JOURNAL OF ENDOUROLOGY
Volume 37, Number 8, August 2023
© Mary Ann Liebert, Inc.
Pp. 948–955
DOI: 10.1089/end.2023.0066

Fully Automated Longitudinal Assessment of Renal Stone Burden on Serial CT Imaging Using Deep Learning

Pritam Mukherjee, PhD,[1] Sungwon Lee, MD, PhD,[1] Daniel C. Elton, PhD,[1]
Stephen Y. Nakada, MD, FACS, FRCS,[2] Perry J. Pickhardt, MD,[2,*] and Ronald M. Summers, MD, PhD[1,*]

Abstract

Purpose: Use deep learning (DL) to automate the measurement and tracking of kidney stone burden over serial CT scans.

Materials and Methods: This retrospective study included 259 scans from 113 symptomatic patients being treated for urolithiasis at a single medical center between 2006 and 2019. These patients underwent a standard low-dose noncontrast CT scan followed by ultra-low-dose CT scans limited to the level of the kidneys. A DL model was used to detect, segment, and measure the volume of all stones in both initial and follow-up scans. The stone burden was characterized by the total volume of all stones in a scan *(SV)*. The absolute and relative change of *SV*, (*SVA* and *SVR*, respectively) over serial scans were computed. The automated assessments were compared with manual assessments using concordance correlation coefficient (CCC), and their agreement was visualized using Bland–Altman and scatter plots.

Results: Two hundred twenty-eight out of 233 scans with stones were identified by the automated pipeline; per-scan sensitivity was 97.8% (95% confidence interval [CI]: 96.0–99.7). The per-scan positive predictive value was 96.6% (95% CI: 94.4–98.8). The median *SV*, *SVA*, and *SVR* were 476.5 mm^3, −10 mm^3, and 0.89, respectively. After removing outliers outside the 5th and 95th percentiles, the CCC measuring agreement on *SV*, *SVA*, and *SVR* were 0.995 (0.992–0.996), 0.980 (0.972–0.986), and 0.915 (0.881–0.939), respectively

Conclusions: The automated DL-based measurements showed good agreement with the manual assessments of the stone burden and its interval change on serial CT scans.

Keywords: renal stone burden, interval change, longitudinal analysis, deep learning

Introduction

The prevalence of kidney stones in the United States is said to be 10.6% in men and 7.1% in women.[1] Recurrent stones are also common, exceeding 30% to 40% at 5 years.[2] A key component in the management of symptomatic urolithiasis is serial assessment of the patient's stone burden. The American Urological Association guidelines suggest using renal imaging for periodic follow-up and assessment.

Nonenhanced CT can be used to accurately diagnose and quantify kidney stones.[3] Although other imaging modalities, such as kidney, ureter, and bladder radiography, and ultrasound are often used due to cost or availability constraints, they suffer from much lower sensitivity and/or specificity.[4–6] Furthermore, these modalities also fail to assess stone size accurately.[7–9] Therefore, noncontrast CT is the best modality for detecting stones and assessing their size.

Despite its accuracy, CT imaging incurs high cost and carries a radiation risk. To mitigate both cost and radiation risks, an ultra-low-dose (ULD) limited kidney protocol has been developed and validated for the follow-up of kidney stones.[10] Compared to a standard low-dose (SLD) CT, the

[1]Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, Maryland, USA.
[2]Department of Radiology, The University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA.
*Co-senior authors.

ULD scan reduces radiation dosage significantly (target ~90%) by adjusting tube current range, slice thickness, and noise index, and costs less due to the limited scan area.

The CT-based follow-up allows the accurate assessment of stone sizes. However, linear measurements, either in one-dimensional (axial or coronal plane) or two-dimensional (reported as length × width), typically used in the clinic may have substantial interobserver variability and may be sensitive to window/level settings.[11] Automated computer-aided volumetric measurements mitigate this issue and are reproducible.[12,13] High interobserver variability (~15%–20%)[13] can make it difficult to assess changes in stone size over time. In addition, volumetric measurements are more sensitive to interval change than linear measurements (for a Δr change in radius, the change in volume $\Delta V \propto r^2 \times \Delta r$) and can enable better assessment of stone size change in follow-up.

The total stone burden, defined as the sum of the volumes of all stones in the kidneys, is a key metric of interest in evaluating kidney stones and assessing interval change between the initial and follow-up scans. Indeed, in predicting future symptomatic stone events, the total volume of all stones in a scan was found to be the most predictive, more so than other features such as the largest stone diameter or the number of stones.[14]

In this article, we modify and use the deep learning (DL)-based model proposed and validated in previous work to assess and track kidney stone burden between initial SLD scans and subsequent follow-up ULD CT scans.[15] While DL models have been developed and used for various applications, including detecting or segmenting kidney stones, inferring stone types, or predicting outcomes, to the best of our knowledge, the current work is the first that evaluates their usage in assessing stone burden and its interval change in serial CT scans.[16–19]

Materials and Methods

Patient population and CT protocol

This was a retrospective study from a single medical center, was Health Insurance Portability and Accountability Act-compliant, and was approved by the Institutional Review Board. The need for additional signed informed consent was waived.

