Computational Statistics and Data Analysis 56 (2012) 1052–1060

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Assessment of observer agreement for matched repeated binary measurements

Jingjing Gao [a,*], Yi Pan [b], Michael Haber [b]

[a] Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, United States

[b] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, United States

ARTICLE INFO

Article history:
Received 29 September 2010
Received in revised form 31 August 2011
Accepted 3 November 2011
Available online 12 November 2011

Keywords:
Coefficient of individual agreement
Observer agreement
Binary data
Repeated measurements
Generalized linear mixed model

ABSTRACT

Agreement is a broad term simultaneously covering evaluations of accuracy and precision of measurements. Assessment of observer agreement is based on the similarity between readings made on the same subject by different observers. The assessment of agreement on categorical observations is traditionally based on kappa or weighted kappa coefficients. However, kappa statistics have been criticized because they attain implausible values when the marginal distributions are skewed and/or unbalanced. New scaled indices called the coefficients of individual agreement (CIAs) have been developed for the assessment of individual observer agreement by comparing the observed disagreement between two observers to the disagreement between replicated observations made by the same observer on the same subject. This is based on the notion that under a good agreement, the disagreement between the two observers is usually not expected to exceed the disagreement between replicated observations of the same observer, and hence, a satisfactory agreement is established if these quantities are similar. This idea is extended and a new method utilizing the generalized linear mixed model is proposed to estimate the CIAs for binary data which consist of matched sets of repeated measurements made by the same observer under different conditions. The conditions may represent different time points, raters, laboratories, treatments, etc. The new approach allows the values of the measured variable and the magnitude of agreement to vary across the conditions. The reliability of the estimation method is examined via a simulation study. Data from a study aiming at determining the validity of diagnosis of breast cancer based on mammography are used to illustrate the new concepts and methods.

1. Introduction

Accurate and precise measurements constitute an important component of any proper study design. Ideally, a quantity or trait should be measured without an error. However, in many cases it is impossible to come up with an exact measurement of the "true value" of the quantity being measured. For example, in a study on depression, patients can be classified as "very depressed", "somewhat depressed" or "not depressed". Since these assessments are subjective, there is no "true value" of the magnitude of depression. Even when the true value can be defined, it may be very difficult to measure it with perfect accuracy. In situations where it is difficult or impossible to determine the true value, usually more than one measurement is made on each subject, preferably by more than one observer, device or measurement method. In

* Corresponding author. Fax: +1 4047120109.
E-mail address: jgao@emory.edu (J. Gao).
URL: http://www.sph.emory.edu/observeragreement (J. Gao).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.11.005

these cases, it is very important to evaluate the agreement between these measurements. In the most ideal situation, two observers are said to be agreeing with each other only if they could produce identical results. However, it is often unrealistic to require readings from different observers to be the same or requiring each measurement to be identical to the truth due to unavoidable measurement errors and the fact that the ground truth might not be available. Therefore, agreement studies aim at quantifying the "closeness between readings", which covers both the accuracy assessment and the precision evaluation.

Observer agreement is traditionally assessed using either unscaled or scaled agreement measures (Barnhart et al., 2007b). The unscaled agreement metrics measure the absolute difference between readings. For continuous data, the unscaled agreement indices mainly include the mean squared deviation (MSD) (Lin et al., 2002, 2007), the coverage probability (CP) (Lin, 2000b; Lin et al., 2002, 2007) and the total deviation index (TDI) (Lin, 2000b; Lin et al., 2002, 2007). On contrast, the difference between measurements made by the same or distinct observers can be scaled. The scaled agreement indices for continuous measurements include the intraclass correlation coefficient (ICC) (Bartko, 1966, 1974; Shrout and Fleiss, 1979; Eliasziw et al., 1994; Muller and Buttner, 1994; McGraw and Wong, 1996), the concordance correlation coefficient (CCC) (Lin, 1989, 1992, 2000a; Lin et al., 2002, 2007; King and Chinchilli, 2001a,b; King et al., 2007; Barnhart et al., 2002, 2005, 2007c), the coefficient of inter-observer variability (CIV) (Haber et al., 2005), and the coefficient of individual agreement (CIA) (Barnhart et al., 2007a; Haber et al., 2007; Haber and Barnhart, 2008). An overview on assessing agreement with continuous measurements was published by Barnhart et al. (2007b).

