Validating AI generated radiotherapy organ at risk contours using a secondary AI model

Dr Jonathan Smith – Clinical Fellow in AI
Matthew Ward – Scientific computing developer
Virgiliu Craciun – Head of Clinical Scientific Computing

The Clinical Problem & Project Aim

- Radiotherapy is a key cancer therapy treating 140,000 patients in the UK each year[1].
- Each treatment plan requires carefully contouring the organs at risk (OAR) from a planning CT scan to minimize the risk of radiation damage (Fig.1).
- Manually this takes between 1–3 hours per scan. AI models have been developed to auto-segment the OAR contours which clinicians then apply final adjustments to.
- AI OAR segmentation reduces inter/intra-operator variability and cuts contouring time by 70%[2]. UHS uses the commercial AI model MVision (MV) for OAR auto-segmentation.
- 16–40% of AI segmented OARs are accepted by clinicians without edits, and a further 40% need only minor mostly stylistic edits[2,3].
- This suggests the verification of the AI generated contours could be partially automated thus further standardising contours and saving clinicians more time.
- **Project Aim: Develop a locally trained AI model to assist and automate the validation of the adequacy of AI generated OAR contours.**

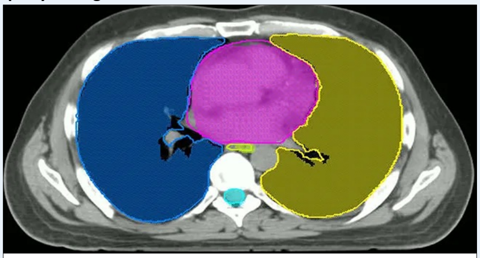

Fig.1 – An example of thorax CT scan organ-at-risk (OAR) segmentation[4]. This contouring needs to be completed for every CT slice.

AI Model Architecture

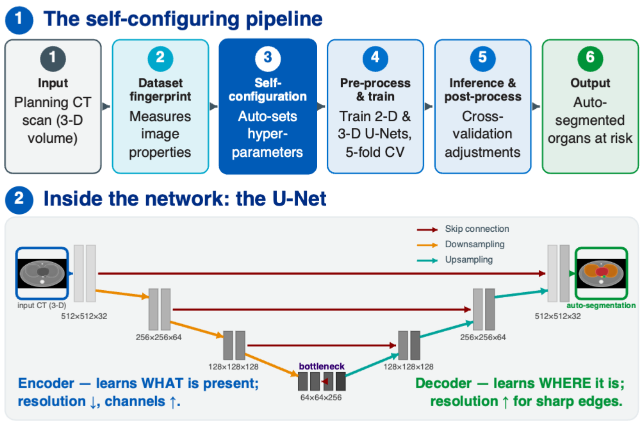

Fig.2 – nnU-Net AI model architecture. The model reads the digital fingerprint of the training data then auto-configures hyperparameters and perform 5-fold cross validation training with automatic post-process adjustments. UNet encodes the images in lower resolution by applying multiple feature filters then decodes the compressed images using the above layer encoder level as a reference.

- The local model (LM) was trained using an nnU-Net architecture (Fig.2) and trained with 100 chest CT scan images with OAR previously segmented by clinicians. These cases were from a period was when MV AI segmentation was being rolled out and it is unclear how widely MV was being used by clinicians at the time.
- A pilot region of lung-radiotherapy was used initially to evaluate the concept before expanding into other anatomical locations.
- Workflows automation developed via in-house ARTMEMIS platform to allow seamless integration into patient care.
- Model performance was evaluated using various metrics (Fig.3).

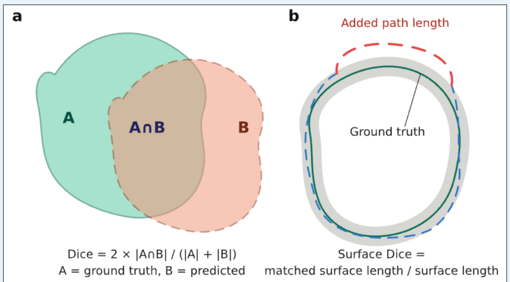

Fig.3 – AI segmentation accuracy metrics. A: Dice score assess the proportion of overlapping volume. B: Surface dice identifies the proportion of the surface cross over within tolerance zone and added path length measures the surface length outside the tolerable zone - grey area = tolerance zone.