A summary of the study plan is illustrated in Figure 1. The cohort comprised 113 patients evaluated for urolithiasis between 2006 and 2019 and followed up with an ULD CT scan between 2017 and 2019 (259 total CT scans). Every patient in the cohort had an initial SLD noncontrast CT scan. The median follow-up interval was 770 days, (intraquartile range [IQR]: [294, 1768] days). Fifteen, seven, and one patients had two, three, and four follow-up scans, respectively. The ULD follow-up scan coverage was limited to the level of the kidneys. Aggressive dose reduction of the ULD series was achieved by adjusting tube current range and noise index.

All ULD CT scans were performed without intravenous contrast on a 64-detector-row Discovery CT750 HD scanner (GE Healthcare) at 120 kV with variable tube current modulation (Smart mA). The mean effective dose was 4.1 mSv for the ULD compared to 13.4 mSv for the SLD ($p < 0.01$).

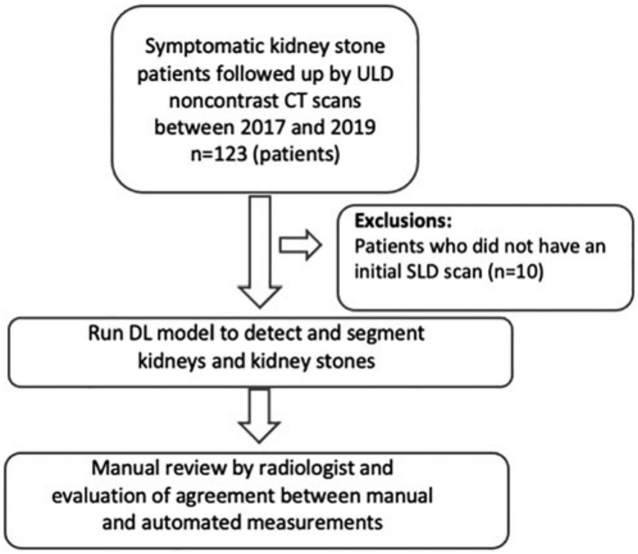
Symptomatic kidney stone patients followed up by ULD noncontrast CT scans between 2017 and 2019 n=123 (patients)
Exclusions: Patients who did not have an initial SLD scan (n=10)
Run DL model to detect and segment kidneys and kidney stones
Manual review by radiologist and evaluation of agreement between manual and automated measurements

FIG. 1. Outline of the study. DL = deep learning; SLD = standard low-dose scan; ULD = ultra-low-dose scan.

CT reconstruction and preprocessing

All images were reconstructed in 2.5 mm slice thickness at 1.25 mm intervals in the transverse (axial) plane. The SLD series was reconstructed using filtered back-projection, and the ULD series was reconstructed using model-based iterative reconstruction. One axial supine scan that included the entire kidney was selected from each study and resampled to 1 mm slice spacing.

Kidney stone detector algorithm

The stone detection and segmentation pipeline is a slightly modified version of the one presented in previous work.[15] The pipeline segments the kidneys using a three-dimensional

TABLE 1. COHORT CHARACTERISTICS

Characteristics	*Cohort*
Number of subjects (*n*)	113
Number of scans per patient (median, IQR)	2, (2, 2)
Time interval between scans (days) (median, IQR)	770, (294, 1727)
Number of kidney stones per scan (median, IQR)	2, (2, 5)
Total stone burden per scan *SV* (mm^3) (median, IQR)	476.5, (196.5, 2043)
Volume of individual stones (mm^3) (median, IQR)	40, (13, 160)[a]
Maximum diameter of individual stones (mm) (median, IQR)	6, (3.3, 10.2)[a]
Sex, *n* (%)	
Male	43 (38.1)
Female	70 (61.9)
Unknown	0
Age at initial scan (years) (median, IQR)	59.5, (48.25, 68)[b]

[a] Scans with numerous stones, nephrocalcinosis, or with stones confluent with adjacent stones were excluded.

[b] Age information was unavailable for four patients.

IQR = intraquartile range.

(3D) U-Net, and then candidate stones are identified within the kidney using thresholding at 130 HU, followed by connected components analysis on the thresholded binary mask. Finally, a convolutional neural network (CNN)-based classifier then predicts if the ‘‘candidate’’ is a stone.[15] In this work, we trained a new kidney segmenter using the nnU-Net[20] framework (a self-configuring method for deep learning-based biomedical image segmentation), using several publicly available datasets which contain segmentations of the kidney: Beyond the Cranial Vault, Multi-Modality Abdominal Multi-Organ Segmentation Challenge 2022,[21] the 2021 Kidney and Kidney Tumor Segmentation Challenge,[22] and Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation challenge (FLARE) 2021[23] and 2022. For each dataset, we only used the kidney segmentations from the labeled examples in the respective training sets to use as ground truth for training the kidney segmenter.

After training the nnU-Net model, the kidney segmenter was quantitatively validated on 20 validation cases from the FLARE 2022 dataset for which the ground truth was available. Since the stone detection and segmentation pipeline was primarily trained and used for small stones, it rejected large stones as false positives using a volume threshold criterion;[15] however, we removed the volume threshold to account for the large stones in our cohort. The CNN classifier, being trained on primarily small stones also tended to reject large stones as false positives—to mitigate that, the CNN was not used on stone candidates with volume greater than or equal to 250 mm^3. We also impose a minimum stone size criterion since the high noise in the ULD scans causes single-voxel false-positive detections.

