Automated Machine Learning Segmentation and Measurement of Urinary Stones on CT Scan

Rilwan Babajide*, Katerina Lembrikova*, Justin Ziemba, James Ding, Yuemeng Li, Antoine Selman Fermin, Yong Fan, and Gregory E. Tasian

OBJECTIVES To evaluate the performance of an engineered machine learning algorithm to identify kidney stones and measure stone characteristics without the need for human input.

METHODS We performed a cross-sectional study of 94 children and adults who had kidney stones identified on non-contrast CT. A previously developed deep learning algorithm was trained to segment renal anatomy and kidney stones and to measure stone features. The performance and speed of the algorithm to measure renal anatomy and kidney stone features were compared to the current gold standard of human measurement performed by 3 independent reviewers.

RESULTS The algorithm was 100% sensitive and 100% specific in detecting individual kidney stones. The mean stone volume segmented by the algorithm was smaller than that of human reviewers and had moderate overlap (Dice score: 0.66). There was substantial variation between human reviewers in total segmented stone volume (Jaccard score: 0.17) and volume of the single largest stone (Jaccard score: 0.33). Stone segmentations performed by the machine learning algorithm more precisely approximated stone borders than those performed by human reviewers on qualitative assessment.

CONCLUSION An engineered machine learning algorithm can identify and characterize stones more accurately and reliably than humans, which has the potential to improve the precision and efficiency of assessing kidney stone burden. UROLOGY 169: 41–46, 2022. © 2022 Elsevier Inc.

Selecting the appropriate management strategy for patients with urolithiasis is heavily dependent on findings from diagnostic imaging, particularly non-contrast CT (NCCT). NCCT remains the gold standard for identification and assessment of clinically important stone features due to its high sensitivity and specificity for detecting urinary stones.[1] The size and location of the calculi identified on diagnostic imaging form the foundation for many of the evidence-based guidelines for the management of urolithiasis.[1-3] For example, trial of passage is commonly recommended for small, distal ureteral stones, while percutaneous nephrolithotomy is often recommended for large kidney stones, the success of which is dependent on the size and complexity of the stone and renal anatomy.[1-3] However, stone characterization currently requires manual measurement, which is laborious, making clinical decision-making more challenging, and introduces inter-rater and intra-rater variation, making analyzing the large number of imaging studies performed for clinical trials problematic. These limitations can produce inaccurate ascertainment of stone burden, which may lead to inappropriate management strategies and suboptimal outcomes, particularly for clinicians who do not personally review images. For example, many radiologists report only the largest or most clinically salient stone in a kidney and no other clinically important features such as increased density of renal papillae, which are associated with stone activity but laborious to manually measure.[4]

Machine learning (ML) algorithms, which have been developed to detect urologic tumors, have the potential to automate the identification and characterization of urinary stones.[5-8] However, prior software programs to

Abbreviations: *ICC, intraclass correlation; IRR, inter-rater reliability; ML, machine learning; NCCT, non-contrast computed tomography*

* *These authors contributed equally to the work and share first co-authorship.*

Funding support: *This study was supported by P20DK127488 from the National Institute for Diabetes and Digestive Diseases and Kidney Diseases (NIDDK) of the National Institutes of Health (NIH). The NIH had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Declarations of interest: *None.*

From the University of Chicago Pritzker School of Medicine, Chicago, IL; the SUNY Downstate College of Medicine, Brooklyn, NY; the University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; the Department of Surgery, Division of Urology, Hospital of the University of Pennsylvania, Philadelphia, PA; the Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; the The Center for Biomedical Image Computing and Analytics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; the Department of Surgery, Division of Pediatric Urology, The Children's Hospital of Philadelphia, Philadelphia, PA; and the Department of Biostatistics, Epidemiology, and Informatics; Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Address correspondence to: Gregory E. Tasian M.D., M.S.c., M.S.C.E., Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104. E-mail: tasiang@chop.edu

