

THE ROLE OF MULTIDISCIPLINARY EXPERTISE IN AI TRAINING DATA: QUANTIFYING VARIABILITY IN PROSTATE AND URETHRAL MRI SEGMENTATION

Szilvia Totin^{1,*}, Andras Negyessy^{2,*}, William Holmlund³, Attila Simko³, Istvan Elod Kiraly², Adam Hodoniczki², Benedek Danko¹, Eniko Koos¹, Viktor Rogowski⁴, Christian Jamtheim Gustafsson⁵, Miklos Zrinyi⁶, Angelika Szatmari⁶, Peter Palasti¹, Linda Varga⁷, Tufve Nyholm³, Zoltan Bajory² and Zsuzsanna Fejes^{1,**}

¹University of Szeged – Department of Radiology
Szeged, Hungary

²University of Szeged – Department of Urology
Szeged, Hungary

³Umeå University – Department of Diagnostics and Intervention
Umeå, Sweden

⁴Skåne University Hospital – Department of Hematology, Oncology, and Radiation Physics
Lund, Sweden

⁵Lund University – Department of Translational Medicine, Medical Radiation Physics
Malmö, Sweden

⁶University of Szeged Faculty of Health Sciences and Social Studies, Department of Nursing
Szeged, Hungary

⁷University of Szeged – Department of Oncotherapy
Szeged, Hungary

DOI: [10.7906/indexs.24.6.9](https://doi.org/10.7906/indexs.24.6.9)
Regular article

Received: 16 June 2026.
Accepted: 24 July 2026.

ABSTRACT

Introduction: Accurate segmentation of prostate mpMRI scans is essential for oncological therapy planning and the development of AI-based decision support systems. The aim of our study was to validate the annotation reliability of observer groups with different levels of professional experience in prostate and urethral segmentation to establish an expert AI database.

Methods: A total of 35 mpMRI cases were processed using 3D Slicer software. Seven participants were categorized in three groups: urologists, radiology and urology residents, and medical students. Interobserver variability was quantified using Dice Similarity Coefficient (DSC), Surface Dice (at 1 mm and 3 mm tolerances), and Central Line Distance (CLD) metrics, with validation performed via Dunn's post-hoc test.

Results: Prostate DSC values were statistically equivalent between urologists, urologists, and residents ($p_{holm} = 1,000$; $z = 0,543$). Midline determination of the urethra (CLD) showed

*both authors should be considered as first authors owing to their overall contribution to the article

**Corresponding author, *η*: fejes.zsuzsanna.02@szte.hu; -;

University of Szeged, Department of Radiology, Semmelweis Street 6, H-6725 Szeged, Hungary

specialty-specific differences between radiologists and urologists ($p = 0,039$), with an average difference of 3,3-3,7 mm (with min 2,9 mm; max 4,4 mm). While the 1 mm surface precision of medical students was significantly lower than that of experts ($z = 4,639$; $p_{\text{holm}} < 0,001$), no significant difference was observed within the 3 mm clinical tolerance margin. Notably, certain student subgroups reached expert-level performance even in urethra segmentation ($p_{\text{holm}} = 1,000$).

Conclusion: Residents and uroradiologists from related fields generate expert-level annotation. The involvement of trained non-experts (students) represents a valid crowdsourcing alternative, provided that quality assurance is aligned with the 3 mm clinical tolerance threshold.

KEY WORDS

segmentation, interobserver variability, artificial intelligence, prostate mpMRI, 3D slicer

CLASSIFICATION

JEL: E59

INTRODUCTION

Prostate cancer remains one of the most significant global health challenges, being the second most common malignancy among men worldwide, with an incidence rate steadily increasing as the population ages [1]. The success of modern oncological therapies, particularly precision radiotherapy and focal treatments, is fundamentally dependent on the accurate anatomical localization of the prostate and surrounding critical structures [2].

