Assessment of Disagreement: A New Information-Based Approach

CRISTINA COSTA-SANTOS, MSc, LUÍS ANTUNES, PhD, ANDRÉ SOUTO, MSc,
AND JOÃO BERNARDES, PhD

PURPOSE: Disagreement on the interpretation of diagnostic tests and clinical decisions remains an important problem in medicine. As no strategy to assess agreement seems to be fail-safe to compare the degree of agreement, or disagreement, we propose a new information-based approach to assess disagreement.

METHODS: The sum over all logarithms of possible outcomes of a variable is a valid measure of the amount of information contained in the variable. The proposed measure is based on the amount of information contained in the differences between two observations. This measure is normalized and satisfies the flowing properties: it is a metric, scaled invariant and it has differential weighting.

RESULTS: Two real examples and a simulation were used to illustrate the usefulness of the proposed measure to compare the degree of disagreement.

CONCLUSIONS: Used as complement to the existing methods, we believe our approach can be useful to compare the degree of disagreement among different populations.

Ann Epidemiol 2010;20:555–561.

KEY WORDS: Observer Variation, Information Theory, Reproducibility of Results.

INTRODUCTION

Measurement is essential both for clinical care and for epidemiologic research; however measurement always implies some degree of variability (1). Ideally the only source of variability in measurements should be the variability within subjects, however, observer variability and other sources of variability are also often present.

In clinical care, diagnosis often depends on measurements, and disagreement in these measurements may have obvious implications for clinical practice and may also have medico-legal consequences (2). Before the introduction, in clinical practice, of a new diagnostic method, it is essential to assess the agreement between the new method and the traditional one and it is also fundamental to know whether the new method can be reproduced by a second observer. In research, disagreement in measurements may lead to discrepant results in validity or randomized controlled studies with the same objectives and with the same methods, consequently misleading and heterogeneous results in meta-analysis will be found (2). Because it is impossible to control the various sources of variation of a measurement, agreement studies have a very important role. Despite the importance of agreement studies, misleading and sometimes inappropriate measures of agreement have been used in leading medical literature (2, 3). Considering the limitations of current strategies to assess agreement and as no strategy seems to be fail-safe to compare the degree of agreement among different populations, agreement studies should be interpreted with caution and possibly combined with the use of different strategies to assess agreement always with the limitations of these strategies in mind (2, 4).

In this article, we propose a new approach to assess agreement based on information theory.

From the Biostatistics and Medical Informatics Department (C.C.-S.) and Obstetrics and Gynecology Department (J.B.); Faculty of Medicine, and the Computer Science Department, Faculty of Science (L.A.), University of Porto, Portugal; the Center for Research in Health Technologies and Information Systems, Porto, Portugal (C.C.-S.); the Instituto de Telecomunicações, Porto, Portugal (L.A., A.S.); and the Instituto de Engenharia Biomédica, Porto, Portugal (J.B.).

Address correspondence to: Cristina Costa-Santos, MSc, Biostatistics and Medical Informatics Department, Faculty of Medicine, University of Porto, Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal. Fax: +351-22-551-3623. E-mail: csantos@med.up.pt.

Received July 7, 2009; accepted February 17, 2010.

METHODS

Most Used Strategies to Assess Agreement

The intraclass correlation coefficient (ICC) is claimed to be suitable for observer agreement assessment (5). The ICC is defined as the ratio of the variance between subjects, to the total variance (6, 7). These variances are derived from analyses of variance (ANOVA). Fleiss and Shrout present different kinds of ICC derive from different ANOVA models, and the ANOVA model depends on the study design (6). One-way random effects model should be used when the subjects are deemed a random sample of subjects to be evaluated by the observers. The focus of interest is the difference of the individual observer's rating from the average rating of the observers for the i^{th} subject (6).

