LETTER TO THE EDITOR

Reply: The Humeral Head Circle-Fit Method Greatly Increases Reliability and Accuracy When Measuring Anterior-Posterior Radiographs

Dear Editor:

We are happy to respond to Dr. Sabour's criticism that is aimed at a portion of the statistical methods that we used in our recent study (Mears et al., 2017 JOR). We agree with Dr. Sabour that Pearson correlation coefficients are not actually a measure of intra-observer and inter-observer reliability, and that we should have used intraclass correlation coefficients (ICCs), which are specifically reliability measures. We have since computed these and display them in tables below. The results of this analysis lead to the same conclusions stated in our study.

We disagree with Dr. Sabour's advice for assessing accuracy. He suggested comparing our method (CFM, circle fit method) with a gold standard or reference method using a Pearson r or Spearman correlation coefficient. If a reference method was even available to do that, which is not, a correlation coefficient would actually assess validity, not accuracy. For example, one method could consistently give a score that was higher by five points, and so inaccurate, and the correlation would still be perfect ($r = 1.0$). Accuracy is correctly assessed with a Bland–Altman analysis.[1,2] A Bland–Altman analysis determines if two measurement methods provide measurements that are similar on average, and with small variability in the differences between them, since if so, they could be used interchangeably. However, if a method fails to have reliability, it cannot possibly have accuracy. Since the other two methods (Tingart and Mather) failed to have reliability, we cannot compare our CFM to either of them to assess accuracy using a Bland–Altman analysis. The CFM we presented, where we quantified how different on average each method was between observers and how variable these differences were, was sufficient to rule out accuracy of the other two methods (Tingart and Mather).

To assess intra-observer reliability, or test–retest reliability, three of our observers measured the same image ($n = 10$) on two occasions in a random order, one week apart. These results, now expressed in terms of ICCs, are shown in Table 1. It is clear to see that even the same observer cannot get consistent results with the other two methods.

Table 1. Intra-Observer Reliability ($n = 10$ Images)

Observer	CFM ICC (95%CI)[a]	Tingart	Mather
A	0.99 (0.96, 1.00)	0.27 (0.00, 0.72)	0.58 (0.04, 0.87)
B	0.98 (0.92, 1.00)	0.22 (0.00, 0.66)	0.57 (0.00, 0.87)
C	0.96 (0.50, 0.99)	0.35 (0.00, 0.77)	0.31 (0.00, 0.75)

[a]ICC, intraclass correlation coefficient; CI, confidence interval.

Table 2. Inter-Observer Reliability ($k = 5$ Observers, $n = 33$ Images)

CFM ICC (95%CI)[a]	Tingart	Mather
0.87 (0.80, 0.93)	0.35 (0.17, 0.55)	0.39 (0.19, 0.59)

[a]ICC, intraclass correlation coefficient, CI, confidence interval.

To assess inter-observer reliability, all five of our observers measured the same images ($n = 33$ images). For inter-observer reliability, all five raters were included in the computation, so the reliability is among fiver raters. These are shown in Table 2. Applying the Cicchetti and Sparrow guideline,[3] which is frequently quoted in the reliability literature for interpreting the ICC coefficient, CFM achieved "excellent agreement" ($0.75 \leq \text{ICC} \leq 1.00$), while the other two methods only achieved "poor agreement" ($\text{ICC} < 0.40$).

John G. Skedros[1]
Chad S. Mears[1]
Tanner D. Langston[1]
Colton M. Phippen[1]
Wayne Z. Burkhead[2]
Gregory Stoddard[3,†]
[1]*Utah Orthopaedic Specialists, Salt Lake City, Utah*
[2]*W.B. Memorial Carrell Clinic, Dallas, Texas*
[3]*University of Utah School of Medicine, Department of Family and Preventative Medicine, Salt Lake City, Utah*
E-mail: jskedrosmd@uosmd.com

Received 15 March 2017
Accepted 4 May 2017

Correspondence to: John Skedros, (T: 8017252168; F: 8017471023; E-mail: jskedrosmd@uosmd.com)

Published online 16 June 2017 in Wiley Online Library
(wileyonlinelibrary.com).
DOI 10.1002/jor.23612

REFERENCES

1. Bland JM, Altman DG. 1999. Measuring agreement in method comparison studies. Stat Methods Med Res 8:135–160.

2. Rankin G, Stokes M. 1998. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 12:187–199.

3. Cicchetti DV, Sparrow SA. 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 86:127–137.