Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond geometric accuracy: Evaluation of an MRI-based autocontouring prototype for brain and prostate radiotherapy

Nazanin Rahnama

nazanin.rahnama@usz.ch

University Hospital of Zurich

Stephanie Tanadini-Lang

University Hospital of Zurich

Vaisakh Nappady Joy

Siemens Healthineers (Germany)

Alina Sophie Elter

Siemens Healthineers (Germany)

Camilla von Wachter

University Hospital of Zurich

Matthias Guckenberger

University Hospital of Zurich

Nicolaus Andratschke

University Hospital of Zurich

Riccardo Dal Bello

University Hospital of Zurich

Research Article

Keywords: Magnetic resonance imaging (MRI), Autocontouring, Artificial Intelligence, Clinical validation, brain metastases, prostate cancer

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10421670/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The University Hospital Zurich has research and teaching agreements with Siemens Healthineers.

Abstract

Background

MRI-based autocontouring has the potential to improve radiotherapy planning and adaptive workflows by reducing contouring time while maintaining clinical quality. We evaluated a prototype deep learning–based MRI autocontouring tool for organ-of-interest (OOI) and target delineation in brain metastases and prostate cancer.

Methods

Fifty-seven patients (27 brain metastases and 30 prostate cancer) were retrospectively analysed. Automatically generated contours were compared with clinically approved reference contours using the Dice Similarity Coefficient (DSC) and Mean Distance to Agreement (MDA). Two radiation oncologists independently assessed clinical usability using a four-point scoring system. In addition, target detection sensitivity and false-prediction rate were evaluated for brain metastases.

Results

A total of 988 structures were generated. In the brain cohort, the prototype achieved a target detection sensitivity of 81.5% with a false-prediction rate of 20.9%. Mean DSC was 0.58 for OOI and 0.66 for gross tumour volumes, while median clinical usability was 3 (minor editing required). In the prostate cohort, mean DSC was 0.73 and median clinical usability was also 3. Performance varied substantially across structures, with large, well-defined anatomy showing the highest accuracy, whereas small or anatomically complex structures such as the cochlea, and seminal vesicles required more extensive correction. Lower DSC values were significantly associated with poorer clinical acceptability for 10 of 19 evaluated structures.

Conclusions

The prototype demonstrated promising performance as a clinician-in-the-loop contouring aid for MRI-guided radiotherapy. Although it can reduce manual contouring workload, expert review remains essential, particularly for target delineation and small or anatomically complex structures. A combined evaluation using geometric, detection, and clinical metrics provides a more comprehensive assessment than geometric measures alone.

Introduction

Accurate and reproducible target and organ-of-interest (OOI) delineation is critical for high-quality radiotherapy planning[1,2]. While Computed Tomography (CT) imaging remains the clinical standard for

treatment planning due to its established electron density information, Magnetic Resonance Imaging (MRI) offers superior soft tissue contrast that could improve delineation accuracy and reduce inter-observer variability.

The introduction of MR-guided radiotherapy (MRgRT) with hybrid MR-linacs [3,4,5] or de-coupled MR+linac[6,7] represents a major step forward in adaptive radiotherapy. These systems enable daily visualisation of anatomical changes and adaptive treatment strategies, particularly beneficial for abdominal and pelvic malignancies where inter-fraction variation is substantial. Nevertheless, a key bottleneck remains the time-consuming nature of daily contouring required for adaptive workflows [8,9].

Recent advances in machine learning and deep-learning-based auto-segmentation offer a promising solution [10,11].MRI-based autocontouring tools, when proven accurate and reliable, can streamline MRI-based treatment planning including adaptive workflows, reducing on-table time and facilitating broader MRgRT implementation. Ensuring the clinical robustness of these tools, however, requires thorough evaluation by multidisciplinary expert teams.

Here, we present the evaluation of a prototype MRI-based auto contouring tool, assessing contour quality and clinical applicability in brain metastases and prostate cancer.

Methods

A prototype MRI-based autocontouring tool based on 3D-U Net architecture developed by Siemens Healthineers was presented to our multidisciplinary team for technical and clinical evaluation. A total of 57 patients (27 brain metastases, 30 prostate cancer) were retrospectively selected. All imaging data was acquired on a 1.5T MR Scanner (MAGNETOM Sola RT Pro Edition, Siemens Healthineers, Forchheim) and were anonymized prior to analysis. This study was conducted under the ethical approval of the local ethics committee (BASEC No. 2023 – 01508). Statistical analysis was performed in R 4.5.0.

