Automated vs. manual segmentation for small renal mass imaging

Kristen McAlpine[1], Nikhil Mirajkar[2], Dominik Deniffel[3,4], Andres Kohan[5], Satheesh Krishna[5], Girish Kulkarni[1], Emily Seto MSc[6,7], Antonio Finelli[1], Masoom A. Haider[5]

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; [2]Department of Radiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdon; [3]Department of Diagnostic and Interventional Radiology, Cantonal Hospital Frauenfeld, Frauenfeld, Switzerland; [4]Technical University of Munich, TUM School of Medicine and Health, Munich, Germany; [5]Joint Department of Medical Imaging, Sinai Health System, Princess Margaret Hospital, University of Toronto, Toronto, ON, Canada; [6]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada; [7]Centre for Digital Therapeutics, Toronto General Hospital Research Institute, University Health Network, Toronto, ON, Canada

Cite as: McAlpine K, Mirajkar N, Deniffel D, et al. Automated vs. manual segmentation for small renal mass imaging. *Can Urol Assoc J* 2026 February 13; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.9476

Published online February 13, 2026

Funding: CUASF-KCC-KCRNC Grant, Sinai Health Foundation.

Corresponding author: Dr. Kristen McAlpine**,** Division of Urology, Department of Surgery, University of Toronto**,** Toronto, ON, Canada; kristen.mcalpine@uhn.ca

***

ABSTRACT

Introduction: Automated segmentation using artificial intelligence (AI) has the potential to rapidly perform three-dimensional (3-D) segmentation of small renal masses (SRM). The objective of this study was to test for clinically and statistically significant differences in time spent segmenting, accuracy, and reliability when comparing manual and automated segmentation of computed tomography (CT) scans with SRM.

Methods: Patients with a CT scan, SRM <4 cm, and renal neoplasm were identified through an institutional database. Of the 854

KEY MESSAGES

- Automated segmentation of small renal masses on CT scans is efficient, accurate, and acceptable when compared to manual segmentation completed by experienced clinicians.
- Automated segmentation run with consumer-grade software allows for accurate identification and isolation of small renal masses on CT scans in a clinical setting.
- Automated segmentation has the potential to dramatically improve the ability to predict growth rate of small renal masses being monitored with small renal masses and to assist with surgical planning of these tumors.

patients identified, 184 were excluded. Forty test cases were randomly selected. There were 630 cases for training (using nnU-Net) to which 488 cases from the KiTS23 open-source data set were added. Each of the test cases was segmented by a radiologist, a urologist, and the AI model. Time to segment and Dice coefficients were compared. Deidentified segmented CTs were provided to two independent radiologists who attempted to identify the segmentor and rated the acceptability of the segmented images on a five-point Likert scale.

Results: There were 39 cases with complete timing data. The median time for the AI model to segment was one third of the radiologist's (152.4 s, interquartile range [IQR] 120.9–177.8 vs. 450 s, IQR 318.8–551.2) and about one-fifth of the urologist's (800.0 s, IQR 492.0–1538.0) ($p<0.001$). There was a high degree of inter-rater reliability (median Dice coefficients 0.86–0.90, $p=0.09$). The scoring radiologists were able to correctly identify the true segmentor in 61.6% of cases ($p <0.001$). The AI segmentations were scored highest among the three segmentors (median score 4.1/5, standard deviation [SD] 1.0) compared to 3.8 (SD 0.7) for the radiologist, and 3.3 (SD 0.7) for the urologist.

Conclusions: Automated segmentation of CT scans for patients with SRM was efficient, accurate, and acceptable in this study. This approach has the potential to greatly improve the clinical use of radiomics to assess medical images for these patients.

INTRODUCTION

Artificial intelligence, specifically convolutional neural networks, have achieved human-level performance in segmenting pathology on CT scans[1,2]. This opens the door to application of AI for 3D segmentation of renal masses in clinical settings.