In diagnostic imaging, multiparametric magnetic resonance imaging (mpMRI) has become the primary modality due to its superior soft-tissue contrast resolution, enabling more precise segmentation than other imaging techniques [3]. However, in clinical practice, organ contouring is still predominantly a manual process, which is both time-consuming and highly subjective. Segmentation is further complicated by the anatomical fact that the prostate lacks a true capsule and is separated from surrounding tissues by loose connective tissue [4]. A major limitation of manual segmentation is inter-observer variability, arising from differing anatomical interpretations among urologists [5]. Studies confirm that contouring inaccuracies directly impact dose planning and the protection of adjacent healthy tissues [6].

When analysing inter-observer variability, the discrepancy between different medical specialties – specifically radiological and urological perspectives – deserves particular attention. While radiologists primarily define contours based on tissue signal intensity variations and radiological diagnostic guidelines (e.g., PI-RADS), urologists' interpretations are significantly influenced by surgical anatomical knowledge and the visualization of tissue planes experienced during radical prostatectomy [7]. Literature indicates that even between experienced radiologists and radiation oncologists, significant differences can occur in defining Gross Tumour Volumes (GTV), directly affecting dose distribution accuracy. This multidisciplinary variability is a critical factor in the segmentation of the urethra and the prostatic apex, where precise boundary definition is crucial for both surgical margins and radiotherapy dose distribution [8]. Quantifying the variability between these two specialties is therefore essential for creating an adequate database that covers all aspects of clinical practice.

The segmentation of the prostatic part of the urethra represents a particularly critical point, as its accurate delineation is essential for preserving urinary continence [9]. Since the urethra is often difficult to distinguish on MRI, segmentation accuracy depends heavily on the investigator's clinical experience and anatomical knowledge. The application of artificial intelligence, specifically Convolutional Neural Networks (CNN), promises a breakthrough in automated segmentation by eliminating human subjectivity [10]. However, developing reliable algorithms requires large-scale, expert-validated training databases [11].

The aim of the present study was to create such a segmented database using the 3D Slicer (v.5.8.1) software [12], while analysing the impact of professional experience and clinical background (radiology vs. urology) on contouring accuracy.

MATERIALS AND METHODS

DATASET AND IMAGING MATERIAL

For this study, 35 cases were randomly selected and processed from the publicly available Annotated Image Dataset of Anatomy database. The dataset provides high-resolution, T2-weighted axial magnetic resonance (MR) images in DICOM format, featuring detailed voxel-level segmentations of pelvic organs. This database was selected for its rigorous acquisition protocols, consistent image quality, and the diverse range of anatomical variability essential for robust analysis.

To assess inter-observer variability, a representative subset of 35 cases was selected from the total pool. These 35 cases were independently segmented by every participant across all experience groups (uroradiologists, residents and medical students), providing a common reference for statistical comparison. The remaining cases were distributed among the investigators to ensure broad coverage of anatomical variations for the overall construction of the Artificial Intelligence database.

Image data processing and structure delineation were performed using the 3D Slicer (version 5.8.1) open-source medical image analysis platform. The Segment Editor module was used for segmentation. During the process, all investigators followed an identical protocol, manually contouring the target areas slice-by-slice. First, the outer contour of the prostate gland was defined, followed by the segmentation of the prostatic urethra (pars prostatica urethrae). To ensure continuity, investigators could use the 3D visualization window for spatial verification of the contours.

The nine participants performing the segmentation were divided into three groups based on professional experience to measure inter-observer variability:

- Group 1 – Two consultant uroradiologists (3 years of post-graduate clinical experience specializing in prostate MRI interpretation).
- Group 2 – Two radiology residents and two urology residents (with 1-2 years of clinical and diagnostic imaging experience).
- Group 3 – Three medical students who received brief, standardized training on prostate MR anatomy, imaging of prostate diseases, and software usage prior to the study.

The database samples were distributed to ensure significant overlap between groups, allowing for a multi-sided comparison of the same anatomical structures.