Submitted: April 15, 2022, accepted (with revisions): July 17, 2022

measure urinary stones have yet to incorporate the advanced automation and analytics of ML.[9] These programs are slow,[10] proprietary,[11] require manual input,[10,12] or have multiple steps.[13] Additionally, these programs often segment (partition) the stone from the surrounding structures without considering other clinically important anatomical features such as hydronephrosis[11,14-19] and do not contain all the functionality needed for analysis and interpretation of images at a standard adequate for clinical care.[20]

A critical barrier to increasing the efficiency and efficacy of individualized care is the lack of fully automated analyses of diagnostic imaging that can precisely identify and measure kidney stone characteristics. Accordingly, the aim of this study was to evaluate the performance of an engineered ML algorithm to identify kidney stones and measure important stone characteristics (including size, volume, location, and density) without the need for human input.

MATERIALS AND METHODS

Study Design and Patient Population

A cross-sectional study of 94 randomly selected abdominal NCCT images from patients with confirmed urolithiasis from 2007 to 2020 at Children's Hospital of Philadelphia and the University of Pennsylvania was conducted (Fig. 1). Adult and pediatric patients were included in the study if they had at least 1 kidney or ureteral stone confirmed by NCCT imaging. NCCT scans were performed on both adult and pediatric patients as part of routine clinical care for symptomatic nephrolithiasis.

Patients were identified by retrospective query of the Kidney Stone Registry at both institutions. Patient characteristics including sex, age at diagnostic imaging, race, ethnicity, and BMI were abstracted from the electronic health record. Radiologic reports and physician notes were reviewed for stone size and location.

Manual Image Analysis

All NCCT images were evaluated independently by 3 research scientists with expertise in imaging analysis. Reviewers measured the same stones in 10 randomly selected scans, which were then reviewed by 2 attending urologists for accuracy, to establish the ground truth used to train the algorithm. Reviewers were blinded to each other's measurements. Human evaluation of NCCT conducted included manual segmentation and measurement of stone and renal anatomic characteristics as follows:

1. Reviewers identified and manually contoured (segmented) urinary stones and each kidney using the MRIcroGL image slicing program.[21] A 3-dimensional image of each area of hyperattenuation considered to be a stone was formed by segmenting the area in sequential 2-dimensional NCCT slices. The total volume of each stone in voxels, density of each stone, and the position of the central stone voxel were recorded. Similarly, 3-dimensional images of the kidney were formed from segmentations of every fifth NCCT slice.
2. Reviewers manually measured renal pelvis diameter, ureter diameter, stone size in 3 orthogonal dimensions (anteroposterior, transverse, and craniocaudal), and average stone density. Anteroposterior and transverse lengths were measured on the axial NCCT slice with the greatest total in-plane area on visual inspection and craniocaudal length was measured on the coronal slice with the greatest in-plane length on

