

ABSTRACT

Quantitative Digital Subtraction Radiography (DSR) is used to detect alveolar bone changes between serially taken radiographs. In practice image analysts may use different software, hardware and procedures to accomplish this task. The purpose of this study was to evaluate inter-analyst variability in quantifying known changes in alveolar bone with DSR using bone chip standards. Measurements were obtained from two analysts, one who has experience measuring bone in multiple clinical trials (Analyst 1) and a second analyst who has focused on DSR method development (Analyst 2). Each analyst digitized 30 radiograph pairs, subtracted and quantitated changes in bone height of 81 chip standards ranging in size from 1.2-3.0 mm. A Mixed Effects Model was fit to the data and Student's t-tests were used to test the slope and intercept terms from each fit. A linear relationship between actual and calculated chip height was demonstrated for both analysts; however the height readings were statistically significantly different from one another ($p < 0.001$). Analyst 2 consistently reported changes which were smaller than Analyst 1, with Analyst 1 being closer on average to the actual chip height. The mean difference observed between the height measurements of Analysts 1 and 2 was 0.348 mm, with a SD of 0.468 mm. The sensitivity for detection of bone chips differed, with Analyst 1 making 5 and Analyst 2 making 1 false positive measurements. Mass measurements were not compared between the analysts due to an inability of Analyst 2's existing analysis software to compensate for variations from the perpendicular of the x-ray beam for the reference wedge as evidenced in some radiographs. This study demonstrates that prior calibration may aid in the interpretation of DSR outcomes.

INTRODUCTION

Numerous authors over the past two decades have reported the ability of digital subtraction radiography (DSR) to provide maximum diagnostic precision in detecting small changes in alveolar bone correlated with worsening or improvement of periodontal disease (e.g., Hausmann, *J Periodontol*, 1985). However; variability between independent DSR analysts evaluating a common dataset of radiographs has not

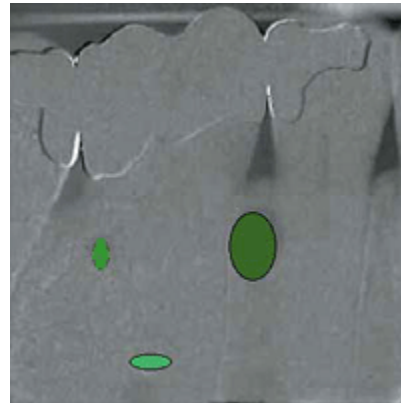
previously been established. A better understanding of inter-analyst measurement error is desirable for determining reliable thresholds to assess meaningful bone change in clinical trials, and for supporting professional and regulatory acceptance of DSR.

OBJECTIVE

The study was conducted to assess inter-analyst variability in quantifying known changes in alveolar bone height using DSR and bone chip standards.

MATERIALS AND METHODS

Thirty paired bilateral radiographs, one of each pair impregnated with up to 3 enamel-equivalent test chips (CaPstrate, Bio-Interfaces, Inc., San Diego, CA) to mimic alveolar bone changes, were exposed in a total of 15 human volunteers in a randomly assigned sequence. Processed, blinded films were forwarded for DSR analyses to 2 distinct experienced analysts.



Colorized subtraction image.
Darker green indicates deeper depth.

While both analysts utilize accepted DSR methods, the approach for generating the final quantitation of bone changes differs, including utilizing different hardware, software and procedures. For example, Analyst 1 consistently uses a warping step to correct for errors in planar geometry between film pairs prior to subtraction. Analyst 2 does not use warping due to strict adherence to achieving reproducible projection geometry prior to analysis of films. For this study, the affine algorithm from Analyst 1 was provided to Analyst 2 and it was used for 53% of the film pairs.

Bone height changes were derived from the most superior and inferior borders of subtracted images. To determine measurement accuracy, observed vs. actual chip height values were calculated by analyst. A Mixed Effects Model (MEM) was fit to the data and Student's t-tests were used to test the slope and intercept terms from each fit. To evaluate inter-analyst variability, data from Analysts 1 and 2 were compared to assess linear regression fits.