STATISTICAL ANALYSIS AND VALIDATION

The agreement between segmentations and the extent of deviations were quantified using objective metrics. The first metric, the DSC, is a statistical measure used to gauge the similarity between two sets. The coefficient ranges from 0 to 1, where 0 represents no overlap and 1 indicates perfect agreement. The second metric was the Surface Dice Measurement (SD), which assesses the accuracy of the boundaries within a predefined tolerance threshold. In this study, tolerance thresholds were set at 1 mm, 3 mm. The 3 mm threshold was specifically chosen as it aligns with the clinically accepted “Planning Target Volume” margins used in most modern image-guided radiotherapy protocols, thereby defining the limit of “clinically acceptable” error. The score represents the percentage of the surface area that falls within the allowed error margin. Similar to DSC, its value ranges from 0 to 1, where a value of 1 indicates that the entire estimated surface is within the specified distance from the reference. For the urethra, the CLD was used to examine the average deviation from the midline. CLD is a geometric and statistical metric primarily utilized to evaluate the segmentation of tubular or branching structures (e.g., vessels, airways). To determine statistical differences between groups, Welch’s t-test and Dunn’s post hoc test (for non-parametric comparisons) were applied, with a p-value smaller than 0,05 considered statistically significant. To account for multiple comparisons, p-values were adjusted using the Holm-Bonferroni method (p_{holm}). Group differences were evaluated using these tests and the corresponding z-scores are reported to indicate the distance from the mean in standard deviation units.

RESULTS

Statistical analysis of the Dice coefficients, representing volumetric overlap, examined segmentation accuracy across groups with different levels of experience and clinical backgrounds. The results highlight a strong correlation between morphological delineation and

clinical expertise. Specialty-Specific Consensus in Urological and Radiological Practice. A key finding of the study is that the comparative analysis between consultant radiologists and urologists showed no significant difference ($p = 0,629$). This confirms that representatives of both specialties apply a unified nomenclature and visual interpretation in the volumetric definition of the GTV. Earlier CLD measurements reflected conceptual differences regarding the urethral pathway; however, the homogeneity of DSC values confirms a consistent recognition of the global volumetric extent. Regarding the second progression of Clinical Competence (uroradiologist vs. resident), the comparison between the resident group and the consultant radiologist control group showed a statistically non-significant difference ($p = 0,844$). This high degree of concordance validates the involvement of residents in large-scale annotation workflows, as their segmentation accuracy reaches the expert gold standard. This is critical for the scalability of validated datasets required for artificial intelligence development. Heterogeneity of Student Segmentation and the Learning Curve Analysis of the student subgroups revealed significant inter-observer variability due to a lack of clinical routine: The Stud 1-2 and Stud 1-3 groups produced significantly lower overlap metrics compared to the expert groups ($p < 0,001$), indicating uncertain detection of anatomical boundaries. The outstanding performance of the Stud 2-3 subgroup (compared to RAD: ($p = 0,062$) suggests that with targeted instructions and adequate use of the 3D Slicer platform, non-expert observers can achieve accuracy approaching expert levels.

Surface Dice (1 mm Tolerance) Surface Dice measurements confirmed that precise boundary delineation is the most sensitive indicator of clinical routine. Expert precision: no significant difference in surface agreement was found between radiologists, residents, and urologists ($p > 0,05$). This demonstrates that clinical experience ensures consistent contouring discipline and 1 mm-accurate recognition of interfaces (e.g., prostate apex), regardless of specialty. Student limitations: Most students (Stud 1-2, Stud 1-3) significantly underperformed compared to all expert groups ($p < 0,001$) or ($p = 0,002$). High effect sizes (rank-biserial correlation equals 0,827) highlight that without clinical routine, manual segmentation accuracy falls short of clinical expectations. Teachability: The Stud 2-3 group's results ($p = 0,137$) suggest that targeted training can significantly improve non-expert accuracy.

In the measurement of Surface Dice (3 mm tolerance), the results of the Dunn's post-hoc analysis exploring inter-group segmentation variability confirmed the decisive role of clinical experience, while simultaneously highlighting a high degree of consistency across expert levels. Statistical analysis revealed no significant difference between the performance of board-certified radiologists and residents ($z = 0,543$; $p_{\text{holm}} = 1,000$), validating the involvement of residents in high-volume annotation workflows. A particularly significant finding is that no justifiable discrepancy was found between the contouring accuracy of radiologists and urologists ($z = 0,322$; $p_{\text{holm}} = 1,000$); this interprofessional concordance proves that representatives of these two specialties apply a uniform visual interpretation when defining prostate boundaries, despite their differing clinical perspectives.