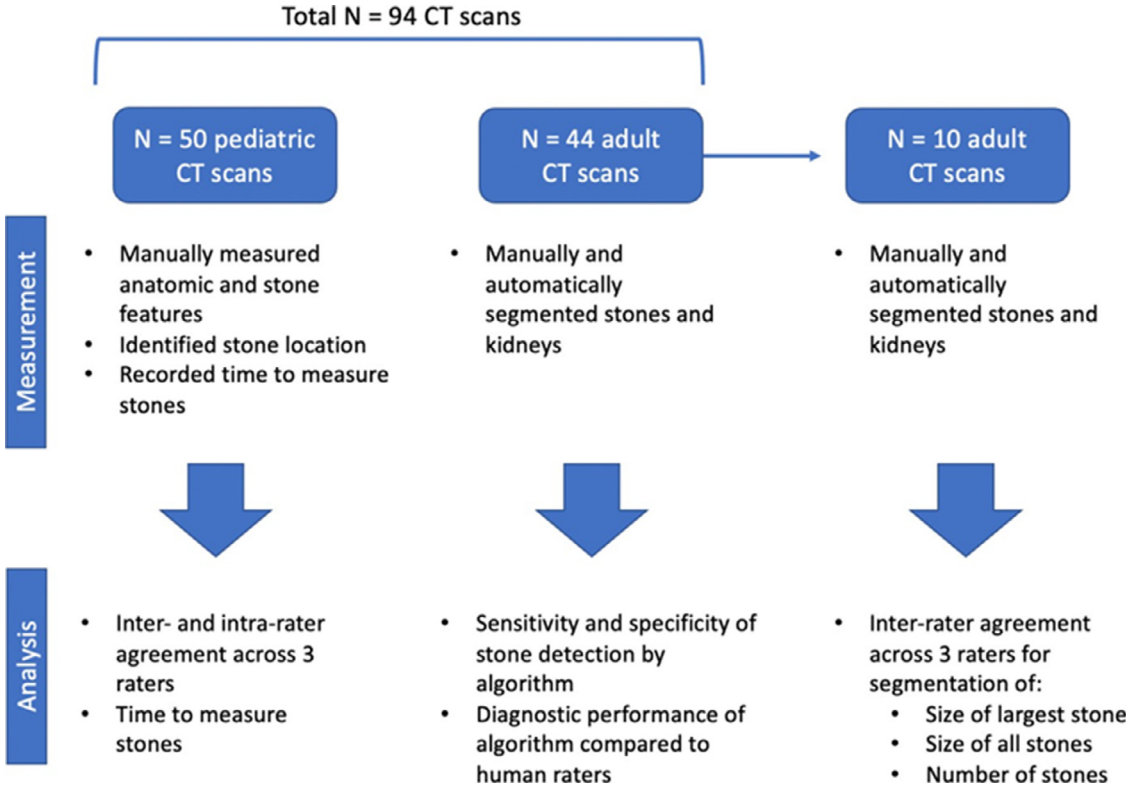

Figure 1. Composition of cohort by pediatric or adult status and purpose for analysis. (Color version available online.)

visual inspection. Stone location was reported as upper pole, mid pole, lower pole, renal pelvis, ureteropelvic junction, proximal ureter, mid ureter, distal ureter, or ureterovesical junction. Ureteral diameter was measured in the coronal view, distal to the most distal stone. If the most distal stone was in the ureterovesical junction, ureteral diameter was measured proximal to the stone. Reviewers also measured the greatest renal pelvis width visualized in the coronal view to quantify hydronephrosis. The time to measure stone size in 3 dimensions once identified was recorded. Anatomic and stone dimensions were measured twice by each reviewer to assess inter-rater and intra-rater agreement. All data were recorded in REDCap (Research Electronic Data Capture) hosted at the Children's Hospital of Philadelphia.[22]

Automated Image Analysis

We developed a ML algorithm for the identification and measurement of urinary stones on NCCT images via refinement of a deep-learning method previously used to segment the brain in MRI images.[23] The original algorithm was developed to automatically segment fine-grained brain structures from magnetic resonance images and achieved state-of-the-art brain structure segmentation performance.[23] Particularly, the deep learning model for segmenting stones was developed using PyTorch with a U-shaped network architecture, consisting of 5 encoder blocks, 1 bottleneck block, and 5 decoder blocks. The algorithm was distinct from the imaging program used to view CT images. Each block contained 2 convolution layers with a kernel size of 3. A max-pooling function with down-sampling was applied at the end of each encoder block and an up-sampling function was applied at the end of each decoder block. A classification loss for distinguishing stones from other structures was utilized at the bottleneck to enforce the model to capture the global information.[23] All the CT scans were resampled to have a spatial dimension of (3.22, 1.62, 1.62) in axial, sagittal, and coronal views. The CT scan intensity was clipped to the range of [-79, 304]. The network was trained based on input image patches of (96, 160, 160). The model was trained in 1000 epochs to automatically segment kidneys and stones using a set of 37 manually segmented NCCT images. This trained model was then applied to a testing set of 7 NCCT images to evaluate its performance for automatically identifying, segmenting, and outputting kidney stone measurements, including size and volume. These sample sizes were chosen based on the success of prior studies looking at machine learning technology for kidney stones and brain MRI.[23,24] The total run time for the algorithm required to perform these tasks was recorded.