Milestones, Progress & Project Support

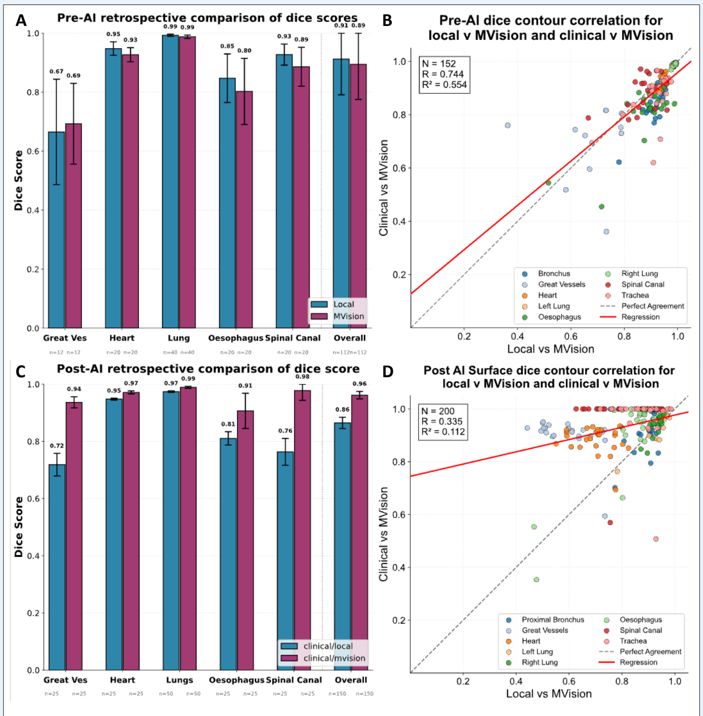

Fig.4 – A: Comparison of LM and MV dice score performance against the ground truth of 20 clinical contours drawn before the introduction of AI assistance. B: Dice score correlation of LM vs MV and MV vs clinical contours with clinical contours drawn using AI assistance. C: Comparison of LM and MV dice score performance against the ground truth of 25 clinical contours drawn after introducing AI assistance. D: Surface dice correlation between LM vs MV and MV vs clinical contours with clinical contours drawn using AI assistance.

- After model training a retrospective analysis of 20 scans contoured prior to the introduction of MV was performed to compare the accuracy of the LM and MV.
- This showed the LM performed equivalent or better compared to MV on dice accuracy when using the clinical contours as ground truth (Fig.4a).
- When correlating the dice score for LM vs MV contours against MV vs clinical contours this was found to predict the adequacy of MV generated contours with moderate accuracy (Fig.4b, r = 0.834, r2 = 0.695).
- 25 recent test cases, after AI auto-contouring was well established, were then analysed.
- Here the LM performed worse on dice score (Fig.3c). When correlating the surface dice score for LM vs MV contours against MV vs clinical contours the LM showed little predictive capacity for the adequacy of MV generated contours across dice, added path length or surface Dice (Fig.3d). This depreciation in performance suggests clinicians altered their contouring technique following the introduction of MV AI support.
- Excellent project support from Virgiliu Craciun and Matthew Ward (SciCom team, UHS).

Future Directions

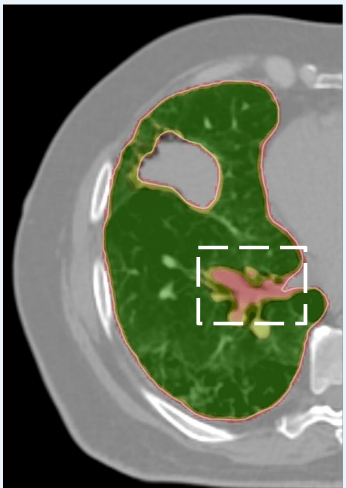

Fig.5 – LM classification probability heatmap using the MV right-lung contour mask. Red = low probability (<0.3). The white box marks where the LM correctly flagged low-probability of right lung but MV incorrectly contoured it as right lung.

- Try to improve MV generated contour adequacy predictability by retraining the model on 200 recent patients to ensure the ground truth contours were edited from MV baselines rather than pure manual inputting.
- Explore probability mapping (Fig.5) from the model's softmax function as a measure of contour adequacy.
- For example the white box in Fig.5 marks a red region where the LM correctly flagged a region as a low-probability of being right lung but MV incorrectly contoured the region as right lung.
- If the LM is proven to reasonably predictive the adequacy of MV drawn contours we will train models for other anatomical regions and undertake a prospective analysis correlating model prediction with clinician editing time and contour Likert scores.
- Develop clinical safety case / risk management as increasing automation may jeopardise clinician oversight and endanger patients.

1. Atkinson TG Claire. "Five stories that show the power of radiotherapy". Cancer Research UK - Cancer News [Internet]. 2025 Mar 13 [cited 2026 Jul 10].; 2.Poel, Robert et al. "A Comprehensive Multifaceted Technical Evaluation Framework for Implementation of Auto-Segmentation Models in Radiotherapy." *Communications Medicine* 5 (July 2025): 319.; 3. Inverarity, Catriona et al. "A Multi-Centre Real-World Evaluation of AI-Assisted OAR Contouring on Radiotherapy Treatment Planning Workflows." *British Journal of Radiology*, May 15, 2026, tqag111.; 4. Zhang, Fuli et al. "Geometric and Dosimetric Evaluation of the Automatic Delineation of Organs at Risk in Non-Small-Cell Lung Cancer Radiotherapy Based on a Modified DenseNet Deep Learning Network." *Frontiers in Oncology* 12 (March 2022).

NHS University Hospital Southampton NHS Foundation Trust
NHS Fellowship in Clinical AI NHS Digital Academy Programme
NHS England