Measurement of Posterior Acoustic Stone Shadow on Ultrasound Is a Learnable Skill for Inexperienced Users to Improve Accuracy of Stone Sizing

Jessica C. Dai, MD,¹ Barbrina Dunmire, MS,² Ziyue Liu, PhD,³ Kevan M. Sternberg, MD,⁴ Michael R. Bailey, PhD,² Jonathan D. Harper, MD,¹ and Mathew D. Sorensen, MD, MS^{1,5}

Abstract

Introduction: Studies suggest that the width of the acoustic shadow on ultrasound (US) more accurately reflects true stone size than the stone width in US images. We evaluated the need for training in the adoption of the acoustic shadow sizing technique by clinical providers.

Methods: Providers without shadow sizing experience were recruited and assigned in a stratified, alternating manner to receive a training tutorial (“trained”) or no intervention (“control”). Each conducted a baseline assessment of 24 clinical US images; where present, shadow width was measured using custom calipers. The trained group subsequently completed a standardized training module on shadow sizing. All subjects repeated measurements after ~1 week. Group demographics were compared using Fisher’s exact test. Measurements were compared to clinically reported stone sizes on corresponding CT and US using mixed-effects models. One millimeter concordance between shadow and CT size was compared using a generalized linear mixed-effects model.

Results: Twenty-six subjects were included. There was no significant difference between groups in demographics, clinical role, or US experience. Mean reported CT and US stone sizes were 6.8 ± 4.0 mm and 10.3 ± 4.1 mm, respectively. At baseline, there was no difference in shadow size measurements between groups ($p=0.18$), and shadow size was no more accurate than US stone size ($p=0.28$ trained; $p=0.81$ control), compared to CT. After training, overestimation bias of shadow size in the trained group decreased to 1.6 ± 0.5 mm ($p<0.01$), relative to CT. This was not significantly associated with clinical rank, US experience, or stone-measuring experience. One millimeter concordance with CT size significantly increased from 23% to 35% of stones after training ($p=0.01$). No significant improvement occurred in the control group.

Conclusion: Acoustic shadow sizing was readily adopted by inexperienced providers, but was not more accurate than reported US stone sizes without training. Education on shadow sizing may be warranted before clinical adoption.

Keywords: sizing, ultrasound, nephrolithiasis, posterior acoustic shadow, kidney stone, accuracy

Introduction

THE ROLE OF ULTRASOUND (US) in the management of nephrolithiasis remains limited by its user dependency and inaccuracy in stone sizing. Prior studies have demonstrated that stone size is overestimated on US by 2.2 mm on average, with even greater average overestimation of 3.3 mm for stones ≤ 5 mm.^{1–3} Consequently, over 20% of patients

could be inaccurately counseled regarding management options based on US stone size alone.⁴

Proposed strategies to improve stone sizing on US include minimizing gain, removing spatial compounding, placing the imaging focus at the stone depth, and employing harmonic imaging.^{5,6} These settings also improve visualization of the acoustic shadow, which has been proposed as an adjunctive measure of stone size on US. This has been demonstrated

¹Department of Urology, University of Washington, Seattle, Washington.

²Center for Industrial and Medical Ultrasound, University of Washington, Seattle, Washington.

³Department of Biostatistics, Indiana University, Indianapolis, Indiana.

⁴Division of Urology, University of Vermont, Burlington, Vermont.

⁵Division of Urology, Veterans Affairs Medical Center, Seattle, Washington.

in vitro and in human subjects to be more accurate than direct measurement of the stone on US images.^{6–8}

Prior studies evaluating this measure have included only reviewers familiar with the shadow sizing technique. Although thought to be easily adoptable in practice, the accuracy of this technique when used by clinicians has not yet been examined. We sought to determine if clinical providers familiar with US, but naive to the acoustic shadow measurement technique, could readily perform measurements with reasonable accuracy, with respect to CT stone size. We further assessed the effect of a brief training intervention on the accuracy of these measurements.