In the field of urology, there are an increasing number of patients with incidentally detected small renal masses (SRM)[3,4]. These masses (<4 centimetres [cm]) are often found on abdominal imaging performed for another indication. Many of these masses are benign or contain low-grade malignant cells and are unlikely to cause patients harm over their lifetime. However, once detected there remains a clinical dilemma about whether the mass needs to be treated, observed or biopsied[5–7]. Patients and clinicians are faced with challenging decisions regarding how to best manage these incidental SRMs. Options include nephron sparing surgery, ablation and surveillance each of which cross-sectional imaging is helpful to guide. Manual segmentation of CT scans is a time-consuming and labour-intensive process. An AI model could overcome the barriers of manual segmentation and open the door to large scale studies to develop radiomics based biomarkers that could radically improve the care of these patients by better assessing volumetric growth rates, improving histology prediction compared to clinical nomograms and improving patient selection for observation versus surgery. If this could be done on consumer grade hardware and work on single phase CT, it could be used directly in the clinic.

The objective of this study was to test for clinical and statistical differences in time-spent segmenting, accuracy and reliability when comparing manual radiologist and urologist segmentation versus AI-based segmentation of CT scans of SRMs.

METHODS

Case identification

Institutional ethics board approval from University Health Network was obtained for this retrospective observational study (REB 20-5942). Patients were identified using an institutional kidney cancer database (January 2007 – December 2021). Patients were included if they were >18 years of age, had pathology proven renal neoplasm, had at least one renal mass that was <4 cm and had a contrast enhanced CT scan prior to any treatment of the SRM and not part of a test set in any other study which the same AI model was to be used. A total of 854 patients were identified. From this cohort, 83 patients were initially excluded (Figure 1). Cases were excluded if they had missing or incorrect imaging files, renal factors that complicated the segmentation of the renal mass including prior renal transplant, polycystic kidneys, atrophic kidneys, lesions <1 cm, horseshoe kidney or prior renal surgery. This left 670 local CT's which were randomly split into 630 for training and 40 for testing. When analyzing these 40 cases, 1 case was later excluded due to missing/corrupt data when distributing cases to the segmentors.

Manual segmentation

Manual segmentation of each of the 39 test cases was performed independently by an early career urologist who was familiar with SRM CTs (KM) and by an abdominal imaging fellow (NM). Time to segment each case (seconds [s]) was recorded. Only one phase of the CT was used in this study which was the portal phase. If the portal phase was not available, the corticomedullary phase was used. If the corticomedullary phase was unavailable, then the nephrographic phase was used.

AI model training

As the model was being used for other studies, a larger test set size was needed to assess the model's diagnostic performance. To the internal training cohort of 630 cases, 488 CT's from the KiTS23 open source data set (database of CT scans with renal masses from open segmentation challenge) were added for a total training set size of 1118[8]. For the KiTS23 data set, the provided segmentation masks were used for training and for our internal data set, segmentations were performed by an abdominal imaging fellow (NM) under the supervision of a senior abdominal radiologist (MH) using ITK-snap software[8,9]. The nnU-Net v2 command line interface (7) was used to train our model with 5-fold cross-validation and 2000 epochs using the '3d_fullres' configuration. The model was run using a single GPU machine with the following specification: GPU-Nvidia GTX 4090-24GB VRAM, CPU-Intel i9 3.2GHz, 64GB RAM, Windows 11 Pro, Pytorchv2.3.1, CUDA 12.1, CUDNN 8907. A hierarchical class structure was used with 'SRM' and 'cysts' being subclasses of 'lesion' which was any kind of lesion. Other parameters were left

as defaults. Predictions were then executed on the test cases. Only the SRM segmentation was assessed. If the AI model only partially segmented a lesion, the case was still included.

Assessment of segmentation acceptability

The completed automated and manual segmentation images were then deidentified, randomized and provided to two scoring radiologists to assess. If a case had multiple SRMs, each segmented lesion was counted as its own deidentified segmentation case for the scoring radiologists to evaluate. One radiologist was a highly experienced staff radiologist (SK) who specializes in abdominal imaging and the other was another abdominal imaging fellow (AK).

A Turing test was performed with each of the scoring radiologists attempting to determine whether an individual case had been segmented by: 1) a urologist, 2) a radiologist, 3) an AI-program.

Finally, each of the scoring radiologists assigned a score for the acceptability of the images using a previously published scoring system[10]:

- 5: Strongly agree: Use-as-is; clinically acceptable and could be used without change.
- 4: Agree: Minor edits that are not necessary. Stylistic differences but not clinically important. Current contours are acceptable.
- 3: Neither agree or disagree: Minor edits that are necessary. Minor edits are those that the reviewer judges can be made in less time than starting from scratch or are expected to have minimal effect on ‘treatment outcome’.
- 2: Disagree: Major edits. This category indicates that the necessary edits are required to ensure ‘appropriate treatment’ and sufficient significant that the used would prefer to start from scratch.
- Strongly disagree: Unusable. This category indicates that the quality of the contours are so bad that they are unusable.

