Radiotherapy and Oncology 214 (2026) 111231

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Urethra contours on MRI: Multidisciplinary consensus educational atlas and reference standard for artificial intelligence benchmarking

Yuze Song [a,b], Lily Nguyen [a,c], Anna M. Dornisch [a], Madison T. Baxter [a], Tristan Barrett [d], Anders M. Dale [e], Robert T. Dess [f], Mukesh Harisinghani [g], Sophia C. Kamran [h], Michael A. Liss [i], Daniel J.A. Margolis [j], Eric P. Weinberg [k], Sean A. Woolen [l], Tyler M. Seibert [a,e,i,m,*]

[a] Department of Radiation Medicine and Applied Sciences, University of California San Diego, La Jolla, CA, USA
[b] Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA
[c] University of California San Diego School of Medicine, La Jolla, CA, USA
[d] Department of Radiology, University of Cambridge, Cambridge, United Kingdom
[e] Department of Radiology, University of California San Diego, La Jolla, CA, USA
[f] Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
[g] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[h] Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[i] Department of Urology, University of California San Diego, La Jolla, CA, USA
[j] Department of Radiology, Cornell University, Ithaca, NY, USA, Ithaca, NY, USA
[k] Department of Clinical Imaging Sciences, University of Rochester Medical Center, Rochester, NY, USA
[l] Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA
[m] Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

A B S T R A C T

Introduction: The urethra is a recommended avoidance structure for prostate cancer treatment. However, even subspecialist physicians often struggle to accurately identify it on available imaging. Automated segmentation tools show promise, but a lack of reliable ground truth or appropriate evaluation standards has hindered validation and clinical adoption. This study aims to establish a reference-standard dataset with expert consensus contours, define clinically meaningful evaluation metrics, and assess the performance and generalizability of a deep-learning-based segmentation model.

Materials and Methods: A multidisciplinary panel of four experienced subspecialists in prostate MRI generated consensus urethra contours on MRI data for 71 patients from 6 centers, establishing a reference standard. Four of these patients were previously used in an international study (PURE-MRI) where 62 physicians contoured the prostate and urethra. Using an independent training dataset (n = 151 patients, 1 center), we developed a deep-learning AI model for urethra segmentation. We evaluated the AI tool in the consensus reference dataset and compared it to human performance using Dice, percent urethra coverage, and maximum 2D (axial, in-plane) Hausdorff Distance (HD) from the reference standard.

Results: The AI model outperformed most physicians, achieving median Dice of 0.41 (vs. 0.33 for physicians), Coverage of 81 % (vs. 36 %), and Max 2D HD of 1.8 mm (vs. 1.6 mm) in the four PURE-MRI cases. In the full reference dataset, AI performance remained consistent, with Dice of 0.40, Coverage of 89 %, and Max 2D HD of 2.0 mm, indicating strong generalizability across a broader patient population and more varied imaging conditions.

Conclusion: We established a multidisciplinary consensus benchmark for segmentation of the urethra. The deep-learning model performs comparably to specialist physicians and demonstrates consistent results across multiple institutions. It shows promise as a clinical decision-support tool for accurate and reliable urethra segmentation in prostate cancer radiotherapy planning and studies of dose-toxicity associations.

Introduction

Magnetic Resonance Imaging (MRI) has become an essential tool in the early detection, diagnosis, and treatment of prostate cancer, playing a crucial role in guiding targeted biopsies and facilitating treatment planning [1–7]. Among the anatomical structures visualized on MRI, accurate contouring of the urethra has attracted increasing attention due to its relationship with treatment-related toxicity [8–10]. However, accurate contouring of the urethra on MRI remains a significant clinical challenge for human experts, due to the urethra's small caliber,

* Corresponding author at: at: University of California San Diego, 9500 Gilman Drive #0861, La Jolla, CA 92093, USA.
E-mail address: tseibert@health.ucsd.edu (T.M. Seibert).

https://doi.org/10.1016/j.radonc.2025.111231
Received 28 July 2025; Received in revised form 12 September 2025; Accepted 11 October 2025
Available online 25 October 2025

anatomical variability, and the limited visibility of its upper half on MRI [11]. Moreover, there is currently no universally accepted guideline or consensus among clinicians regarding the optimal approach to urethral contouring.

Several automated methods based on deep-learning and radiomics have been proposed for urethra segmentation, demonstrating promising potential [12–15]. Their reported performance remains difficult to evaluate because there is no accepted contouring guideline, and there are large variations in human physician contours that would serve as the "ground truth" for these models [11]. Moreover, the commonly used Dice may not be an ideal metric for assessing urethra segmentation. While a Dice of 1.0 indicates perfect spatial overlap, the metric heavily penalizes small deviations and over-segmentation, which is problematic given the urethra's small size and low contrast on MRI. These characteristics can lead to disproportionately low Dice even for clinically acceptable segmentations. For example, expert consensus supports sparing the urethra for radiation therapy (RT), but this could be achieved with a urethra avoidance structure that encompasses most of the urethra, even if the boundaries of the contours do not have the submillimeter precision that could be required to have a very high Dice [9,16].

To address these challenges, we constructed multidisciplinary consensus urethra contours for a multi-center dataset, following a similar process to that used in prior work on prostate contouring [11,17]. By incorporating interactive feedback and careful review from multiple experienced clinicians with diverse expertise, we aim to establish a more reliable and confident reference standard. We also proposed an evaluation strategy that combines coverage of the urethra and distance-based deviation metrics, in addition to the traditional Dice, which may better reflect clinical relevance and performance in real-world scenarios.

