future work can recruit residents on a broader scale to capture more diverse teaching and learning strategies. In addition, iConTutor leverages an existing dataset of crowdsourced contouring data [61]; as larger datasets of expert contours become available, future systems can prioritize selecting and presenting diverse feedback.
(2) The study design does not measure learning benefits of iConTutor compared to existing educational resources: this decision aimed to maximize time spent with highly-specialized physicians on novel contouring education techniques, and as such, we relied on residents reflecting on their past experiences. Follow-up works can conduct comparative studies with existing educational atlases like eContour. Also, the ordering of manual feedback after automatic feedback may have influenced participants' perceptions of usability and learnability. We presented automatic feedback first to ensure participants had sufficient exposure to all three feedback types before judging which to retain in the manual phase. Regardless, future studies could counterbalance the order to more accurately assess perceptions of the tool and its feedback features, especially in studies that directly compare manual and automatic feedback which was not an objective of this work. Future research can also directly measure the three design goals presented in this study, by for instance, employing NASA-TLX [32] for measuring cognitive load (DG3).
(3) While instrumenting identical case and target structure for pre- and post-contouring aimed to measure exact learning gains, there is also potential for unaccounted carry-over and practice effects. These effects may have also been amplified by the short duration of the overall contouring session and the relatively brief interval between the pre and post steps. To mitigate this bias, we used different tasks for the feedback steps than ones used in the pre-/post-tests and incorporated an in-depth familiarization session before the pre-test (to reduce tool-related learning effects). However, future research can more robustly measure learning gain by employing between-subject study with a control condition, or increase tool exposure via longitudinal studies and further examine the real-world impact of interactive learning tools. The higher tool exposure can also lessen the novelty bias that might have existed in this study. Lastly, while the study participants (comprising almost the entire cohort) varied in residency years, everyone had at least one rotation with breast cancer. Future work can recruit senior medical students to examine how total novices interact with iConTutor.

7 Conclusion

This work presents the design and development of iConTutor, a learning tool for medical contouring that augments apprenticeship training (commonly seen in residency programs) with timely and diverse feedback from expert physicians. This paper first offers three design goals by leveraging dynamics of medical apprenticeship and educational technology, such as balancing between consensus guidelines and broader tendencies among experts. Lastly, we introduce three feedback mechanisms based on these goals to facilitate the highly specialized and critical task of contouring in radiation oncology. We evaluated iConTutor in a lab study with nine residents (out of a cohort of 10) and found that contouring accuracy increased by 32% when comparing pre- and post-tests. iConTutor's mix of consensus contours and expert tendencies further enabled physicians to reflect on the underlying structures when adjusting boundaries. Lastly, the three design goals (evaluated as part of iConTutor)

provide a pathway for transforming other residency programs with timely and diverse expert feedback, especially image-based practices like pathology and dentistry.

References

[1] Ross A Abrams, Kathryn A Winter, William F Regine, Howard Safran, John P Hoffman, Robert Lustig, Andre A Konski, Al B Benson, John S Macdonald, Tyvin A Rich, et al. 2012. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704—a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. *International Journal of Radiation Oncology* Biology* Physics* 82, 2 (2012), 809–816.

[2] Sahaja Acharya, Samantha Hsieh, Jeff M Michalski, Eric T Shinohara, and Stephanie M Perkins. 2016. Distance to radiation facility and treatment choice in early-stage breast cancer. *International Journal of Radiation Oncology* Biology* Physics* 94, 4 (2016), 691–699.

[3] Oke James Ajogbeje. 2023. Enhancing classroom learning outcomes: The power of immediate feedback strategy. *International Journal of Disabilities Sports and Health Sciences* 6, 3 (2023), 453–465.

[4] Tayyaba Gul Malik Rabail Alam. 2019. Comparative analysis between pre-test/post-test model and post-test-only model in achieving the learning outcomes. *Pakistan Journal of Ophthalmology* 35, 1 (2019).

[5] Vincent Aleven. 2001. Helping students to become better help seekers: Towards supporting metacognition in a cognitive tutor. *German-USA Early Career Research Exchange Program: Research on Learning Technologies and Technology-Supported Education, Tubingen, Germany* (2001).