Methods

This was a prospective cohort study granted exempt status by the University of Washington Institutional Review Board (IRB). IRB approval was also obtained from Vermont Medical Center to share the clinical images used in this study.

Stone image sets

Forty-four sets of de-identified, clinically indicated US and low-dose noncontrast CT images of renal stones from a single institution (University of Vermont) were obtained as a subset from a previously published study by Sternberg and colleagues.⁵ The studies were performed within 24 hours of each other and gathered as part of a retrospective review for stone sizing. No clinical information about the patient images was provided. Of these, 24 were previously identified by 5 expert reviewers to have a visible posterior acoustic shadow, and were included in this study.⁸ Reference CT stone size was defined as the largest dimension of the largest stone in axial, sagittal, or coronal section. All measurements were made by a single reviewer. Reference US stone size was the largest measured dimension in the formal radiology report.

Shadow sizing protocol

The technique for posterior acoustic shadow measurement on US has been previously described.⁶ Each image was displayed on a MATLAB™ (MathWorks, Natick, MA) platform with the stone in question marked by an arrow. The interface also included two moveable guidelines and calipers. Subjects were instructed to use the guidelines as needed to help delineate the shadow borders and to use the caliper to measure the width of the posterior acoustic shadow ~1 cm posterior to the stone (Fig. 1). Subjects were blinded to their measurements. If no shadow was deemed present, this could be denoted and the image was skipped.

Study population

Clinical providers familiar with interpretation of US images, but inexperienced with the shadow measuring technique, were recruited. Subjects included urologists (attending, fellows, and residents), emergency medicine providers (attending and residents), and registered sonographers.

Study protocol

Subjects were queried regarding their weekly US exposure, comfort with US interpretation, experience measuring

FIG. 1. Example of one US image presented for shadow sizing. The stone in question is marked with an arrow. Black guidelines extending from the stone to the edge of the image are used by the subject to delineate the boundaries of the posterior acoustic shadow. The white caliper extending across the guidelines records the width of the shadow. US = ultrasound.

stones on US, and prior exposure to the posterior acoustic shadow. All received a standardized introduction to the shadow sizing display interface, use of the guidelines and caliper, and the characteristic appearance of the posterior acoustic shadow on US. Each US image was then assessed by all subjects for the posterior acoustic shadow and baseline measurements were made.

Given the small size and heterogeneity of the cohort, subjects were stratified by clinical role and assigned in an alternating manner to receive either no additional training (“control group”) or a standardized, narrated, 15-minute Powerpoint®-based training module on the posterior acoustic shadow measurement (“trained group”). For both groups, shadow measurements on the same set of images were repeated about 1 week after baseline measurements.

Training module

The training module highlighted several considerations in the shadow sizing technique (Supplementary Data S1; Supplementary Data are available online at www.liebertpub.com/end). The importance of adequately identifying the shadow was emphasized, as it may be present over only part of the image and may not be present immediately behind the stone. In addition, other nearby structures could introduce shadowing that may mimic the stone shadow. Helpful techniques include tracing the shadow along the path of the US beam and examining where the shadow would cross the renal capsule. Detailed instructions were also given on use of the guidelines to project the shadow back to the stone, where to measure the shadow width relative to the stone, and orientation of the measurement caliper such that it is perpendicular to the US beam direction.

Statistical analysis

Average CT stone size, US stone size, and shadow size for both groups were calculated. Intraclass correlation (a measure of the consistency across subjects) for shadow size

