MPBS2-16 RELIABLE KIDNEY STONE VOLUMES: SEGMENTATION PROTOCOL VALIDATION

Kalon L. Morgan[1], Rohit Bhatt[1], Seyed Hossein Hosseini Sharifi[1], Amanda McCormac[1], Channon Chantaduly[2], Sohrab N. Ali[1], James Shi[3], Pengbo Jiang[1], Roshan M. Patel[1], Peter Chang[2], Jaime Landman[1,3], Ralph V. Clayman[1]

[1]Department of Urology, University of California, Irvine, Orange, CA, USA
[2]Center for Artificial Intelligence in Diagnostic Medicine, University of California, Irvine, Irvine, CA, USA
[3]Department of Radiology, University of California, Irvine, Orange, CA, USA

INTRODUCTION

Computed tomography (CT) based volumes more accurately quantify the 3D nature of kidney stones compared to traditional linear measurements; however, human error in the manual segmentation process may limit their reliability. We evaluated kidney stone volumes determined using segmentations from individuals with varying levels of medical training and expertise in order to develop a segmentation protocol for rapid and consistent (inter-observer), and reproducible (intra-observer) kidney stone volume data.

METHODS

- A board-certified radiologist, a fellowship trained endourologist, a postgraduate endourology fellow, and two fourth-year medical students, segmented the same set of stone-containing kidneys (N=98). The radiologist then repeated the segmentation of the entire set to determine intra-observer reproducibility.
- Stones were segmented using a web-based software program created at UCI called Akila Viewer which utilizes a "paintbrush" allowing for identification of scan portions within a fixed Hounsfield Unit range.
- Agreement between the 5 sets of volumes was assessed using Interclass Correlation Coefficient (ICC).
- Dice score, a measure of virtual object 3D overlap, was used to compare segmentations for each pairing of individuals (Figure 1 and Table 1).
- Volumes created by the radiologist and attending endourologist were used as a standard against which the results of the endourology fellow and the two medical students were compared. Bland-Altman plots were used to visualize volume agreement between the standard and the other three individuals (Figure 2).

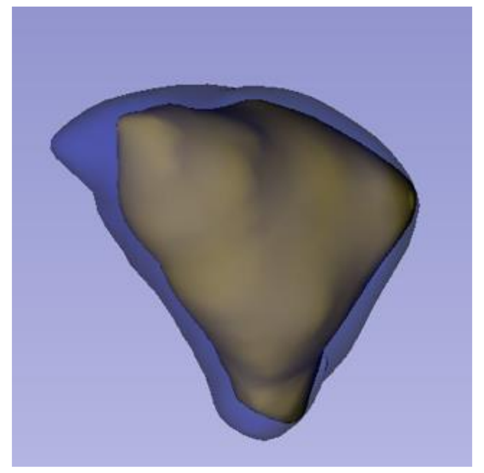


Figure 1: Visual representation of Dice score calculation. The 3D overlap between two virtually-rendered stones are superimposed. The translucent blue represents one observer's stone volume, and the dark yellow represents another observer. The Dice scores for each 2-way comparison of observers are found in Table 1.