We reject all stone candidates with volume <3 mm^3 (equivalent spherical diameter of 1.8 mm) since smaller stones are often not considered to be clinically significant, and it helps us reduce the false-positive rate in the noisy ULD scans.[24,25] The software outputs the volumetric size (mm^3), location (XYZ coordinates, upper/lower pole, left/right kidney), and attenuation (mean, median, standard deviation, maximum HU, HU) of kidney stones.[15]

Manual screening of the automatically detected stones

A board-certified radiologist (13 years of experience) manually reviewed the results of the kidney stone detector to identify false-positive detections and missed stones. Ground truth segmentations were obtained by correcting the automated segmentations: missed stones were segmented using a 130 HU threshold paint brush module, and false-positive

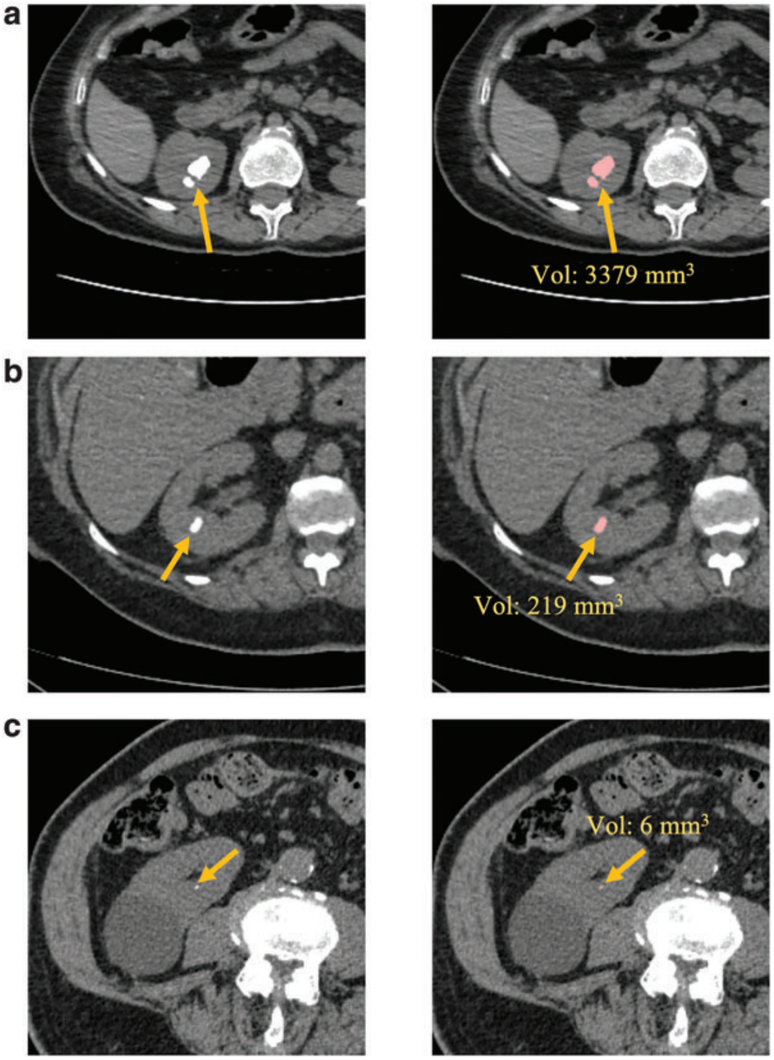

FIG. 2. Examples of stone detection and segmentation by the automated pipeline. The *left column* shows the original CT (windowed) while the *right column* shows the stone segmentations overlaid on the CTs. The *yellow arrows mark* the position of the stones Subfigures **(a)** – **(c)** show correctly detected stones of various sizes, from large **(a)** to very small **(c)**. Color images are available online.

stone segmentations were removed using 3D Slicer (version 4.10.2) software. The stone burden per scan was assessed based on the ground truth segmentation, as described in the following section.

Assessing stone burden

We consider the total volume of all stones *(SV)* as the primary measure of stone burden. An individual stone in the initial scan may remain stable, grow, or disappear in the follow-up scan, or new stones can appear on the follow-up scan. For assessing the change in stone burden, we considered two statistics: the absolute and relative change in total stone volume. Given a pair of consecutive scans (initial and follow-up) of the same patient, we computed the (signed) absolute difference $SVA = SV_{FU} - SV_{initial}$ and the relative change $SVR = SVA/SV_{initial}$ for the manual and the automated measurements. If $SV_{initial}$ was zero, *SVR* was set to ''undefined'' and dropped in subsequent statistical analysis.

Evaluating agreement between manual and automated measurements

First, we evaluate the agreement between the manual and automated measurements of *SV* considering every scan in a cohort as an individual sample. For *SVA* and *SVR*, every pair of consecutive scans from each patient is considered as a sample. We use Lin's concordance correlation coefficient (CCC) to evaluate the agreement between manual and automated measurements of *SV*, *SVA*, and *SVR*. We also trim our data to remove outliers beyond the 5th and 95th percentiles while calculating the *SV*, *SVA*, and *SVR*, since CCC is sensitive to outliers and non-normality of data due to its reliance on the squared distance function

We also created Bland–Altman[26] and scatter plots to visualize the agreement between manual and automated measurements for *SV*, *SVA*, and *SVR*.