TABLE 1. DEMOGRAPHIC CHARACTERISTICS AND PRIOR ULTRASOUND EXPERIENCE AMONG STUDY GROUPS

	Trained (n)	Untrained (n)	p
Total no.	13	13	
Clinical roles (attending vs nonattending)			1.0
Urology attending	2	3	
Urology fellow	1	1	
Urology resident	5	5	
Emergency medicine attending	1	0	
Emergency medicine resident	3	2	
Sonographer	1	2	
Mean age	35	35	0.96
Gender			0.69
Male	7	9	
Female	6	4	
No. of days between sizing sessions (standard deviation)	8±5.9	9±5.5	0.66
Mean number of ultrasounds performed, reviewed, or interpreted per week (standard deviation)			1.0
0–5	8	6	
>5	5	7	
Degree of comfort with ultrasound reading			0.7
“Not at all” or “Somewhat”	8	6	
“Fairly” or “Very”	5	7	
Prior experience measuring kidney stone size	9	6	0.43
Heard of posterior acoustic shadow	11	13	0.48

within groups was calculated. Demographics were compared between groups using Fisher's exact test and Student's *t*-test. Shadow size was compared to reported US and CT stone sizes using mixed-effects models to account for within-stone correlations and within-rater correlations. A generalized linear mixed-effects model was used to determine the 1 mm concordance between shadow and CT sizes. All analyses were performed in SAS 9.4 (SAS Institute, Cary, NC).

Results

Twenty-six subjects were recruited, with 13 subjects in each study group. There were no significant differences between groups in clinical rank, prior US experience, level of comfort with US interpretation, experience with stone sizing on US, age, gender, or time interval between shadow sizing sessions (Table 1).

Mean reported CT stone size, US stone size, and baseline (before training) shadow width measurements are listed in Table 2. Mean stone depth on US was 6.6 ± 2.3 cm. There was no significant difference in shadow measurements between the two groups ($p=0.18$). Mean overestimation bias in shadow size for both the control and trained groups was not significantly better than US stone size ($p=0.81$ and $p=0.28$, respectively) with respect to CT at baseline.

The effect of training on shadow size accuracy is summarized in Table 3. For the repeat (after training) session, shadow size measurements among the trained group were, on average, significantly more accurate than both shadow measurements from the control group ($p=0.005$) and reported US stone sizes ($p=0.01$), relative to CT. For both groups, intraclass correlation for shadow size also significantly improved from baseline (0.55–0.58 for control group, $p=0.02$; 0.64–0.67 for trained group, $p=0.01$).

Figure 2 shows size measurements for each stone within the two groups. Overall, shadow measurements from the trained group more closely approached the CT stone size. Within the trained group, 84.6% of subjects (11/13) showed a 1.0 ± 0.7 mm mean decrease in bias (relative to a mean baseline bias of 3.3 ± 1.6 mm) with respect to CT stone size, whereas only two subjects showed a mean increase in bias of 0.7 ± 0.5 mm after training (relative to a mean baseline bias of 1.6 ± 1.2 mm). Notably, one of these had the most accurate shadow size measurements of all subjects at baseline, and decreased in accuracy by only 0.3 mm, on average, after training. The proportion of stones with ≤ 1 mm concordance between shadow size and CT stone size significantly increased from 23% at baseline to 35% after training ($p=0.01$). Improvement in shadow size accuracy from baseline was not significantly associated with clinical rank, prior US experience, or experience measuring stones on US.

In the control group, there was no change in the mean bias of shadow size ($p=0.35$) or proportion of stones with ≤ 1 mm concordance between shadow size and CT stone size

TABLE 2. BASELINE (BEFORE TRAINING) SHADOW SIZE MEASUREMENTS COMPARED WITH THE MEAN REPORTED CT AND ULTRASOUND STONE SIZES

	CT stone size (mm)	US stone size (mm)	Shadow size	
			Trained group (mm)	Control (mm)
Mean size	6.8 ± 4.0	10.3 ± 4.1	9.3 ± 4.9^a	10.2 ± 5.0^a
Mean overestimation bias		3.8 ± 2.4	2.7 ± 0.5^b	3.6 ± 0.63^c

Shadow size was not significantly different between the control and trained groups. Mean overestimation bias of shadow measurements with respect to CT size was not significantly different for either group, or significantly different from bias of US stone size.

^a $p=0.18$.

^b $p=0.28$, compared to US stone size.

