The ROYAL MARSDEN NHS Foundation Trust

NHS Fellowship in Clinical AI
NHS Digital Academy Programme

Evaluating auto-contouring systems in radiotherapy

Katherine Mackay, David Bernstein, Alexandra Taylor
The Royal Marsden Hospital and The Institute of Cancer Research

Background

- Radiotherapy planning requires "contouring" of structures on CT scans
  - **Target volumes** (*e.g. gross tumour volume, clinical target volumes (CTV)*)
    - Visible tumour/ nodes AND areas of potential microscopic spread
    - *Ensure a tumoricidal dose of radiation is given here to* ***treat the cancer***
  - **Organs at risk (OAR)**
    - Healthy surrounding tissues
    - *Aim to minimise dose of radiation here to* ***reduce the risk of side effects***

- Contouring can be:
  - Time-consuming
  - Vulnerable to inter-observer variation

- Target volume contouring is **more challenging**
  - Varies depending on the tumour type
  - Relies on clinical decisions (using examination findings, other imaging modalities, histology and co-morbidities)
  - Cervical cancer target volumes are complex; inter-observer variation in contouring is well recognised

- **AI based contouring (auto-contouring) proposed as a solution**
  - Save time
    - Manual gynae contouring takes median 120 minutes per case[1]
  - Improve consistency of contouring

1. Evaluation of auto- contour clinical acceptability for cervical cancer CTVs

Methods:
- An inter-observer contouring study was undertaken with 6 observers contouring CTV structures on 6 retrospective cervical cancer cases. Auto-contours were also generated.
- Dice similarity coefficients were calculated for the auto-contour in relation to each manual observer.
- Protocol compliant contours were amalgamated to make an "inter-observer range". The proportion of auto-contours falling outside of this range was calculated and expressed as a percentage of the volume of the structure.
- The delineation uncertainty for the auto-contours and manual contours was calculated using established methodology[3].

Results
- Dice Similarity Coefficients varied depending on the observer (Fig 1) and structure
- >96% auto-contours fitted within the inter-observer range. Deviations from the range were areas needing editing to reflect clinical decisions (Fig 2).
- The delineation uncertainties were consistently lower for auto-contours compared to manual contours (Fig 3).

Conclusions
- The nnU-Net produces auto-contours that fit within the usual range of inter-observer variation, suggesting they are clinically acceptable.
- The delineation uncertainty for auto-contours was lower than for manual contours, suggesting auto-contours may need smaller treatment margins.
- The Dice Similarity coefficient is not a useful metric, as the value varies depending on the ground-truth used. A clinically useful cut-off value cannot also be defined.

Next steps
- The nnU-Net appears to produce clinically acceptable contours and current work is ongoing to embed this into a shadow clinical workflow.

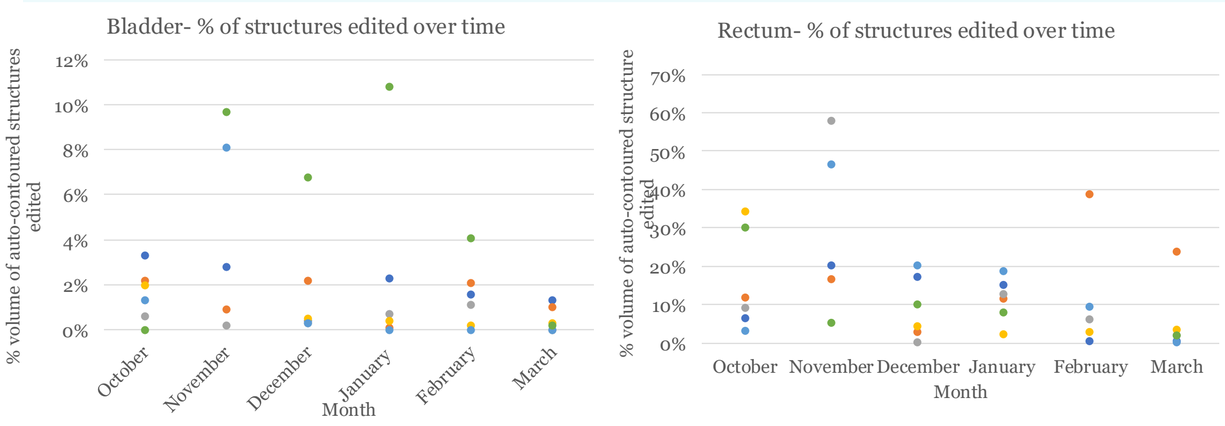

Fig 4. The percentage volume of each auto-contoured structure edited, where each point represents a sampled case in that month

