WJM World Journal of Methodology

Online Submissions: http://www.wjgnet.com/2222-0682office
wjm@wjgnet.com
doi:10.5662/wjm.v1.i1.12

World J Methodol 2011 September 26; 1(1): 12-14
ISSN 2222-0682 (online)

OBSERVATION

Challenges in estimating reproducibility of imaging modalities

Giovanni Di Leo

Giovanni Di Leo, Radiology Unit, IRCCS Policlinico San Donato, Piazza E. Malan, 20097 San Donato Milanese, Italy

Author contributions: Di Leo G solely contributed to this paper.

Correspondence to: Giovanni Di Leo, Assistant Professor, Radiology Unit, IRCCS Policlinico San Donato, Piazza E. Malan, 20097 San Donato Milanese, Italy. gianni.dileo77@gmail.com
Telephone: +39-2-52774468 Fax: +39-2-52774626

Received: August 12, 2011 Revised: September 5, 2011
Accepted: September 19, 2011
Published online: September 26, 2011

Peer reviewer: Domenico Rubello, MD, Professor, Head, Nuclear Medicine, PET Unit, S. Maria della Misericordia Hospital, Istituto Oncologico Veneto (IOV)-IRCCS, Viale Tre Martiri 140, 45100 Rovigo, Italy

Di Leo G. Challenges in estimating reproducibility of imaging modalities. *World J Methodol* 2011; 1(1): 12-14 Available from: URL: http://www.wjgnet.com/2222-0682/full/v1/i1/12.htm DOI: http://dx.doi.org/10.5662/wjm.v1.i1.12

Abstract

Estimating reproducibility is often wrongly thought of as basic science. Although it has a significant clinical relevance, its importance is underestimated. It was Alexander Pope in 1732 who was first to understand the value of reproducibility, with his famous comment “Who shall decide when doctors disagree?”. Pope’s question concerns the medical doctors’ opinion on a patient’s status, which from a statistical point of view may be considered a categorical variable. However, the same question may be posed for continuous quantitative variables. Reproducibility is complementary to variability: the larger the variability, the lower the reproducibility, and vice versa. Thus, we can think at them as interchangeable, even thought statistical methods have been developed for the estimation of variability. The question now is “Why do we need to know the reproducibility of measurements? ”. The most important and simplest answer is that we need to know how reliable a measured value or a subjective judgment is before taking clinical decisions based on this measurement/judgment. Integrating this knowledge in clinical practice is a key aspect of evidence-based medicine.

Key words: Reproducibility; Intraobserver; Interobserver; Imaging

“Who shall decide when doctors disagree?” This question, raised by Alexander Pope in 1732, must have been a very common one in Pope’s day, since medical practice at that time was based largely on tradition and opinion, not science. In the 21st century, medicine should be considered at least a combination of art and science. Consequently, careful clinical research should provide clear answers that stand the test of time and the scrutiny of additional investigations. This is the theory behind evidence-based, data-driven scientific medicine[1-3].

In scientific terms, when focused strictly on the evaluation of clinical variables, Pope’s question challenges reproducibility, in particular interobserver reproducibility[4-8]. It relates to the common experience where two independent observers provide different results, with this disagreement implying a sort of uncertainty about the truth. From the patient’s point of view, it may appear that his/her condition is not an objective one and that each clinician is allowed to have his/her own opinion. This may be very frustrating and cause the patient to lack trust in medicine.

In addition to interobserver reproducibility there is also intraobserver reproducibility, i.e. the ability of a single observer to provide the same opinion regarding a patient’s condition if he/she is questioned again later. In fact, self-disagreements occur more frequently than might be expected, in particular if the question posed has more than two mutually exclusive answers (categorical variables).

An example of efforts to better clarify intra- and interobserver reproducibility is the BI-RADS score system for breast lesions[9]. Based mainly on the appearance at mammography, radiologists may apply one of the following scores: (0) Incomplete, when mammograms do not give the radiologist enough information to make a clear diagnosis; (1) Negative, when there is nothing to comment on; (2) Benign, in presence of a definite benign finding; (3) Probably benign, in presence of findings that have a high probability of being benign; (4) Suspicious abnormality, in presence of a lesion not characteristic of breast cancer, but with reasonable probability of being malignant; and (5) Highly suspicious of malignancy, in presence of a lesion that has a high probability of being malignant.

Because of their different experience in reading mammograms, two independent observers may apply two different scores to the same image (lack of interobserver reproducibility). On the other hand, the learning curve of an individual radiologist, may mean that he/she will apply a score to a single mammogram different to that applied during a previous reading (lack of intraobserver reproducibility).

Intra- and interobserver reproducibility not only apply to categorical and ordinal variables but also, and more strictly, to quantitative (continuous) variables. Examples include cardiac ventricle volumes, a vessel diameter, arterial blood pressure, and body temperature. From the observer's point of view, the numerical values observed for such variables are obtained by mean of "instruments", i.e. technical systems, based on a physical principle, that are sensitive to the quantity to be measured. Many of these instruments are now available as software algorithms implemented on computers used for imaging techniques.

