Real-world Evaluation of an AI Auto-contouring Software in Radiotherapy – A Multi-Dimensional Assessment

A. Nanda[1], D. Sandys[1], G. Bell[1], D. Brand[1,2]

[1]University College London Hospitals NHS Foundation Trust
[2]University College London University

Clinical Problem

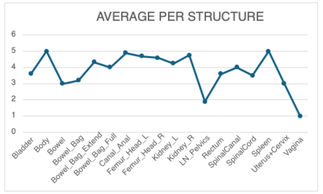
- When planning radiotherapy, oncologists must contour target volume (TV) and organs at risk (OAR) on the planning CT
- CTs have hundreds of images/slices so contouring is time consuming
- Delays in contouring may contribute to delays in patient pathways

AI Solution

- AI-based auto-contouring is increasingly used in radiotherapy (RT) planning.
- Auto-contouring involves a software, Limbus-AI in this case, drawing the contours automatically and then the clinician edits and approves them.
- Most evaluations rely solely on geometric evaluation or time-saving metrics obtained under controlled conditions. These do not reflect real-world operational or patient impact.
- NICE and RCR guidance recommends multi-domain evaluation.
- This study provides a real-world, multi-domain assessment of Limbus AI® for Urology radiotherapy and examines the strengths and limitations of evaluation methods.

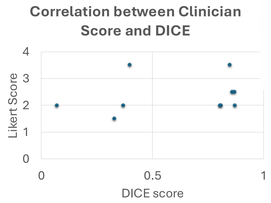
1. Clinician Usability Feedback

- Prospective subjective review by clinicians scoring the auto-contours compared to their opinion using Likert score (1-5)
- Two oncologists scored contours for 10 patients using a 5-point Likert scale.
- Median clinician usability was 3/5.



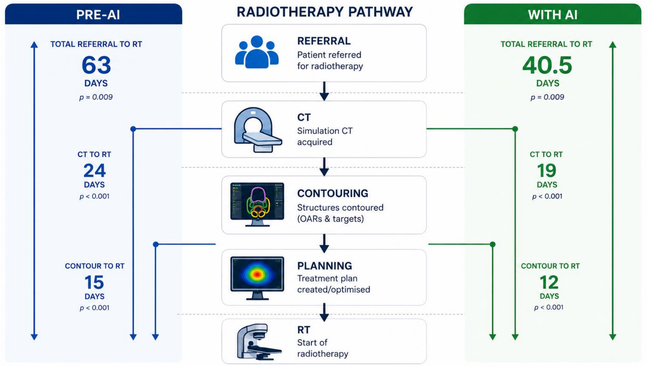
2. Geometric Accuracy

- Median DSC was 0.71 (IQR 0.64–0.75) and median HD 30.6 mm (IQR 23.4-34.62).
- Comparison of clinician scores to objective Dice score and Hausdorff distance (objective measures of difference) was weak (DSC Spearman 0.46 p=0.18; HD Spearman =-0.02 p=0.957).
- No correlation between the two showing that Likert scores by clinicians may not be an accurate way of analysing AI contours



4. Workflow Analysis

- Workflow analysis demonstrated reduced referral-to-RT (63→40.5 days, p=0.009), CT-to-RT (24→19 days, p<0.001) and contour-to-RT intervals (15→12 days, p<0.001).
- Limbus ALSO reduced the no. of replans needed due to plan quality (plans sent back to clinicians by physics team).
- Significantly (p = 0.002) reduced the number of patients that took >31 days from CT to receiving radiotherapy



3. Dosimetry

- For 32 patients, plans were re-optimised using AI-generated organs at risk (OAR) while retaining clinician targets.
- Mandatory constraint violations and optimal variations were assessed using clinician contours as reference.
- Across 32 AI-optimised plans, 75% showed identical constraint attainment to clinical plans. All rectum, bladder and femoral head constraints were met. A small increase in bowel violations was observed (5 AI vs 0 clinical, p=-0.052).
- See further details in poster "*How reliably can AI-generated OAR contours be used for prostate radiotherapy without compromising clinical dose constraints?*"

5. Cost Analysis

(Cost of Limbus for a year/ Total cost of contourer contouring) * 100 = % of time per person needed to be saved to make Limbus viable

- Median = 6 patients/month/consultant
- Economic modelling showed that a 6.1% reduction in contouring time was required for cost-neutrality
- Assuming each oncologist countours 1/5th of their time (2PAs), need to save 130/mins/month per oncologist
- Clinicians self-reported AI saves them medial 45 minutes per case
- Each clinical oncologist in Urology contours median 6 patients a month
- 6 * 45 = **270 minutes/month**
- Almost DOUBLE the required time saving to be viable

Conclusion

Each method demonstrated the safety and benefit of Limbus AI but had its strengths and limitations:

- Clinician usability is quick and highlights structure-specific errors but is subjective.
- Geometric metrics are objective but did not correlate with clinical scores.
- Dosimetric evaluation reflects clinical safety not just stylistic differences.
- Workflow analysis and economic evaluation capture operational impact but are influenced by local practice.

Although some workflow and economic gains were observed, this study highlighted opportunities to improve efficiency by adjusting workflows or expanding automation for high-contrast pelvic organs and it confirms the need for continued oversight of bowel segmentation.