[6] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.

[7] VA Andrianarison, M Laouiti, O Fargier-Bochaton, G Dipasquale, X Wang, NP Nguyen, R Miralbell, and V Vinh-Hung. 2018. Contouring workload in adjuvant breast cancer radiotherapy. *Cancer/Radiotherapie* 22, 8 (2018), 747–753.

[8] Anet Aselmaa, Richard HM Goossens, Ben Rowland, Anne Laprie, Yu Song, and Adinda Freudenthal. 2014. Medical factors of brain tumor delineation in radiotherapy for software design. In *5th International conference on applied human factors and ergonomics (AHFE)*. 4865–4875.

[9] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[10] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[11] Esdras Jorge Santos Barboza and Márcia Terra da Silva. 2016. The importance of timely feedback to interactivity in online education. In *Advances in Production Management Systems. Initiatives for a Sustainable World: IFIP WG 5.7 International Conference, APMS 2016, Iguassu Falls, Brazil, September 3-7, 2016, Revised Selected Papers*. Springer, 307–314.

[12] Hana Baroudi, Kristy K Brock, Wenhua Cao, Xinru Chen, Caroline Chung, Laurence E Court, Mohammad D El Basha, Maguy Farhat, Skylar Gay, Mary P Gronberg, et al. 2023. Automated contouring and planning in radiation therapy: what is 'clinically acceptable'? *Diagnostics* 13, 4 (2023), 667.

[13] Hansen C Bow, Jonathan R Dattilo, Andrea M Jonas, and Christoph U Lehmann. 2013. A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine* 88, 6 (2013), 766–770.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.

[16] Kristy K Brock. 2013. *Image processing in radiation therapy*. CRC press.

[17] Chen Chen, Matin Yarmand, Varun Singh, Michael V Sherer, James D Murphy, Yang Zhang, and Nadir Weibel. 2022. VRContour: bringing contour delineations of medical structures into virtual reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 64–73.

[18] Ran Cheng and Julita Vassileva. 2005. Adaptive Reward Mechanism for Sustainable Online Learning Community.. In *AIED*. 152–159.

[19] Jessica C Dai, Thomas S Lendvay, and Mathew D Sorensen. 2017. Crowdsourcing in surgical skills acquisition: a developing technology in surgical education. *Journal of Graduate Medical Education* 9, 6 (2017), 697–705.

[20] Indra J Das, Julia J Compton, Amishi Bajaj, and Peter A Johnstone. 2021. Intra-and inter-physician variability in target volume delineation in radiation therapy. *Journal of Radiation Research* 62, 6 (2021), 1083–1089.

[21] Janet de Groot, Richard Tiberius, Joanne Sinai, Aileen Brunet, Peter Voore, David Sackin, Susan Lieff, and Susan Reddick. 2000. Psychiatric Residency. *Academic Psychiatry* 24, 3 (2000), 139–146.

[22] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R Simon Sherratt. 2023. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society* 4, 1 (2023), 68–75.

[23] Frances Duane, Marianne C Aznar, Freddie Bartlett, David J Cutter, Sarah C Darby, Reshma Jagsi, Ebbe L Lorenzen, Orla McArdle, Paul McGale, Saul Myerson, et al. 2017. A cardiac contouring atlas for radiotherapy. *Radiotherapy and Oncology* 122, 3 (2017), 416–422.

[24] Jeanne M Farnan, Lindsey A Petty, Emily Georgitis, Shannon Martin, Emily Chiu, Meryl Prochaska, and Vineet M Arora. 2012. A systematic review: the effect of clinical supervision on patient and residency education outcomes. *Academic Medicine* 87, 4 (2012), 428–442.

[25] C Ailie Fraser, Tricia J Ngoon, Ariel S Weingarten, Mira Dontcheva, and Scott Klemmer. 2017. CritiqueKit: A mixed-initiative, real-time interface for improving feedback. In *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 7–9.

[26] Peter N Furness. 1997. The use of digital images in pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 183, 3 (1997), 253–263.

