Readers are invited to submit letters for publication in this department. Submit letters online at http://joem.edmgr.com. Choose "Submit New Manuscript." A signed copyright assignment and financial disclosure form must be submitted with the letter. Form available at www.joem.org under Author and Reviewer information.

Reliability of Common Provocative Tests for Shoulder Tendinitis by Doxey et al—Letter to the Editor

To the Editor:

The article by Doxey et al[1] is stated to be:

> "A cross-sectional evaluation of a large population of workers to systematically assess the sensitivity, specificity, and reproducibility of common provocative shoulder maneuvers against a case definition of shoulder tendinitis has not been reported and is the purpose of this report."[1]

There are a number of methodological issues related to the estimation of "sensitivity, specificity, and reproducibility" in this article that merit comment.

Estimations of sensitivity and specificity (and subsequent calculation of positive predictive value and negative predictive value) are most often based on comparison of results obtained with a particular test to those obtained with an independent "gold standard" test.[2,3] In the present investigation, a case of shoulder tendinitis "required a history of glenohumeral joint shoulder pain in the 30 days before enrollment and a positive Jobe test."[1] Results of four other physical examination maneuvers, plus range of motion testing, were then compared with this case definition (see Table 3). However, reporting inconsistencies and methodological problems lead us to question the validity of the reported comparisons.

First, the origin of the positive examination counts shown in Table 3 is unclear. For example, in Table 2 it is shown that examiners 1 and 2 found 17 and 31 positive results, respectively, for "abnormal range of motion" for the right shoulder, yet in Table 3 the corresponding number of positive examination results is 34. It is not stated, and it is unclear how the latter result was derived from the former (the same question applies to all results shown in Table 3).

Second, the independence of the case definition from the examination maneuvers being tested is unclear, since the associations of the four physical examination maneuvers with the Jobe test are not shown. Also, critically, the Jobe test was based on the results of the second examiner and not an independent examiner. Furthermore, the second examination is described in an earlier paper reporting results of the same study as follows:

> "The second physical examination is performed by board-certified Occupational Medicine physicians. There are two main information sources used to begin that examination: the symptoms summary page from the structured interview (above) and the results of the first examiner's physical examination tests. The purpose of this examination is to confirm positive findings and to evaluate pertinent negatives."[4]

Such an unblinded process likely would serve to bias estimates of sensitivity, specificity, positive predictive value, and negative predictive value, thus making comparison of results of Doxey et al[1] with other studies problematic.

For assessing the reliability of physical examination test of shoulder tendonitis, Doxey et al[1] stated the following:

> "Two physical examinations were performed on all workers and were conducted with knowledge of the results of the structured interview... The first examination was comprehensive and included all physical examination maneuvers regardless of symptoms. The second examination was performed by an experienced, board-certified occupational medicine physician and assessed both the positive and pertinent negatives from the first examination."[1]

Notably, the authors fail to mention that, as stated in Garg et al,[4] the second examiner was fully informed of the results of the first examiner. Although the lack of blinding of the second examiner to the results of the first examiner is noted by Doxey et al[1] as a "limitation," it is, in fact, a fatal design flaw that the examiners failed to "operate independently,"[5] and thereby serves to overestimate inter-rater reliability by a substantial but ultimately unknown amount.

Third, a major discrepancy was noted between the authors' description of their statistical methods for assessing inter-rater reliability and the reporting of such results. Specifically, although the Methods Section noted that "kappa reliability coefficients" were used to assess inter-tester reliability, "inter-tester reliability percentage" was reported in Table 2. These are not the same. Kappa is a dimensionless statistic that ranges from −1 (ie, perfect disagreement) to +1 (ie, perfect agreement).[5] It is not reported as a percentage. This is not a mere error in labeling. Given the data shown in Table 2, the "Inter-tester Reliability %" values shown in most instances are not mathematically feasible for kappa. For example, the first row in the table shows results for "abnormal range of motion" for the right shoulder involving 317 subjects, with 17 positive results for examiner 1, and 31 positive results for examiner 2. Mathematically, given these marginal totals, the maximum possible value for kappa is 0.687, not 0.936.[5] Similar discrepancies exist for the other reliability results shown in Table 2. It is not clear, but, the "Inter-tester Reliability %" values shown in Table 2 appear to be raw percent agreement values, not kappa values. This is problematic for a number of reasons. First, raw percent agreement is an inadequate metric of inter-rater reliability.[5] Second, since none of the results shown in Table 2 are kappa values, the conclusions related to reliability (which appear to assume that they are kappa values) are misguided. Finally, comparison of the results shown in Table 2 (which are not kappa values) with kappa results from Michener et al[6] is not valid.

Fourth, the authors' fail to account for so-called *spectrum bias*.[7] Specifically, spectrum bias occurs when the sensitivity and specificity of a diagnostic test is evaluated among persons who are unlikely to be administered such a test in clinical practice. Doxey et al[1] reported that about 76% of workers did not report right shoulder pain during the month prior to examination and about 80% of workers did not report left shoulder pain during the month prior to examination. In clinical practice, how often would a patient free of shoulder pain for at least 1 month prior to examination be administered a test for shoulder tendonitis? The answer is, of course, rarely or never. The reliability of a test among such persons is not relevant to clinical practice, although it likely to be very high and the observed specificity artifactually inflated. That raters agree on a negative test among persons free of shoulder pain for a

Dr Franzblau has provided expert testimony on behalf of the US EEOC. Drs Gerr and Werner have no conflicts.

Address correspondence to: Alfred Franzblau, MD, Department of Environmental Health Sciences, University of Michigan School of Public Health, University of Michigan, Ann Arbor, MI (afranz@umich.edu).

DOI: 10.1097/JOM.0000000000001575

month or longer is not of relevance to clinical practice where the tests being evaluated are used typically among person suspected clinically of having shoulder tendonitis, a suspicion most often resulting from a report of recent shoulder pain.

Overall, multiple serious methodological flaws lead us to believe that the results of the study by Doxey et al[1] are neither reliable nor valid.

Alfred Franzblau, MD
Department of Environmental Health Sciences
University of Michigan
School of Public Health
University of Michigan
Ann Arbor, Michigan

Fred Gerr, MD
Department of Occupational and Environmental Health
College of Public Health
University of Iowa
Iowa City, Iowa

Robert A. Werner, MD, MS
University of Michigan Health System
Ann Arbor Veterans Health System
Ann Arbor, Michigan

REFERENCES

1. Doxey R, Thiese MS, Hegmann KT. Reliability of common provocative tests for shoulder tendinitis. *J Occup Environ Med*. 2018;60:1063–1066.
2. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ*. 1994;308:1552.
3. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia. *Crit Care Pain*. 2008; 8:221–223.
4. Garg A, Hegmann KT, Wertsch JJ, et al. The WISTAH hand study: a prospective cohort study of distal upper extremity musculoskeletal disorders. *BMC Musculoskeletal Disord*. 2012; 13:90.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure*. 1960;20:37–46.
6. Michener LA, Walsworth MK, Doukas WC, Murphy KP. Reliability and diagnostic accuracy of 5 physical examination tests and combination of tests for subacromial impingement. *Arch Phys Med Rehabil*. 2009;90:1898–1903.
7. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrim effect in tests for risk prediction, screening, and diagnosis. *BMJ*. 2016;353:i3139.