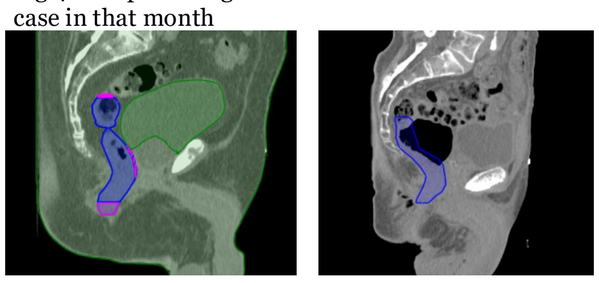

Fig 5. Images taken from error library showing failure to contour bladder in a post-prostatectomy case (left) and under-contouring of a gaseous rectum (right)

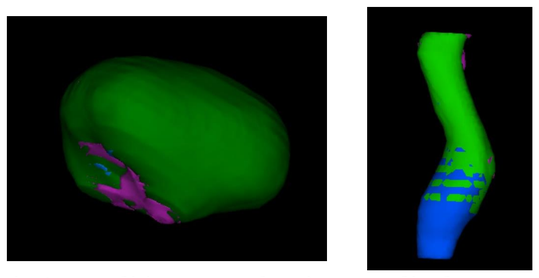

Fig 6: 3D editing maps showing un-edited contour (green), over-contouring (pink) and under-contouring (blue) for a bladder (left) and rectum (right)

The current landscape

- NICE have approved 9 technologies for use in the NHS providing contours are reviewed by a clinician
  - Mostly OAR
  - *Some simple single organ CTV structures*
- Commercial OAR auto-contouring is being increasingly deployed
  - *Raystation auto-contouring has been adopted at RMH*
  - Methods for ongoing monitoring/ quality assurance may vary but are needed
  - **Automation bias is a risk[2]**
- There has been limited success in *commercial* cervical cancer CTV auto-contouring
  - Promising early results seen from an in-house nnU-Net
    - Not previously used in a clinical pathway
    - Demonstrating **clinical acceptability** is challenging- there is **no perfect ground-truth contour** and there may be a range of acceptable contours

Project Aims:

1. Evaluate the **clinical acceptability** of cervical cancer CTV structures generated by an **in-house nnU-Net** in the context of acceptable inter-observer variation.

2. Review approved auto-contours for organs-at-risk generated by a commercial system (Raystation) and **monitor editing patterns to detect automation bias**

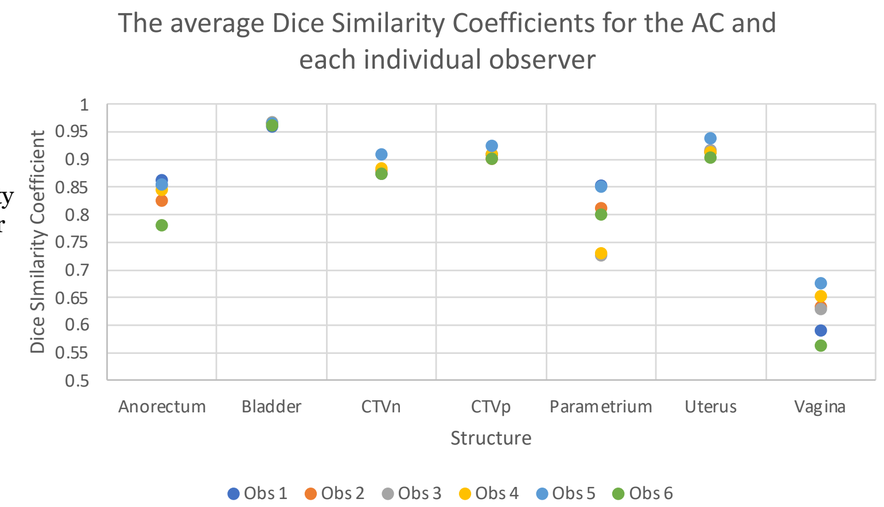

Fig 1: The average Dice Similarity Coefficients for the auto-contour compared to different manual observers