[27] Yuxiao Gao, Yang Jiang, Yanhong Peng, Fujiang Yuan, Xinyue Zhang, and Jianfeng Wang. 2025. Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. *Tomography* 11, 5 (2025), 52.

[28] Erin F Gillespie, Neil Panjwani, Daniel W Golden, Jillian Gunther, Tobias R Chapman, Jeffrey V Brower, Robert Kosztyla, Grant Larson, Pushpa Neppala, Vitali Moiseenko, et al. 2017. Multi-institutional randomized trial testing the utility of an interactive three-dimensional contouring atlas among radiation oncology residents. *International Journal of Radiation Oncology* Biology* Physics* 98, 3 (2017), 547–554.

[29] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

[30] Hao Guan and Mingxia Liu. 2021. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering* 69, 3 (2021), 1173–1185.

[31] Maaike Haak, Menno De Jong, and Peter Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour IT* 22 (09 2003), 339–351. doi:10.1080/0044929031000

[32] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[33] Sophie Hieke and Petra Wilczynski. 2012. Colour Me In–an empirical study on consumer responses to the traffic light signposting system in nutrition labelling. *Public health nutrition* 15, 5 (2012), 773–782.

[34] Michelle Ichinco. 2014. Towards crowdsourced large-scale feedback for novice programmers. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 189–190.

[35] Yuchao Jiang, Daniel Schlagwein, and Boualem Benatallah. 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice. *PACIS* (2018), 180.

[36] Lisa A Kachnic, Kathryn Winter, Robert J Myerson, Michael D Goodyear, John Willins, Jacqueline Esthappan, Michael G Haddock, Marvin Rotman, Parag J Parikh, Howard Safran, et al. 2013. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. *International Journal of Radiation Oncology* Biology* Physics* 86, 1 (2013), 27–33.

[37] Manu Kapur. 2008. Productive failure. *Cognition and instruction* 26, 3 (2008), 379–424.

[38] Malcolm S Knowles, Elwood F Holton III, and Richard A Swanson. 2014. *The adult learner: The definitive classic in adult education and human resource development*. Routledge.

[39] David L Kok, Sathana Dushyanthen, Gabrielle Peters, Daniel Sapkaroski, Michelle Barrett, Jenny Sim, and Jesper Grau Eriksen. 2022. Virtual reality and augmented reality in radiation oncology education–A review and expert commentary. *Technical Innovations & Patient Support in Radiation Oncology* 24 (2022), 25–31.

[40] Elizabeth A Krupinski. 2000. The importance of perception research in medical imaging. *Radiation medicine* 18, 6 (2000), 329–334.

[41] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 75–84.

[42] Georgios Lappas, Cecile JA Wolfs, Nick Staut, Natasja G Lieuwes, Rianne Biemans, Stefan J van Hoof, Ludwig J Dubois, and Frank Verhaegen. 2022. Automatic contouring of normal tissues with deep learning for preclinical radiation studies. *Physics in Medicine & Biology* 67, 4 (2022), 044001.

[43] Elinda Ai-Lim Lee, Kok Wai Wong, and Chun Che Fung. 2010. How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education* 55, 4 (2010), 1424–1442.

[44] Heather B Leisy and Meleha Ahmad. 2016. Altering workplace attitudes for resident education (AWARE): discovering solutions for medical resident bullying through literature review. *BMC medical education* 16 (2016), 1–10.

[45] Chun Chieh Lin, Suanna S Bruinooge, M Kelsey Kirkwood, Dawn L Hershman, Ahmedin Jemal, B Ashleigh Guadagnolo, B Yu James, Shane Hopkins, Michael Goldstein, Dean Bajorin, et al. 2016. Association between geographic access to cancer care and receipt of radiation therapy for rectal cancer. *International Journal of Radiation Oncology* Biology* Physics* 94, 4 (2016), 719–728.

[46] Pasquale Loiacono and Luca Pascoletti. 2012. *Photography in dentistry: theory and techniques in modern documentation*. Quintessenza Edizioni Milan.

[47] Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, and Andre Dekker. 2018. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology* 126, 2 (2018), 312–317.