1047-2797/$–see front matter
doi:10.1016/j.annepidem.2010.02.011

Selected Abbreviations and Acronyms

ANOVA = analyses of variance
CCC = concordance correlation coefficient
ICC = intraclass correlation coefficient

In two-way models the observers are deemed an important factor in the ICC computation. In two-way random effects model, not only the subjects are deemed random, but the observers are deemed a random effect as well. In two-way mixed model each subject is evaluated by k observers, who are the only observers of interest, in this case, the observers are a fixed effect whereas the subject ratings are a random effect (6).

The ICC (from two-way models) that should be used for assessing agreement was defined by McGraw and Wong as the "ICC for agreement." We obtain the "ICC for consistency" or the "ICC for agreement" excluding or not the observer variance from the denominator mean square, respectively (8). The systematic variability due to observers is irrelevant for "ICC for consistency" but it is not for "ICC for agreement."

The ICC ranges from 0 (no agreement) to 1 (perfect agreement), however, it can be negative, how such negative ICC values should be interpreted is quite unclear. The ICC assumptions, multivariate normal distributions and equality of variances, should be checked.

An important limitation of ICC is that it is strongly influenced by the variance of the trait in the population in which it is assessed (9). This limitation can be illustrated with the following example: suppose that we aim to assess whether a depression scale can be reproduced by a second observer when applied to a random sample of the adult population (a heterogeneous population, with high variance) the scale's ICC may be higher than when the scale is applied to a very homogeneous population (with low variance), such as patients hospitalized for acute depression. This ICC characteristic has also been considered by some authors as an advantage, for it would make the discordance relative to the magnitude of measurements (10), however comparability across populations is not possible with ICC. Consequently, ICC values have no absolute meaning, and the cut-off value of 0.75 proposed by Burdock et al. (7) and Lee et al. (11), and often reproduced to signify a good agreement, has limited justification.

Lin's concordance correlation coefficient (CCC) is the Pearson correlation of coefficient, which assesses the closeness of the data to the line of best fit, modified by taking into account how far the line of best fit is from the 45º line through the origin (12). Lin objected to the use of the ICC as a way to assess agreement between methods of measurement and developed the CCC. However, there are similarities between certain specifications of the ICC and the CCC (13). Some limitations of ICC, like the limitation of comparability of population described above, are also present in CCC (14).

Bland and Altman (15) propose the limits of agreement to assess agreement between methods of measurement. Limits of agreement can be calculated based on the mean difference between the measurements of two methods in the same subjects and the SD of these differences. Approximately 95% of these differences will lay between the mean differences ± 1.96 standard deviation of these differences. The limits of agreement approach depends on some assumptions about the data: the mean and SD of the differences are constant throughout the range of measurement and these differences are approximately normally distributed (15). Limits of agreement are expressed in terms of the scale of measurement and the decision whether a degree of agreement is acceptable or not is always a clinical, not statistical, judgment.

Other approaches (16, 17) have been proposed for assessing agreement; however, all of them are also limited when the aim is to compare the agreement in different populations with different trait characteristics.

The New Approach: Information-Based Measure of Disagreement

Entropy, introduced by Shannon (18), can be described as the average amount of information contained in a variable. A high value of entropy means that a large amount of information is needed to describe an outcome variable about which we have high uncertainty. The sum over all logarithms of possible outcomes of the variable is a valid measure of the amount of information, or uncertainty, contained in a variable. Consider that we aim to measure disagreement between measurements obtained by Observer A (variable A) and Observer B (variable B). Also, consider for variable A, a vector A that can take the range of non-negative values $(a_1,\ldots,a_n)$ and for variable B, a vector B that can take the range of non-negative values $(b_1,\ldots,b_n)$. The intuition behind our definition is that the disagreement between A and B is related to the differences between them, with the minimum reached when A and B are identical. So, it is then natural to consider $\sum_{i=1}^{n} \log_2|a_i - b_i|$ the amount of information contained in the differences between Observers A and B. By adding 1 to the differences, we avoid the behavior of the logarithmic function between 0 and 1, while keeping a quite natural close relation to the notion of entropy where Shannon considered the log of the inverse of a probability, i.e., the log of a value always greater or equal to 1. To get a value between 0 and 1 we normalize the amount of information contained in the differences to obtain the following measure of disagreement.