RESULTS

A comparison of analyst measurement error revealed a linear relationship between observed and actual chip height for both analysts. Generally Analyst 1 overestimated and Analyst 2 underestimated actual values, but on average Analyst 1 was closer to the actual height. The estimate of slope and intercept from the MEM was similar for both analysts and did not differ significantly from the theoretical line with a slope of 1 ($p > 0.225$) and y-intercept of 0 ($p > 0.251$). Table 1 shows measurement error results by analyst for representative small, mid-range and large height chips.

Table 1. Measurement Error-Analyst vs. Analyst 2

Actual Dimension (mean) (mm)	Analyst	Observed Value (mean±Sd) (mm)	Mean Measured Relative Error ¹ (%)
1.23	1	1.280±0.110	4.3
	2	1.175±0.139	-4.0
2.15	1	2.296±0.201	6.8
	2	2.000±0.141	-7.2
2.98	1	3.127±0.350	4.8
	2	2.750±0.169	-7.9

¹[Observed-Actual]*100

The difference in observed chip height measurements was highly significant (p<0.001), with a mean difference of 0.348 mm (SD 0.468). Figure 3-A illustrates the relationship between observed height values by Analyst 1 and Analyst 2. As depicted in Figure 3-B, 95.8% of the differences in the two analysts' height readings were within 1.0 mm of zero. There appeared to be no obvious trend in the differences as a function of actual chip height.

Figure 3 A:

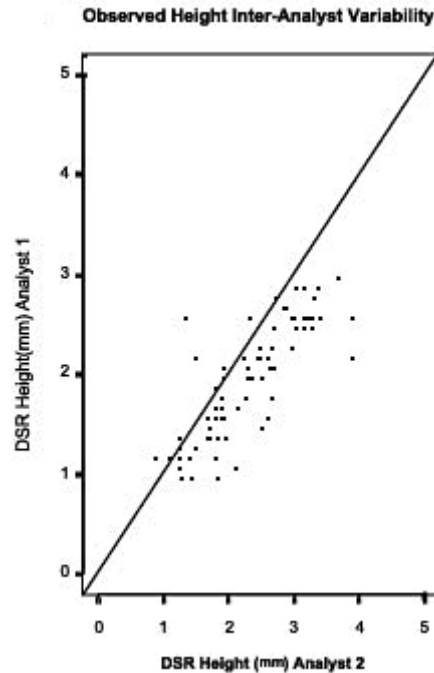
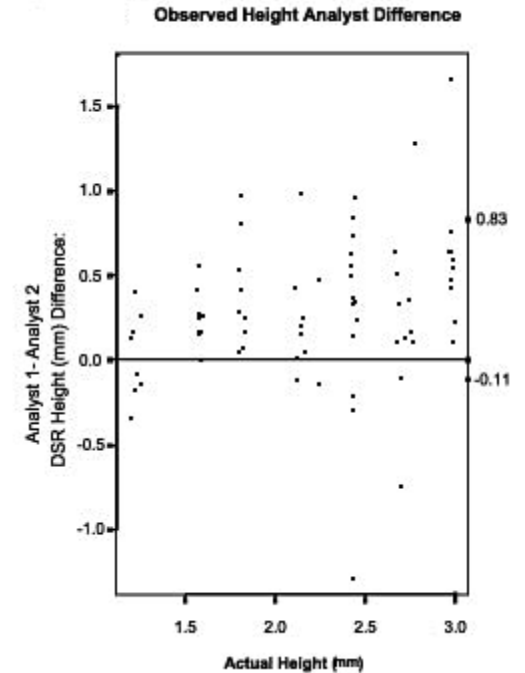


Figure 3 B:



Inter-analyst chip detection sensitivity and specificity differed negligibly, with Analyst 1 and Analyst 2 identifying 78/81 chips and 77/81 chips, respectively. Analyst 1 had 5 false positive chip identifications, while Analyst 2 had 1.

CONCLUSION

While DSR calculated height measurements of a common dataset by two analysts were close to actual, significant inter-analyst variability existed in this study. If multiple analysts are to be used for analysis of DSR data, calibration would be of value in order to more accurately interpret DSR outcomes.