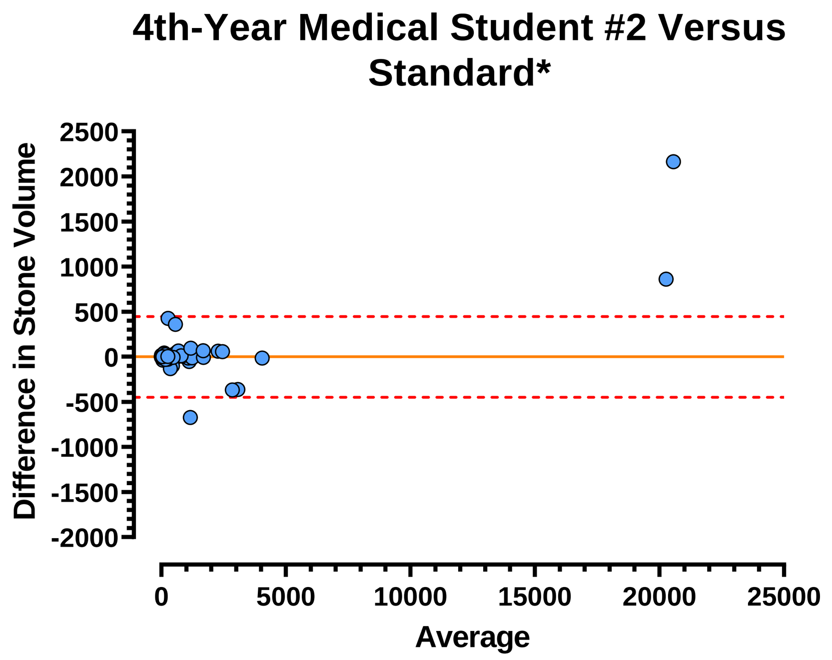
RESULTS

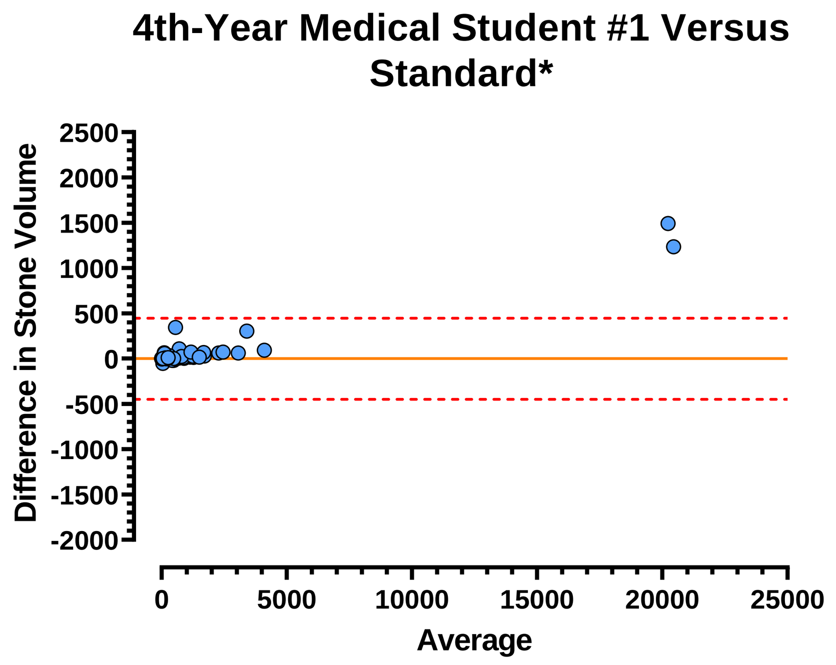
- ICC for volumes from all five participants was 0.994.
- Mean Dice scores for every pairing of individuals were ≥ 0.90 (Table 1).
- Bland-Altman plots comparing the endourology fellow and the two medical students to the standard showed strong consistency and agreement (Figure 1).

Dice Scores for Paired Observers					
Observer	MS #1	MS #2	Fellow Urologist	Attending Endourologist	Attending Radiologist
Attending Radiologist	0.91	0.93	0.92	0.91	0.97*
Attending Endourologist	0.90	0.93	0.91	x	
Endourology Fellow	0.93	0.96	x		
MS4-2	0.95	x			

Table 1: Average Dice scores comparing each pairing of segmentations. Dice score ranges from 0.0 (no overlap) to 1 (perfect overlap).

*The attending radiologist segmented the entire data set, and then repeated the segmentations a week later to evaluate intra-observer variability. This average Dice score compares the first and second round segmentations; there was no significant difference between the two separate observations.





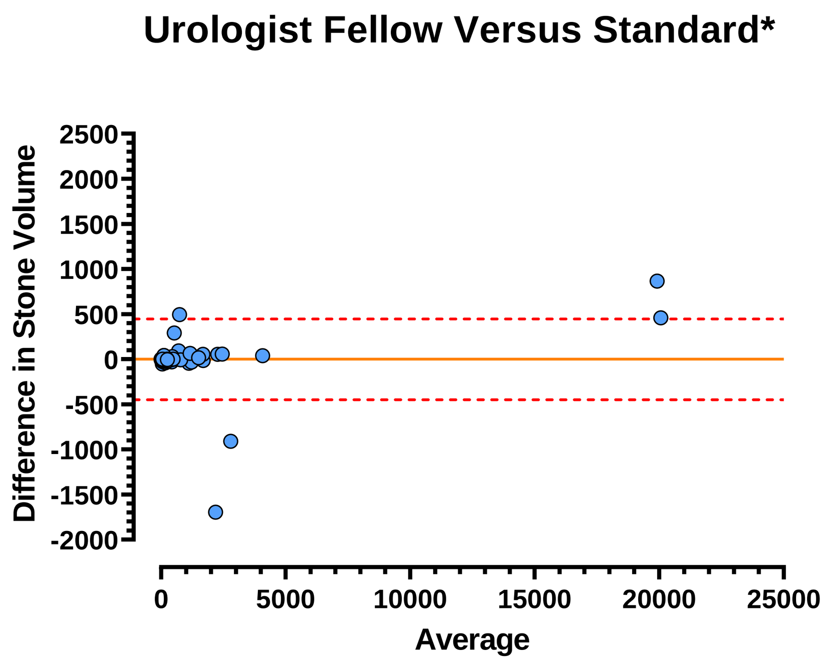


Figure 2:
Bland-Altman Plots depicting kidney stone volume degree of agreement: *Top Left* – Fellow Urologist versus Experts; *Top Right* – 4th-year medical student #1 versus Experts; *Bottom* – 4th-year medical student #2 versus Experts; Solid Orange line – average mean difference (mm^3); Dotted Orange Lines - Upper and Lower limits (95%).

*Standard is defined as the averaged volumes generated by an attending endourologist and an attending radiologist.

CONCLUSION

- Segmentation and kidney stone volumes for all five participants had strong agreeance (ICC=0.998 / mean Dice scores ≥ 0.9).
- The use of web-based software with fixed segmentation parameters yielded low inter-observer variability, regardless of years of medical training.

UCI Department of Urology