Recognizing the difficulty of the urethra contouring task for many physicians [11], we also developed a deep-learning–based model for urethra segmentation, which we evaluated using the newly constructed expert consensus dataset and the proposed evaluation metrics. We conducted two key comparisons to assess model performance. The first involves a smaller dataset previously used to evaluate human expert annotations [11], allowing for a direct comparison between the AI model and clinical experts. The second evaluation utilized a larger and more diverse dataset to examine the model's robustness and generalizability across varied imaging conditions and patient populations. Through this dual evaluation strategy, our goals are to provide a more comprehensive and clinically meaningful assessment of AI performance and to contribute to the development of reliable, AI-assisted tools for prostate cancer RT planning and future studies of RT-related toxicity.

Methods

Cases

We obtained patient cases from six institutions from the Quantitative Prostate Imaging Consortium (QPIC) as part of a study of prostate cancer detection approved by the institutional review board (IRB) at each institution [7,18]. *Two datasets were utilized for this study. A multicenter dataset was used to create a reference standard of urethra contours; this comprised* $n = 71$ *patient cases with MRI acquired on one of eleven 3T scanners (manufactured by two vendors) at six imaging centers* (Table 1; 68 *of these patient cases were used in a prior study of prostate segmentation* [17]). *The PURE-MRI* [11] *study included a subset of the reference standard dataset* ($n = 4$ *patient cases from* 1 *center*); 62 *physician participants of that study generated* 110 *urethra contours* [11]. *An independent training dataset* ($n = 151$ *patients from* 1 *center*) *was used to develop an AI model for urethra contours; the urethra was segmented on these* 151 *prostates by a single radiation oncologist. All patients included in these datasets underwent prostate multi-parametric MRI (mpMRI) for the evaluation of known or suspected prostate cancer. Cases with substantial imaging artifacts (hip implants, excessive bowel gas, or large body habitus interfering with image quality) were excluded* [17]. *None of the cases used an endorectal coil.*

Table 1
Characteristics of the patient cases included in the full reference standard dataset. Data for the PURE-MRI study is a subset of 4 cases from UC San Diego (UCSD), with images acquired on two scanners (GE Healthcare Signa Premier).

Center and Scanner Platforms for Patient Cases of the Full Reference Standard Dataset (n = 71)		
Cohorts		
UC San Diego (UCSD)		13
Harvard University's Massachusetts General Hospital (MGH)		14
University of Rochester Medical Center (URMC)		12
UC San Francisco (UCSF)		10
UT Health Sciences Center San Antonio (UTHSCSA)		10
University of Cambridge (Cambridge)		12
Institution	**MRI scanner models**	**Number of scanners**
UCSD	GE Healthcare Discovery MRI750, GE Healthcare Signa Premier	3
URMC	SIEMENS Skyra	2
MGH	GE Healthcare Signa Premier	1
UCSF	GE Healthcare Signa Premier	2
Cambridge	GE Healthcare Discovery MRI750	1
UTHSCSA	SIEMENS Skyra	2
Total	**3 models**	**11 scanners**

Reference standard contour development.

A panel of four experts was convened to develop the reference standard urethra segmentation dataset [11,17]: two genitourinary (GU) radiologists (with 12 and 21 years' experience, each surpassing the number of MRIs reported to qualify as prostate MRI experts [19]) and two GU radiation oncologists specializing in MRI-guided prostate radiotherapy (10 years' experience each). One of the GU radiation oncologists generated initial contours on high-resolution axial T_2-weighted slices. During the panel meeting, all three planes were always visualized, and high-resolution coronal and sagittal T_2-weighted acquisitions were available. Each slice case was reviewed at least twice. All members of the panel in attendance agreed on the final contour for each use.

Urethra contouring ground rules.

The ground rules for the contouring of the urethra were agreed by all the members in the panel [11]. Both the prostatic and membranous urethra were included in the delineation. In cases where the urethra was not clearly visible on a given slice—most commonly at the mid-gland and portions of the prostatic base—the panel reviewed all available anatomic information across the three imaging planes to estimate the most likely location. The urethra was consistently contoured as a continuous structure extending from the bladder neck through the prostatic apex and membranous urethra, terminating at the most superior axial slice where the penile bulb was fully visualized.

Physician participant eligibility in PURE-MRI study.

Eligible participants included radiation oncologists, radiologists, and urologists involved in the diagnosis and/or treatment of prostate cancer. Clinical residents or fellows in these specialties were also eligible if they had completed at least one prostate-focused clinical rotation. All study materials, recruitment communications, and procedures were approved by the IRB at UC San Diego (UCSD). Participants were recruited through social media, email outreach, and word-of-mouth within the genitourinary oncology community [11].

Auto-segmentation (AI) model development.

Training data for the deep-learning AI model included another 151

cases described in a previous publication [20]. One of the radiation oncologist panelists had previously contoured the urethra on these 151 patient cases in 2023 to create sample data for understanding anatomic patterns. These cases were obtained through routine clinical care at UCSD on a single 3T scanner (GE Healthcare Discovery 750). For the AI model, because the exact boundaries of the urethra are often not precisely visible—even when the location of the urethra is discernible—we applied a 2 mm dilation to the manual contours with OpenCV [21]. The idea was to train the AI model to identify the urethra's position and course through the prostate rather than necessarily reproduce the exact contours as drawn by the human. The training process utilized nnU-Net [22], a robust and well-established model architecture.