Statistics

All statistics were done with R version 4.2. For computing CCC, we used the ''DescTools'' (version 0.99.46) libraries in R. The ''BlandAltmanLeh'' (0.3.1) and ''ggplot'' (version 3.3.6) libraries were used to construct the Bland–Altman plots.

Results

Table 1 summarizes key characteristics of the cohort. The median number of stones per scan was 2 (IQR: [2, 5]) and the median *SV* was 476.5 mm^3 (IQR: [196.5, 2043] mm^3) per scan.

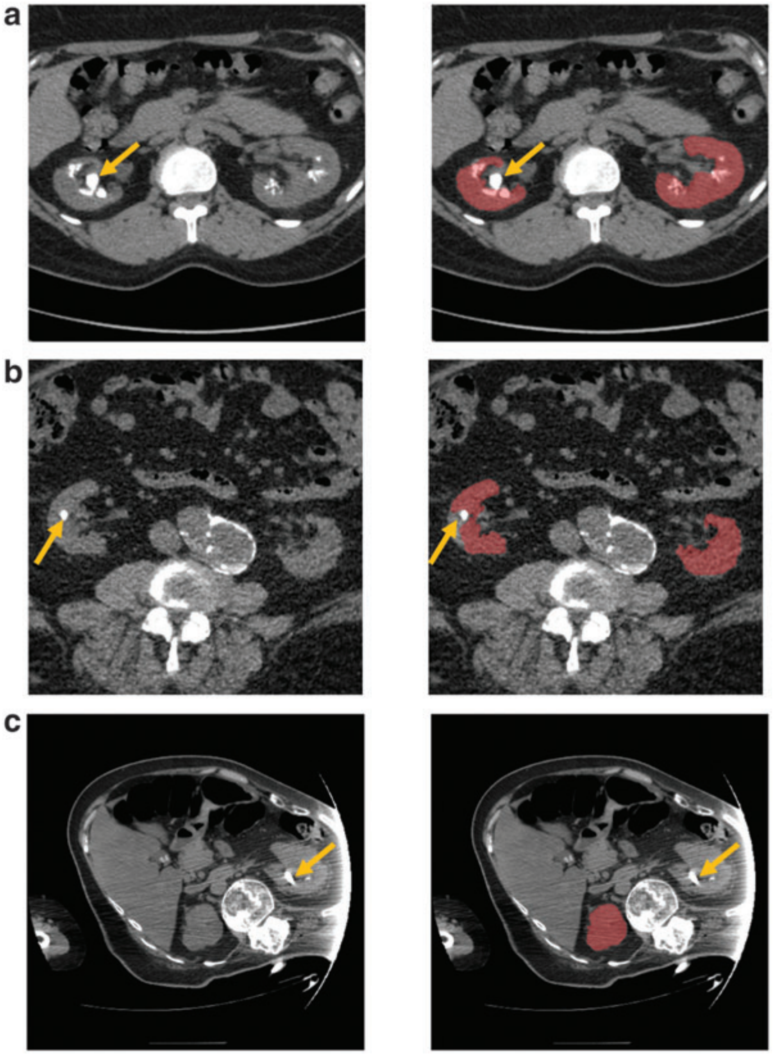

FIG. 3. Examples of stones missed by the automated pipeline. The *left column* shows the original CT (windowed) while the *right column* shows the kidney segmentations overlaid on the CTs. The *yellow arrows mark* the position of the missed stones. In **(a)** and **(b)**, the kidneys are undersegmented resulting in missed stones. In **(c)**, the left kidney is missed completely, possibly due to severe imaging artifacts, resulting in missed stones in the left kidney. Color images are available online.