^c $p=0.81$, compared to US stone size.

US=ultrasound.

TABLE 3. REPEAT SHADOW SIZE MEASUREMENTS

	Trained group	Control group	p
Mean bias for the repeat measurement (vs CT stone size) (improvement in mean bias from baseline)	1.6 ± 0.5 mm (1.0 mm)*	3.4 ± 0.6 mm (0.2 mm)	0.02 (0.019)
≤1 mm size concordance for the repeat measurement (vs CT) (improvement in ≤1 mm size concordance from baseline)	34.7% (11.8%)*	18.8% (0.01%)	0.004 (0.07)

For the training group, there was significant reduction in mean overestimation bias from baseline and a significant increase in the proportion of stones that had 1 mm concordance between shadow and CT stone size. No significant changes were observed with the control group.

*Significant change from baseline, $p < 0.01$.

($p = 0.99$) with repeated measurements. Overall, 54% of (control) subjects (7/13) showed a decrease in overestimation bias compared to CT size with repeated experience (by 0.8 ± 0.5 mm on average). However, 46% (6/13) of subjects in this group demonstrated an increase in bias on repeated shadow measurements (by 0.4 ± 0.7 mm on average).

For several stones, there was variation among reviewers in whether a shadow was deemed present or absent, and this was not necessarily consistent for the same reviewer from baseline to the second measurement. Figure 3 is one example of an equivocal case, where 4 of 26 subjects deemed there to be a shadow at baseline, and 7 of 26 subjects deemed there to be a shadow during the second session; of these, only 3 were consistent with their baseline assessments. Seventy-nine percent (19/24) of stones were noted to have a shadow by the majority of subjects, but only 29% (7/24) stones were unanimously assigned a shadow size by all subjects. There were no cases in either group where everyone reported “no shadow.” Of note, there was no significant increase in the unanimous identification of shadow in either the trained or control group after intervention. These differences were accounted for in the mixed-effects model.

Discussion

Although the posterior acoustic shadow measurement has previously been shown to improve the accuracy of stone

sizing on US, the ease of use, adoption, and accuracy of this technique among inexperienced clinical providers was previously unknown. We found that baseline shadow measurements of novice users significantly overestimated stone size and were no more accurate than reported US stone size, when compared to CT stone size. However, after a brief training intervention, there was significant improvement in the accuracy of measured shadow sizes, with a mean overestimation bias of 1.6 ± 0.5 mm (vs 3.4 ± 0.6 mm for controls) and 35% of shadow measurements demonstrating ≤1 mm concordance with CT measurements (vs 19% for controls). This is comparable to published results from experienced users, who sized 42% of stone shadows within 1 mm of the CT size.⁸ Nearly all who received training improved, suggesting that this effect was not driven by just a few individuals. Although about half of the control group also improved in accuracy with repeated practice, the average magnitude of improvement was fourfold less; there was also no consistency in improvement across the entire group or in the magnitude of improvement. This suggests that repeated practice alone is insufficient to facilitate more accurate size measurements among clinicians.

Our results suggest that basic instruction on the shadow measuring technique may be sufficient to allow novice users to generate reasonably accurate shadow measurements on US. The techniques presented in the training module—identifying the shadow within the image, evaluating for

FIG. 2. Average shadow size per stone from the repeat measurement session, compared to clinical US (gray circle) and CT measurements (black circle). Data are displayed separately for the trained group (blue square) and control group (red triangle); corresponding arrow bars are displayed in same color. The average shadow sizes measured by the trained group are closer to the CT values.

FIG. 3. Example of an US image where presence of stone shadow was equivocal. The stone of interest is marked with an arrow. Four out of 26 subjects deemed there to be a measurable shadow during the baseline session and 7 out of 26 subjects deemed there to be a measurable stone shadow during the repeat session, which may have been defined by projecting the shadow from the renal capsule onto the stone using the guidelines. Others may have deemed there to be no stone-associated shadow present because there is no clear extension of a shadow for the marked stone through the underlying perinephric fat within the image. The more prominent shadowing within the image is not attributable to the stone in question, but is thought to be diffraction artifact around the cyst within the image.

confounding artifacts, tracing the entire path of the shadow, projecting the shadow using guidelines, and measuring the shadow perpendicular to the direction of the US beam—help to structure an assessment of the shadow for each US image. These tools may be particularly useful when the shadow appearance is equivocal on clinical US images.