Analysis of data

Time to segment was assessed for normal distribution and differences between segmentors were assessed using Mann-Whitney U tests. Inter-rater reliability was assessed using Dice-Sorensen coefficients (Dice coefficient) to assess the overlap between the segmented area for each case by the different segmentors. To test for any difference in the distribution of the Dice coefficients, a Kruskal-Wallis test was used.

Data from the scoring radiologists was also analyzed. To determine if there was a statistically significant ability for the radiologists to identify the correct segmentor (Turing test), a chi-square test was used. Finally, the average score and standard differentiation (SD) for the acceptability score provided by the scoring radiologists were reported and assessed for significant differences using Mann-Whitney U tests.

As only SRM patients were in the test set, real world model performance for SRMs was not assessed. However, if the AI model did not detect the SRM this was reported.

RESULTS

Complete time to segment from all three segmentors was available for 39 cases. The median patient age in this cohort was 62 years old (IQR 52-68 years). The cohort consisted of 22 males (56%) and 17 females (44%). The median tumour size was 3.37 cm (IQR 2.37-4.16 cm). The majority of the lesions were clear cell renal cell carcinoma (21/39, 54%), however there were also a number of chromophobe renal cell carcinoma (9/39, 23%), papillary renal cell carcinoma (6/39, 15%), and other lesions (3/39, 8%). The median time to segment each case was 800.0 s (interquartile range [IQR] 492.0-1538.0), 450 s (IQR 318.8 – 551.2) and 152.4 s (IQR 120.9 – 177.8) for the urologist, radiologist, and automated segmentation respectively. This difference was statistically significant ($p < 0.001$)(Table 2).

Inter-rater reliability was similar between each of the assessed segmentor pairs (Table 3). The median Dice coefficient between the urologist and the radiologist was 0.86 (IQR 0.82 – 0.90). The median Dice coefficient between the urologist and the automated segmentation was 0.89 (IQR 0.85 – 0.93). The median Dice coefficient between the radiologist and the automated segmentation was 0.90 (IQR 0.86 – 0.92). The difference in Dice coefficients between the 3 segmentors was not significant (p-value = 0.09).

The Turing test results found that the scoring radiologists were able to correctly identify the segmentor in 61.6% of cases ($p < 0.001$)(Table 4). The scoring radiologists were best able to correctly identify images segmented by the urologist (67.8%) and the automated segmentor (67.4%). The scoring radiologists were more likely to mistaken the automated segmented images for radiologist segmented images (23.3%) rather than urologist segmented images (9.3%).

The median acceptability score was assessed for each segmentor. The segmentation images completed by the AI tool were scored the highest (mean score 4.1, SD 1.0) (Figure 2). The segmentations completed by the radiologist were the next highly ranked (mean score 3.8, SD 0.7), followed by the segmentations completed by the urologist (mean score 3.3, SD 0.7) (Figure 3). The overall distribution in these scores between segmentors was statistically significant (p-value <0.001).

The number of automated segmentations given a perfect score of "5" ('use-as-is') was also better than the human readers. When combining the scores of the two scoring radiologists, the number of perfect scores for the automated segmentations was 36 (n=86, 42%). This compares to 13 perfect scores for the segmentations completed by the radiologist (n=82, 16%) and 6 perfect scores for the urologist segmentor (n=90, 7%). There was 1 case in which segmentation images were lost in the analysis process resulting in unequal number of cases in the acceptability test results.