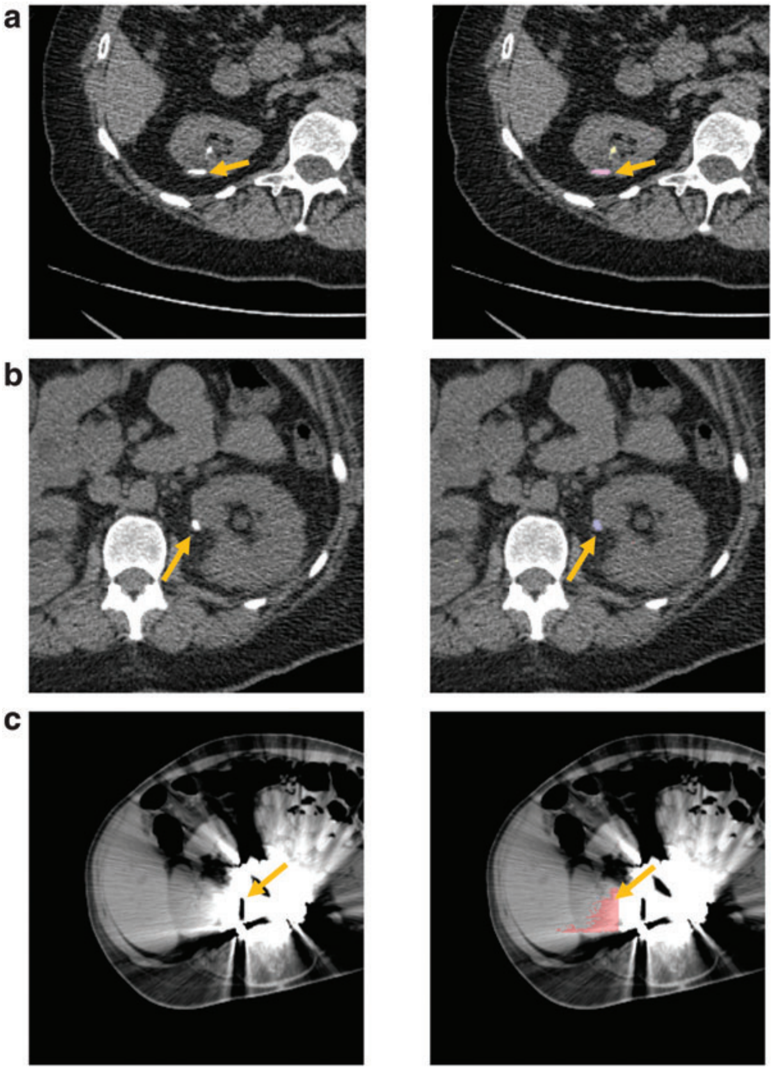

FIG. 4. Examples of false-positive stone detection by the automated pipeline. The *left column* shows the original CT (windowed) while the *right column* shows the stone segmentations overlaid on the CTs. The *yellow arrows mark* the position of the stones. In **(a)** and **(b)**, the detections are likely due to calcification in the wall of a cyst and a proximal ureteral calculus (not truly a false positive), respectively. In **(c)**, the false positive is caused due to severe artifacts in the image. Color images are available online.

The kidney segmenter achieved a mean Dice coefficient of 0.968 (std: 0.030) on the 20 test cases obtained from the FLARE validation set. Qualitative evaluation also showed adequate performance even in the presence of large stones, compared to the previously validated method.

Manual assessment identified 233 (out of 259) scans with at least one stone, of which 228 were identified by the DL pipeline (per scan sensitivity 97.8% [95% confidence interval {CI}: 96.0–99.7]). Two hundred twenty-eight out of 236 scans identified by the DL pipeline had a true stone (positive predictive value per scan was 96.6% [95% CI: 94.4–98.8]). Fifty-one out of 233 scans had numerous stones, nephrocalcinosis, or had stones that were confluent with adjacent stones, making it difficult to discern individual stones. Barring these cases, 726/830 stones greater than or equal to 3 mm^3 in volume were detected by the DL with a per-stone sensitivity of 87.5% (95% CI: 85.2–89.7); the median number of false positives per scan was 0 (IQR: [0, 1], mean: 0.72).

False-positive stone detections were due to dense renal calcification (in 23 scans), cyst-related calcification (15), coarse noise (14), metal artifacts (6), foreign bodies such as metal or catheter (6), atherosclerotic plaque (4), and adjoining vertebra, rib, or other extraneous features outside the

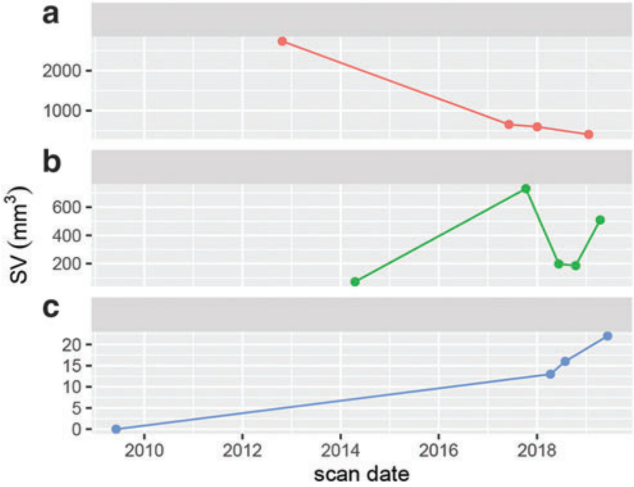

FIG. 5. Total stone burden (SV) trajectories over time for three selected patients. Patient **(a)** is male, age 79 years on the initial scan date, **(b)** is female, age 46 years on the initial scan date, and **(c)** is female, age 79 on the initial scan date. The SV decreases steadily for **(a)**, fluctuates for **(b)** and increases steadily for **(c)**. Color images are available online.