[48] Stephen MacNeil, Celine Latulipe, and Aman Yadav. 2015. Learning in distributed low-stakes teams. In *Proceedings of the eleventh annual International Conference on International Computing Education Research*. 227–236.
[49] Renuka Malik, Julia L Oh, John C Roeske, and Arno J Mundt. 2005. Survey of resident education in intensity-modulated radiation therapy. *Technology in cancer research & treatment* 4, 3 (2005), 303–309.
[50] Brent Martin, Antonija Mitrovic, Kenneth R Koedinger, and Santosh Mathan. 2011. Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction* 21 (2011), 249–283.
[51] Lukasz M Mazur, Prithima R Mosaly, Gregg Tracton, Marjorie P Stiegler, Robert D Adams, Bhishamjit S Chera, and Lawrence B Marks. 2017. Improving radiation oncology providers' workload and performance: Can simulation-based training help? *Practical radiation oncology* 7, 5 (2017), e309–e316.
[52] Janet Metcalfe and Nate Kornell. 2003. The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General* 132, 4 (2003), 530.
[53] Camil Ciprian Mireștean, Roxana Irina Iancu, and Dragoș Petru Teodor Iancu. 2022. Education in radiation oncology—current challenges and difficulties. *International Journal of Environmental Research and Public Health* 19, 7 (2022), 3772.
[54] Naomi Miyake and Donald A Norman. 1979. To ask a question, one must know enough to know what is not known. *Journal of verbal learning and verbal behavior* 18, 3 (1979), 357–364.
[55] Jan L Plass, Roxana Moreno, and Roland Brünken. 2010. Cognitive load theory. (2010).
[56] Kate Rassie. 2017. The apprenticeship model of clinical medical education: time for structural change. *The NZ Medical Journal* 130, 1461 (2017), 66.
[57] Michael V Sherer, Diana Lin, Kartikeya Puri, Neil Panjwani, Zhigang Zhang, James D Murphy, and Erin F Gillespie. 2019. Development and usage of eContour, a novel, three-dimensional, image-based web site to facilitate access to contouring guidelines at the point of care. *JCO Clinical Cancer Informatics* 3 (2019), 1–9.
[58] Maarten Someren, Yvonne Barnard, and Jacobijn Sandberg. 1994. *The Think Aloud Method - A Practical Guide to Modelling Cognitive Processes*.
[59] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2645–2656.
[60] Shalini K Vinod, Michael G Jameson, Myo Min, and Lois C Holloway. 2016. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiotherapy and Oncology* 121, 2 (2016), 169–179.
[61] Kareem Wahid, Diana Lin, Onur Sahin, Michael Cislo, Benjamin Nelms, Renjie He, Mohammed Naser, Simon Duke, Michael Sherer, et al. 2023. Large scale crowdsourced radiotherapy segmentations across a variety of cancer anatomic sites. *Scientific data* 10, 1 (2023), 161.
[62] Simon K Warfield, Kelly H Zou, and William M Wells. 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging* 23, 7 (2004), 903–921.
[63] Robert Woolson. 2008. *Wilcoxon Signed-Rank Test*. doi:10.1002/9780471462422.eoct979
[64] Evan J Wuthrick, Qiang Zhang, Mitchell Machtay, David I Rosenthal, Phuc Felix Nguyen-Tan, André Fortin, Craig L Silverman, Adam Raben, Harold E Kim, Eric M Horwitz, et al. 2015. Institutional clinical trial accrual volume and survival of patients with head and neck cancer. *Journal of Clinical Oncology* 33, 2 (2015), 156–164.
[65] Matin Yarmand, Chen Chen, Kexin Cheng, James Murphy, and Nadir Weibel. 2024. " I'd be watching him contour till 10 o'clock at night": Understanding Tensions between Teaching Methods and Learning Needs in Healthcare Apprenticeship. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
[66] Matin Yarmand, Chen Chen, Michael V Sherer, Yash N Shah, Peter Liu, Borui Wang, Larry Hernandez, James D Murphy, and Nadir Weibel. 2024. Enhancing Accuracy, Time Spent, and Ubiquity in Critical Healthcare Delineation via Cross-Device Contouring. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 905–919.
[67] Matin Yarmand, Srishti Palani, and Scott Klemmer. 2021. Adjacent Display of Relevant Discussion Helps Resolve Confusion. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–11.
[68] Matin Yarmand, Courtney N Reed, Udayan Tandon, Eric B Hekler, Nadir Weibel, and April Yi Wang. 2025. Towards Dialogic and On-Demand Metaphors for Interdisciplinary Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
[69] Matin Yarmand, Michael Sherer, Chen Chen, Larry Hernandez, Nadir Weibel, and James D Murphy. 2022. Evaluating Accuracy, Completion Time and Usability of Everyday Touch Devices for Contouring. *International Journal of Radiation Oncology, Biology, Physics* 114, 3 (2022), S96.
[70] Matin Yarmand, Borui Wang, Chen Chen, Michael Sherer, Larry Hernandez, James Murphy, and Nadir Weibel. 2023. Design and development of a training and immediate feedback tool to support healthcare apprenticeship. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
[71] Lian Zhang, Zhengliang Liu, Lu Zhang, Zihao Wu, Xiaowei Yu, Jason Holmes, Hongying Feng, Haixing Dai, Xiang Li, Quanzheng Li, et al. 2024. Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology. *Medical physics* 51, 3 (2024), 2187–2199.