All T_2-weighted images were processed in 3D. Each volume was cropped into patches of size $16 \times 320 \times 320$ voxels and resampled to a uniform voxel size of $3.0 \times 0.5 \times 0.5$ mm^3. Model development followed a 5-fold cross-validation strategy, whereby all 151 cases were iteratively used for both training and validation across the folds. For final inference, an ensemble of the five trained models was employed by averaging the softmax outputs, consistent with the strategy described in the nnU-Net framework [22].

Data augmentation also followed the nnU-Net [22]. Spatial transformations included random rotations, random scaling (range 0.7–1.4), and mirroring along different axes. Intensity transformations included gamma correction, Gaussian noise injection, Gaussian blur, random brightness and contrast adjustment, and simulation of low-resolution acquisitions through downsampling and upsampling. These augmentations were applied to improve robustness and reduce overfitting.

There was a total of 1,000 training epochs in line with the nnU-Net default setting [22]. The initial learning rate was 0.1 and gradually decreased to 0.00002 by the final epoch. Training was conducted on a single NVIDIA Tesla V100 GPU at San Diego Supercomputer Center [23], with a batch size of 8. Model selection was based on validation performance within each fold, using the exponential moving average (*EMA*) of the Dice as the criterion.

$$EMA_{new} = \alpha \bullet Dice_{new} + (1 - \alpha) \bullet EMA_{previous} \# \quad (1)$$

The *EMA* was updated at each epoch (smoothing factor $\alpha = 0.5$), providing a more stable estimate of model performance over time. The epoch yielding the highest *EMA* Dice on the validation data was selected. Together, the use of cross-validation, extensive augmentation, and *EMA*-based model selection was designed to mitigate overfitting and ensure generalizability.

The upper half of the urethra is almost invisible, which can result in missing segments and produce disconnected components in the segmentation. To address this limitation, a linear interpolation approach was implemented to ensure a continuous segmentation consistent with known anatomical prior information and thus improve overall segmentation performance.

Evaluation of AI and human performance.

We compared urethra segmentations by the AI model to those produced by physician participants for patient cases in the PURE-MRI study, using the expert consensus urethra contour as the reference standard. To evaluate the generalizability and robustness of the AI model, we also assessed its performance in the full reference standard (n = 71) dataset. Segmentation performance was quantified using the percent coverage and Dice for the entire urethra. Percent coverage refers to the percentage of the consensus reference urethra that was included in the AI tool's (or physician participants') urethra segmentation. The Dice ranges from 0 (no overlap) to 1 (perfect overlap), accounting for both perfect and no overlap and penalizes over-segmentation as much as under-segmentation. Percent coverage is a conservative metric in the context of a dose constraint like the near-maximum dose to the urethra. On the other hand, an enormously over-segmentation of the prostate (e.g., the whole prostate) would ensure high percent coverage but would also make it difficult to focally boost dose to prostate cancer. To address this limitation, we also incorporated the maximum 2D Hausdorff distance (Max 2D HD), which captures the maximum deviation between the predicted and reference contours across all axial T_2-weighted slices, thereby providing an indication of degree of over-segmentation.

Results

The panel reached consensus for the urethra segmentation on each slice of all 71 cases to create the reference standard dataset. We present a urethra contouring guide with some tips and annotations in Fig. 1; all slices for this patient are presented in Supplementary Fig. 1 to illustrate a complete urethra volume and serve as a contouring atlas (Supplementary Fig. 1). We also present an MRI of a patient with a urinary catheter (not included in any of the datasets used in the study) to illustrate a definitive urethra course, albeit likely altered somewhat by the presence of the plastic device, and for comparison to MRI when no catheter is present (Supplementary Fig. 2).

All the cases are T_2-weighted MRI acquired exclusively on 3T scanners. Our dataset included examinations from two manufacturers (GE Healthcare and Siemens Healthineers) and three scanner models: GE Healthcare Discovery MR750, GE Healthcare Signa Premier, and Siemens Skyra (Table 1). Detailed acquisition parameters, including sequence specifications, are provided in Supplementary Tables 2 and 3, which summarize the imaging characteristics and ensure clarity regarding the applicability of the method.

A total of 62 physicians from 11 countries participated in the PURE-MRI study, including radiation oncologists, radiologists, and urologists, with a range of sub-specialization and varying years of experience (Supplementary Table 1) [11]. In all, physician participants generated 110 urethra contours (as compared to 114 prostate contours from PURE-MRI, as there were four instances where the physician participant stated they were unable to identify the urethra).

In patient cases from the PURE-MRI study, the AI model generated contours that covered a median 81 % of the reference urethra [IQR: 80, 84] for the four cases (Table 2). Physician participants' contouring attempts, in comparison, covered a median 36 % of the reference urethra [IQR: 25, 59]. The best-performing group of human participants were those with 5–10 years of experience, who achieved median 53 % coverage [IQR: 31, 71].

The AI model achieved a median Dice of 0.41 [IQR: 0.37, 0.45] across the four cases (Table 2), while the physician participants achieved a median Dice of 0.33 [IQR: 0.23, 0.42]. The subgroup of physicians with < 5 years of experience performed best by this metric with median Dice of 0.37 [IQR: 0.32, 0.45].

The AI model achieved a median Max 2D Hausdorff Distance (HD) of 1.8 mm [IQR: 1.6, 1.8], meaning a typical AI urethra contour never included prostate tissue more than 1.8 mm from the reference standard (Table 2). Physician participants performed similarly on this metric with median Max 2D HD of 1.6 mm [IQR: 1.2, 1.7]. The best-performing subgroup was those still in training, who achieved a median of 1.3 mm [IQR: 1.1, 1.6]. In terms of overall performance range, the AI model produced values for Max 2D HD between 1.3 mm and 2.1 mm, while the range for physician participants extended from 1.0 mm to 2.3 mm.