In contrast, the performance of the expert radiologist group significantly and substantially outperformed the accuracy of the majority of medical students (Rad-Stud 1-2: $z = 4,639$, $p_{\text{holm}} < 0,001$; Rad-Stud 1-3: $z = 5,694$, $p_{\text{holm}} < 0,001$), clearly indicating the importance of clinical routine in precise segmentation. However, the heterogeneity within the student group is a noteworthy observation: while most subgroups lagged significantly behind the expert control, the results of the Stud 2-3 group showed no statistically justifiable difference from those of the board-certified radiologists ($z = -1,008$, $p_{\text{holm}} = 1,000$). This phenomenon suggests that with targeted training, a subset of students is capable of approaching expert-level performance; nevertheless, due to the high variability observed among them, rigorous and continuous uroradiologist supervision remains essential for annotations performed by medical students.

CLD: The statistical analysis of the prostatic urethral midline (Dunn's *post hoc* test) revealed significant discrepancies. Radiologists vs. Urologists: A significant difference was observed between the midlines defined by radiologists and urologists ($z = -2.833$; $p = 0,005$; $p_{\text{holm}} = 0,039$), proving that clinical background (radiological signal intensity vs. surgical anatomy) substantially influences urethral localisation. Uroradiologists vs. Residents: No significant difference was found ($p_{\text{holm}} = 0,793$), suggesting that diagnostic skills acquired early in residency are sufficient for expert-level CLD contouring. Systematic Deviations: The largest systematic differences occurred between expert groups and the majority of students. The significant gap between urologists and a subset of students ($z = -4.263$; $p < 0,001$) further emphasizes that urethral localization follows two distinct "interpretive schools" even among experienced observers.

DISCUSSION

Our study demonstrated that inter-observer variability in the manual segmentation of the prostate and urethra can be maintained at a low level, regardless of experience, by implementing a standardized protocol, Table 1. The significant 3,3-3,7 mm discrepancy in urethral CLD between uroradiologists and urologists highlights a fundamental divergence in anatomical interpretation rather than a simple error. While uroradiologists adhere to the low-signal-intensity tubular structure visible on T2-weighted images, urologists may subconsciously bias their contours based on the expected functional position and surgical experience of the prostatic bed. For AI development, this suggests that training sets must be

Table 1. Interobserver variability and clinical significance of prostate and urethral segmentation across different levels of expertise (NS – non-significant, p_{holm} indicates p-values adjusted using the Holm-Bonferroni method to account for multiple comparisons, z represents the test statistic from Dunn's post-hoc analysis, Rad – Board-certified uroradiologists (with more than 3 years post-graduate experience), Urol – Urology residents, Res – Radiology residents, and Stud – Medical students).

Parameter	Statistical Result (Dunn's Post Hoc)	Clinical/Practical Significance	Explanation and Conclusion
DSC	Rad-Urol $p = 0,629$ (NS)	High volumetric agreement	Specialists (radiologists/urologists) consistently judge the overall extent of the prostate.
CLD	Rad-Urol $p = 0,039^*$	Professional bias phenomenon	The two specialties localize the urethral axis differently (3,3-3,7 mm deviation), but this remains within radiotherapy margins.
Surface Dice (1 mm)	Rad-Res $p = 0,949$ (NS)	Expert-level precision	Residents are capable of specialist-level contouring within 1 mm accuracy, making them suitable for generating AI training data.
DSC	Stud-Res $p > 0,05$ (NS)	Validity of "Crowdsourcing"	Students (non-experts) are able to identify the main anatomical area, allowing their involvement in high-data-demand research.
Surface Dice (1 mm)	Rad-Stud 1-2 $p < 0,001^{***}$	Contouring uncertainty	Although students identify volume well, they handle 1 mm boundaries (e.g., apex) with uncertainty; expert supervision is required.
Surface Dice (3 mm)	Rad-Stud 2-3 ($p_{\text{holm}} = 1,000$)	Learning curve and affinity	Following focused training, certain student subgroups are capable of expert-level segmentation even in the case of structures with low visibility (e.g., the urethra).