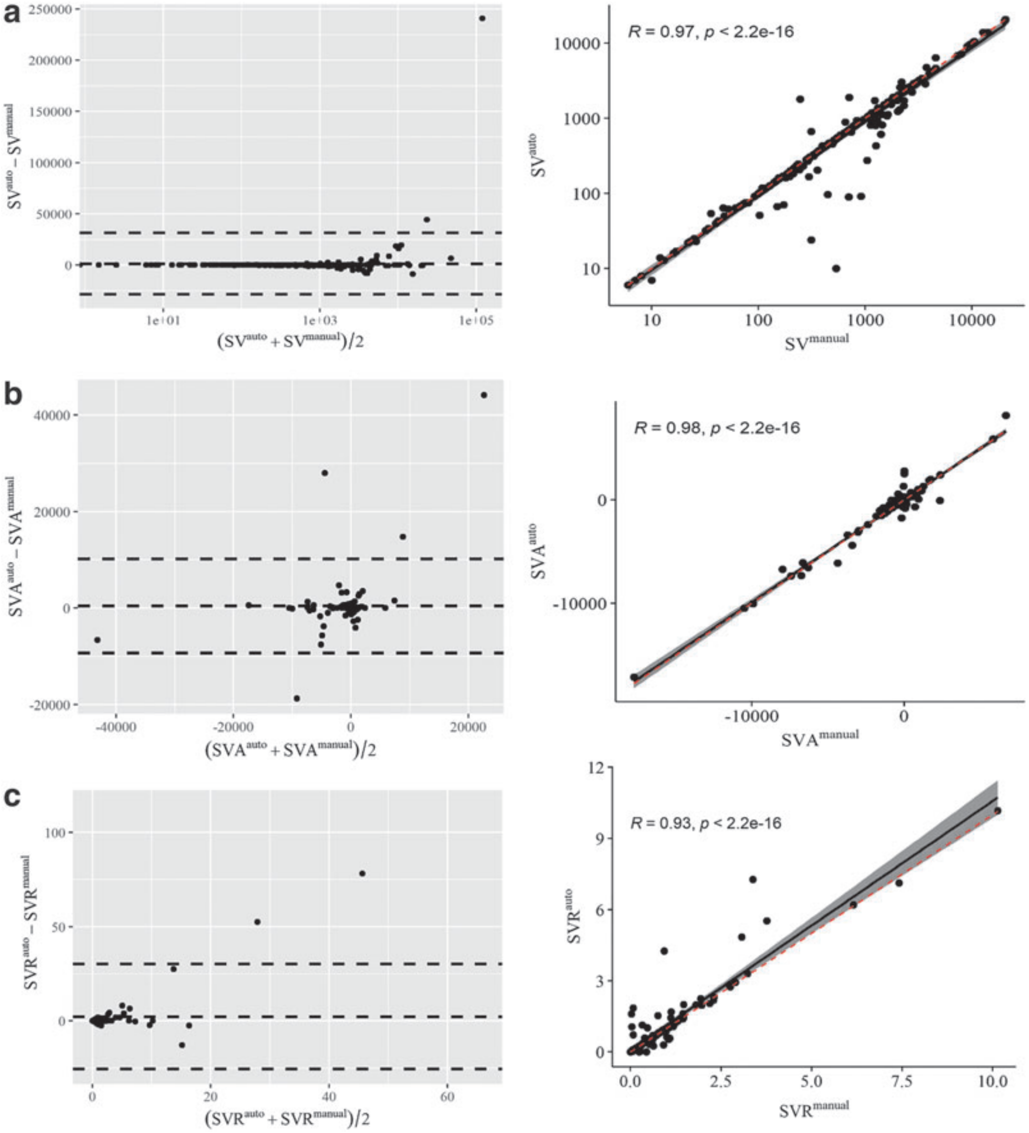

FIG. 6. Agreement between the automated and the manual measurements for **(a)** SV, **(b)** SVA, and **(c)** SVR. The *left column* shows the Bland–Altman plots while the *right column* shows scatter plots with automated and manual measurements. The *upper* and *lower horizontal lines* in the Bland–Altman plots represent the 95% CI. The *gray bands* in the scatter plots represent the 95% CI for the best fit regression line (the *solid black line*). The *red dashed line* denotes perfect agreement. Please note that outliers were removed for the scatter plots on the *right*. CI = confidence interval.

kidney caused by oversegmentation of the kidneys (3). Figure 2 shows a few examples of stone segmentation, while Figures 3 and 4 show some examples of missed stones and false-positive stone detections, respectively.

Figure 5 shows the trajectory of *SV* over multiple scans for three selected patients. The median *SVA* was –10 mm^3 (IQR: [–521, 99]) at an annualized rate of –5.2 mm^3/year. The median *SVR* was 0.89 (IQR: [0.33, 1.28]) with an annualized rate of 0.95 (IQR: [0.62, 1.15]). The CCC between the manual and automated measurements for *SV* was 0.995 (95% CI: 0.993–0.996). For *SVA* and *SVR*, the CCC was 0.980 (0.972–0.986) and 0.914 (0.881–0.939), respectively. The median (signed) difference between the manual and the automated measurement of *SV* was 0 mm^3 (IQR: [–1, 36.75] mm^3). For *SVA* and *SVR*, the median differences were 0 mm^3 (IQR: [–59, 19] mm^3) and 0 (IQR: [–0.04, 0.01]), respectively.

Figure 6 shows Bland–Altman and scatter plots showing the agreement between manual and automated measurements for *SV*, *SVA*, and *SVR*. The extreme outliers on Figure 6a were due to mis-segmentation of the kidneys and subsequent false positive stone detections outside the kidneys. Despite the presence of a few outliers, Figure 6 shows the strong agreement between the manual and automated measurements for *SV*, *SVA*, and *SV*.