Despite being novices, there was a fair degree of intraclass correlation for shadow measurements among both control and trained groups (0.55 and 0.64, respectively), with significant improvement during the repeat sizing session. However, such improvement may not be clinically significant. Although this is markedly lower than the intraclass correlation of 0.86 seen among experienced shadow sizers,⁸ these findings suggest that perhaps the consistency of shadow size measurements across multiple users may improve with experience. Even for CT stone measurements, as much as 25% interrater variability and 1.3 mm stone size discrepancies have been reported when the same CT studies are read by different radiologists.^{9,10} This margin of error approaches that of the shadow size.

Clinical seniority, prior US experience, and comfort with US interpretation were not significantly associated with the accuracy of shadow measurement. One potential explanation is that inherent visual-spatial ability was a potential unmeasured confounder. Indeed, among radiology residents, performance on standardized visual-spatial tests correlated with clinical performance and appeared to be independent of experience level among trainees and faculty.¹¹ Although no such evaluation was made in this study, assessment of these inherent abilities might provide further prognostic informa-

tion about the likelihood of accurate shadow measurements among users. Regardless, it remains encouraging that improvement was demonstrated after training across clinical ranks and varying levels of experience.

Although it has previously been shown that not all stones shadow, 35%–42% of those that do may be sized to within 1 mm of the CT measurement.⁸ Moreover, as many as 83% of stones <5 mm in size do not shadow on US.^{7,8} Taken together, the information provided by assessment of the posterior acoustic shadow can be used to better interpret stone images on US and provide additional understanding about true stone size, which may inform clinical decision making. Shadow measurements may be particularly informative in the acute setting, such as the Emergency Department, where US has been suggested as first-line imaging for patients with nephrolithiasis.^{12–14}

This was a small prospective study involving a limited number of clinical US images from a single institution, and these results may not be fully generalizable to other institutions. Given the small size of our cohort, we did not randomize our groups, allowing the potential for unmeasured confounders. Our institution has previously developed and described this technique, and some participants may have been aware of previous findings; all, but two had heard of the posterior acoustic shadow, which could have biased their measurements. However, our approach of a stratified alternating group assignment did not result in significant differences in demographic or experiential factors between groups, suggesting that other potential confounders may have also been equally distributed. A larger sample size would have allowed for greater power to detect a significant difference between groups.

US images in this study were also not optimized to reveal the shadow, as they were captured as part of routine clinical care. Therefore, system settings were adjusted to user preference and the acoustic shadow was not necessarily sought out or highlighted in these images, potentially leading to more equivocal images for which reviewers disagreed on the presence of a posterior acoustic shadow. As previous work in phantoms has demonstrated, specific B-mode US settings may impact the appearance of the posterior acoustic shadow.⁶ For example, avoidance of spatial compounding, particularly at greater depths, may better delineate the shadow borders and limit the lateral splay of the shadow at deeper depths. The use of novel beam-forming techniques may further impact the appearance of the shadow on US.¹⁵ However, we did account for discrepancies in subjects' assessments of shadow presence using mixed-effects models. Moreover, our results represent a pragmatic outcome, as in clinical practice, US is not typically optimized to display the shadow.

Conclusions

The posterior acoustic shadow of a stone is consistently recognizable and readily measurable by clinicians familiar with US. Compared to CT stone size, accuracy of shadow size is no better than reported US stone size when measured by novices. However, a brief training module significantly decreased overestimation of shadow size. Careful identification of the shadow, assessment of potential shadow artifacts, judicious utilization of the guidelines, and measurement of the shadow perpendicular to the direction of the US beam are all

nuances that can improve the accuracy of shadow measurement by novices.

