Gold Standards

Kevin O'Brien

September 16, 2015

Contents

?, 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' (**?**). **?** 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 (**?**).

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 **?**. 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% (**?**).

In literature they are, perhaps more accurately, referred to as 'fuzzy gold standards' (**?**). Consequently when one of the methods is essentially a fuzzy gold 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 Gold and Bronze Standards

?, 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* (**?**). **?** 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. (**?**).

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 **?**. 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%) (**?**)

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

? 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.

? 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.* **?** 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% . (**?**) 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.

3.1 Repeatability and gold standards

Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a 'bronze standard' exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a 'gold standard'. For example, by determining the ratio of CR to the sample mean $\bar{X}$. Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value of $\lambda = CR/\bar{X}$ with $\lambda < 0.1\%$. Similarly, a silver standard may be defined as the method with the lowest value of λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.