PM R 9 (2017) 622-628

www.pmrjournal.org

Statistically Speaking

Reliability Statistics

Kristin L. Sainani, PhD

Introduction

Measurement tools used in clinical studies need to be sufficiently reliable. Reliability means that the tool gives consistent results when administered by different people (interrater reliability) or at different time points (test-retest reliability). Statistical tests of association are not appropriate for assessing reliability—reliability statistics assess agreement rather than association. Reliability statistics include the Kappa statistic for categorical scales and the intraclass correlation coefficient (ICC) for continuous scales. Multiple versions of both the Kappa statistic and the ICC exist, and researchers need to know how to choose the right type for their problem. Additional statistics are needed to gauge how useful an instrument is for measuring changes over time. This article reviews statistics for reliability and also gives tips for designing reliability studies. The statistics discussed work the same whether applied to raters or time points.

Kappa for Categorical Scales

Categorical scales include binary variables such as whether an athlete has a concussion; ordered categories such as whether it's probable, uncertain, or unlikely that an athlete has a concussion; and unordered categories, such as race. Researchers commonly will report absolute agreement as a measure of interrater (or test-retest) reliability for categorical scales. For example, if 2 physicians evaluate 100 athletes and agree on the presence or absence of a concussion for 70 athletes, the absolute agreement is 70%. However, this statement ignores the fact that agreement can occur due to chance. If 2 physicians each just flipped a coin to decide, they would agree 50% of the time. The Kappa statistic corrects for this chance agreement.

There are multiple versions of Kappa—choosing the right one depends on how many raters (or time points) are involved, whether the categories are ordered or unordered, and other factors. Kappa is always less than or equal to 1, where 1 implies perfect agreement and 0 implies no better than chance. Kappa can be negative if raters agree less often than expected by chance.

Two Raters or Two Time Points

Cohen's Kappa is used when comparing only 2 raters or 2 time points. The formula for Cohen's Kappa is fairly intuitive. The numerator is the observed percent agreement minus the expected percent agreement due to chance. The denominator represents the maximum agreement possible when subtracting out chance:

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

Figure 1 gives some example calculations. If 2 physicians independently evaluate 100 athletes for concussion and agree in 50 cases—25 concussions and 25 nonconcussions—the Kappa value is 0 (Figure 1A). This is because the physicians are expected to agree by chance 50% of the time. If, however, the physicians agree on 35 concussions and 35 nonconcussions, then Kappa is (70% − 50%)/(1 − 50%) = 0.40 (Figure 1B), which is considered "fair agreement" (Table 1). Cohen's Kappa can be applied to measures with more than 2 categories, such as race or probable/uncertain/unlikely concussion.

Cohen's Kappa can underestimate agreement if the sample is too homogenous. If most subjects are concussed or most are not concussed, the expected agreement due to chance is high. For example, if raters rate 5% of subjects as concussed and 95% as not concussed, the expected agreement due to chance is a whopping 90.5% (see the Sidebar for more details on how to calculate the expected chance agreement). Figure 1C and D illustrate this issue. Statisticians recommend reporting a prevalence-adjusted Kappa alongside Cohen's Kappa when the one category predominates (see Hallgren [1] for more details).

For ordered categories, Cohen's weighted Kappa gives raters "partial credit" for being close even if they don't agree exactly. The weighted Kappa counts probable versus uncertain as a partial match, because this

http://dx.doi.org/10.1016/j.pmrj.2017.05.001

comes closer than probable versus unlikely. Figure 2 shows a hypothetical example: The unweighted Kappa is .39, but the weighted Kappa is .67 (using quadratic weights).