For the **brain metastases dataset**, each case included the planning CT scan, clinically approved RT structures, and contrast-enhanced T1w MPRAGE MRI with 0.45 mm voxel size in-plane and 0.9 mm slice thickness. The prototype was applied to this sequence of MRI datasets to generate automatic contours for both OOIs and gross tumour volumes (GTVs). All imaging and structure sets were imported into the MIM software platform (MIM Software Inc., Cleveland, OH). The Dice Similarity Coefficient (DSC) and Mean Distance to Agreement (MDA) between clinically acceptable contours and auto segmented structures were extracted from MIM. Planning CT scans were rigidly registered to MRI. For MRI-specific structures (pituitary, retina, cornea, hippocampus), expert contours were generated directly on MRI. Geometric agreement was evaluated for the 185 correctly matched GTV pairs. Missed clinical lesions and false-positive GTV detections were excluded from the DSC and MDA analyses because no corresponding contour was available for comparison.

The clinical usability of each auto-generated contour was independently assessed by two radiation oncologists. A 4-point scale (1 = not usable, 2 = usable after major edits, 3 = usable after minor edits, 4 =

clinically usable without edits) was used to rate the quality of each contour.

For the **prostate cancer dataset**, the planning CT scan with clinically approved RT structures was registered to both T1w VIBE Dixon (1.5 mm in-plane, 2 mm slice thickness) and T2w TSE (0.3 mm in-plane, 3 mm slice thickness) MRI sequences. To minimise registration uncertainty, bladder, rectum, and anus were contoured directly on T2 MRI.

The tool segmented femoral heads on T1 MRI, and pelvic structures (bladder, rectum, anus, penile bulb, prostate, seminal vesicles) on T2 MRI. Clinical usability scoring followed the same 4-point scale.

Lastly, contours were classified as acceptable (scores 3–4) or not acceptable (1–2). Mean DSC values between groups were compared using independent samples t-tests to assess whether geometric metrics predict clinical usability.

Results

Quantitative Performance of the Prototype

A total of 988 structures (754 OOIs, 234 targets) were generated.

Brain metastases dataset

769 structures were generated in the brain subset (535 OOIs and 234 GTVs). Of 234 targets, 185 correctly matched clinical GTVs (paired), while 42 clinical lesions were missed and 49 detections were false positives, corresponding to a sensitivity of 81.5% and a false prediction rate of 20.9% [Fig. 1].

Median clinical score was 3 [IQR 2–3] for both OOIs and GTVs. Lens and cornea performed best (median 4), while cochlea and lacrimal gland performed worst (median 2).

Mean MDA was 1.33 ± 3.03 mm (OOIs) and 0.81 ± 4.95 mm (GTVs). Retina showed the lowest MDA (0.44 mm), lacrimal gland the highest (2.33 mm).

Mean DSC was 0.58 ± 0.21 for OOIs and 0.66 ± 0.18 for GTVs. Brainstem achieved the highest DSC (0.87), cochlea the lowest (0.32). [Fig. 2]

Prostate cancer dataset

219 structures were generated. Median clinical score was 3 [IQR 3–3]. Prostate performed worst (median 2.5); all other structures had median 3.

Mean MDA was 2.72 mm. Anus had the lowest (1.65 mm), prostate the highest (3.69 mm). Mean DSC was 0.73, with femoral heads highest (0.87) and seminal vesicles lowest (0.54). [Fig. 3]

Association Between Metrics and Clinical Evaluation

As shown in the figure [Fig. 4], the mean Dice coefficient of the not-acceptable contours is significantly lower than that of the acceptable contours for 10 out of 19 structures, indicating that geometric performance correlates with clinical usability for a majority, but not all, structures.

In one instance, Anus, the mean Dice for not-acceptable contours was higher than acceptable rated contours.

Discussion

This study evaluated a prototype MRI-based autocontouring tool for the brain and pelvis. Although many contours were clinically usable, performance was structure-dependent and expert review remained necessary. Unlike many autocontouring solutions that are limited to OOI delineation, this prototype is designed to also contour brain metastases as targets, addressing a clinically critical and technically more demanding task.

