Quality assurance of large-scale retrospective CT auto-segmentation for breast cancer radiotherapy

Nicholas Santoso[1,3]; Amy Frederick[2], PhD; Alanah Bergman[1,3], PhD; Tania Karan[1], MSc; Alan Nichol[1,3], MD
[1]BC Cancer - Vancouver, [2]BC Cancer - Abbotsford, [3]University of British Columbia - Vancouver

BACKGROUND

- In some patients, breast cancer can spread to the **internal mammary lymph nodes (IMNs)**, making **IMN irradiation (IMNI)** an effective treatment option.
- For patients with only a low risk of IMN involvement, IMNI is **controversial** due to the potential risks associated with **increased radiation exposure** of the **heart and lungs** and potential for **secondary cancer** in the contralateral breast.
- This study is part of a retrospective population study investigating the impact of primary tumor location and the radiation dose delivered to the IMNs on **patient survival outcomes.**
- This study focuses on the development of a **quality assurance (QA)** tool aimed at evaluating the fidelity of IMN contours generated by **Limbus Contour** (Radformation, v. 1.8.0-B3), a deep learning-based auto-segmentation software.

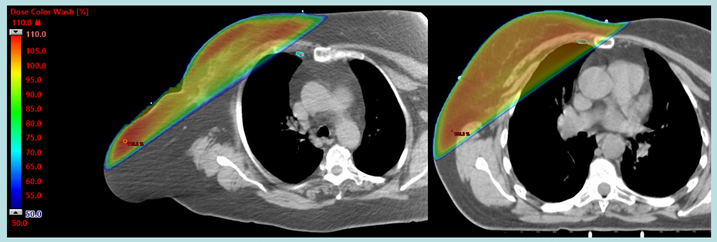

Figure 1: *Example dose distributions for plans that (a) exclude the internal mammary nodes (IMNs, cyan ovals) and (b) include the IMNs.*

MATERIALS AND METHODS

- A QA tool was developed to evaluate AI-generated IMN contours on CT images from a **cohort of about 19,000 women** referred to BC Cancer for breast cancer radiotherapy (2005-2014). In-house scripts were developed using **C#** and the **Eclipse Scripting API** to assess contour fidelity according to two evaluation methods:
  1. **Method A** compared AI-generated contours with **clinical contours** as the "ground truth."
  2. **Method B** compared AI-generated contours with a subset of **100 AI-generated IMN contours** that were manually reviewed by the authors, as the "ground truth".
     - Manual review involved checking for gross errors: missing contours, incorrect location, incorrect contour extent (e.g., number of intercostal spaces).
- Limbus Contour supports the application of different contouring guidelines for IMNs, specifically ESTRO and RTOG. For clarity, all subsequent references to IMN contours will follow the **RTOG guidelines**.

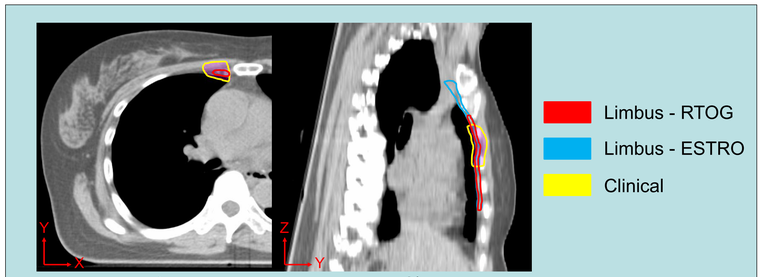

Figure 2: *Example IMN contours following RTOG and ESTRO guidelines generated by Limbus Contour and a clinical contour on (a) axial and (b) sagittal CT slices.*

- **Geometric data** (volume, surface area, and dimensions) from the clinical contours (Method A) or the subset of screened AI-contours (Method B) were used to define a **clinically acceptable range: +/- 2 standard deviations** from the median.
- The QA tool flags patients with AI-generated contours that fall outside the clinically acceptable range (Fig. 3) or have missing slices/contours.
- The QA tool's performance was tested on an **additional, independent subset of 100 patients**, using AI-generated IMN contours **reviewed by clinicians**. Each method was evaluated by comparing the contours flagged by the QA tool to those flagged during manual review.