Discussion

We used fully automated DL software to detect and measure kidney stones, assess the stone burden and track its change between initial and follow-up scans. We considered a cohort of patients who underwent a noncontrast SLD scan and were followed up by one or more noncontrast ULD scans limited to the level of the kidneys, targeted to reduce radiation dose by about 90%. While we focused on the total stone volume across both kidneys (*SV*) as the measure of stone burden in this article, our method can be used to measure and track stone burden for individual kidneys as well.

Requiring no human intervention, our method can be cheap and fast. Such automated tracking of stone burden can be useful for following-up individual patients with kidney stones, making treatment decisions, and for assessing the efficacy of stone-related interventions in individuals as well as large cohorts; stones are often prone to recurring—even after interventions such as extracorporeal shock-wave lithotripsy or percutaneous nephrolithotomy—and follow-up with ULD scans and the proposed automated assessment method may improve outcomes in the clinic.[19,25]

We found that the agreement of the automated measurements with manual measurements in terms of *SV* is high. The disagreement between manual and automated measurements mainly stemmed from either missed stones or from false-positive detections. In the presence of large stones or artifacts in the imaging, the kidney segmenter sometimes undersegmented the kidneys, resulting in missed stones in the undersegmented kidney regions (Fig. 3). False-positive detections were typically bright spots in the kidney caused by dense renal calcifications (which may be considered prestones and not true false positives), cyst-related calcification, CT imaging artifacts caused by noise, atherosclerotic plaque in renal vessels, or foreign bodies such as catheter or metal (Fig. 4).

While detecting, measuring, and tracking kidney stones can be a tedious task with significant intrareader and interreader variability, the automated method presents a viable alternative. Normally, the size of a kidney stone is measured by the maximal transverse diameter. However, this diameter is a poor indicator of the stone's volume because stones are rarely spheres[27] and because the linear measurement varies between readers[12] and window settings,[11] especially when the stones have complex 3D shapes. The automated DL-based method provides accurate volumetric measurements (validated in a prior study[15]) that can be more informative in the clinical setting. Further, small size changes over time are easier to assess with 3D volumetric measurements compared to linear size measurement. Focusing on the total stone volume has additional benefits when tracking stone burden change.

An individual stone in an initial scan can move, grow, split into two, or be passed, and new stones may form during the follow-up interval. Thus, tracking individual stones can be difficult and error-prone. Obtaining corresponding ground truth may be tedious, time-consuming, and even impossible to ascertain, except for relatively small cohorts. Furthermore, tracking individual stones may have limited value in clinical practice. Besides providing prognostic and clinical value, focusing on the total stone volume, and using the automated method allows us to examine large cohorts that would have been otherwise out of reach.

Some limitations should be noted. The cohort was obtained from a single institution in the United States, and the robustness of results to variations in CT acquisition or patient characteristics across different institutions could not be assessed. We plan to validate our method on external cohorts in the future. Second, although we used thin slice spacing CT scans (1.25 mm), partial volume averaging can affect the stone volume measurements—we have not considered its effect here since the ground truth assessments performed by the radiologist did not consider it either. Third, we have also not considered stone composition when computing the stone burden.[18] Finally, in computing *SVA* and *SVR*, we have assumed that *SV* measurements on SLD and ULD scans are directly comparable. While we do not expect any significant discrepancies between *SV* measurements made on SLD and ULD scans, this has not been validated in this study.

In summary, we used fully automated DL software to detect and measure kidney stones and track stone burden over multiple ULD follow-up scans. The automated measurements of stone burden and its change over follow-up scans show good agreement with their manual counterparts.

Authors' Contributions

P.M. and S.L.: conceptualization, data curation, formal analysis, methodology, software, writing–original draft, and writing–review and editing. D.C.E.: conceptualization, data curation, methodology, software, and writing–review and editing. S.Y.N.: data curation and writing–review and editing. P.J.P.: conceptualization, data curation, writing–review and editing, supervision, and project administration. R.M.S.: conceptualization, data curation, supervision, resources, project administration, and writing–review and editing.

Author Disclosure Statement

R.M.S. receives royalties from iCAD, Philips, ScanMed, PingAn, and Translation Holdings and has received research support from Ping An (CRADA). P.J.P. is an adviser or consultant for Zebra Medical Vision and Bracco Diagnostics, and shareholder in Cellectar, Elucent, and SHINE.

Funding Information

This research was supported by the Intramural Research Program of the National Institutes of Health, Clinical Center, Z01 CL040003 and Z01 CL040004 and we utilized the computational resources of the National Institutes of Health high-performance computing Biowulf cluster.

References

1. Scales CD, Smith AC, Hanley JM, et al. Prevalence of kidney stones in the United States. Eur Urol 2012;62(1): 160–165; doi: 10.1016/j.eururo.2012.03.052.
2. Johnson CM, Wilson DM, O'Fallon WM, et al. Renal stone epidemiology: A 25-year study in Rochester, Minnesota. Kidney Int 1979;16(5):624–631.
3. Vrtiska TJ. Quantitation of stone burden: Imaging advances. Urol Res 2005;33(5):398–402; doi: 10.1007/s00240-005-0490-6.
4. Chan VO, Buckley O, Persaud T, et al. Urolithiasis: How accurate are plain radiographs? Can Assoc Radiol J 2008; 59(3):131–134.
5. Fowler KA, Locken JA, Duchesne JH, et al. US for detecting renal calculi with nonenhanced CT as a reference standard. Radiology 2002;222(1):109–113; doi: 10.1148/radiol.2221010453.
6. Narepalem N, Sundaram CP, Boridy IC, et al. Comparison of helical computerized tomography and plain radiography for estimating urinary stone size. J Urol 2002;167(3):1235–1238.