In the brain metastases cohort, the prototype achieved a target detection sensitivity of 81.5%, demonstrating a solid capability to identify the majority of lesions on contrast-enhanced T1-weighted MRI acceptable for a clinician-in-the-loop workflow. This performance reflects the advantage of MRI-based input, where lesion enhancement is high and tumor–brain contrast is superior compared to CT. The observed false positive rate of 20.9% indicates that the algorithm occasionally generated spurious targets, emphasizing the fact the model needs further tuning. We speculate that performance could improve further if the prototype were applied to a patient cohort whose histological composition closely matches the data used during model training, as lesion appearance and enhancement patterns vary across tumor types; notably, brain metastases from different histologies may exhibit distinct enhancement kinetics, reaching peak conspicuity at different times after contrast administration [12]. In addition, consistency in the timing between contrast media administration and image acquisition is likely to play a critical role, as suboptimal timing may reduce lesion conspicuity and thereby impair segmentation accuracy [13]. However, such histology-specific and protocol-sensitive optimisation would be more difficult to implement in routine clinical practice, where robust tools applicable across heterogeneous patient populations and variable imaging workflows are preferable.

Moreover,in clinical applications, false positives increase review burden and introduce a risk of misinterpretation if not rigorously checked.

Quantitatively, the mean DSC for targets was 0.67 and the mean MDA was 0.81 mm. This finding is consistent with the results reported by Doshi et al.[14] and Wang et al.[15], who evaluated multiple auto-contouring tools for GTV delineation across a range of tumour sites. Evidently more accurate boundary localization for GTVs is mostly driven by the use of contrast-enhanced MRI and the more distinct intensity gradients at tumor borders. Clinically, targets achieved a median usability score of 3 [IQR 2–3], indicating that most GTV contours were suitable for use with minor edits.

Despite this encouraging performance, the detection sensitivity below 85% underscores that automated target contouring cannot be used in isolation [16]. Missed lesions in stereotactic or high-precision radiotherapy workflows carry direct clinical consequences. Therefore, expert review remains mandatory, particularly in patients with multiple small metastases or atypical lesion appearances. However, as a starting point for target delineation, the prototype substantially reduces manual workload and provides consistent initial contours that can be efficiently refined by clinicians. This is also reflected in Ziyaee et al.'s work that dives deeper into the logic behind false-positives in an auto segmentation model[17].

On the other hand, marked inter-structure variability was observed. High-contrast, anatomically well-defined structures such as the brainstem, lens, and cornea achieved strong geometric agreement and high clinical ratings. In contrast, small, low-contrast structures including the cochlea and lacrimal glands consistently performed poorly across both quantitative and clinical metrics. These findings are also in line with several previous publications indicating worse performance of auto segmentation models in smaller structures[18]. Unsurprisingly, these structures are challenging even for expert manual contouring and are particularly susceptible to limited spatial resolution, partial volume effects, and registration uncertainty [19 20].

The prototype demonstrated marked underperformance in 5 cases with respect to GTV detection. In 2 cases with disseminated brain metastases, the algorithm failed to identify 37 GTVs out of a total of 108 lesions, whereas across all remaining cases only five GTVs were missed. This suggests that leptomeningeal disease characteristics may represent a contributing factor to detection failure.

In the other three cases, the prototype substantially oversegmented, generating a total of 33 false-positive GTVs, in all other cases only 12 false positive GTVs were contoured.

In the prostate cancer cohort, geometric performance was higher (mean DSC 0.73), but most contours still required editing (median clinical score 3). Femoral heads performed best. In contrast, the prostate and seminal vesicles were consistently the weakest performers, both geometrically and clinically. These structures are inherently challenging due to shape variability, indistinct borders, and strong dependence on bladder and rectal filling[21,22]. The elevated MDA for the prostate highlights that even when DSC appears acceptable, boundary accuracy may be insufficient for high-precision planning[23].

More broadly, these findings indicate that geometric similarity metrics capture clinically relevant information only incompletely[24]. For just over half of the evaluated structures (10 of 19), unacceptable auto-contours exhibited significantly lower DSC than clinically acceptable contours, supporting the notion that poorer geometric agreement often translates into reduced clinical usability. However, the absence of this relationship in the remaining structures underscores an important limitation of relying on Dice alone. The paradoxical observation of higher mean DSC values in contours rated as not acceptable (i.e. Anus) suggests that volumetric overlap can remain high despite clinically meaningful local errors[25]. Such errors, including boundary misplacement or systematic shifts in regions of high dosimetric relevance, may have minimal impact on DSC while still rendering a contour unusable in practice. Overall,

these results reinforce that geometric metrics are necessary but not sufficient surrogates for clinical acceptability and should be interpreted in conjunction with structure-specific considerations and expert review rather than as standalone indicators of contour quality.