[72] Qichang Zheng, Tianjun Mo, Xu Wang, et al. 2023. Personalized Feedback Generation Using LLMs: Enhancing Student Learning in STEM Education. *Journal of Advanced Computing Systems* 3, 10 (2023), 8–22.

A Individual DSC Over Time for Automatic Phase

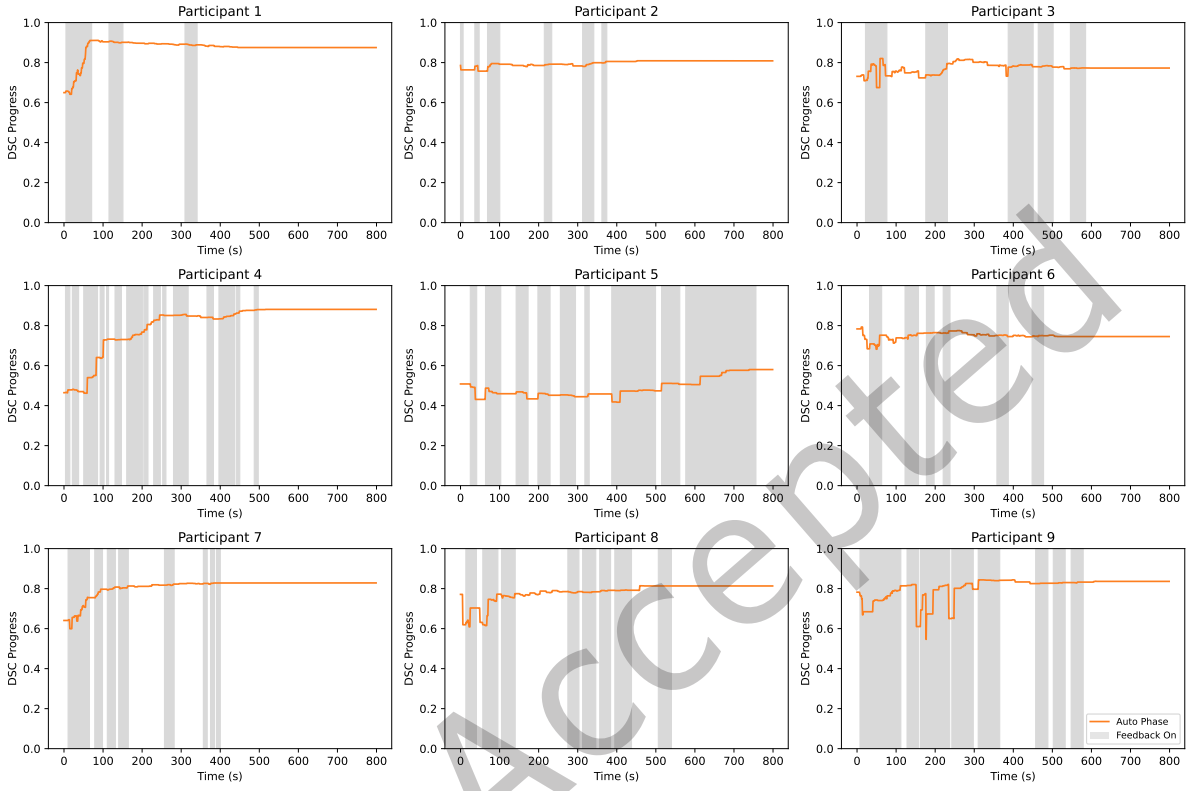


Fig. 6. Individual DSC progress across participants during the iContutor session with automatic feedback. This multi-panel figure displays the DSC progress over time for nine individual participants. Each subplot shows the DSC progression (orange line) over time (in seconds). Shaded regions represent intervals during which the system's feedback mechanism was activated.