Agreement between observers making qualitative observations is usually evaluated via the kappa statistic (Cohen, 1960) and the weighted kappa statistic (Cohen, 1968). However, several researchers have argued that the kappa statistics may not perform satisfactorily under certain situations, especially when the marginal distributions are skewed and/or unbalanced (Feinstein and Cicchetti, 1990; Kraemer, 1979). In addition, the role of kappa statistics for analyzing replicated and repeated measurements is limited because it can be applied when each observer makes a single reading on each subject.

Recently, Barnhart et al. (2007a), Haber et al. (2007) and Haber and Barnhart (2008) proposed new coefficients called the coefficients of individual agreement (CIAs) for assessing observer agreement in studies involving unmatched replications which are made under the "same condition". By the "same condition", it is assumed that the subject's true value does not change between the replicated readings. By "unmatched", it is assumed that permuting the replications within one observer would not affect the order of the replications of the other observer. Frequently, agreement studies are designed in a way that multiple observers make multiple readings on each subject, where the readings are matched on a factor whose levels are considered as conditions and where the subjects' true values may change across conditions. These observations are then considered as matched repeated measurements.

Haber et al. (2010) extended CIAs for assessing observer agreement for matched repeated continuous measurements. In this paper, we further extend the concepts and ideas of the CIAs for assessing observer agreement for data consisting of matched repeated dichotomous observations made by the same observer under different conditions. These conditions may correspond to different time points, laboratories, devices, treatments and so forth. Our approach allows the values of the measured variables and the magnitude of agreement to vary across the conditions.

For binary measurements, we assume that the readings follow specific generalized linear mixed models (GLMMs). Then, the parameter and variance estimates from the mixed models can be adapted to calculate the inter- and intra-observer disagreements and hence estimate new CIAs for this type of data.

2. Definition of coefficients of individual agreement

Instead of deriving a direct index for agreement, the concept of the complement of agreement, which is disagreement, is considered. A strong disagreement certainly indicates a poor agreement, and vice versa. We denote the measurements of two observers by Y_1 and Y_2. Let $G(Y_1, Y_2)$ denote the inter-observer disagreement. The disagreement between the observers under condition h can be quantified by the mean squared deviation (MSD), defined as $G_h(Y_1, Y_2) = \mathrm{MSD}_h(Y_1, Y_2) = E[(Y_1 - Y_2)^2|h]$, where the expectation is over the joint distribution of (Y_1, Y_2). Particularly, for binary observations, the possible values for $(Y_1 - Y_2)^2$ are zero when $Y_1 = Y_2$; or one when $Y_1 \neq Y_2$, as a result, it reduces to $G_h(Y_1, Y_2) = \mathrm{MSD}_h(Y_1, Y_2) = E[(Y_1 - Y_2)^2|h] = 0 \cdot \Pr(Y_1 = Y_2|h) + 1 \cdot \Pr(Y_1 \neq Y_2|h) = \Pr(Y_1 \neq Y_2|h)$. Also, let $G_h(Y_j, Y_j')$ indicate the disagreement between two (hypothetical) replicated observations of observer j $(j = 1, 2)$ under the same condition h. To measure the intra-observer disagreement $G_h(Y_j, Y_j')$, we use the mean squared deviation between two replicated readings made by observer j $(j = 1, 2)$ under the same condition h $(h = 1, \ldots, H)$, i.e., $G_h(Y_j, Y_j') = \mathrm{MSD}_h(Y_j, Y_j') = E[(Y_j - Y_j')^2|h]$. For binary data, it is equal to $\Pr(Y_j \neq Y_j'|h)$.

The CIAs are based on the comparison of the probability of disagreement between two observers to the probability of disagreement between replicated observations made by a single observer. The rationale for this approach, which is commonly used in individual bioequivalence studies (Barnhart et al., 2007a), lies in that when two or more observers can be used interchangeably, then we can expect the variability of observations made by different observers to be similar to the variability of observations made by the same observer. Therefore, we summarize the magnitude of agreement as a ratio comparing intra-observer disagreement to inter-observer disagreement. We define different CIAs for the cases of comparing two observers without the presence of a reference and for comparing a new observer to an established "gold standard".