Tendencies Matter: Augmenting One-to-One Medical Apprenticeship with Crowdsourced Expert Contours

Matin Yarmand
The Design Lab, UCSD
La Jolla, USA
myarmand@ucsd.edu

Joyce Lu
Data Science, UCSD
La Jolla, USA
jol072@ucsd.edu

Yunhao Luo
Computer Science, UCSB
Santa Barbara, USA
yunhaoluo@ucsb.edu

Megan Orr
Radiation Medicine, UCSD
La Jolla, USA
m1orr@health.ucsd.edu

Michael Sherer
Radiation Medicine, UCSD
La Jolla, USA
msherer@health.ucsd.edu

Nadir Weibel
The Design Lab, UCSD
La Jolla, USA
weibel@ucsd.edu

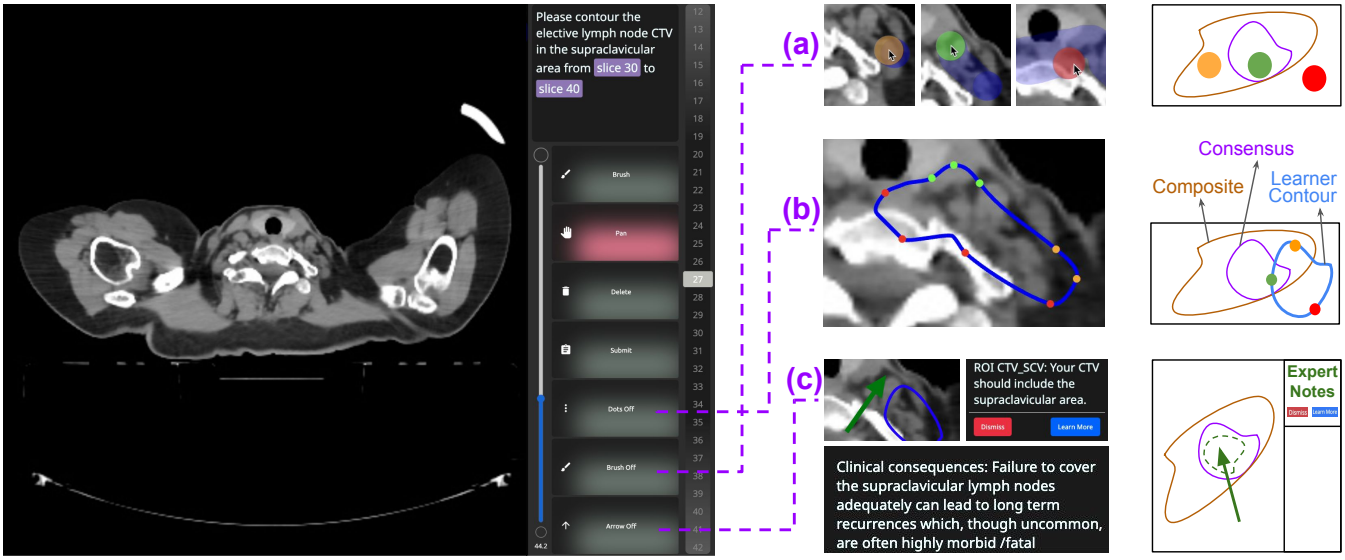

Figure 1: A learning platform that leverages contours from expert physicians to not only present a single consensus contour, but also show the inter-expert variability in contouring that is commonly hidden from learners due to the one-to-one apprenticeship model of residency programs. This tool offers three types of 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. (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. (c) The system displays targeted regions to be included and avoided via an arrow and a description which can be expanded for more details.

Abstract

The one-to-one apprenticeship in medical residency programs, while providing direct supervision from an expert physician, lacks timely and diverse feedback especially from other experts with unique tendencies. This often contributes to subpar training for high-stakes medical tasks and further diminishes patient safety. This paper first offers three design goals aimed to prevent gaming, balance between expert consensus and tendencies, and minimize cognitive load. Focused on contouring education in radiation oncology, this work then designs and develops a learning platform that not only provides just-in-time feedback, but also incorporates crowdsourced contours from eight expert physicians. The feedback strategies in this platform can inform computer-supported learning tools in other healthcare domains that aim to augment apprenticeship training with timely and diverse expert feedback.

CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan

CCS Concepts

• **Human-centered computing** → **Interaction techniques**.

Keywords

Contouring Feedback, Healthcare Apprenticeship, Crowdsourcing

1 Introduction and Background

Apprenticeship training in healthcare residency programs (especially radiation oncology) is the underlying structure for transferring highly specialized and critical medical skills from expert to novice physicians [6]. At its core, residents learn these skills by observing the clinical practices of an assigned faculty and later re-create their processes [10].

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 [7]. As such, feedback exchange between residents and faculty can take substantial time; this time pressure can degrade the quality of residency training [3]. 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 correct "gold standard" solution even among experts [5]. While this training model provides the direct supervision needed to carry out complex tasks with significant consequences, facilitating broader feedback can improve the quality and robustness of specialized medical practices.