Statistical Analysis

Inter-rater reliability (IRR) of manual measurement of stone and renal characteristics was assessed by calculating a 2-way mixed intra-class correlation (ICC) score. Intra-rater test-retest reliability was assessed using a 2-way random effects ICC score. We use the cutoffs proposed by Cicchetti[25] to determine level of agreement between reviewers, whereby IRR is considered to be poor for ICC values less than 0.40, fair for values from 0.40-0.59, good for values from 0.60-0.74, and excellent for values greater than 0.74.

The sensitivity and specificity of the ML algorithm for stone detection was calculated at the stone and voxel level. The number and location of detected stones were used as ground truth to determine the sensitivity and specificity of the ML algorithm for individual stone detection. Moreover, the shape and volume in voxels of stone segmentations done by reviewers were compared to those done by the ML algorithm for agreement and accuracy.

Image segmentation performance of the ML algorithm was evaluated using Dice score, Jaccard index, and global competitiveness index calculations between the manual measurement and the automatically segmented urinary stones and kidneys. Additionally, a qualitative visual analysis of stone segmentations performed by reviewers and the ML algorithm was performed to assess for accuracy of approximation of stone borders. Statistical analysis was performed using SPSS 25 (IBM Corporation 2017, United States). A P-value of $<.05$ was considered statistically significant. This study was approved by the CHOP IRB and informed consent was waived.

RESULTS

Ninety-four patients were included in this study, of which 50 were pediatric and 44 were adult. Forty-two patients (44.7%) had bilateral urinary stones. Characteristics of the patient population are reported in Table 1.

Supplemental Table 1 reports ICC scores for the anatomic and stone measurements reported by the 3 reviewers. IRR was poor for measurements of renal pelvis width (ICC: 0.30) and ureteral diameter (ICC: 0.46), good for transverse stone length (ICC: 0.72), and excellent for anteroposterior stone length (ICC: 0.78), craniocaudal stone length (ICC: 0.87), and stone location (ICC: 0.95). On the other hand, it was found that the intra-rater test-retest reliability for a single reviewer was excellent for all parameters: renal pelvis width (ICC: 0.97), ureter diameter (ICC: 0.97), transverse stone length (ICC: 0.81), anteroposterior stone length (ICC: 0.98), and craniocaudal stone length (ICC: 1.00). (Supplemental Table 2).

The sensitivity and specificity of the ML algorithm for detection of the central voxel of urinary stones were both 100%. However, the mean stone volume segmented by the algorithm was smaller than that of the human reviewers (744 vs 589 voxels, P = .5), but this difference was not statistically significant. Moreover, there was only moderate overlap between the manual and algorithm-based stone segmentations (Dice score: 0.66 ± 0.16). A qualitative comparison of stone segmentations performed by the reviewers with the ML algorithm revealed that

Table 1. Characteristics of study population

Characteristic	
Sex, n (%)	
Female	42 (44.7)
Male	52 (55.3)
Race, n (%)	
Black	17 (18.1)
White	70 (74.5)
Other	7 (7.4)
Ethnicity, n (%)	
Hispanic or Latino	7 (7.4)
Not Hispanic or Latino	87 (92.6)
Age, median (IQR), years	22 (15, 57)
BMI, median (IQR), kg/m^2	24.7 (20.1, 30.0)
CT slice thickness, median (IQR), mm	2.00 (2.00, 2.03)
Density, median (IQR), Hounsfield units	266.2 (206.4, 363.2)

BMI, body mass index.

