

Beyond ground truth: Validating MRI-based auto-segmentation of cervical brachytherapy targets within interobserver variation



Sootha Kim^{1,2}, Omar Todd³, Arvind Sathyamurthy¹, Lisa Barraclough⁴, Asif Muzamil⁵, Kate Lankester⁶, Lorna Kviat⁷, David Bernstein^{1,2}, Ben Glocker³, Alexandra Taylor^{1,2}



¹The Royal Marsden NHS Foundation Trust ²The Institute of Cancer Research ³Imperial College London, ⁴The Christie NHS Foundation Trust, ⁵Imperial College Healthcare NHS trust, ⁶University Hospitals Sussex NHS Foundation Trust, ⁷Maidstone and Tunbridge Wells NHS trust

Introduction

- In radiotherapy contouring, variability between experts reflects inherent uncertainty in defining true anatomical boundaries.
- Despite this, deep learning auto-contouring (AC) models are typically evaluated against a single reference contour.
- This limitation is particularly relevant in cervical cancer brachytherapy, where target volume delineation demonstrates substantial interobserver variability (IOV) (1).
- In such settings, validation against a single reference may misrepresent true model performance.
- We developed an MRI-based AC model for intra-uterine brachytherapy, for both target volumes and organs at risk (OARs) and evaluated its performances within the context of IOV.

Aims

- To develop an auto-contouring model for MRI-based intra-uterine brachytherapy
- To evaluate its performance against expert inter-observer variation

Methods and Materials

- Target volumes: high-risk clinical target volume (HR-CTV) and residual gross tumour volume (GTVres) and OARs: bowel, sigmoid, bladder and rectum were contoured on 50 MRI planning scans from cervical cancer patients treated with intra-uterine brachytherapy, using an EMBRACE II-based protocol (2).
- The AC model was developed using the nnUNet framework with five-fold cross validation, where each fold consists of 40 training and 10 validation cases (3). Predictions on test data were generated through ensembling of all five models.
- An independent test set of 10 cases was contoured by experts from five UK centres using the same contouring protocol.
- IOV of HR-CTV and GTVres was quantified using the Dice Similarity Coefficient (DSC) between expert manual contours (MCs). ACs were compared against all MCs and evaluated relative to the distribution of expert variation. The proportion of AC volume lying within the bounds of MCs was assessed.

Results

- ACs demonstrated agreement with MCs comparable to IOV, with DSC values falling within the range observed between experts (Figure 1).
- For HR-CTV, median DSC between ACs and MCs ranged from 0.76 to 0.87 across the cases, compared with median DSC values of 0.59 to 0.85 between MCs.
- In cases with 6 or more MCs, ACs consistently lay within the bounds of IOV, with no volume outside the outer bounds and only a small proportion lying within the inner bounds of IOV (mean 1.5%).

Results

- Performance for GTVres was more variable, with median DSC between ACs and MCs ranging from 0.23 to 0.89 in cases with three or more expert observers contouring GTVres.
- In the lower-performing cases (median DSC < 0.6), inter-observer agreement was also poor: median DSC between ACs and MCs was 0.23, 0.51 and 0.59, with corresponding median DSC between MCs of 0.38, 0.21 and 0.43, respectively.