We present representative T_2-weighted mid axial, sagittal and coronal slices illustrating contours generated by both the AI model and physician participants (Fig. 2).

To evaluate the performance of the AI model in a broader range of patients, scanners, and imaging centers, we applied the automated tool to all 71 cases with a consensus reference standard urethra (Table 3). Percent coverage of the reference standard urethra was median 89 % [IQR: 78, 95] across the 71 cases. The worst example covered only 34 % of the reference urethra (Fig. 3), while the best example had 100 % coverage. Median Dice was 0.40 [IQR: 0.35, 0.44]. Median Max 2D Hausdorff Distance was 2.0 mm [IQR: 1.5, 2.3], indicating that the AI model's contour for a typical case strayed no more than 2 mm from the

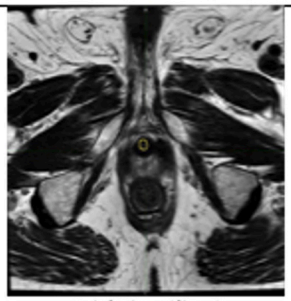
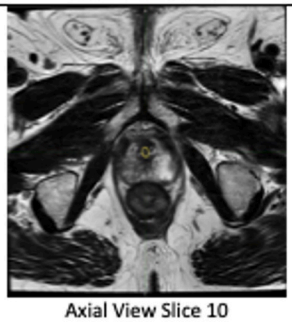
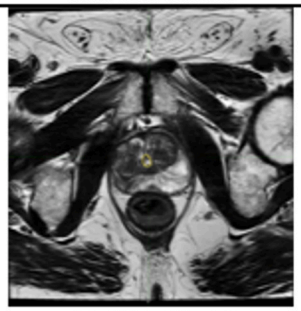
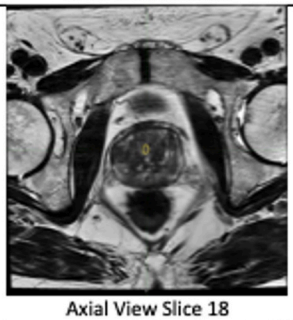
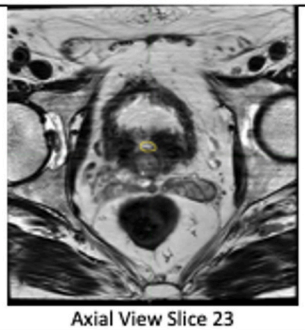
Axial T_2-weighted Slices	Annotations for Contouring
 Axial View Slice 8	Membranous urethra extends from prostatic urethra to inferior aspect of urogenital diaphragm, ending just superior to the penile bulb.
 Axial View Slice 10	Inferior prostatic urethra: continuous with membranous urethra and often moderately hyperintense, just anterior to an inverted "V" of hyperintensity that can extend superiorly to the verumontanum.
 Axial View Slice 13	Example of urethra at level of verumontanum. Superior to this, the urethra often becomes difficult to visualize.
 Axial View Slice 18	Midgland prostatic urethra: The urethra can be difficult to visualize in the upper midgland. It can be helpful to start at apex and track as far superiorly as possible then skip to the prostate base to identify the bladder neck and track the urethra as far inferiorly as possible before returning to the midgland. Consider viewing all three planes while delineating the urethra in the midgland. The urethra may not remain in the mid-sagittal plane because of benign prostatic hyperplasia that causes asymmetry of the prostate.
 Axial View Slice 23	Superior prostatic urethra: identify region of marked hyperintensity (urine) where the bladder neck inserts into the prostate base and trace inferiorly. The superior origin of the urethra must be continuous with the bladder neck; it is often not at the superior tip of the prostate but is on the anterior aspect of the gland.

Fig. 1. Annotations for urethra contouring. The expert contour is shown as yellow contour on axial T_2-weighted slices on the right column of this figure and annotations for urethra contouring on the left column.

Table 2
Dice, Coverage (%) and Max 2D HD (mm) of the AI model and physician participants with different year of experience on cases from the PURE-MRI study.

Model	Dice			Coverage (%)			Max 2D HD (mm)		
Model	median			median			median		
	min	IQR	max	min	IQR	max	min	IQR	max
AI Model	0.41			81			1.8		
	0.35	0.37–0.45	0.48	80	80–84	91	1.3	1.6–1.8	2.1
All physician (n = 62)	0.33			36			1.6		
	0.03	0.23–0.42	0.69	3	25–59	96	1.0	1.2–1.7	2.3
Still in Training (n = 14)	0.30			26			1.3		
	0.03	0.21–0.42	0.69	3	19–44	85	1.0	1.1–1.6	2.0
Less than 5 years (n = 12)	0.37			37			1.7		
	0.15	0.32–0.45	0.62	11	30–60	89	1.0	1.5–1.8	2.3
5 to 10 years (n = 11)	0.34			53			1.6		
	0.08	0.25–0.44	0.65	9	31–71	84	1.1	1.4–1.7	2.0
Greater than 10 years (n = 25)	0.28			34			1.6		
	0.06	0.23–0.38	0.62	3	26–53	96	1.0	1.3–1.8	2.3

reference standard. In the worst case, the AI model's contour strayed 2.9 mm from the reference standard in one slice.