B Individual DSC Over Time for Manual Phase

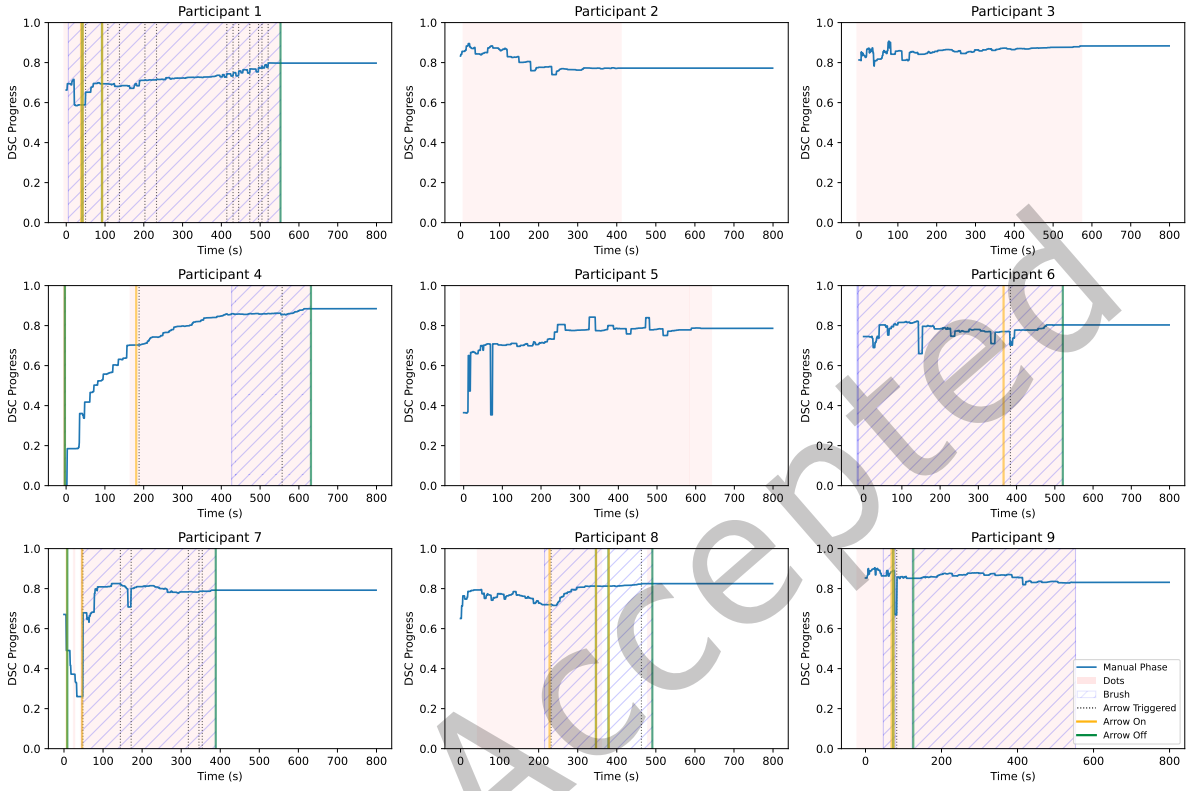


Fig. 7. Individual DSC progress across participants during the iConTutor session with manual feedback. This multi-panel figure displays the change in DSC over time for nine individual participants. Shaded red regions indicate time intervals during which the participant enabled the dots feature, and blue diagonal lines indicate where brush feature was turned on. Vertical dashed lines indicate when the arrow was triggered, and yellow and green vertical lines mark when the arrow feature was turned on and off, respectively.