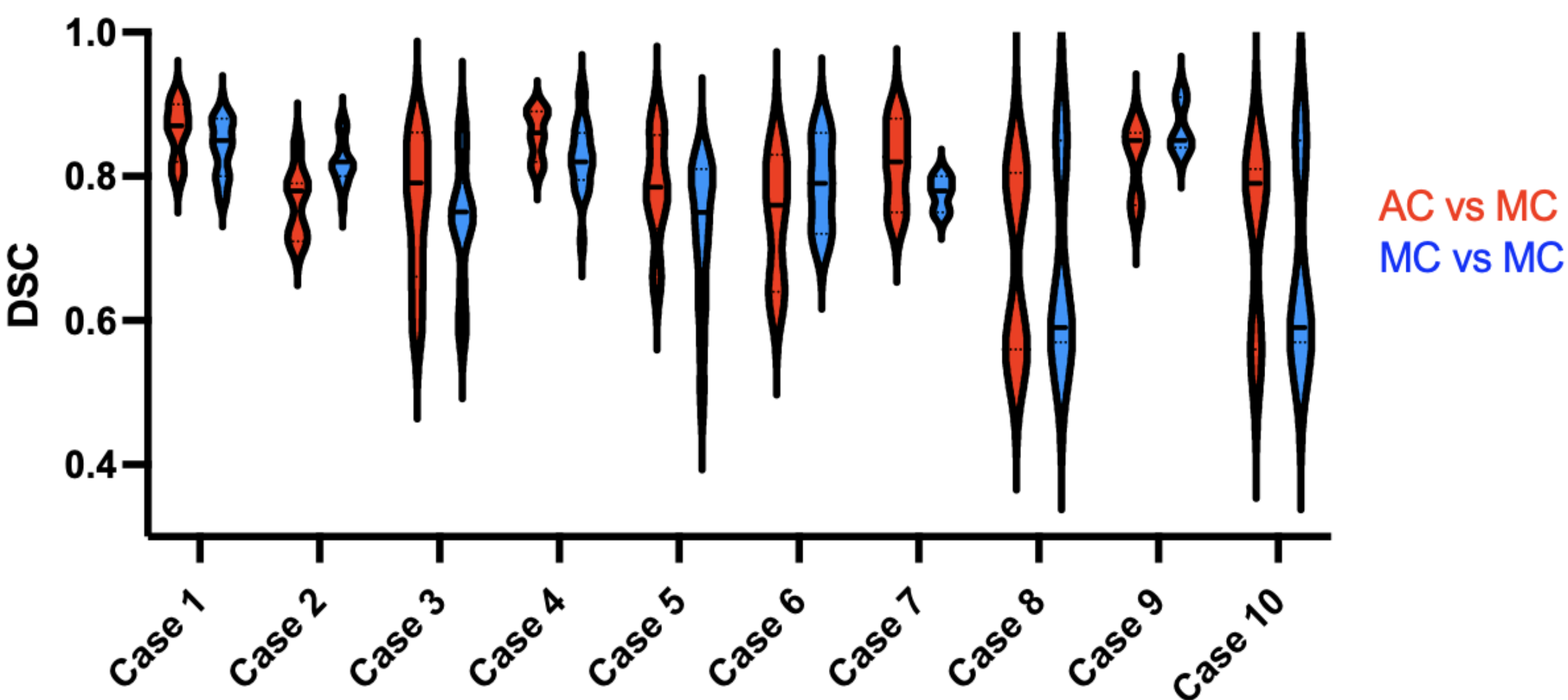


Figure 1. Violin plot of DSC comparing auto-contours (AC) with expert manual contours (MC) (red) and MC vs MC (blue). The plots display the median (solid line), interquartile range (dotted lines) and minimum-maximum range. The number of expert observers varied by case: 7 for Cases 1-4, 6 for Case 5 and 3 for Cases 6-10.