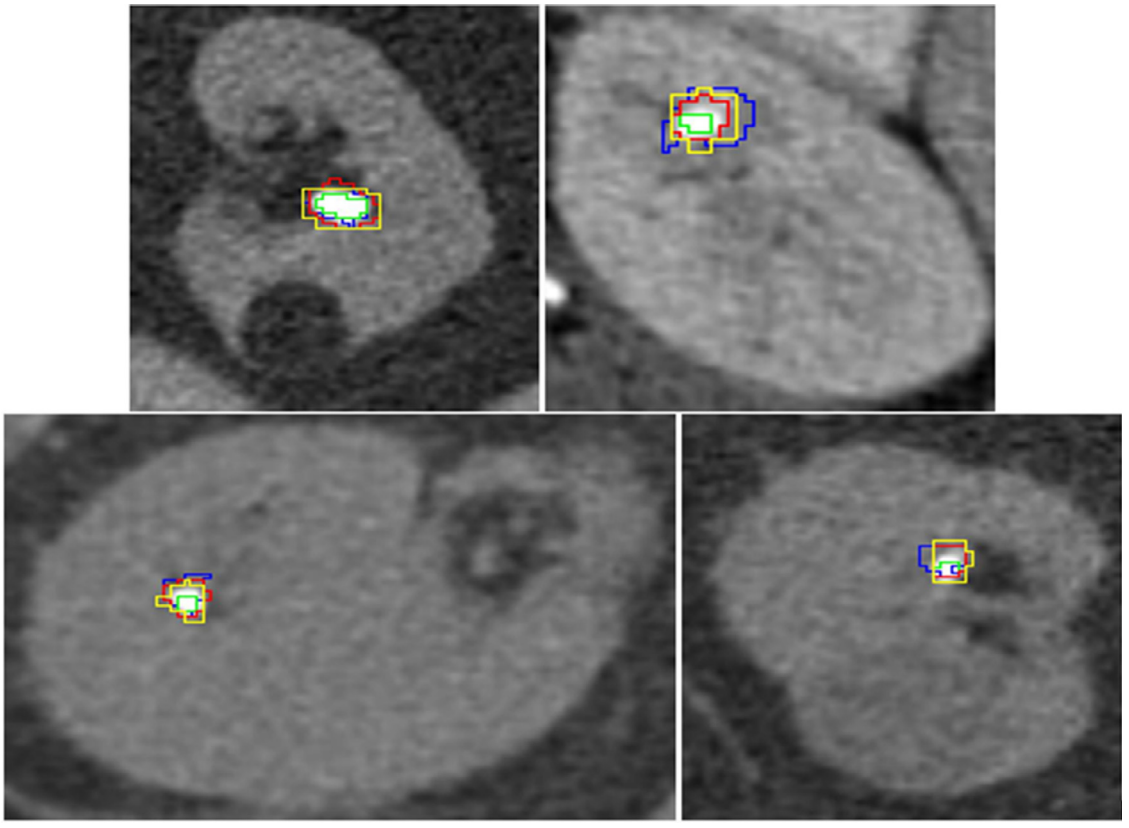

Figure 2. Manual segmentation results of three reviewers (red, green, and blue) and deep learning segmentation results (yellow) for four randomly selected cases.

the algorithm more accurately approximated stone borders and volume (Fig. 2). Furthermore, there was substantial variation between reviewers in total segmented stone volume (Jaccard score: 0.17) and volume of the single largest stone (Jaccard score: 0.33). Despite this variation in volume segmentation, the IRR for the number of stones detected by each reviewer was good (ICC: 0.87).

Additionally, the time required to measure stone dimensions was less for the ML algorithm than the humans. The median time for reviewers to measure each stone was 10 seconds (IQR 6, 20). There was a strong positive correlation between the number of stones present and the total time required for reviewers to analyze a NCCT image (Fig. 3). In contrast, the ML algorithm analyzed images in a mean time of 12 seconds and showed no variation in computation time based on the number of stones present.

COMMENT

In this study, we found that human reviewers had poor ability to reliably measure renal anatomy, good ability to accurately measure stone size, and excellent ability to identify stone location. However, the ML algorithm was nearly perfect at rapid stone identification, independent of the location, shape, size, or number of the stone(s). Notably, the mean stone volume segmented by the algorithm was smaller than that of the human reviewers, although not statistically significant, which was attributed to the superiority of the machine at identifying stone borders. These results demonstrate the success of an ML algorithm in identifying urinary stones faster, more accurately, and more reproducibly than humans.

Treatment decisions for patients with urolithiasis are heavily informed by stone size and location, making accurate characterization of stones and renal anatomy an issue of great clinical importance.[2,3] NCCT imaging provides useful information about these parameters, but extracting this information currently requires manual measurement. We demonstrated that human measurements are inefficient and exhibit substantial variability. In our cohort, we found that IRR was moderate or good for measurements of stone size and excellent for stone location. However, there was little agreement among reviewers for important anatomical measurements (ie, renal pelvis width and ureteral diameter). Studies on segmentation of gliomas on MRI have also shown this variation amongst physician-raters,[26,27] suggesting that this discordance is not unique to our study.