IN-DEPTH: CALCULATING EXPECTED CHANCE AGREEMENT

The expected chance agreement is calculated based on how frequently each rater picks each category. Imagine that 2 raters assign subjects to the concussed and nonconcussed categories based on the flip of a coin (heads = concussed, tails = nonconcussed). We expect them to both get heads (concussed) 50% × 50% = 25% of the time, and to both get tails (nonconcussed) 50% × 50% = 25% of the time—so the total expected chance agreement is 50%. Now imagine that the raters flip a biased coin to decide—they are still deciding randomly but now they have a preference for one category or the other. For example, say rater 1 picks the concussed category 70% of the time and rater 2 picks the concussed category 30% of the time. In this case, we expect them to both flip heads (concussed) 70% × 30% = 21% of the time, and to both flip tails (nonconcussed) 30% × 70% = 21% of the time. Therefore, the total expected chance agreement is 42%. Similarly, if rater 1 and rater 2 both pick the concussed category 95% of the time (still randomly guessing), then a given athlete will have a 95% × 95% = 90.25% chance of being placed in the concussed category by both raters; and a 5% × 5% = 0.25% chance of being placed in the non-concussed category by both raters. So, the expected chance agreement is 90.5%.

Three or More Raters or Time Points

Cohen's Kappa only works for 2 raters. For 3 or more raters, researchers may calculate pairwise Cohen's Kappas. They calculate a separate Cohen's Kappa for each pair of raters and then report the average Kappa or range of Kappas. For example, researchers in one study sought to examine the interrater reliability of physicians' diagnosis of mild traumatic brain injury in a patient after a military blast exposure [2]. Five physicians each classified 66 patients as brain injured or not. Researchers calculated Cohen's Kappas for all 20 possible physician pairs and found a range of 0.19-0.83, suggesting high variation in physician agreement. The mean was 0.43, suggesting fair agreement overall. Alternatively, Fleiss's Kappa is an extension of Cohen's kappa for 3 raters or more (see Hallgren [1] for more details), although a drawback is that it does not have a version for ordered categories.

For binary or ordered categories, researchers may opt to calculate an ICC (see section below: ICC for Continuous (or Ordinal) Scales) instead of a Kappa. Although the ICC was conceived for continuous variables, it can be applied to categorical variables when the categories are coded numerically (eg, 0/1 or 0/1/2/3). In fact, the Cohen's Kappa for binary variables and the quadratic-weighted Cohen's Kappa for ordered categories are mathematically equivalent to the ICC under certain conditions [3]. The advantage of using the ICC is that there is no limit on the number of raters or time points.

Interpreting Kappa or ICC

Statisticians have proposed varying guidelines for how to qualitatively interpret Kappa or the ICC (see Table 1 for examples). Although there is considerable disagreement about the lower categories (eg, what constitutes fair versus poor), researchers generally agree that measurement tools need ICC or Kappa values in the 0.75-0.80 or greater range to be useful for clinical research, and values greater than 0.90 are considered ideal.

Researchers should always report confidence intervals around both Kappa and the ICC. Take the example in Figure 1B. Here, Kappa is 0.40, which indicates fair agreement (according to the Cicchetti interpretation [4], which I prefer). However, the 95% confidence interval for Kappa is 0.22-0.58, which means that agreement could be as low as 0.22, or poor agreement. With small samples, the confidence interval for Kappa can easily span all the way from poor to excellent.

ICC for Continuous (or Ordinal) Scales

The ICC assesses reliability for continuous or ordinal scales. The ICC can handle any number of raters or time points. Like Kappa, the ICC will be 1 if all raters assign the exact same score to the same subject (perfect agreement), and it will be 0 if raters come no closer than expected by chance. When only chance is at work, the scores assigned to any one subject will be as variable as the scores assigned to different subjects. As with Kappa, it is important to always report confidence intervals for the ICC.

Take a simple example. Imagine that 3 raters apply a concussion severity score of 0 to 10 (with 10 meaning the most severe) to each of 5 subjects. Variation in the observed concussion scores comes from 2 sources: true differences in concussion severity between the subjects and measurement error. The ICC will be high when measurement error is low and low when measurement error is high relative to true subject variability.

Measurement error can be further subdivided. Some error is due to systematic difference between raters—one rater's "4" may equate to another rater's "2," for example. Raters also make random errors—scoring some subjects too high and others too low. Random