*significant at the level $p < 0,05$ (Holm-corrected p-values)

***significant at the level $p < 0,001$ (Holm-corrected p-values)

labelled with the specific clinical intent (e.g., surgical planning versus radiotherapeutic sparing) to avoid confusing the neural network with conflicting “expert truths”.

Furthermore, our results regarding the “top-tier” student group (Student 2-3) are highly encouraging for crowdsourcing efforts. The finding that their 3 mm Surface Dice scores reached statistical equivalence with the expert group suggests that high-level anatomical affinity and focused training can compensate for the lack of clinical years. This holds true provided the required precision is defined by clinical margins – such as those used in radiotherapy planning – rather than absolute geometric identity. However, the high variability in other student groups emphasizes that AI databases cannot rely on unvetted student work; a multi-tier validation system is required where students provide the initial pre-segmentation, which is then refined by residents and finally approved by consultants.

Particular attention was paid to the discrepancies between radiologist and urologists. Although the two specialties approach the target volume with different anatomical perspectives – radiologists focusing on signal intensity variations and urologists on surgical planes – the DSC and Surface Dice (1 mm) measurements showed no statistically significant differences between the two groups ($p = 0,629$ and $p = 0,330$, respectively). This indicates that T2-weighted MRI images provide an objective basis for multidisciplinary collaboration, and boundary delineation remains consistent at the expert level. The significant difference measured in the determination of the urethral midline (CLD, $p = 0,039$) confirms the presence of specialty-specific bias. Nevertheless, this average deviation of 3,3-3,7 mm remain clinically manageable, as the safety margins typically applied in radiotherapy fall within this range [13].

CROWDSOURCING AND THE INVOLVEMENT OF NON-EXPERTS IN ARTIFICIAL INTELLIGENCE DEVELOPMENT

A key finding of our study is that there was no statistically significant difference in DSC scores between medical students and radiology residents during the volumetric segmentation of the prostate gland. This aligns with international literature advocating for crowdsourcing or the involvement of trained non-experts in building large-scale datasets for AI [11]. However, the Surface Dice (1 mm tolerance) analysis revealed that most students significantly trailed expert precision ($p < 0,001$). This suggests that while non-experts identify the primary organ mass effectively, precise delineation of fine contour boundaries still requires clinical experience. Interestingly, although some students approached the expert gold standard, their intra-group agreement was lower than that of the experts. This variability reflects differences in individual segmentation styles among novice observers; however, when combined, these annotations may still produce accurate anatomical boundaries [14].

THE ROLE OF SOFTWARE AND URETHRAL SEGMENTATION

The lack of clinical experience was effectively mitigated by the intuitive interface and high-resolution visualization capabilities of the 3D Slicer software environment, which facilitated the practical application of theoretical anatomical knowledge. The focused task-performance of the students – in contrast to the complex diagnostic load of uroradiologists – may have enhanced manual precision along well-defined anatomical boundaries. The segmentation of the urethra is particularly critical, as this structure is often difficult to distinguish on MRI. Despite the statistical uncertainty among students compared to experts at high-precision levels, their measured error margin remained within the 3 millimetre clinical tolerance. This indicates that this portion of our database is suitable for training CNN, as automated segmentation can subsequently eliminate the individual subjectivity experienced during manual contouring [15], Tables 2-4.

LIMITATIONS OF THE STUDY

When interpreting the results of our study, several limiting factors should be considered:

Reliance on a Single Sequence: The research relied exclusively on T2-weighted axial MR images. Although this is the primary sequence for segmentation in clinical practice, the absence of other multiparametric MRI (mpMRI) components (e.g., DWI, ADC maps) may have limited the precise delineation of tumour infiltration, potentially influencing contouring decisions.

Software Specificity: Segmentation was performed solely within the 3D Slicer environment. While the software provides high-resolution visualization, results – particularly the Surface Dice, which reflects manual precision – may differ when using other commercial oncological dose-planning or radiological diagnostic software.

Sample Size and Training: Although some individuals among the students demonstrated expert-level performance, the relatively small sample size of the student groups and the brevity of the standardized training provided do not allow for a steady-depth analysis of the individual learning curve. The “Ground Truth”

Challenge: As in all studies investigating inter-observer variability, defining an absolute “gold standard” remains challenging. Although consultant radiologists’ contours were used as the reference, the significant difference between specialties (in the case of CLD) highlights that the definition of “perfect” segmentation may remain clinically subjective.