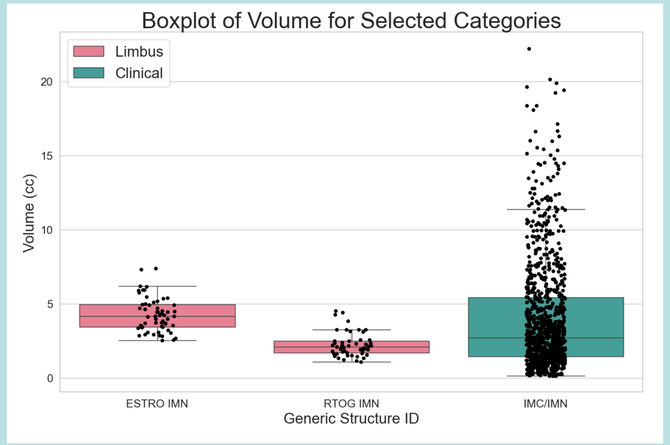

Figure 3: *Distribution of IMN volumes for AI-generated vs clinical contours (Boxplot shows 1st, 2nd and 3rd quartiles).*

RESULTS

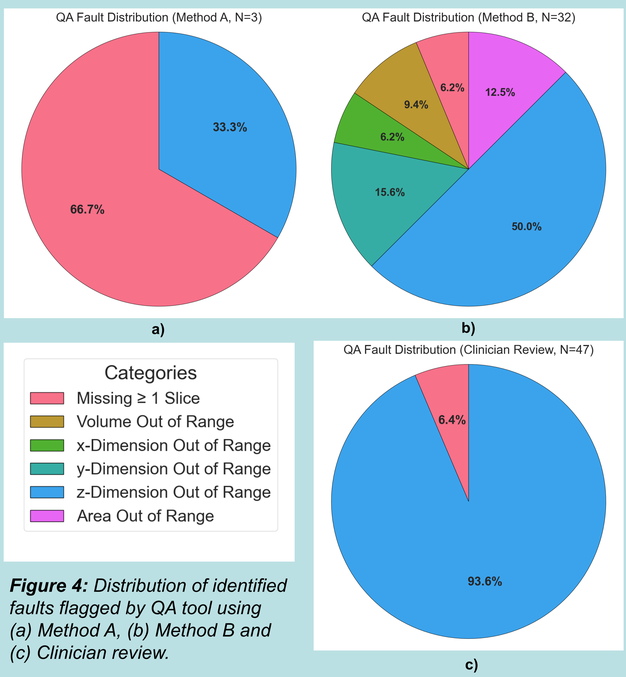

Figure 4: *Distribution of identified faults flagged by QA tool using (a) Method A, (b) Method B and (c) Clinician review.*

- Method A flagged **3/100** RTOG IMN contours, with the most frequent fault being a **missing slice fault.** (Fig. 4a)
- Method B flagged **32/100** RTOG IMN contours, with the most frequent fault being the **z-dimension** (i.e., the number of intercostal spaces; Fig. 4b).
- Clinician review flagged **47/100**, with the most frequent fault detected being **z-dimension.** (Fig. 4c)

DISCUSSION

- **Method A** detected fewer faults than Method B and Manual Review due to:
  - **Significant variability** in clinical contours (Fig. 5).
  - **Inability to differentiate** between ESTRO and RTOG contouring guidelines
- Range-based filtering using clinical data proved ineffective in flagging poor quality IMN contours

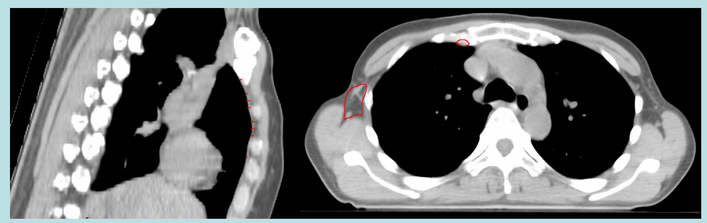

Figure 5: *Examples of (a) under contoured IMN and (b) misidentified IMN + axilla contour (red contours).*

- **Method B** detected faults that were more comparable to manual review due to:
  - Limbus Contour generating **consistent contours** that closely follow current contouring guidelines.
    - This reduces variability in the dataset, enabling more precise filtering and analysis.
- **Clinician review** results more closely align with Method B's distribution, highlighting its superior ability to flag contours relative to Method A. This is due to both Method B and clinicians identifying the **z-dimension length** as the most common fault in IMN AI-contouring, either over- or under-contouring intercostal spaces.
- **Future work** for this tool would include:
  - **Expanding and refining the dataset** used to inform the filtering process to improve the QA tool's accuracy.
  - **Addressing a key limitation** by considering the relative distance between structures within the CT scan. This would help detect faults such as structures being contoured in the wrong location which are not accounted for in the current iteration of the QA tool.

CONCLUSION

A QA tool using range-based filtering based on reviewed AI contours had less variability than one based on clinical contours. Future work will involve expanding and refining the QA tool's ranged-based filter to improve its accuracy, assessing the QA tool in a larger patient cohort, and addressing contouring faults that are not currently accounted for. The QA tool will play a crucial role in future studies by ensuring the high quality of AI-generated contours.

ACKNOWLEDGMENTS

This project is supported by the BC Cancer Foundation's Sprakkar Award and a research agreement with Limbus AI/Radformation.