Even if the use of a technical instrument may lead an observer the believe the measurement to be an objective process without uncertainty, we must remember that this process does not proceed by itself and that it needs the observer's intervention. This intervention may apply at any level and certainly impacts on the final observed value. For example, the measurement of a vessel diameter based on a magnetic resonance image needs the observer to place a ruler between two distant points (the vessel boundaries) and the repetition of this action rarely provides the same value as that previously obtained. Furthermore, an independent observer may perform this measurement by placing the ruler at another part of the vessel course, i.e. on another slice of the magnetic resonance scan. Therefore, as for categorical variables, the measurement of continuous variables also is characterized by intra- and interobserver variability.

Reproducibility and variability are two complementary concepts: the larger the variability, the lower the reproducibility, and vice versa. Thus, we may think of them as interchangeable, even though statistical techniques have been developed for estimating variability. Moreover, intra- and interobserver variability are only two of the possible sources of the total variability of a measurement obtained using imaging techniques. In general, if an examination on a patient is repeated after a treatment, the total variability associated with the measurements will consist of the following components: (1) The intraobserver variability of the radiologist who performed the measurement prior the treatment; (2) Intraobserver variability of the radiologist who performed the measurement after treatment; (3) The interobserver variability between those radiologists; (4) The interstudy variability, due to the repetition of the examination; (5) The inter-instrumentation variability, due to the possible use of two different machines; and (6) The biological variability, due to changes in the patient's health status during the time elapsed between the two examinations (the effect of treatment may also be a part of this variability).

Why do we need to know the variability of measurements of categorical and continuous variables? The most important and simplest answer is because we need to know the reliability of measured values before taking decisions based on those measurements! Recalling the previous example, if we observe a difference between the values measured before and after the treatment, can we establish that the patient's health status is changed, or is that difference within the overall variability? Of course, the only way to answer that question is to know the overall variability.

In theory, one way to estimate the measurement variability is to repeat a measurement many times, to calculate the mean value and the 95%-confidence interval. However, this approach has three important limitations. Firstly, it no longer holds if the measurements are taken by different observers, adding interobserver variability. Secondly, in clinical practice there is little or no time available for repeating the same measurement. Thirdly, although this allows estimation of the variability associated to a particular value, that variability cannot be applied to all possible values. Therefore, it is more practical to perform a preliminary analysis of at least intra- and interobserver variability.

The statistical techniques suitable for the estimation of the intra- and interobserver variability depend only on the type of the measured variables. Two main methods are available: Cohen *k* statistics for categorical variables[5] and Bland-Altman statistics for continuous variables[7,8]. Here, I will not go into the mathematical details of these methods (a complete description may be found in references[4]), but I would like to highlight the main difference between them. The Cohen *k* method provides a coefficient of agreement that lies within the range (-1, 1), where $k = 1$ indicates perfect agreement, $k = 0$ absolutely no agreement, and $k = -1$ the "perfect disagreement". Conversely, Bland-Altman analysis results in a value expressed with the same measurement units as the measured variable.

The estimation of the intra- and interobserver variability may be performed in parallel. In clinical settings, a suitable protocol would include two observers with different experience in the measurement under evaluation. The less experienced observer should measure the variable of interest twice for each patient, with the more experienced observer making only one measurement per patient. The intraobserver variability may be estimated using the pairs of values obtained by the first observer, while the interobserver variability may be estimated using the first value of the first observer and the single value obtained by the second observer.

Let me conclude with an example taken from my own experience as an author. In 2008 we demonstrated that the interobserver variability in the measurement of the left ventricle ejection fraction on magnetic resonance imaging may be as large as 17%, in absolute units[10]! This means that if an observer obtains a value of, for example, 50% for a patient's ejection fraction, a second observer may obtain a value of between 33% to 67% for the same patient. Considering such variability, I can only smile when I see continuous variables expressed to two or three decimals places.

REFERENCES

1 **Sardanelli F**, Hunink MG, Gilbert FJ, Di Leo G, Krestin GP. Evidence-based radiology: why and how? *Eur Radiol* 2010; **20**: 1-15

2 **Malone DE**. Evidence-based practice in radiology: an introduction to the series. *Radiology* 2007; **242**: 12-14

3 **Sackett DL**, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; **312**: 71-72

4 **Sardanelli F**, Di Leo G. Biostatistics for radiologists. Milan: Springer, 2009

5 **Cohen J**. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**: 37-46

6 **Di Leo G**, Di Terlizzi F, Flor N, Morganti A, Sardanelli F. Measurement of renal volume using respiratory-gated MRI in subjects without known kidney disease: Intraobserver, interobserver, and interstudy reproducibility. *Eur J Radiol* 2010; Epub ahead of print

7 **Bland JM**, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135-160

8 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307-310

9 **American College of Radiology**. ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas. Reston, VA: American College of Radiology, 2003

10 **Sardanelli F**, Quarenghi M, Di Leo G, Boccaccini L, Schiavi A. Segmentation of cardiac cine MR images of left and right ventricles: interactive semiautomated methods and manual contouring by two readers with different education and experience. *J Magn Reson Imaging* 2008; **27**: 785-792

S- Editor Wang JL **L- Editor** Hughes D **E- Editor** Zheng XM