This work aims to improve one-on-one healthcare apprenticeship by integrating timely and diverse feedback processes. Specifically, we explore contouring education in radiation oncology training; contouring describes meticulous outlining of tumor and nearby organs at risk, producing a blueprint for high-dose radiation. Informed by prior work in learning science and medicine, we offer three design goals to prevent gaming, balance between consensus and tendencies, and minimize cognitive load. These goals contributed to the design and development of 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 aim to evaluate this tool in lab studies with residents, and later conduct a longitudinal study with senior medical students to validate which feedback mechanisms are most useful and whether this varies with the learner's level of experience. This paper has the potential to inform learning tools in healthcare that provide timely and diverse feedback, contributing to higher quality in healthcare training and ultimately, patient care.

2 Design Goals

The design of this tool followed three goals unique to learning tools, medical apprenticeship, and contouring education.

[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 [2]. As such, we aimed to design (intentionally) vague feedback mechanisms that 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: instances in which learners don't know what they don't know, and hence, are incapable of seeking explicit guidance [12].

[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: for instance, varying contouring styles stem from applying radiation to tumour while potentially damaging nearby healthy organs. While consensus guidelines represent the commonalities among different experts (i.e., near a "gold-standard" solution), there is value in learning about granular tendencies within experts [4]. The feedback mechanisms of this tool aim to distinguish and balance between consensus feedback and the broader expert tendencies.

[DG3] Minimize 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 [1]. As such, just-in-time feedback mechanisms should balance between the provided learning resources and the main contouring actions. Excessive and obtrusive feedback can not only diminish learning gains, but also prolong contour delineation and putting patients at risk. Our learning tool aims to facilitate this balance by addressing the *what* (i.e., content snippets), the *when* (i.e., during idle times), and the *how* (i.e., easily dismissible feedback).

3 Feedback Mechanisms

In addition to core contouring features for delineation, navigation, and image manipulation [11, 13], our learning platform contains three feedback mechanisms (Fig. 1) according to the design goals.

The feedback content used in this work comes from two sources. First, we worked closely with two experts (one attending faculty and one senior resident) to develop regions and the descriptions. Second, we leveraged the crowdsourced expert feedback collected as part of the C3RO project (Contouring Collaborative for Consensus in Radiation Oncology) [8]. 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 [9], and *composite*, the union of all eight contours, designed to help learners understand the degree of acceptable variation in contours across multiple experts.

Feedback (a): Changing Brush Color when Outlining Contours — This feedback initiates as early as the learner starts delineating by dragging their cursor on the image (Fig. 1–a). The brush (that is used for contouring) changes colors according to where the center of the brush lies. 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 minimizes cognitive load by

solely changing color of an existing element [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 promote learning the underlying anatomy [DG1].

Feedback (b): Color-coded Dots placed On Learner Contours — The second type of feedback (Fig. 1–b) activates immediately after feedback (a), 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), and outside composite (red). To calculate the frequency and placement of the dots, the prototype first calculates the number of distinct segments within the contour, multiply this number by 0.7 while upper-bounding by 10 (validated via trial and error with physicians), and randomly select segments that encapsulate the dots. In this method, the areas that are refined more (resulting in more segments) are likely to contain more dots; meaning, learners are likely to receive more feedback on more complex and clinically significant regions. Limiting the number and 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 lessen crowding the interface, and hence, the cognitive load [DG3]. Lastly, the same color-scheme as feedback (a) distinguishes and balances between contouring guidelines and tendencies [DG2].

Feedback (c): Targeted Avoid and Include Regions — This feedback aims to place emphasis on nearby Regions of Interest (ROIs) that (according to consensus guidelines) should always be included (i.e., unique areas subject to recurrence) or fully avoided, such as lungs. Overall, we included five avoidance regions and three inclusion areas, and for each region, we added a short and long explanations of why the region should be avoided (or included). When a 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. 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. 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 description highlights core topics based on consensus guidelines, more so than the existing tendencies [DG2]. Lastly, displaying text-based content in-situ of the main workspace is a common technique used to minimize cognitive load of learning tasks [12].

4 Discussion and Future Work

This work presents the design and development of a learning environment that augments apprenticeship training (common in residency programs) with diverse feedback from expert physicians. This paper further 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 (Fig. 1) to facilitate the highly specialized and critical task of contouring in radiation oncology.

We plan to extend this line of work by not only evaluating the three feedback mechanisms, but also measuring the impact of expertise on each type of feedback (and specifically, consensus guidelines vs subjective tendencies). We aim to recruit UCSD residents for a lab study, and later delve into longitudinal studies with both residents and senior medical students, covering a broad range of contouring expertise. We further aim to facilitate two types of workflows: (1) *Automatic Feedback* in which the tool detects when and how to present support, and (2) *Manual Feedback* that gives agency to participants to control the type of feedback they receive (i.e., dots, brush, and arrow). Lastly, we plan on collecting log data (including timestamps and accuracy scores) to capture the usability and learnability of the tool.

Beyond contouring education, this work has the potential to provide a pathway for transforming other residency programs, specifically by introducing learners to diverse expert tendencies complementary to the core guidance from the main supervisor. The design goals and feedback mechanisms introduced in the paper can inform educational tools for other healthcare domains, especially image-based domains like pathology and dentistry.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] Kate Rassie. 2017. The apprenticeship model of clinical medical education: time for structural change. *The NZ Medical Journal* 130, 1461 (2017), 66.

[8] 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.

[9] 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.

[10] 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 CHI Conference on Human Factors in Computing Systems*. 1–19.

[11] 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.

[12] 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.

[13] 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.