We also demonstrated that the ML algorithm we developed characterized urinary stones on NCCT more accurately than human reviewers. The algorithm was 100% sensitive and 100% specific in detecting individual kidney stones. While there was perfect agreement with human detection of stones and the algorithm's detection of stones, there was variance between the 2 in determining

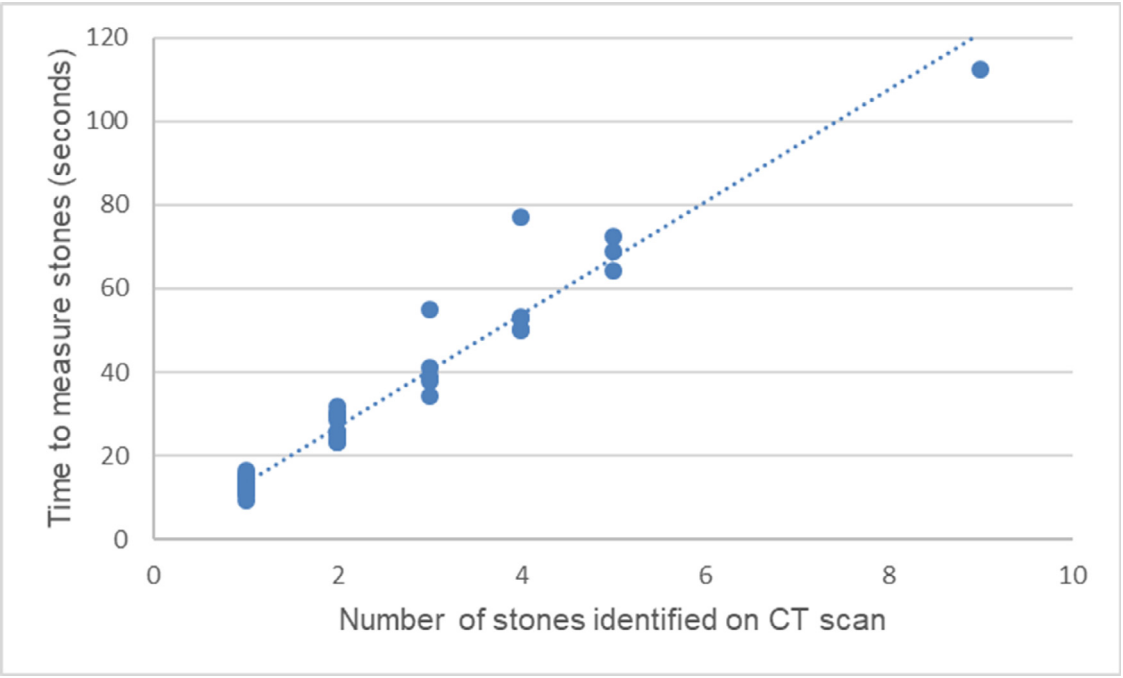

Figure 3. Total time to measure stones for human reviewers by number of stones identified on CT scan. Time to measure stones for the machine learning algorithm (not represented here) did not vary by number of stones. (Color version available online.)

stone volumes at the voxel level (Dice score: 0.66), with the ML algorithm identifying smaller stone volumes on average. The fact that this difference did not reach statistical significance may be explained in part because the ML algorithm was trained to measure stones using manual segmentations as ground truth. Nonetheless, visual inspection revealed that the ML algorithm showed greater accuracy in 3-dimensional segmentation of stones than human reviewers, demonstrating its ability to trace stone borders at the voxel level. Over time, as the algorithm, which will be made publicly available, learns and improves with additional data, it may demonstrate even better efficiency and reproducibility compared to human segmentations. Our results suggest not only that humans are poor at reliably measuring stone size, but that automating this process could produce rapid, yet highly accurate and reproducible results.