Taken together, the results position the prototype firmly as a contouring aid. It performs reliably for large, well-defined structures and can provide a useful starting point for many OOIs, with potential time-saving benefits. However, for small OOIs, anatomically complex pelvic targets, and lesion detection tasks, expert review and correction remain mandatory. The observed variability across structures and anatomical sites further indicates that site-specific training, optimisation, and validation are essential prerequisites for broader clinical adoption.

Finally, this study highlights the importance of a multimodal evaluation framework. Clinical usability scores, detection sensitivity and false positive rates, and spatial agreement metrics each capture distinct and complementary aspects of performance. Reliance on any single metric would have produced an incomplete and a possibly overly optimistic assessment. Future development should prioritise improved robustness for small and anatomically variable structures and reduction of false positives to better support safe and efficient clinical workflow.

These quantitative results demonstrate that the prototype autocontouring tool performs robustly for larger, clearly defined anatomical structures in both brain and pelvis. Performance decreases for smaller, low-contrast, or anatomically complex structures, and certain clinically critical targets require careful review and editing. While further developments are required if fully autonomous workflows are envisioned, the tool provides a solid foundation for automated contouring and it can be used as an assistive tool for expert delineation.

Declarations

Ethics approval and consent to participate

This retrospective study involving imaging data from human participants was approved by the **Kantonale Ethikkommission Zürich (Cantonal Ethics Committee of Zurich), Switzerland**, which operates under the Gesundheitsdirektion (Health Directorate) of the Canton of Zurich. The approval number is BASEC No. 2023-01508. The study was conducted in accordance with the ethical principles of the **Declaration of Helsinki** and applicable Swiss regulations. In accordance with the decision of the ethics committee and Swiss national regulations, the requirement for obtaining individual informed consent to participate was waived.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are not publicly available due to institutional and patient privacy restrictions but are available from the corresponding author on reasonable request and with permission of the University Hospital Zurich.

Competing interests

The University Hospital Zurich has research and teaching agreements with Siemens Healthineers.VNJ and ASE are employees of Siemens Healthineers. The authors declare that they have no other competing interests.

Funding

This work was supported by Siemens Healthineers. Employees of the funding organization contributed to the development of the prototype software, technical support, manuscript review, and editing as described in the Authors' contributions. The funding body had no role in the clinical evaluation, interpretation of the results, or the decision to submit the manuscript for publication.

Authors' contributions

NR conceived the analysis, performed the investigation, curated the data, conducted the formal analysis, prepared the visualizations, and wrote the original draft of the manuscript. **STL** contributed to the study conceptualization and methodology, acquired funding, validated the results, and critically reviewed and edited the manuscript. **VNJ** developed and supported the prototype software and reviewed and edited the manuscript. **ASE** developed and supported the prototype software and reviewed and edited the manuscript. **CvW** contributed to the investigation and reviewed and edited the manuscript. **MG** supervised the study, acquired funding, and reviewed and edited the manuscript. **NA** supervised the study, acquired funding, and reviewed and edited the manuscript. **RDB** conceived the study, developed the methodology, supervised the project, validated the results, administered the project, and contributed to writing the original draft as well as reviewing and editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Jin D, Guo D, Ge J, Ye X, Lu L. Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era. J Natl Cancer Cent. 2022;2:306–13.
2. Mir R, et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. Radiother Oncol. 2020;150:30–9.
3. Ng J, et al. MRI-LINAC: A transformative technology in radiation oncology. Front Oncol. 2023;13:1117874.