The automated segmentation missed the lesion in 1 case. There were an additional 2 cases in which the automated segmentation had large errors when attempting to identify the region-of-interest. These cases had acceptability scores of 1 or 2 from both radiologist raters. One of these lesions was adjacent to a large renal cyst and the second lesion was mixed solid-

cystic in density. The automated program performed an incomplete segmentation, missing a large proportion of the cystic component (Figure 4)

DISCUSSION

Segmentation of SRMs on CT has the potential to improve clinicians' ability to assess volumetric size change and plan nephron sparing treatments as well as to allow development of imaging biomarkers in the field of radiomics. Manually segmenting the region-of-interest on CT is a time-consuming, labour-intensive process that limits the clinical applicability of using radiomics programs during patient encounters. Our study assessed the role of AI segmentation of SRMs on routine single-phase CT scans using consumer grade GPU hardware potentially useable in clinic. We compared the efficiency and acceptability of these automated segmentations to manual segmentation performed by two experienced clinicians. We found that the automated segmentations were significantly faster (median time 152.4 s vs 450 s vs 800 s, p <0.001), equally accurate (median Dice coefficients 0.86 – 0.90, p 0.09) and had higher acceptability scores (4.1 vs 3.8 vs 3.3, p <0.001) when compared to manual segmentations performed by two independent clinicians.

Clinicians caring for patients with SRMs often rely on cross-sectional abdominal imaging to assess the tumour and to counsel patients on the next steps in their care. Images, patient demographics, and histology of the mass are all used to risk stratify patients when determining if active treatment or observation is most appropriate. By employing automated SRM segmentation models, further studies into quantitative imaging features are also possible for assessment of tumor histology, prognosis and volumetric growth rate for surveillance. The feasibility of auto-segmenting a tumor on a standard post-contrast CT, using consumer grade hardware, further opens the door to local applications in clinic without the training or time needed to manually segment a tumor by a urologist or waiting for a radiologist to perform this task.

Automated segmentation of renal mass imaging has been reported by other groups using different approaches to creating an automated model and comparing automated segmentations to manual segmentations[8,11–13]. We found similarly high Dice coefficients between our automated model and manual segmentation when compared to other studies in the literature[8,11–13]. With previous publications citing values of >0.7-0.8 as excellent for inter-rater reliability, our values of 0.86-0.90 are well above the acceptable range[14,15]. Unique to our study is the inclusion of a radiologist and a urologist as manual segmentors which simulates real-world use of this technology occurring in hospitals internationally. We also focused only on SRMs and CT scan images. Our results show that automated segmentation of SRMs on CT scans is time-efficient, accurate and acceptable to staff radiologists. Using this automated process, a significantly greater number of scans can be included in future surveillance or risk model development studies as person-power is not a limiting factor. With this automated process, further models will be created to assess the diagnostic and prognostic ability of radiomics for patients with SRMs.

The strengths of this study include the multi-disciplinary approach to a challenging clinical situation for urologists and radiologists, the focus on SRM CTs, the use of consumer

grade hardware to run the model and the large number of cases included in the training cohort. However, there are limitations to our study. Firstly, CT scans performed at different institutions were included in our training and test cohorts. Varying CT protocols and contrast timings at the different sites introduced variability in the appearance of the images to both the automated and manual segmentors yet this is reflective of real-world practice. Second, we excluded certain renal disease types that can confound an AI tool such a polycystic kidney disease, renal failure and transplant cases. In our test cohort there were no cases where the AI missed a tumor however this may not be the case in the real world and in those cases, segmentation would need to be done manually. By excluding these patients, we likely exaggerated the accuracy of the automated model. In a real-world setting in which these patients are present, the automated system may not perform as well. It should be noted that there were 3 cases in which the automated program had large misses when segmenting the SRM. For these reasons, we recommend that clinicians using an automated program to segment SRMs, routinely review the images and the radiology report to ensure that the annotations match the lesion of concern. Finally, the radiologist and urologist segmentors were not blinded to the study design which may have introduced bias into their segmentation time and accuracy. However, with excellent Dice coefficients between all three segmentors, this goes to highlight that a high degree of care was taken when performing these segmentations.

CONCLUSIONS

Automated segmentation of CT scan with SRMs can improve the clinical assessment and diagnostic work-up for patients. Our study shows automated segmentation of SRM imaging is efficient, accurate and highly acceptable. These findings set the ground work for future diagnostic and prognostic studies using radiomics with automated segmentation.