In addition to having greater accuracy than human reviewers, the ML algorithm identified and measured stones with greater efficiency. The algorithm identified and segmented all stones in an NCCT image in a mean of 12 seconds, regardless of stone location, size, or number. This was substantially less than the 10 seconds per stone that humans needed for measurements. Indeed, the time required for human measurement of stones is an underestimate as we recorded only the time to measure a stone once identified on NCCT and not the time that it took to open the file, load the image, and then visually locate the stone. Consequently, for patients with numerous urinary stones, use of the ML algorithm could save several minutes of evaluation time and decrease the resultant burden of a laborious task. These results have implications for improving the current standard of care by automating what is now a laborious and tedious process, thereby allowing radiologists and clinicians to focus on higher level applications of the more accurate measurements. Future refinement and analysis of this machine algorithm will incorporate additional renal and ureteral anatomy measurements, such as ureteral diameter for enhanced accuracy, as well as other clinically important features such as stone density.

This study has several limitations. First, our sample size was modest. However, it was chosen based on the success of prior studies looking at machine learning technology for kidney stones and brain MRI.[23,24] It is expected that better segmentation performance would be achieved if a larger training dataset is used to train the deep learning model. Second, NCCT images were assessed by research scientists who were not radiologists or urologists. It is possible that the performance of clinicians would be different. Finally, measurements by human reviewers were considered the ground truth for comparison, which we demonstrated are subject to inter-rater variability. Measurement of the actual stone passed or removed by surgery would provide the theoretical best gold standard. However, this is not practical as not all stone material is perfectly and uniformly removed at the time of surgery and removed samples are routinely sent for analysis. Furthermore, we chose not to use phantom models of stones as we preferred to develop and assess this algorithm under pragmatic real-world conditions. It is reasonable that in the future, ML algorithm measurements may be the ground truth for stone identification and measurement given its ability to detect stone borders at the voxel level.

CONCLUSION

Manual measurement of kidney stones is slow and inaccurate. An engineered machine learning algorithm can characterize stones more accurately, reliably, and quickly than humans, which has the potential to improve the precision and efficiency of assessing kidney stone burden.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.urology.2022.07.029.