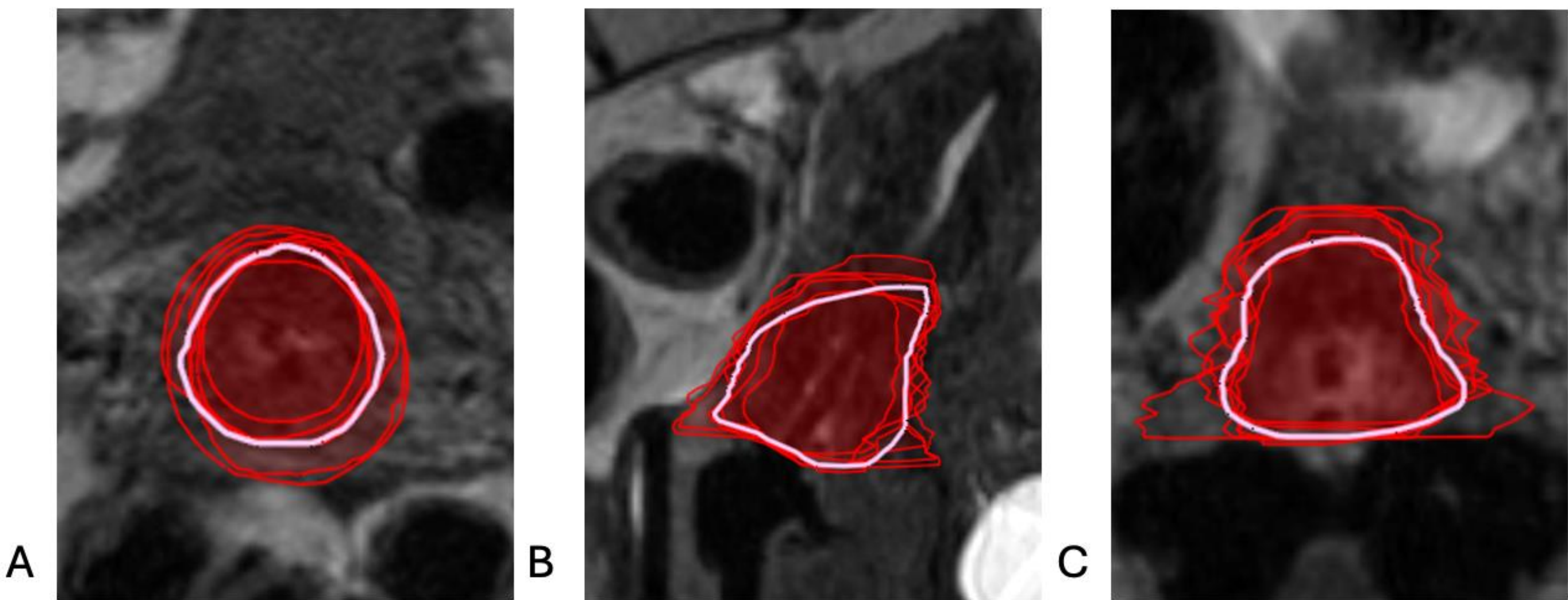


Figure 2. Axial (A), sagittal (B) and coronal (C) views of the HR-CTV. Expert contours are shown in red and the auto-contoured HR-CTV is shown in light pink.

Conclusions

- AC of HR-CTV for MRI-guided cervical brachytherapy achieved performance comparable to expert observers and consistently operated within the bounds of IOV.
- This was achieved using a relatively small but carefully curated, protocol-compliant dataset, supporting the value of high-quality, expertly curated training data over dataset size alone in developing clinically relevant models.
- Model performance varied according to the level of expert agreement. Lower DSC values for target volumes were observed in cases with greater IOV, suggesting that reduced AC performance metrics may reflect underlying delineation uncertainty rather than poor model performance.
- These findings highlight the limitations of validating auto-contouring models against a single reference contour, particularly for complex target volumes where no true ground truth exists.
- Validation against expert IOV provides a more clinically meaningful framework for assessing auto-contouring, helping to distinguish clinically acceptable variation from poor model performance.

References

1. Hellebust TP, Tanderup K, Lervag C, Fidarova E, Berger D, Malinen E, et al. Dosimetric impact of interobserver variability in MRI-based delineation for cervical cancer brachytherapy. *Radiother Oncol.* 2013;107(1):13-9.
2. Potter R, Tanderup K, Kirisits C, de Leeuw A, Kirchheiner K, Nout R, et al. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. *Clin Transl Radiat Oncol.* 2018;9:48-60.
3. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203-11.