Dynamic Chiropractic

CHIROPRACTIC (GENERAL)

Thinking About Cohen's Kappa

Keith Charlton, DC, MPhil, MPainMed

Let's think about some notions of reliability and validity, and about what it means for diagnostic examiners to agree in meaningful ways. Diagnostic tests must obviously be both reliable and valid. *Reliability* means the extent to which two or more examiners agree when using the same test on the same population of patients. *Validity* means the extent to which the test reveals what it is supposed to reveal. A test is no use without both these attributes.

Typically, for so many other things, too, truth comes to us in 2x2 contingency tables. Across the top, left to right, we have observer #1: positive in the left column and negative in the right column; and on the left, observer #2: positive in the top row, and negative in the bottom row. (Draw a little 2x2 contingency table like the one below as you read this – it's a very simple concept.)

	Observer #1	Observer #1
Observer #2	Positive	Negative
Positive	a	b
Negative	c	d

The number of patients in the sample is N, which is the sum of a, b, c and d. Box a is where both examiners agree the test is positive, and so on. Note that at first sight, it appears both examiners agreed in boxes a and d, and both disagreed in boxes b and c. If really so, it seems that a crude (but illegitimate) estimate of agreement is (a+d)/N. However, what if one or both are simply guessing? Results might be no more than random, and not telling us anything useful about reality. So, a correction is required for chance agreements.

The true strength of such an analysis lies not in its apparent agreement, but in its agreement beyond chance. No test should be endowed with credit for finding those agreements it would have found by chance alone. If two observers agree in 40 percent of cases by chance alone, the test must tell us what is happening in the other 60 percent of cases to be useful.

I'll leave the math alone here (it's not that heavy, if you want to, look it up on the Net or in a stats text), for it's the ideas, the thinking, that are most important. You should know Kappa values from 0.8 to 1.0 are very good, from 0.6 to 0.8 good, from 0.4 to 0.6 moderate, from 0.2 to 0.4 slight, and from 0.0 to 0.2 very poor. The test suffers some technical glitches in very high or very low prevalence conditions, and this is taken into account with some modifications.

The need, then, is to ask for the Kappas when someone describes outcomes of diagnostic tests. If two observers cannot agree on whether a test is positive or negative (low Kappas), the test simply cannot serve any useful purpose. It's mashed potato, not science.

In that context, the Waddell signs have Kappa scores of <0.3 and may be practically useless. What do you think? Lawyers win cases with this test!

Why not get together with a colleague or two and scrutinize some of our tests? It's easy landmark

research that requires almost no resources other than a couple of good thinks, a bit of time (not much) – and you're on your way. We need this kind of research in chiropractic. Thanks for thinkin' with me!

Editor's Note: This is the second in a short series of articles by Dr. Charlton focused on different aspects of research as applicable to clinical practice. “Mechanism vs. Outcome: Thinking About the Gap Between Research and Clinical Practice" ran in the Sept. 1 issue.

Resources

- Cohen J. A coefficient of agreement for nominal scales. *Educ Psych Meas,* 1960;20:37-46.
- Bogduk N. Truth in diagnosis: reliability. *Aust Musc Med*, 1998 March:21-3.
- Waddell G, McCulloch JA, Kummel E, Venner RM. Non-organic signs in low back pain. *Spine*, 1980; 5:117-25.
- McCombe PF, Fairbank JCT, Cockersole BC, Pynsent PB. Reproducibility of physical signs in low back pain." *Spine*, 1989;14:908-18.

OCTOBER 2015