Discussion

The urethra is a critical organ of interest in prostate cancer treatment because of its relationship to functional outcomes for patients undergoing radiation therapy [8–10]. Measurement of urethra length may also be relevant for predicting outcomes after surgery [24,25]. However, the PURE-MRI study revealed poor accuracy of physician urethra contours and very poor inter-physician agreement. Those results call into question both the validity of urethra dose constraints and the feasibility of introducing standardized urethra contouring in general clinical practice. We provide here a multidisciplinary consensus contouring atlas for the urethra (Fig. 1, Supplementary Fig. 1). We also developed an AI tool that outperforms physicians for urethra coverage, covering a median 81 % of the reference standard (vs. median 36 % across 110 physician contours) without including excessive amounts of non-urethra tissue (measured by Max 2D HD). In a broader multi-center dataset, the AI tool typically covered 89 % of the reference standard urethra while straying no more than 2.0 mm from the reference urethra contour.

Dice is a commonly used metric for evaluating segmentation accuracy but can be poorly suited to some applications [26]. The urethra is small in diameter and has a comparatively long course relative to its diameter. It is also poorly visualized in parts of the prostate, and its exact boundary is difficult or impossible to outline even if its location is agreed upon, as noted by our multidisciplinary panel [11]. We therefore proposed percent coverage and Max 2D HD as more appropriate metrics for this case. Coverage of the urethra is of obvious importance if the goal is to avoid overdosing the true urethra during radiation therapy. Conversely, an overly large urethra avoidance structure would unnecessarily limit the dose achievable to some prostate tumors, especially during focal radiation boost [27,28]. Max 2D HD is helpful as an indicator of how overly conservative a urethra contour is in its least accurate slice. Consistent with our reasoning, we found that the AI model was not impressive for Dice (while still better than physician participants), but the AI tool performed well on percent coverage and Max 2D HD.

Several recent studies have explored the deep-learning-based urethra segmentation on MRI or CT [12,13,29]. One such study by Belue et al. reported a model achieving an overall Dice of 0.61 and typical centerline distance (distance between ground truth and AI segmentation on each slice, averaged across slices of 2.6 mm) [13]. It is not valid to compare summary statistics of AI models applied to distinct datasets, but the Dice in the Belue et al. model is impressive and possibly suggests incorporation of more data for training (Belue et al. had 657 cases for training, compared to the 151 we used), possibly suggesting a larger training dataset would improve our model. Our study, on the other hand, benefitted from including 110 urethra contours made by 62 physicians to directly compare AI performance to human performance on the same patient cases. We also used additional metrics in our study with clinical relevance for radiation treatment planning. Max 2D HD is an indicator of error on the worst slice in a patient case and represents the worst-case scenario for designating non-urethra tissue as an avoidance structure that could interfere with focal boost of a nearby tumor. Percent coverage of the urethra is an indicator of how much of the urethra is correctly labeled for sparing (uncovered urethra could lead to urethra doses higher than intended). We are also planning to integrate our urethra model into the clinical workflow by providing physicians with its auto-segmentations, helping them accurately locate the urethra to avoid exceeding dose constraints. This will first be done in two prospective studies, including NCT06990542.

Our study has several limitations. First, establishing a definitive ground truth for urethral contours is inherently challenging. In particular, the upper half of the urethra is often difficult or even impossible to visualize clearly on MRI, making objective delineation unreliable. As a result, the reference standard used in this study—derived from consensus among experienced experts with complementary clinical backgrounds—represents possibly the best feasible approximation at present. Second, we evaluated only a single AI-based segmentation approach. To gain a more comprehensive understanding of the capabilities and limitations of AI tools for urethral segmentation, future studies should include a broader set of algorithms, encompassing both academic methods and commercial solutions. Third, training data came from only one center and 151 patient cases; larger and more diverse training data might improve performance. Fourth, some centers may employ protocol variations (e.g., specialized 3D acquisitions) or use MRI systems that could not be evaluated with data available for this study. A strength of this study, though, is inclusion of consensus contours on images from two major vendors, three scanner models, and six acquisition protocols from six centers (Supplementary Table 3). Fourth, it is unclear how large an error in urethra segmentation must be to have a clinically meaningful impact on patient outcomes, so this cannot be evaluated here (likewise, we cannot confidently determine an optimal planning risk volume, or PRV). On the other hand, the paucity of clear data on dose-toxicity associations for various urethra-sparing approaches strengthens the rationale for the present study—consistent urethra contours are needed to support robust evidence for better urethra constraints. Our results demonstrate that the AI model developed here already outperforms a large cohort of physicians and offers a promising complementary tool to enhance the accuracy and consistency of urethral contouring in clinical practice.

Conclusion

We established a multidisciplinary consensus reference-standard dataset for the evaluation of urethra segmentation and provide an