4. Seravalli E, et al. The potential role of MR-guided adaptive radiotherapy in pediatric oncology: Results from a SIOPE-COG survey. Clin Transl Radiat Oncol. 2021;29:71–8.
5. de Mol SR, et al. The MOMENTUM study: An international registry for the evidence-based introduction of MR-guided adaptive therapy. Front Oncol. 2020;10:1328.
6. Wilke L, et al. Safety in MR-enhanced daily adaptive SBRT Radiotherapy using a conventional C-arm linear accelerator: An FMEA approach. Z Med Phys. 2025;35:423–7.
7. Kim J-Y, et al. Clinical workflow of cone beam computer tomography-based daily online adaptive radiotherapy with offline magnetic resonance guidance: The modular adaptive radiotherapy system (MARS). Cancers (Basel). 2024;16:1210.
8. Keall PJ, et al. ICRU REPORT 97: MRI-guided radiation therapy using MRI-linear accelerators. J ICRU. 2022;22:1–100.
9. Votta C, et al. Evaluation of clinical parallel workflow in online adaptive MR-guided Radiotherapy: A detailed assessment of treatment session times. Tech Innov Patient Support Radiat Oncol. 2024;29:100239.
10. Eidex Z, et al. Deep learning in MRI-guided radiation therapy: A systematic review. J Appl Clin Med Phys. 2024;25:e14155.
11. Harrison K, et al. Machine learning for auto-segmentation in radiotherapy planning. Clin Oncol (R Coll Radiol). 2022;34:74–88.
12. Sundaralingam R, Abubacker Sulaiman F, Praveenkumar R. MRI Findings in Brain Metastasis. *ECR* 2024 EPOS https://epos.myesr.org/poster/esr/ecr2024/C-23393 (2024).
13. Seo J, Lim C, Lee KY, Koh Y-C, Moon W-J. Time optimization of gadobutrol-enhanced brain MRI for metastases and primary tumors using a dynamic contrast-enhanced imaging. BMC Med Imaging. 2022;22:180.
14. Doshi T, et al. Validation of a magnetic resonance imaging-based auto-contouring software tool for gross tumour delineation in head and neck cancer radiotherapy planning. Clin Oncol (R Coll Radiol). 2017;29:60–7.
15. Wang T, et al. Evaluation of AI-based auto-contouring tools in radiotherapy: A single-institution study. J Appl Clin Med Phys. 2025;26:e14620.
16. Hoebers FJP, et al. Artificial intelligence research in radiation oncology: a practical guide for the clinician on concepts and methods. BJR Open. 2024;6:tzae039.
17. Ziyaee H, et al. Automated brain metastases segmentation with a deep dive into false-positive detection. Adv Radiat Oncol. 2023;8:101085.
18. Andrearczyk V, et al. Automatic detection and multi-component segmentation of brain metastases in longitudinal MRI. Sci Rep. 2024;14:31603.
19. Guzene L, et al. Assessing interobserver variability in the delineation of structures in radiation oncology: A systematic review. Int J Radiat Oncol Biol Phys. 2023;115:1047–60.

20. Chen M, et al. Interobserver variability in organ delineation on radiotherapy treatment planning for nasopharyngeal carcinoma: A dosimetric and prognostic analysis. Front Oncol. 2025;15:1510568.
21. Deurloo KEI, et al. Quantification of shape variation of prostate and seminal vesicles during external beam radiotherapy. Int J Radiat Oncol Biol Phys. 2005;61:228–38.
22. Pinkawa M, et al. Prostate position variability and dose-volume histograms in radiotherapy for prostate cancer with full and empty bladder. Int J Radiat Oncol Biol Phys. 2006;64:856–61.
23. Sherer MV, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol. 2021;160:185–91.
24. Baroudi H, et al. Automated contouring and planning in radiation therapy: What is 'clinically acceptable'? Diagnostics (Basel). 2023;13:667.
25. Vaassen F, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol. 2020;13:1–6.