C Individual DSC Over Time for Pre and Post Phases

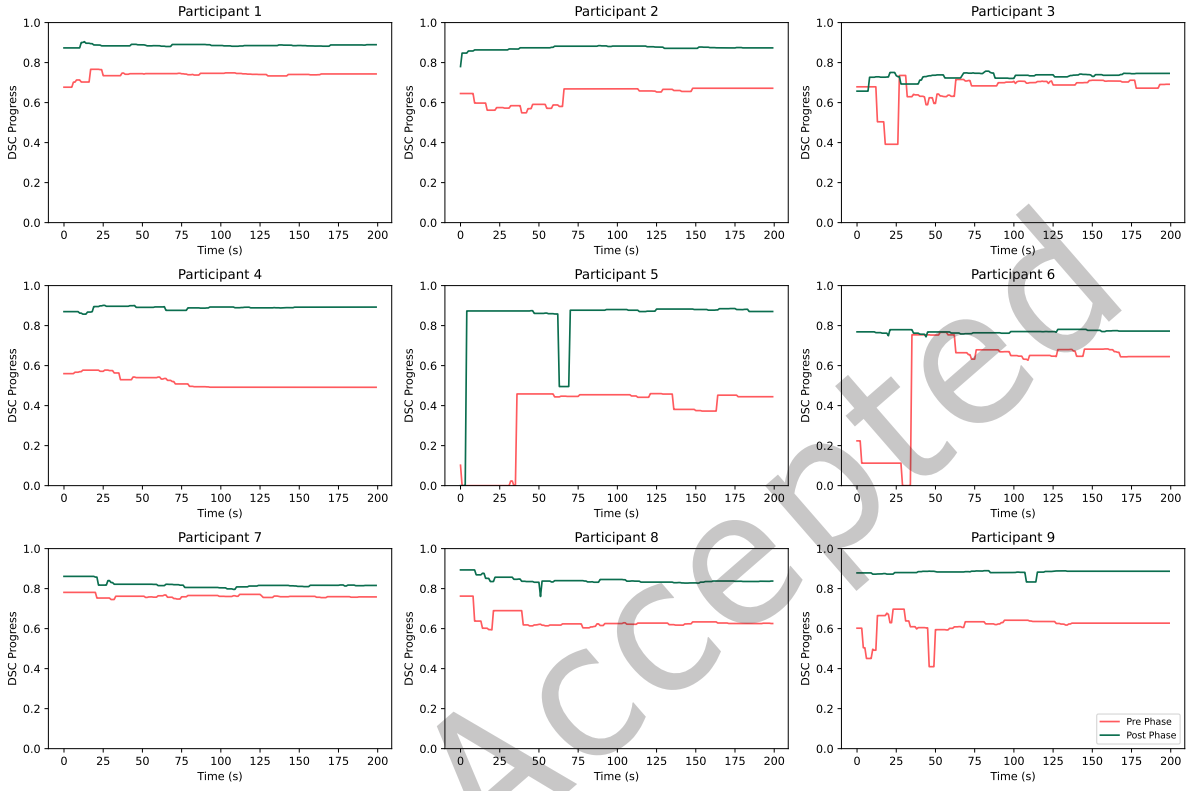


Fig. 8. This figure presents DSC progress over time for nine participants during the pre- (red) and post-tests (green). Each subplot corresponds to one participant and shows how their DSC evolved throughout each phase. The green lines represent performance during the post-contouring phase, conducted after participants had interacted with iConTutor's feedback system. The red lines show performance during the initial pre-test. For all participants, post phase curves remain consistently higher than pre phase curves, indicating improved accuracy after feedback exposure.

Received 31 May 2025; revised 5 December 2025; accepted 8 January 2026