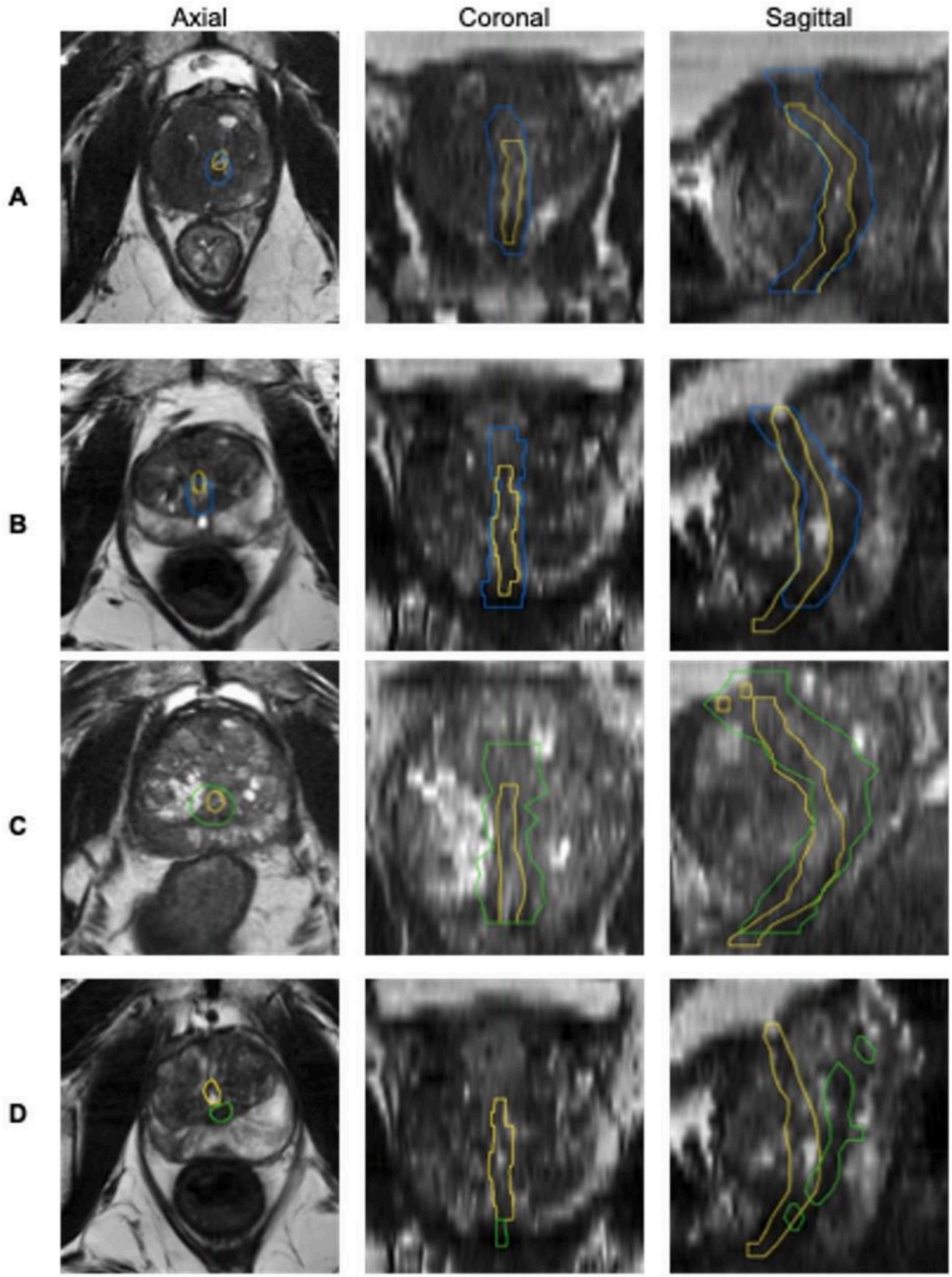


Fig. 2. Examples of AI and physician urethra contours, as well as the reference standard (consensus contour by a multidisciplinary panel of four subspecialists. Contours are overlayed on axial, coronal, and sagittal T_2-weighted MRI. The multidisciplinary consensus expert-defined urethra contour is shown as yellow contour. The AI generated urethra contour is shown as blue contour. The PURE-MRI study physician participants' urethra contours are shown in green. **2A** is an example of AI generated urethra contour with Dice of 0.35, Coverage of Urethra of 91 %, and Max 2D HD of 2.1 mm. **2B** is an example of AI generated urethra contour with Dice of 0.38, Coverage of Urethra of 80 %, and Max 2D HD of 1.3 mm. **2C** is an example of physician participants' urethra contour with Dice of 0.21, Coverage of Urethra of 96 %, and Max 2D HD of 1.7 mm. **2D** is an example of a physician participant's urethra contour with Dice of 0.09, Coverage of Urethra of 8 %, and Max 2D HD of 1.1 mm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Dice, Coverage (%) and Max 2D HD (mm) of the AI model on the full reference standard dataset.

Model	Dice			Coverage (%)			Max 2D HD (mm)		
Model	median			median			median		
	min	IQR	max	min	IQR	max	min	IQR	max
AI Model	0.40			89			2.0		
	0.20	0.35–0.44	0.54	34	78–95	100	1.0	1.5–2.3	2.9