Figures

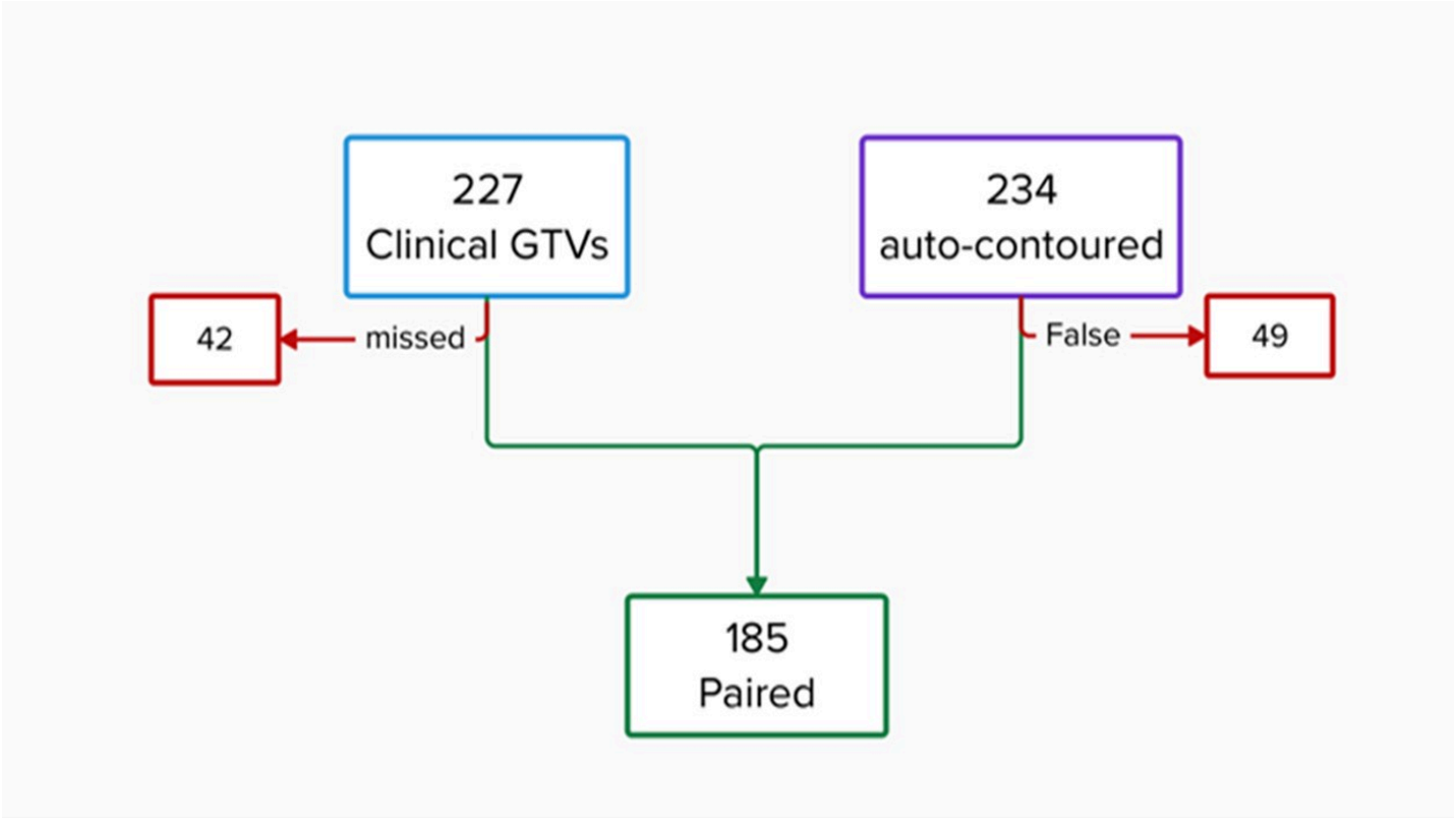


Figure 1

Distribution of contoured GTVs

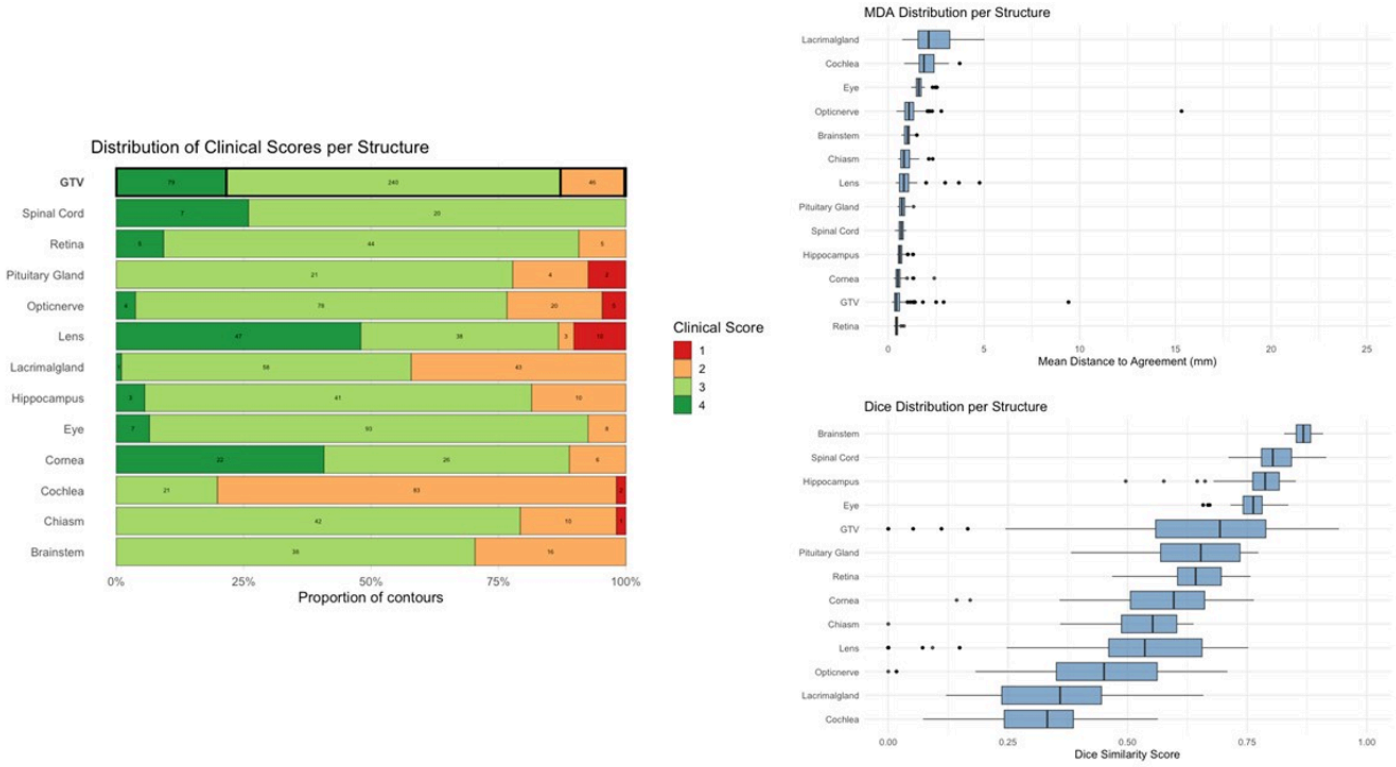


Figure 2

Clinical score, Mean Distance to Agreement (MDA) and Dice score results for the brain subset