7. Olcott EW, Sommer FG, Napel S. Accuracy of detection and measurement of renal calculi: In vitro comparison of three-dimensional spiral CT, radiography, and nephrotomography. Radiology 1997;204(1):19–25; doi: 10.1148/radiology.204.1.9205217.
8. Dundee P, Bouchier-Hayes D, Haxhimolla H, et al. Renal tract calculi: Comparison of stone size on plain radiography and noncontrast spiral CT scan. J Endourol 2006;20(12):1005–1009; doi: 10.1089/end.2006.20.1005.
9. Ray AA, Ghiculete D, Pace KT, et al. Limitations to ultrasound in the detection and measurement of urinary tract calculi. Urology 2010;76(2):295–300; doi: 10.1016/j.urology.2009.12.015.
10. Planz VB, Posielski NM, Lubner MG, et al. Ultra-low-dose limited renal CT for volumetric stone surveillance: Advantages over standard unenhanced CT. Abdom Radiol (NY) 2019;44(1):227–233; doi: 10.1007/s00261-018-1719-5.
11. Danilovic A, Rocha BA, Marchini GS, et al. Computed tomography window affects kidney stones measurements. Int Braz J Urol 2019;45(5):948–955; doi: 10.1590/S1677-5538.IBJU.2018.0819.
12. 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(6):2275–2279; doi: 10.1016/j.juro.2011.07.091.
13. Patel SR, Wells S, Ruma J, et al. Automated volumetric assessment by noncontrast computed tomography in the surveillance of nephrolithiasis. Urology 2012;80(1):27–31; doi: 10.1016/j.urology.2012.03.009.
14. 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(1):45–50.
15. Elton DC, Turkbey EB, Pickhardt PJ, et al. A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans. Med Phys 2022;49(4):2545–2554; doi: 10.1002/mp.15518.
16. Li D, Xiao C, Liu Y, et al. Deep segmentation networks for segmenting kidneys and detecting kidney stones in unenhanced abdominal CT images. Diagnostics 2022;12(8):1788; doi: 10.3390/diagnostics12081788.
17. Babajide R, Lembrikova K, Ziemba J, et al. Automated machine learning segmentation and measurement of urinary stones on CT scan. Urology 2022;169:41–46; doi: 10.1016/j.urology.2022.07.029.
18. Abraham A, Kavoussi NL, Sui W, et al. Machine learning prediction of kidney stone composition using electronic health record-derived features. J Endourol 2022;36(2):243–250; doi: 10.1089/end.2021.0211.
19. Hameed BMZ, Shah M, Naik N, et al. Application of artificial intelligence-based classifiers to predict the outcome measures and stone-free status following percutaneous nephrolithotomy for staghorn calculi: Cross-validation of data and estimation of accuracy. J Endourol 2021;35(9):1307–1313; doi: 10.1089/end.2020.1136
20. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–211; doi: 10.1038/s41592-020-01008-z.
21. Ji Y, Bai H, Yang J, et al. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation; 2022; arXiv:2206.08023.
22. 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;67:101821; doi: 10.1016/j.media.2020.101821.
23. Ma J, Zhang Y, Gu S, et al. Fast and low-GPU-memory abdomen CT organ segmentation: The FLARE challenge. Med Image Anal 2022;82:102616; doi: 10.1016/j.media.2022.102616.
24. Kang HW, Lee SK, Kim WT, et al. Natural history of asymptomatic renal stones and prediction of stone related events. J Urol 2013;189(5):1740–1746; doi: 10.1016/j.juro.2012.11.113.
25. Sorensen MD, Harper JD, Borofsky MS, et al. Removal of small, asymptomatic kidney stones and incidence of relapse. N Engl J Med 2022;387(6):506–513; doi: 10.1056/nejmoa2204253.
26. Bland JM, Altman D. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. Lancet 1986; 327(8476):307–310; doi: 10.1016/S0140-6736(86)90837-8.
27. Demehri S, Kalra MK, Rybicki FJ, et al. Quantification of urinary stone volume: Attenuation threshold–based CT method—A technical note. Radiology 2011;258(3):915–922.

Address correspondence to:
Ronald M. Summers, MD, PhD
Imaging Biomarkers and Computer-Aided Diagnosis Laboratory
Department of Radiology and Imaging Sciences
National Institutes of Health Clinical Center
Building 10, Room 1C224D, 10 Center Drive
Bethesda, MD 20892-1182
USA

E-mail: rms@nih.gov

Abbreviations Used

3D	= three-dimensional
CCC	= concordance correlation coefficient
CI	= confidence interval
CNN	= convolutional neural network
CT	= computed tomography
DL	= deep learning
FLARE	= Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation
HU	= Hounsfield units
IQR	= intraquartile range
SLD	= standard low dose
SV	= total stone burden, sum of volumes of all stones in a scan
SVA	= stone burden change (absolute)
SVR	= stone burden change (relative)
ULD	= ultra-low dose