Table 2. Proposed roles and validation requirements in AI database development based on user proficiency.

User Level	Task Type	Recommended Role in AI Development	Validation Requirement
Uroradiologist/ Urologist	Complex cases, Urethral CLD	Establishing the Ground Truth (Gold Standard)	–
Radiology Resident	Full-scale segmentation	Independent Annotator (Expert Level)	Random spot- checks
Medical Student (Top tier - e.g., Student 2-3)	Volumetric DSC, 3mm Surface	Mass data generation (Crowdsourcing)	Continuous monitoring
Medical Student (Novice)	Pre-segmentation	Sketching initial contours	Full expert review

Table 3. Synthesis of participant competencies and technical error sources for training prostate MRI-based algorithms.

Anatomical Region	Typical Error Type	Most Likely Cause	Recommended Correction
Prostate Apex	Under- or over-segmentation	Blurred boundaries; transition toward the pelvic musculature	3D visualization and coronal plane verification
Urethra (CLD)	Midline displacement	Divergence in professional perspective (surgical vs. radiological)	Centroid-based automated correction
Prostatic Capsule Margin	Deviation exceeding 1 mm	Manual tremor; low level of experience	Software-based smoothing algorithms

Table 4. Detailed post-hoc results of all inter-group comparisons. The table presents the results of the Dunn's post-hoc test, including z-scores and Holm-corrected p-values (p_{holm} indicates p-values adjusted using the Holm-Bonferroni method to account for multiple comparisons, z represents the test statistic from Dunn's post-hoc analysis, Rad – Board-certified urologists (>3 years post-graduate experience), Urol – Urology residents, Res – Radiology residents, and Stud – Medical students).

Comparison	DSC	Surface Dice (1 mm tolerance)	Surface Dice (3 mm tolerance)	Urethral Midline (CLD)
	$z \mid p_{holm}$	$z \mid p_{holm}$	$z \mid p_{holm}$	$z \mid p_{holm}$
Rad-Res	0,197 1,000	-0,064 1,000	0,543 1,000	-1,097 0,793
Rad-Urol	0,483 1,000	-1,839 0,330	0,322 1,000	-2,833 0,039*
Rad-Stud 2-3	-1,867 0,371	-2,336 0,137	-1,008 1,000	1,430 0,611
Urol-Res	-0,679 1,000	-1,775 0,330	-0,221 1,000	-1,737 0,412
Urol-Stud 2-3	1,385 0,831	0,497 1,000	1,329 1,000	-4,263 < 0,001***
Res-Stud 2-3	-2,064 0,273	-2,272 0,139	-1,551 0,847	2,527 0,081
Rad-Stud 1-2	4,169 < 0,001***	3,623 0,002**	4,639 < 0,001***	-4,750 < 0,001***
Rad-Stud 3-1	5,416 < 0,001***	5,166 < 0,001***	5,694 < 0,001***	-3,950 < 0,001***

*significant at the level $p < 0,05$ (Holm-corrected p-values)

**significant at the level $p < 0,01$ (Holm-corrected p-values)

***significant at the level $p < 0,001$ (Holm-corrected p-values)

CONCLUSION

Our study confirms that professional background and experience levels significantly influence the precision of prostate and urethral segmentation, yet these variations remain within clinically acceptable limits when standardized protocols are applied. The identified 3,3-3,7 mm specialty-specific bias between radiologists and urologists in urethral localization highlights the necessity for multi-disciplinary consensus in defining “ground truth” for AI training.

Furthermore, we demonstrated that trained medical students could achieve expert-level volumetric segmentation accuracy within a 3 mm clinical tolerance threshold. This validates the feasibility of crowdsourcing as a reliable strategy for large-scale AI database construction, provided that rigorous quality control and multi-tier validation by residents and consultants are maintained. Our findings suggest that while high-precision boundary delineation still requires clinical expertise, the involvement of motivated non-experts offers a scalable solution to the data bottleneck in medical AI development.