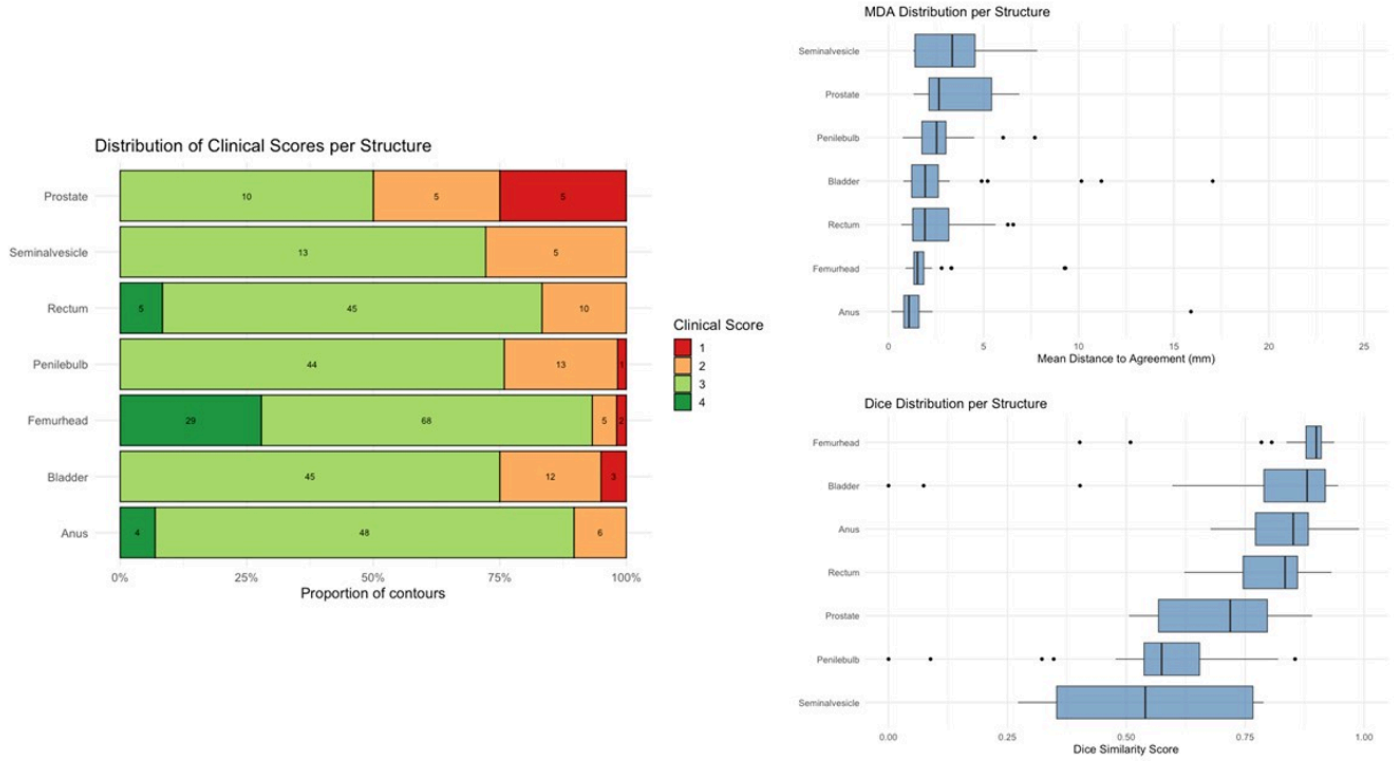


Figure 3

Clinical score, Mean Distance to Agreement (MDA) and Dice score results for the pelvis subset

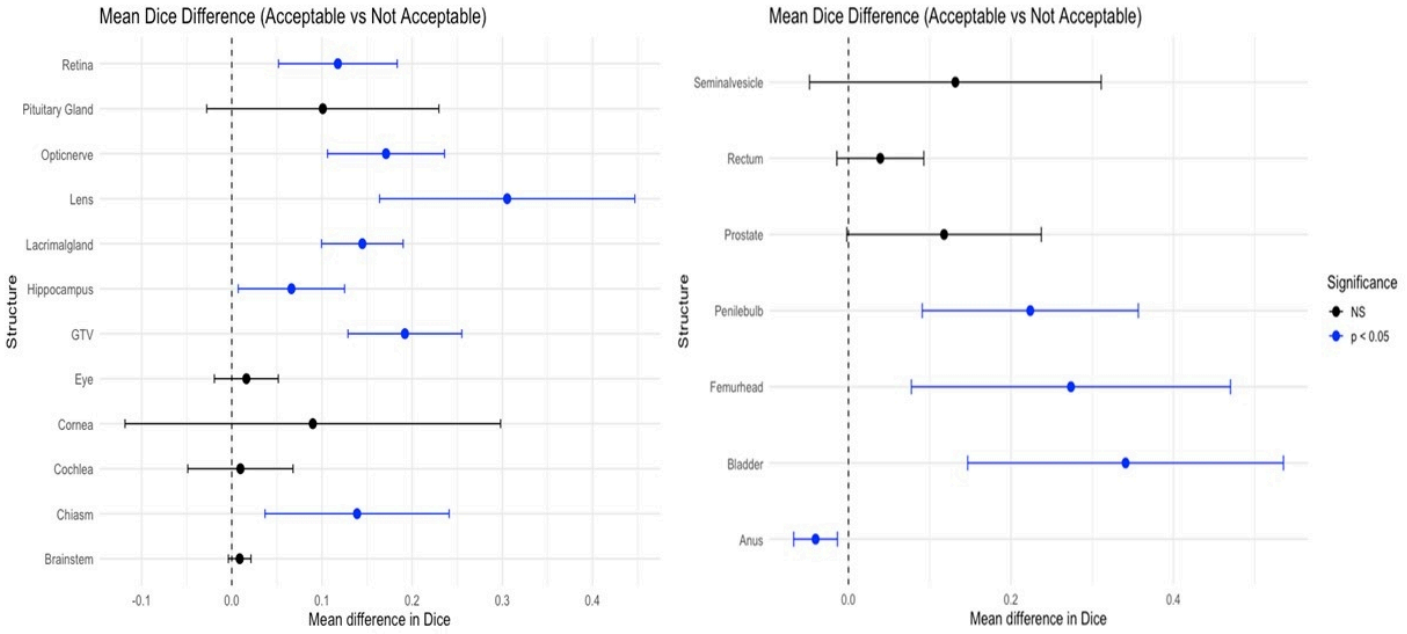


Figure 4

Mean difference in Dice score of Acceptable (clinical score= 3, 4) versus Not-acceptable (clinical score= 1, 2) auto-segmented structures