Acknowledgments

This work was supported by NIH P01 DK043881. This material is the result of work supported by resources from the Veterans Affairs Puget Sound Health Care System, Seattle, Washington.

Author Disclosure Statement

B.D., M.R.B., and M.D.S. have equity in and consult for SonoMotion, Inc.

References

1. Fowler KA, Locken JA, Duchesne JH, Williamson MR. US for detecting renal calculi with nonenhanced CT as a reference standard. *Radiology* 2002;222:109–113.
2. Ray AA, Ghiculete D, Pace KT, Honey RJ. Limitations to ultrasound in the detection and measurement of urinary tract calculi. *Urology* 2010;76:295–300.
3. Sternberg KM, Eisner B, Larson T, Hernandez N, Han J, Pais VM. Ultrasonography significantly overestimates stone size when compared to low-dose, noncontrast computed tomography. *Urology* 2016;95:67–71.
4. Ganesan V, De S, Greene D, Torricelli FC, Monga M. Accuracy of ultrasonography for renal stone detection and size determination: Is it good enough for management decisions? *BJU Int* 2017;119:464–469.
5. Dunmire B, Lee FC, Hsi RS, Cunitz BW, Paun M, Bailey MR, Sorensen MD, Harper JD. Tools to improve the accuracy of kidney stone sizing with ultrasound. *J Endourol* 2015;29:147–152.
6. Dunmire B, Harper JD, Cunitz BW, Lee FC, Hsi R, Liu Z, Bailey MR, Sorensen MD. Use of the acoustic shadow width to determine kidney stone size with ultrasound. *J Urol* 2016;195:171–177.
7. May PC, Haider Y, Dunmire B, et al. Stone-mode ultrasound for determining renal stone size. *J Endourol* 2016;30:958–962.
8. Dai JC, Dunmire B, Sternberg KM, et al. Retrospective comparison of measured stone size and posterior acoustic shadow width in clinical ultrasound images. *World J Urol* 2018;36:727–732.
9. Patel SR, Stanton P, Zelinski N, Borman EJ, Pozniak MA, Nakada SY, Pickhardt PJ. Automated renal stone volume measurement by noncontrast computerized tomography is more reproducible than manual linear size measurement. *J Urol* 2011;186:2275–2279.
10. Lidén M, Andersson T, Geijer H. Making renal stones change size-impact of CT image post processing and reader variability. *Eur Radiol* 2011;21:2218–2225.
11. Smoker WR, Berbaum KS, Luebke NH, Jacoby CG. Spatial perception testing in diagnostic radiology. *Am J Roentgenol* 1984;143:1105–1109.
12. Smith-Bindman R, Aubin C, Bailitz J, et al. Ultrasonography vs computed tomography for suspected nephrolithiasis. *N Engl J Med* 2014;371:1100–1110.
13. Choosing Wisely. American College of Emergency Physicians. Available at: www.choosingwisely.org/wp-content/uploads/2015/02/ACEP-Choosing-Wisely-List.pdf 2015 (Accessed April 13, 2018).
14. Sternberg KM, Littenberg B. Trends in imaging use for the evaluation and follow-up of kidney stone disease: A single center experience. *J Urol* 2017;198:383–388.
15. Tierney JE, Schlunk SG, Jones R, George M, Karve P, Duddu R, Byram BC, Hsi RS. In vitro feasibility of next generation non-linear beamforming ultrasound methods to characterize and size kidney stones. *Urolithiasis* 2018. DOI: 10.1007/s00240-018-1036-z

Address correspondence to:

Jessica C. Dai, MD

Department of Urology

University of Washington

1959 NE Pacific Street, Box 356510

Seattle, WA 98195

E-mail: jcdai@uw.edu

Abbreviations Used

CT = computed tomography
 IRB = Institutional Review Board
 US = ultrasound