References

1. Türk C, Petrik A, Sarika K, et al. EAU guidelines on diagnosis and conservative management of urolithiasis. *Eur Urol*. 2016;69:468–474. https://doi.org/10.1016/j.eururo.2015.07.040.
2. Assimos D, Krambeck A, Miller NL, et al. Surgical management of stones: american urological association/endourological society guideline, PART I. *J Urol*. 2016;196:1153–1160. https://doi.org/10.1016/j.juro.2016.05.090.
3. Assimos D, Krambeck A, Miller NL, et al. Surgical management of stones: american urological association/endourological society guideline, PART II. *J Urol*. 2016;196:1161–1169. https://doi.org/10.1016/j.juro.2016.05.091.
4. Iremashvili V, Li S, Penniston KL, et al. Role of residual fragments on the risk of repeat surgery after flexible ureteroscopy and laser lithotripsy: single center study. *J Urol*. 2019;201:358–363. https://doi.org/10.1016/j.juro.2016.05.091.
5. Kim JK, Yook IH, Choi MJ, et al. A performance comparison on the machine learning classifiers in predictive pathology staging of prostate cancer. *Stud Health Technol Inform*.. 2017;245. https://doi.org/10.3233/978-1-61499-830-3-1273. 1273-1273.
6. Algohary A, Viswanath S, Shiradkar R, et al. Radiomic features on MRI enable risk categorization of prostate cancer patients on active surveillance: preliminary findings. *J Magn Reson Imaging*.. 2018;48:818–828. https://doi.org/10.1002/jmri.25983.
7. Ginsburg SB, Algohary A, Pahwa S, et al. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: Preliminary findings from a multi- institutional study. *J Magn Reson Imaging*.. 2017;46:184–193. https://doi.org/10.1002/jmri.25562.
8. Zheng H, Ji J, Zhao L, et al. Prediction and diagnosis of renal cell carcinoma using nuclear magnetic resonance-based serum metabolomics and self-organizing maps. *Oncotarget*.. 2016;7:59189–59198. https://doi.org/10.18632/oncotarget.10830.
9. Ziemba JB, Li P, Gurnani R, et al. A user-friendly application to automate CT renal stone measurement. *J Endourol*. 2018;32:685–691. https://doi.org/10.1089/end.2018.0326. 2018.
10. Demehri S, Steigner ML, Sodickson AD, et al. CT-based determination of maximum ureteral stone area: a predictor of spontaneous passage. *Am J Roentgenol*. 2012;198:603–608. https://doi.org/10.2214/AJR.11.7276.
11. Bell JR, Posielski NM, Penniston KL, et al. Automated computer software compared with manual measurements for CT-based urinary stone metrics: an evaluation study. *J Endourol*. 2018;32:455–461. https://doi.org/10.1089/end.2017.0787.
12. Jendeberg J, Geijer H, Alshamari M, et al. Prediction of spontaneous ureteral stone passage: automated 3D-measurements perform equal to radiologists, and linear measurements equal to volumetric. *Eur Radiol*. 2018;28:2474–2483. https://doi.org/10.1007/s00330-017-5242-9.
13. Viswanath K, Gunasundari R. Analysis and implementation of kidney stone detection by reaction diffusion level set segmentation using xilinx system generator on FPGA. *VLSI Des*.. 2015;2015:e581961. https://doi.org/10.1155/2015/581961.
14. Duan X, Wang J, Qu M, et al. Kidney stone volume estimation from computerized tomography images using a model based method of correcting for the point spread function. *J Urol*. 2012;188:989–995. https://doi.org/10.1016/j.juro.2012.04.098.
15. Lee SR, Jeon HG, Park DS, et al. Longitudinal stone diameter on coronal reconstruction of computed tomography as a predictor of ureteral stone expulsion in medical expulsive therapy. *Urology*.. 2012;80:784–789. https://doi.org/10.1016/j.urology.2012.06.032.
16. Lidén M, Andersson T, Broxvall M, et al. Urinary stone size estimation: a new segmentation algorithm-based CT method. *Eur Radiol*. 2012;22:731–737. https://doi.org/10.1007/s00330-011-2309-x.
17. Patel SR, Stanton P, Zelinski N, et al. Automated renal stone volume measurement by noncontrast computerized tomography is more reproducible than manual linear size measurement. *J Urol*. 2011;186:2275–2279. https://doi.org/10.1016/j.juro.2011.07.091.
18. Selby MG, Vrtiska TJ, Krambeck AE, et al. Quantification of asymptomatic kidney stone burden by computed tomography for predicting future symptomatic stone events. *Urology*.. 2015;85:45–50. https://doi.org/10.1016/j.urology.2014.08.031.
19. Zorba OÜ, Ogullar S, Yazar S, et al. CT-based determination of ureteral stone volume: a predictor of spontaneous passage. *J Endourol*. 2016;30:32–36. https://doi.org/10.1089/end.2015.0481.
20. Kurtzman JT, Song L, Ross ME, et al. Urology consultation and emergency department revisits for children with urinary stone disease. *J Urol*. 2018;200:180–186. https://doi.org/10.1016/j.juro.2018.02.069.
21. Rorden C, Brett M. Stereotaxic display of brain lesions. *Behav Neurol*. 2000;12:191–200. https://doi.org/10.1155/2000/421719.
22. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95: 103208. https://doi.org/10.1016/j.jbi.2019.103208.
23. Li Y, Li H, Fan Y. ACEnet: anatomical context-encoding network for neuroanatomy segmentation. *Med Image Anal*. 2021;70: 101991. https://doi.org/10.1016/j.media.2021.101991.
24. Thein N, Nugroho HA, Adji TB, et al. An image preprocessing method for kidney stone segmentation in CT scan images. *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*. 2018147–150. https://doi.org/10.1109/CENIM.2018.8710933.
25. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284–290. https://doi.org/10.1037/1040-3590.6.4.284.
26. Visser M, Muller DMJ, van Dujin RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage Clin*. 2019;22: 101727. https://doi.org/10.1016/j.nicl.2019.101727.
27. Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;56:4557–4577. https://doi.org/10.1088/0031-9155/56/14/021.