REFERENCES

- [1] Sung, H., et al.: *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. CA: A Cancer Journal for Clinicians **71**(3), 209-249, 2021, <http://dx.doi.org/10.3322/caac.21660>,
- [2] Nikitas, J., et al.: *The interplay between acute and late toxicity among patients receiving prostate radiotherapy: an individual patient data meta-analysis of six randomised trials*. The Lancet Oncology **26**(3), 378-386, 2025, [http://dx.doi.org/10.1016/S1470-2045\(24\)00720-4](http://dx.doi.org/10.1016/S1470-2045(24)00720-4),

- [3] Turkbey, B., et al.: *Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2*. European Urology **76**(3), 340-351, 2019, <http://dx.doi.org/10.1016/j.eururo.2019.02.033>,
- [4] Magi-Galluzzi, C., et al.: *International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 3: extraprostatic extension, lymphovascular invasion and locally advanced disease*. Modern Pathology **24**(1), 26-38, 2011, <http://dx.doi.org/10.1038/modpathol.2010.158>,
- [5] Patrick, H.M.; Souhami, L. and Kildea, J.: *Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention*. Acta Oncologica **60**(2), 229-236, 2021, <http://dx.doi.org/10.1080/0284186X.2020.1825801>,
- [6] Zhong, A.Y., et al.: *Clinical Impact of Contouring Variability for Prostate Cancer Tumor Boost*. International Journal of Radiation Oncology • Biology • Physics **120**(4), 1024-1031, 2024, <http://dx.doi.org/10.1016/j.ijrobp.2024.06.007>,
- [7] Nensa, F.: *The future of radiology: The path towards multimodal AI and superdiagnostics*. European Journal of Radiology Artificial Intelligence **2**, No. 100014, 2025, <http://dx.doi.org/10.1016/j.ejrai.2025.100014>,
- [8] Shore, N.D.; Morgans, A.K.; El-Haddad, G.; Srinivas, S. and Abramowitz, M.: *Addressing Challenges and Controversies in the Management of Prostate Cancer with Multidisciplinary Teams*. Targeted Oncology **17**(6), 709-725, 2022, <http://dx.doi.org/10.1007/s11523-022-00925-7>,
- [9] Pham, J., et al.: *Evaluation of T2-Weighted MRI for Visualization and Sparing of Urethra with MR-Guided Radiation Therapy (MRgRT) On-Board MRI*. Cancers **13**(14), No. 3564, 2021, <http://dx.doi.org/10.3390/cancers13143564>,
- [10] Patel, H.D., et al.: *Variability in prostate cancer detection among radiologists and urologists using MRI fusion biopsy*. BJUI Compass **5**(2), 304-312, 2024, <http://dx.doi.org/10.1002/bco2.294>,
- [11] Maier-Hein, L., et al.: *Metrics reloaded: recommendations for image analysis validation*. Nature Methods **21**, 195-212, 2024, <http://dx.doi.org/10.1038/s41592-023-02151-z>,
- [12] Fedorov, A., et al.: *3D Slicer as an image computing platform for the Quantitative Imaging Network*. Magnetic Resonance Imaging **30**(9), 1323-1341, 2012, <http://dx.doi.org/10.1016/j.mri.2012.05.001>,
- [13] Navran, A., et al. *The impact of margin reduction on outcome and toxicity in head and neck cancer patients treated with image-guided volumetric modulated arc therapy (VMAT)*. Radiotherapy & Oncology **130**, 25-31, 2018, <http://dx.doi.org/10.1016/j.radonc.2018.06.032>,
- [14] Warfield, S.K.; Zou, K.H. and Wells, W.M.: *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation*. IEEE Transactions on Medical Imaging **23**(7), 903-921, 2004, <http://dx.doi.org/10.1109/TMI.2004.828354>,
- [15] Ronneberger, O.; Fischer, P. and Brox, T.: *U-Net: Convolutional Networks for Biomedical Image Segmentation*. In: Navab, N.; Hornegger, J.; Wells, W. and Frangi, A., eds.: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, Cham, pp.234-241, 2015, http://dx.doi.org/10.1007/978-3-319-24574-4_28.