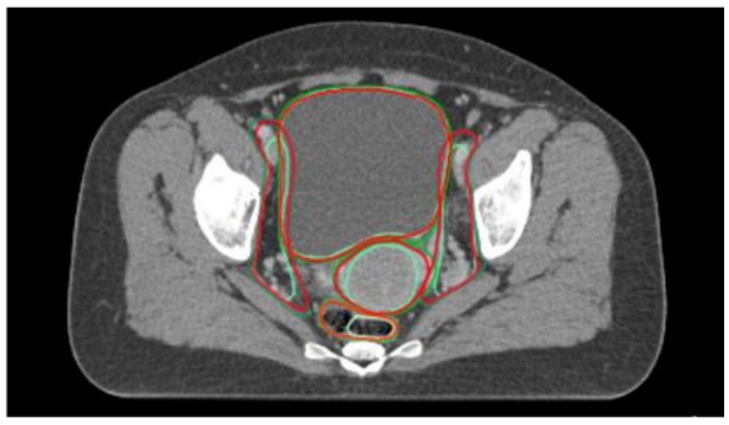

Fig 2: auto-contours from the nnU-Net in red, inter-observer range contours in green. Deviations outside this range need editing

Fig 3: The delineation uncertainty for manual contours and auto-contours from the in-house nnU-Net

Structure	Median (+ IQR) delineation uncertainty (mm)		P values for Wilcoxon Signed Rank Test
	Manual contours	Auto-contours	
CTV nodes	2.7 (+0.925)	1.6 (+0.725)	0.012*
CTV primary	1.8 (+0.425)	1.35 (+0.475)	0.078
Parametrium	3.85 (+0.6)	2.3 (+0.85)	0.045*
Anorectum	1.95 (+1.175)	1.7 (+0.9)	0.715
Bladder	0.95 (+0.55)	0.75 (+0.4)	0.026*

2. Monitoring for automation bias

Methods
- The original auto-contours generated with Raystation have been compared to the final edited contours used for treatment across a number of tumour types. The volume and location of edits for a sample of cases each month was recorded over a 6 month period.

Results
- Clinicians continued to edit contours throughout the 6 month period, for all tumour types, although there appeared to be a trend towards a reduction in editing (see Fig 4). All contours used for treatment were clinically appropriate.
- Locations of edits were similar throughout
- Auto-contouring failed in some situations; e.g. post-prostatectomy, significant gas in organs, abdominal sarcoma, prosthetic hips

Conclusions
- There was no evidence that automation bias occurred during this period. Trends towards reduced editing may relate to increased clinician familiarity with the technology. Automation bias is still a risk, and further monitoring is required.
- Editing occurs in similar places, suggesting these are areas to be prioritized for checking.
- A library of editing patterns and common errors has been generated for clinicians who are learning to review the contours (examples in Fig 5 and 6)

Acknowledgements: Many thanks to my supervisors, the physics team at RMH, my colleagues who completed the inter-observer study, Prof Ben Glocker and his team at Imperial for generating our nnU-Net model and the FCAI for giving me the opportunity to do this work.

References:
1. Montague, E. et al (2024). How Long Does Contouring Really Take? Results of the Royal College of Radiologists Contouring Surveys. Clinical oncology (Royal College of Radiologists (Great Britain)), 36(6), 335–342. https://doi.org/10.1016/j.clon.2024.03.005
2. Nealon, K. A. et al (2024). Monitoring Variations in the Use of Automated Contouring Software. Practical radiation oncology, 14(1), e75–e85. https://doi.org/10.1016/j.prro.2023.09.004
3. Tudor G, et al. Geometric Uncertainties in Daily Online IGRT: Refining the CTV-PTV Margin for Contemporary Photon Radiotherapy. British Institute of Radiology. 2020. 10.1259/geo-unc-igrt.

"This poster represents independent research supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and the Institute of Cancer Research, London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care."