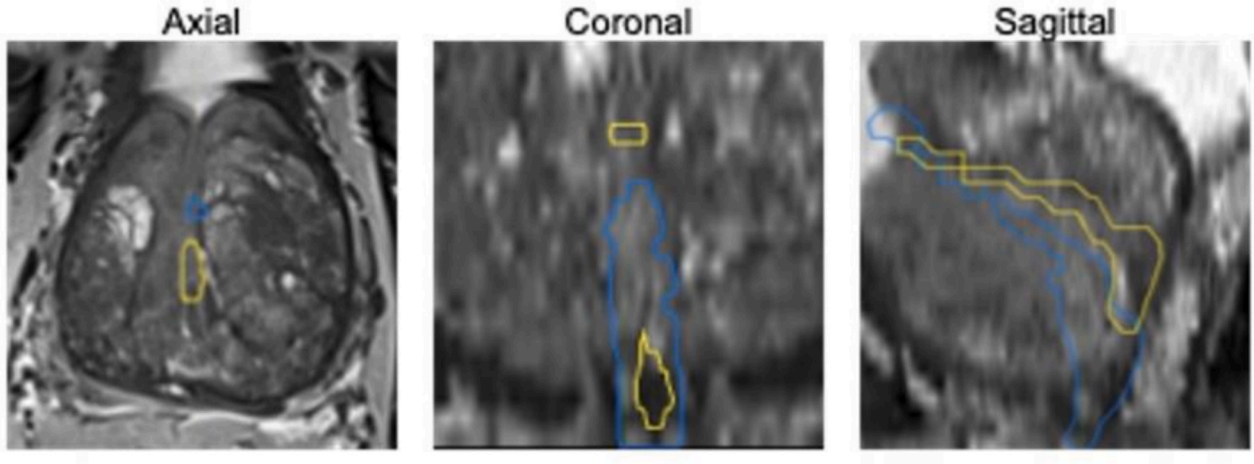


Fig. 3. Example auto-segmentation errors of urethra from AI. Contours are overlayed on axial, coronal and sagittal T_2-weighted MRI. The expert-defined urethra contour is shown as yellow contour. The AI generated urethra contour is shown as blue contour. The AI generated urethra contour with Dice of 0.20, Coverage of Urethra of 34 %, and Max 2D HD of 2.9 mm. This is the patient case where the AI tool performed worst; the discrepancy between the AI-generated urethra and reference standard is greatest in the mid-gland and base of the prostate, where the urethra is most difficult to visualize. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

atlas for contouring the urethra on MRI. We introduced clinically meaningful metrics, including percent urethra coverage and Max 2D HD, to better reflect segmentation quality in real-world clinical settings. We developed a deep-learning model for urethra segmentation, a challenging and time-consuming task. Our AI tool outperformed most physicians and was comparable to even the best-performing physician participants in the PURE-MRI study. Moreover, its consistent performance across a diverse, multi-center dataset demonstrates overall robustness and generalizability, though there were some errors large enough to require correction. These results suggest a strong potential for AI-based auto segmentation of the urethra to assist, or even augment, physician urethra segmentation during radiation therapy planning, ultimately facilitating investigations of valid dose constraints and helping to reduce the risk of urethral injury.

[Funding Statement]

This study has received funding by:

National Institutes of Health, Grant/Award Numbers: NIH/NIBIB K08EB026503, NIHUL1TR000100;

American Society for Radiation Oncology, the Prostate Cancer Foundation: PCF20YOUN01;

Department of Defense, Grant/Award Number: DOD/CDMRPPC220278;

CRediT authorship contribution statement

Yuze Song: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lily Nguyen:** Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Anna Dornisch:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Madison Baxter:** Writing – review & editing, Data curation. **Tristan Barrett:** Writing – review & editing, Data curation. **Anders M Dale:** Writing – review & editing, Data curation. **Robert T Dess:** Writing – review & editing, Data curation. **Mukesh Harisinghani:** Writing – review & editing, Data curation. **Sophia C Kamran:** Writing – review & editing, Data curation. **Michael A Liss:** Writing – review & editing, Data curation. **Daniel JA Margolis:** Writing – review & editing, Data curation. **Eric P Weinberg:** Writing – review & editing, Data curation. **Sean A Woolen:** Writing – review & editing, Data curation. **Tyler M Seibert:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

Daniel J.A. Margolis reports a clinical advisor role for Stratagen Bio and an ad hoc consultant role for Guerbet and Promaxo. Sean A. Woolen was supported by funding from the ARRS Scholarship for professional development and has received investigator-initiated research grants (paid to the institution) from Siemens. Sophia C. Kamran's spouse is employed by Sanof. Anders M. Dale is a founder of and holds equity interest in CorTechs Labs and serves on its scientific advisory board. He is also a member of the Scientific Advisory Board of Healthlytix and receives research funding from General Electric Healthcare (GEHC). Michael A. Liss Founder/President of Oncobiomix with no realation to this manuscript. Tyler M. Seibert reports honoraria from Varian Medical Systems, WebMD, MJH Life Sciences, GE Healthcare, Blue Earth Diagnostics, and Janssen; he has an equity interest in CorTechs Labs, Inc. and serves on its Scientific Advisory Board; he receives research funding from GE Healthcare and Blue Earth Diagnostics, as well as in-kind research support from Quibim, Inc., both through the University of California San Diego. These companies might potentially benefit from the research results. The terms of this arrangement have been reviewed

and approved by the University of California San Diego in accordance with its conflict-of-interest policies. All remaining authors havbe declared no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2025.111231.

References

[1] Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med 2018;378:1767–77. https://doi.org/10.1056/NEJMoa1801993.

[2] Ahmed HU, Bosaily A-E-S, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet 2017;389:815–22.

[3] Ahdoot M, Wilbur AR, Reese SE, et al. MRI-Targeted, Systematic, and combined Biopsy for Prostate Cancer Diagnosis. N Engl J Med 2020;382:917–28. https://doi.org/10.1056/NEJMoa1910038.

[4] Schoots IG, Barentsz JO, Bittencourt LK, et al. PI-RADS Committee Position on MRI without contrast medium in biopsy-naive men with suspected prostate cancer: narrative review. Am J Roentgenol 2021;216:3–19. https://doi.org/10.2214/AJR.20.24268.

[5] Dorfinger J, Ponholzer A, Stolzlechner M, et al. MRI/ultrasound fusion biopsy of the prostate compared to systematic prostate biopsy–Effectiveness and accuracy of a combined approach in daily clinical practice. Eur J Radiol 2022;154:110432.

