ACRM
AMERICAN CONGRESS OF REHABILITATION MEDICINE

Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2013;94:1635-7

INVITED COMMENTARY

Invited Commentary on Intrarater and Interrater Reliability of Select Clinical Tests in Patients Referred for Diagnostic Facet Joint Blocks in the Cervical Spine

Stuart McGill, PhD

From the Spine Biomechanics Laboratories, Faculty of Applied Health Sciences, Department of Kinesiology, University of Waterloo, Waterloo, Ontario, Canada.

Abstract

The authors of an article in this issue of *Archives of Physical Medicine and Rehabilitation* have advanced our understanding of reliability of clinical tests designed to provide insight into suspected facet joint pain generators. Because the issue of reliability influences, and is influenced by, political and clinical issues, my commentary has 2 parts. First is a general commentary on reliability of assessment tests, followed by comments specific to this article.

Archives of Physical Medicine and Rehabilitation 2013;94:1635-7

I thank the editors for their invitation to write this commentary. Schneider[1] and colleagues contributed a study that advanced our understanding of the reliability of clinical tests designed to provide insight into suspected facet joint pain generators. Because the issue of reliability influences, and is influenced by, political and clinical issues, my commentary has 2 parts. The first part is a general commentary on the reliability of assessment tests. This is followed by comments specific to this article.

One of the critical issues raised by this article is whether we should judge the value of an assessment instrument solely on its reliability. Simple tests are reliable. Measuring core temperature in a case of infection, and blood chemistry in a case of diabetes, are examples of reliable tests that can lead to a homogeneous and consistent treatment. Even physicians in the intensive care unit are able to predict survival reliably over 24 hours after admittance. Here, the outcome is binary: patients either survive or they do not.

Musculoskeletal (MSK) disorders are different: their symptoms are highly variable in terms of pain, there is often more than one source of pain, the dosage of intervention is critical (as too much exacerbates and too little has no effect), and the outcomes are highly variable in terms of duration and effectiveness. Why should 2 clinicians obtain the same impression when examining a biological system that is continually in a state of flux? If their skills were equal and the patient remained static, then reliability may be possible and even justifiable. But this is not the case with most MSK disorders. Thus, the typical "rules for reliability" associated with evidence-based medicine need a liberal amount of reflection for logical application in MSK situations. Machines can be extremely reliable, but no diagnostic machine (however reliable) has ever lived up to hyperbole or obtained a better outcome for MSK disorders than a highly skilled clinician. The best pattern recognition system, data integrator, decision processor, and manual applicator of corrective cues remains the skilled clinician.

Clinicians have differing levels of clinical skill. Clinical skill involves perception of touch, interpreting what the patient verbalizes and displays with body language, knowing how much force to apply, knowing how to explore the end range and arc of motion with subtle trajectory variations to interpret joint capsule, bony interaction, ligament spring, and associated muscle tone, to name just a few variables. A simple test, such as one to objectify the range of motion, becomes much more revealing in the hands of skilled clinicians who probe tissues through a range. They interpret changes in perceived tissue texture and note instability catches through the normal movement range. They know how the order of tests will influence the feel and pain provocation of various candidate mechanisms. They are cognizant that time of day influences spinal disk hydration, which influences joint mechanics, and neural sensitivity. They recognize that related

No commercial party having a direct financial interest in the results of the research supporting this article has conferred or will confer a benefit on the authors or on any organization with which the authors are associated.

0003-9993/13/$36 - see front matter © 2013 by the American Congress of Rehabilitation Medicine
http://dx.doi.org/10.1016/j.apmr.2013.04.023

modulators could include the length of time the patient spends in the waiting room and the time spent traveling to the appointment. If the patient drove in heavy traffic, skilled clinicians better understand that "checking the blind spot" while driving has influenced the pain sensitivity of the facet joints in the patient's neck. This self-stretching can radically modulate short-term proprioception. The point is: skilled clinicians account for these sources of valuable information in reaching a conclusion. Less skilled clinicians will overlook modulators of pain and function and should not be expected to replicate the clinical impression. Their interpretation from a simple, but reliable test matters not.

So what is the better test? The simple test to obtain a range-of-motion score that is reliable, or the test that facilitates a branching decision tree using variables and relationships that are nonlinear, that change over time, and are not repeatable between clinicians of differing skill level? Should reliability be used to identify a good test, and by logical extension, should reliability be a metric for inclusion into clinical guidelines? This argument suggests that skilled diagnosticians using less reliable tests but more complex decision trees will obtain more insight into the patient than unskilled ones.

Clinical guidelines create "average practice." Their benefit is that unskilled clinicians will be able to follow the guidelines, incorporating reliable tests, and arrive at an impression, albeit a "junior" one. This will serve the simple cases but fail to help more complex patients. Tragically, the more complex cases, many of which fail to be sorted out by the clinical guidelines, are dismissed by the health care system. The health care system is purported to be evidence-based, implying that it is based on published studies. The choice of which studies get published becomes critical. I submit that given the number of sad cases I see in my consulting, the status quo is far from optimal. I read the charts that document how these patients had been assessed with reliable tests, all with published reliability scores. But these tests were clinically useless.

Studies reporting high reliability scores without an indication of their relation to patient outcome are missing an important point.