REFERENCES

1. Park T, Yoon MA, Cho YC, et al. Automated segmentation of the fractured vertebrae on CT and its applicability in a radiomics model to predict fracture malignancy. *Sci Rep* 2022;12. https://doi.org/10.1038/s41598-022-10807-7
2. Lenchik L, Heacock L, Weaver AA, et al. Automated segmentation of tissues using CT and MRI: A systematic review. *Acad Radiol* 2019;26. https://doi.org/10.1016/j.acra.2019.07.006
3. Patel HD, Gupta M, Joice GA, et al. Clinical stage migration and survival for renal cell carcinoma in the United States. *Eur Urol Oncol* 2019;2:343-8. https://doi.org/10.1016/j.euo.2018.08.023
4. Volpe A, Panzarella T, Rendon RA, et al. The natural history of incidentally detected small renal masses. *Cancer* 2004;100:738-45. https://doi.org/10.1002/cncr.20025
5. Bhindi B, Thompson RH, Lohse CM, et al. The probability of aggressive versus indolent histology based on renal tumor size: Implications for surveillance and treatment. *Eur Urol* 2018;74:489-97. https://doi.org/10.1016/j.eururo.2018.06.003
6. Richard PO, Violette PD, Bhindi B, et al. Canadian Urological Association guideline: Management of small renal masses - Full-text. *Canadian Urological Association Journa*l 2022;16. https://doi.org/10.5489/cuaj.7763
7. McAlpine K, Breau RH, Stacey D, et al. Shared decision-making for the management of small renal masses - development and acceptability testing of a novel patient decision aid. *Canadian Urological Association Journal*. Published online 2020. https://doi.org/10.5489/cuaj.6575
8. Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Med Image Anal* 2021;67. https://doi.org/10.1016/j.media.2020.101821
9. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* 2006;31. https://doi.org/10.1016/j.neuroimage.2006.01.015
10. Baroudi H, Brock KK, Cao W, et al. Automated contouring and planning in radiation therapy: What Is 'clinically acceptable'? *Diagnostics* 2023;13. https://doi.org/10.3390/diagnostics13040667
11. Pandey M, Gupta A. Tumorous kidney segmentation in abdominal CT images using active contour and 3D-UNet. *Ir J Med Sci* 2023;192. https://doi.org/10.1007/s11845-022-03113-8
12. Lin Z, Cui Y, Liu J, et al. Automated segmentation of kidney and renal mass and automated detection of renal mass in CT urography using 3D U-Net-based deep convolutional neural network. *Eur Radiol* 2021;31. https://doi.org/10.1007/s00330-020-07608-9
13. Han JH, Kim BW, Kim TM, et al. Fully automated segmentation and classification of renal tumors on CT scans via machine learning. *BMC Cancer* 2025;25:173. https://doi.org/10.1186/s12885-025-13582-6
14. Bartko JJ. Measurement and reliability: Statistical thinking considerations. *Schizophr Bull* 1991;17. https://doi.org/10.1093/schbul/17.3.483
15. Zhang Y, Paulson E, Lim S, et al. A patient-specific autosegmentation strategy using multi-input deformable image registration for magnetic resonance imaging-guided online

adaptive radiation therapy: A feasibility study. *Adv Radiat Oncol* 2020;5. https://doi.org/10.1016/j.adro.2020.10.004

FIGURES AND TABLES

Figure 1. Study flow chart of cases.