[6] Song Y, Rojo Domingo M, Conlin CC, et al (2024) Deep learning AI and Restriction Spectrum Imaging for patient-level detection of clinically significant prostate cancer on MRI. medRxiv 2024–11.

[7] Domingo MR, Do DD, Conlin CC, et al. Restriction Spectrum Imaging as a quantitative biomarker for prostate cancer with reliable positive predictive value. J Urol 2023:10–1097.

[8] Groen VH, van Schie M, Zuithoff NP, et al. Urethral and bladder dose–effect relations for late genitourinary toxicity following external beam radiotherapy for prostate cancer in the FLAME trial. Radiother Oncol 2022;167:127–32.

[9] Le Guévelou J, Sargos P, Ost P, et al. Urethra-sparing prostate cancer radiotherapy: current practices and future insights from an international survey. Clin Transl Radiat Oncol 2024;51:100907.

[10] Martin JM, Richardson M, Siva S, et al. Mechanisms, mitigation, and management of urinary toxicity from prostate radiotherapy. Lancet Oncol 2022;23:e534–43.

[11] Nguyen L, Song Y, Dornisch A, et al (2025) PURE-MRI: An International Study Assessing Physician Accuracy in Delineating the Prostate and Urethra on Prostate MRI. medRxiv 2025–04.

[12] Rezaeijo SM, Nesheli SJ, Serj MF, Birgani MJT. Segmentation of the prostate, its zones, anterior fibromuscular stroma, and urethra on the MRIs and multimodality image fusion using U-Net model. Quant Imaging Med Surg 2022;12:4786.

[13] Belue MJ, Harmon SA, Patel K, et al. Development of a 3D CNN-based AI model for automated segmentation of the prostatic urethra. Acad Radiol 2022;29:1404–12.

[14] Hambarde P, Talbar SN, Sable N, et al. Radiomics for peripheral zone and intra-prostatic urethra segmentation in MR imaging. Biomed Signal Process Control 2019;51:19–29.

[15] Xu D, Ma TM, Savjani R, et al. Fully automated segmentation of prostatic urethra for MR-guided radiation therapy. Med Phys 2023;50:354–64. https://doi.org/10.1002/mp.15983.

[16] Ratnakumaran R, Mohajer J, Withey SJ, et al. Developing and validating a simple urethra surrogate model to facilitate dosimetric analysis to predict genitourinary toxicity. Clin Transl Radiat Oncol 2024;46:100769.

[17] Song Y, Dornisch AM, Dess RT, et al. Multidisciplinary consensus prostate contours on magnetic resonance imaging: educational atlas and reference standard for artificial intelligence benchmarking. Int J Radiat Oncol Biol Phys 2025.

[18] Do DD, Rojo Domingo M, Conlin CC, et al (2024) Robustness of a Restriction Spectrum Imaging (RSI) quantitative MRI biomarker for prostate cancer: assessing for systematic bias due to age, race, ethnicity, prostate volume, medication use, or imaging acquisition parameters. medRxiv 2024–09.

[19] De Rooij M, Israël B, Tummers M, et al. ESUR/ESUI consensus statements on multi-parametric MRI for the detection of clinically significant prostate cancer: quality requirements for image acquisition, interpretation and radiologists' training. Eur Radiol 2020;30:5404–16. https://doi.org/10.1007/s00330-020-06929-z.

[20] Zhong AY, Digma LA, Hussain T, et al. Automated patient-level prostate cancer detection with quantitative diffusion magnetic resonance imaging. Eur Urol Open Sci 2023;47:20–8.

[21] Gollapudi S. OpenCV with Python. In: Learn Computer Vision Using OpenCV. Berkeley, CA: Apress; 2019. p. 31–50.

[22] Isensee F, Jaeger PF, Kohl SA, et al. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11.

[23] Strande S, Cai H, Tatineni M, et al. Expanse: Computing without Boundaries: Architecture, Deployment, and Early Operations Experiences of a Supercomputer Designed for the Rapid Evolution in Science and Engineering. In: Practice and Experience in Advanced Research Computing. Boston MA USA: ACM; 2021. p. 1–4.

[24] Coakley FV, Eberhardt S, Kattan MW, et al. Urinary Continence after radical retropubic prostatectomy: relationship with membranous urethral length on preoperative endorectal magnetic resonance imaging. J Urol 2002;168:1032–5. https://doi.org/10.1016/S0022-5347(05)64568-5.

[25] Paparel P, Akin O, Sandhu JS, et al. Recovery of urinary continence after radical prostatectomy: association with urethral length and urethral fibrosis measured by preoperative and postoperative endorectal magnetic resonance imaging. Eur Urol 2009;55:629–39.

[26] Sherer MV, Lin D, Elguindi S, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review. Radiother Oncol 2021;160:185–91.

[27] Kerkmeijer LGW, Groen VH, Pos FJ, et al. Focal boost to the Intraprostatic Tumor in External Beam Radiotherapy for patients with Localized Prostate Cancer: results from the FLAME Randomized phase III Trial. J Clin Oncol 2021;39:787–96. https://doi.org/10.1200/JCO.20.02873.

[28] Guricová KM, Pos FJ, Schoots IG, et al. Intra-prostatic recurrences after radiotherapy with focal boost: location and dose mapping in the FLAME trial. Radiother Oncol 2024;201:110535.

[29] Cubero L, García-Elcano L, Mylona E, et al. Deep learning-based segmentation of prostatic urethra on computed tomography scans for treatment planning. Phys Imaging Radiat Oncol 2023;26:100431.