Miscellaneous Topics

Kevin O'Brien

December 22, 2015

Contents

1 Gold and Bronze Standards

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free. *It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard.* The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer*leaves considerable room for improvement* (Dunn, 2002). Pizzi (1999) similarly addresses the issue of glod standards: *well-established gold standard may itself be imprecise or even unreliable.*

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years. (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram ,used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the Angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. (This is reported as sensitivity of 95% and a specificity of 92%) (ACR, 2008)

In literature they are, perhaps more accurately, referred to as 'bronze standards'. Consequently when one of the methods is essentially a bronze standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

1.1 Fuzzy Gold Standards

The Gold Standard is considered to be the most accurate measurement of a particular parameter. But even gold standard raters must be assumed to have some level of measurement error. Fuzzy gold standard are considered by Phelps and Hutson (1994)

2 Fuzzball Agreement

Fuzzball agreement is a case where the correlation coefficient is close to zero. The sample values is restricted to a narrow range. but an examination of a relevant scatter-plot would indicate that there is agreement between the two methods.

Agreement - a numerical measure Hutson et al define a numerical measure for agreement.

For example, suppose the pairs of rater measurements are (1, 1), (1.1, 1), (1, 1.1), and (1.1, 1.1) then the sample Pearson correlation r = .0, yet the two raters or devices are considered to be in good agreement. We will refer to the instance where r is close to 0, yet there may be good agreement as "fuzzball agreement."

Fuzzball agreement occurs quite often in practice when the sample values have very narrow or restricted ranges. Fuzzball agreement is just one instance where the correlation coefficient is a poor measure of agreement.

Furthermore, note that the ICC is also a poor measure of agreement when there is fuzzball agreement. At the other extreme suppose the same raters given in the previous example had pairs of measurements (1, 101), (2, 102), (3, 103), and (4, 104) on the same relative scale as before. In this instance, r = 1.0, yet there is large disagreement between rater.

3 Types of Method Comparisons

Lewis et al. (1991) categorize method comparison studies into three different types, with the first two being of immediate concern. A method that is not considered to be a gold standard is referred to as an 'approximate method'.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard.

2. Comparison problems - When two approximate methods, that use the same units of measurement, are to be compared.

3. Conversion problems - When two approximate methods, that use different units of measurement, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement.

Dunn (2002) makes two important points in relation to these categories. Firstly he remarks that there isn't clear cut differences between each category.

Secondly he comments on the clinician gold standard, the sphygmomanometer, *leaves considerable room for improvement.* Pizzi (1999) also attends to this issue: *well-established gold standard may itself be imprecise or even unreliable.* The Magnetic resonance angiogram is considered to the gold standard for measuring aortic dissection, with a sensitivity of 95% and a specificity of 92% . (ACR, 2008) In literature they are, perhaps more accurately, referred to as 'bronze standards'.

Consequently when one of the methods is essentially a bronze standard, as opposed to a true gold standard, the comparison procedure should be considered as being of the second category.

4 Structural Equation Modelling

This is a statistical technique used for testing and estimating causal relationships using a combination of statistical data and qualitative causal assumptions. This technique was proposed by Lewis et al. (1991) as a method of assessing the reliability of a new measurement technique. It can indicate the presence of bias. However Bland and Altman (1987) have criticized it on the basis that it offers no insights into the variability about the line of equality.

In this paper, the SEM method is used to assess the linear relationship between the new method and the standard method.

Structural analysis is a generalization of regression analysis.

In Hopkins papers, a critique of the Bland-Altman plot he makes the following remark: *What's needed for a comparison of two or more measures is a generic approach more powerful even than regression to model the relationship and error structure of each measure with a latent variable representing the true value.*

Hopkins also adds that he himself is collaborating in research utilising SEM and Mixed Effects modelling.This is a methodology proposed by Kelly (1985).

5 ICC, Reproducibility Index and Passing-Bablok

5.1 Intraclass Correlation Coefficient

This measure of agreement is estimated using variance components from appropriate analysis of variance models. Measures of agreement are variance dependent, and so the ICC can be misleading. The ICC takes a value between 0 and 1, and is based on Analysis of Variance methodologies.

The ICC is a measure of reliability.

Bartko (1994) considers the ICC as just another measure of agreement.

5.2 Passing and Bablok (1983)

Passing & Bablok have described a linear regression model that are without the usual assumptions regarding the distribution of the samples and the measurement errors. The result does not depend on the assignment of the methods (or instruments) to X and Y. The slope and intercept are calculated with their 95% confidence interval.Hypothesis tests on the slope and intercept maybe then carried out.

If the hypothesis of the intercept is rejected, then it is concluded that it is significant different from 0 and both raters differ at least by a constant amount.

If the hypothesis of the slope is rejected, then it is concluded that the slope is significant different from 1 and there is at least a proportional difference between the two raters.

5.3 Lin's Reproducibility Index

Lin proposes the use of a reproducibility index, called the Concordance Correlation Coefficent (CCC).While it is not strictly a measure of agreement as such, we have included it.

References

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.