Skilled clinicians must be encouraged to maintain and develop their skills in the more complex, but more insightful (but by default more unreliable) assessment tests. Over 30 years of observation, my impression is that these complex skills have declined. Are patients better off now? Journals can influence this with political stances that shape the development of clinical skill. I am against the growing emphasis of only developing tests that are reliable, and even more against the practice of journals now requiring a high test reliability score in order to qualify for publication. This practice is retarding expert assessment skill acquisition and development, and optimal patient outcome for all branches of manual medicine.

Instead, I support a 2-phase approach: (1) that we continue to encourage the development of clinical guidelines incorporating reliable clinical tests to assist trainees, educational curriculum development, and newer practicing clinicians; and (2) that we also must continue to encourage excellence and continuing development of expert assessment skill (albeit less reproducible) throughout the life cycle of the clinician.

The article by Schneider et al[1] is thought provoking and provides an excellent forum to discuss the consumption of test reliability data. It contributes to the efforts to produce better trainees and reduce disorganized and ill-founded practice. Comparing different studies' reporting reliability can be tricky to interpret. Comparing 2 clinicians with equal years of experience, and for the sake of argument, equal skill, will provide an impression for the tests reported here. However, comparing new graduates to highly skilled veterans might render poorer reliability scores. Understanding the range of skill level of the clinician is needed to interpret the reported reliability values obtained from different studies. Will this consideration reflect the worthiness of the test?

Schnieder's[1] reported differences in the inter- and intrarater reliability are fascinating. Facet joint issues are often not the primary injury mechanism—rather in low-energy mechanisms they are usually considered either secondary or correlated to endplate and disk damage, and in high-energy mechanisms are usually considered secondary or correlate to ligamentous damage. An inclusion criterion for the participants was that they had longer-term pain. Mechanical neck pain is generally accepted to cascade over time, involving more tissues with migrating pain patterns. Daily pain differences are the norm. Interexaminer reliability scores assume that the clinical presentation is constant and homogeneous. Thus, the very characteristics that distinguish facet pain would disqualify this MSK condition for reliability measures where time has elapsed between assessments. As the authors astutely note, the "reliability" of the clinician may have been 100%, but any shifting patient presentation over 1 week could account for any divergent score.

Conducting clinical tests, particularly for syndromes such as cervical facet disorders, can exacerbate pain symptoms and increase stiffness and guarding. Thus, the use of 1 week for the intrareliability test obscures any control over the consistency of patient presentation whether or not it was influenced by the test or daily pain variance. However, listing the scores of the repeated tests could have assisted interpretation. For example, an increase in a range of motion could be due to increased compliance in the tissues as a function of repeated cycles of motion, or the patient may have gained confidence in the assessor and allowed more motion. For the interreliability scores, only 5 minutes elapsed between assessments, which is well within the latency times for neural stretch modulation and mechanical strain relaxation. However, the data of this study suggest that this consideration is less important, given the better reliability scores obtained in the same session. Further, reporting a single "average score" for reliability from tests with repeated measures would obscure the potential to uncover divergent behavior (ie, better or worse over a week) of individuals. It would be most helpful to examine the data of patients, sorting those with positively and negatively changed scores, with reported symptoms, and order of tests. I am suggesting that the real clinical implication and relevance are contained in the variance of response, not in repeatability.

As a closing thought, tests are intended to assist with identifying pain mechanisms so that clinicians/patients can be guided as to what to do and not do. Probing and provoking pain with various postures, motions, and loads would identify what is exacerbating and what is relieving, together with the levels of each that trigger pain. This would guide the clinician (and thus the patient) in eliminating the cause of the pain/disorder. Then a clinical plan of progressive therapy could begin to address the deficits that allowed the pain. Issues of reliability and validity are then placed at the feet of the clinician—scores would be based on clinical outcome and efficacy. While this will make some clinicians uncomfortable, I submit that fewer of the patients with complex problems would end up classed as "failures." I would not argue

List of abbreviations:

MSK **musculoskeletal**

for this as a replacement approach but rather one to add to the current efforts to judge clinical tests.

In summary, technically flawless studies, assessing reliability of simple tests that do not reveal MSK pain generators or guide a progressive treatment approach, waste journal pages. Unfortunately, in my opinion, they become accepted as examples of "good science." Over the years, such reliability studies have displaced the less reliable but skill-dependent tests. The higher question must weigh their real value in the ability to enhance clinical skill and improve patient outcome. As a reviewer and editorial board member of several journals, I see sophisticated tests with tremendous potential rejected because of poor reliability. I would urge journal editors and reviewers to balance reliability as a benchmark for publication with the potential to encourage skill development, thoroughness of patient investigation, and subsequent treatment efficacy. Then scientists/clinicians will return to the journals with highly clinically relevant and action-oriented submissions. After all, in the end, it is about getting people better.

Keywords

Pain; Rehabilitation; Reliability

Corresponding author

Stuart McGill, PhD, Professor of Spine Biomechanics, Spine Biomechanics Laboratories, Faculty of Applied Health Sciences, Dept of Kinesiology, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1. *E-mail address:* mcgill@uwaterloo.ca.

Reference

1. Schneider GM, Jull G, Thomas K, et al. Intrarater and interrater reliability of select clinical tests in patients referred for diagnostic facet joint blocks in the cervical spine. Arch Phys Med Rehabil 2013;94:1628-34.