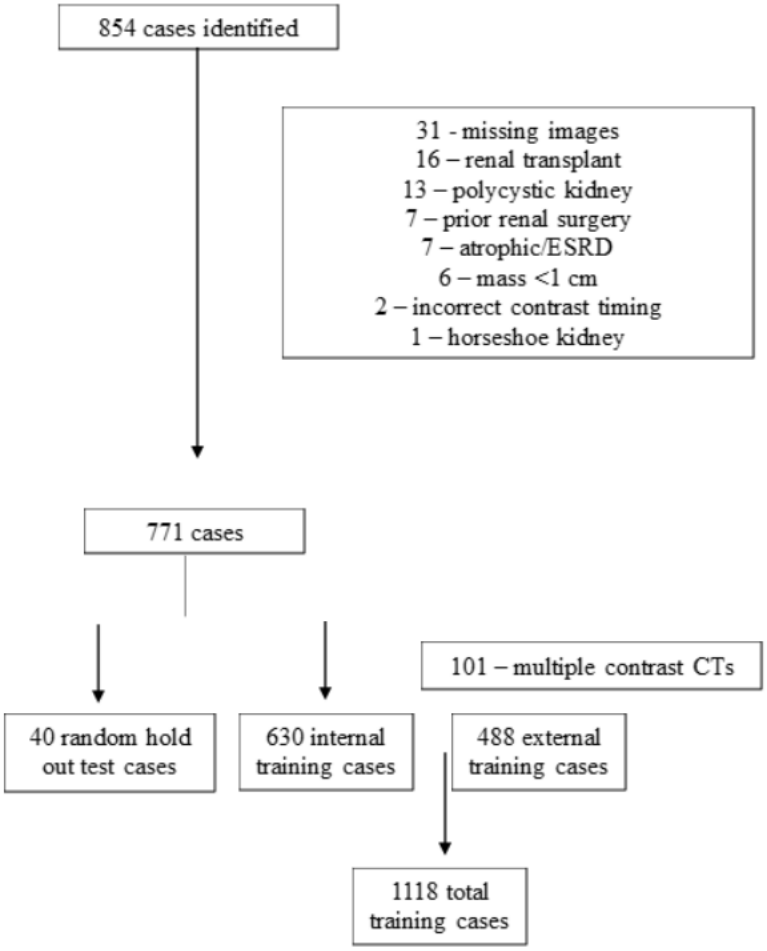


Figure 2. Case with high scores and interrater reliability between automated segmentation and radiologist segmentor. DICE coefficients between segmentors of 0.91–0.93.

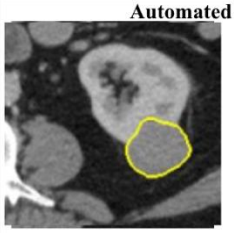
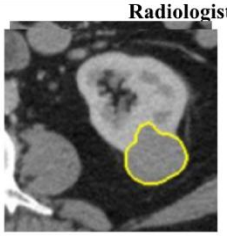
	Automated	Radiologist	Urologist
			
Mean acceptability score (out of 5)	5	4.5	3

Figure 3. Case with automated segmentation outperforming radiologist and urologist segmentors. DICE coefficients between segmentors of 0.74–0.86.

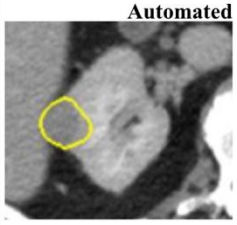
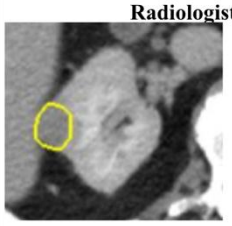
	Automated	Radiologist	Urologist
			
Mean acceptability score (out of 5)	4.5	2.5	3.5

Figure 4. Case of poor segmentation by automated segmentation. DICE coefficients of 0.1 between automated and manual segmentations. DICE coefficient of 0.87 between radiologist and urologist segmentors. The automated model has only segmented the more solid appearing potion of the tumor leaving out the cystic area.

	Automated	Radiologist	Urologist
			
Mean score (out of 5)	1	4	3.5

Table 1. Tumor characteristics	
	Number of tumors n=39
Kidney involved – Left – Right – Bilateral	 19 (49%) 17 (44%) 3 (8%)
Polarity of dominant renal tumor – Upper pole – Interpolar – Lower pole	 12 (31%) 15 (38%) 12 (31%)
Depth of dominant tumor – Exophytic – Endophytic	 18 (46%) 21 (54%)
Location of dominant tumor – Anterior – Poster	 20 (51%) 19 (49%)

Table 2. Time to segment			
Segmentor	**Time (median, s)**	**IQR**	**p**
Urologist	800.0	492.0–1538.0	<0.001
Radiologist	450.0	318.8–551.2	
Automated	152.4	120.9–177.8	

IQR: interquartile range.

Table 3. Inter-rater reliability				
Segmentor 1	**Segmentor 2**	**Dice coefficient (median)**	**IQR**	**p**
Urologist	Radiologist	0.86	0.82–0.90	0.09
Urologist	Automated	0.89	0.85–0.93	
Radiologist	Automated	0.90	0.86–0.92	

IQR: interquartile range.

Table 4. Turing test results			
	True segmentor		
Scoring radiologist choice	**Urologist (%)**	**Radiologist (%)**	**Automated (%)**
Urologist	67.8	29.3	9.3
Radiologist	24.4	48.8	23.3
Automated	7.8	22.0	67.4