ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 963–967

Journal of Clinical Epidemiology

Assessment of agreement of a quantitative variable: a new graphical approach

Ronir Raggio Luiz*, Antonio José Leal Costa, Pauline Lorena Kale, Guilherme L. Werneck

NESC/UFRJ, Prédio do HUCFF, 5º andar Cidade Universitária, Ilha do Fundão 21941-590, Rio de Janeiro, Brazil

Accepted 20 May 2003

Abstract

In clinical or epidemiologic research, the measurement of variables always implies some degree of error. Because it is impossible to control the various sources of variation, the assessment of the reliability of a measurement is essential. Otherwise, concordance analysis must take into account the "clinical" interpretation of the measurement under study, because its practical usefulness is of central importance. In this article, we propose a new approach to assess the reliability of a quantitative measurement. We use a graphical approach familiar to statisticians and data analysts of the biomedical area, associating to it the useful feature of interpretation based on the proportion of concordant cases. We believe that the proposed graphical approach can serve as a complement, or as a alternative, to the Altman-Bland method for agreement analysis. It allows a simple interpretation of agreement that takes into account the "clinical" importance of the differences between observers or methods. In addition, it allows the analysis of reliability or agreement, by means of survival analysis techniques.

Keywords: Agreement; Reliability; Graphical approach; Clinical relevance; Altman-Bland approach

1. Introduction

In clinical or epidemiologic research, the measurement of variables always implies some degree of error. In an idealised investigation, the only source of variation of a certain measure should be the one inherent to the different individuals included in the study. Nevertheless, various other sources of variation—such as within or between observer variation—are common, and imply potential biases or an increase of imprecision in the analysis of the phenomenon [1]. Obtaining reliable measures constitutes, therefore, one of the major challenges in clinical or epidemiologic research. Because it is impossible to control the various sources of variation of a measurement, the assessment of its reliability is essential. Before using a new diagnostic method, for example, it is important to evaluate in which extension its results differ from that obtained by applying a more traditional method. In certain situations, the difference between the two methods may not be sufficient to impair clinical interpretations, then the new one could substitute the traditional, or the two methods may be used indistinctly [2].

When one wishes to assess the agreement of a measurement in relation to a "gold standard," conformity studies are usually performed [3]. On the other hand, studies of consistency are used when within as well as interobserver reproducibility of measurements are of major interest. In other words, we use the concept of consistency when referring to the concept of agreement when none of the measurements is taken as "correct." In general, these two types of studies can be aggregated in a more inclusive statistical approach, commonly known as agreement studies, although these studies may refer to different concepts [4]. Conformity may be designated, also, as accuracy and validity, and consistency as reliability, repeatability, and reproducibility [3].

Otherwise, to analyze adequately a measurement, researchers should not restrict themselves to standard statistical procedures [2,5]. For instance, as in the situation in which constructing reference limits is the major interest the researcher is called to evaluate the "clinical" relevance of the obtained values [6,7]. Analogously, all concordance analysis must take into account the "clinical" interpretation of the measurement under study, because the practical usefulness of the measurement is of central importance.

From the practical standpoint, an analysis of agreement is dependent upon the scale of measurement of the variable [4], notwithstanding the proposals of a unifying approach by means of a generalized coefficient [8]. For categoric variables the utilization of the kappa coefficient of agreement is classic. Although some of its characteristics may impair its interpretation—as, for example, the prevalence bias [4,9]—this coefficient has the advantage of being defined as the proportion of concordant cases, discounted those

* Corresponding author: Tel.: +55-21-25626235; fax: +55-21-25626220.
E-mail address: ronir@acd.ufrj.br (R.R. Luiz).

0895-4356/03/$ – see front matter
doi: 10.1016/S0895-4356(03)00164-1

that are in concordance due to mere chance. This definition makes easier the evaluation of the "clinical reliability" of a measure. For continuous or discrete quantitative variables, however, the analytic approaches are different, and not immediately associated to the interpretation of this definition of the kappa coefficient.

In this article, we propose a new approach to assess the reliability of a quantitative measurement. We use a graphical approach familiar to statisticians and data analysts of the biomedical area associating to it the useful feature of interpretation based on the proportion of concordant cases.

2. Approaches to the evaluation of agreement of a quantitative variable

The usual statistical approach used to evaluate the agreement of a clinical measure is the estimation of coefficients that quantify the degree of agreement [10]. Due to statistical reasons or to the lack of immediate "clinical" interpretation of these statistics [4], this approach, even when using an appropriate coefficient, does not seem sufficient to describe the reliability of any measure, in particular, a quantitative one.

Pearson's correlation coefficient is inadequate, in various instances, to assess agreement, because it evaluates only the association of two sets of observations [3,11]. The intraclass correlation coefficient (ICC), otherwise, has been considered appropriate for the evaluation of both consistency and conformity studies, because it is capable of estimating the proportion of the total variation due to the variability between independent units of analysis. There are many variants of the ICC, and in consequence, its calculation is strongly influenced by the study scene. Müller and Bütter [3] propose the use of a decision tree to choose which variant of ICC should be preferred. Another limitation of this coefficient is its dependence upon the degree of variability within and between observations. For the same degree of variation within observations, the greater the variation between observations, the greater will be the ICC [12]. But this same characteristic has been considered an advantage, for it would make the discordance relative to the magnitude of the measurement [13]. Bartko [14] considers the ICC as just another measure of agreement, and proposes the use of statistical tests—such as the paired *t*-test—in the evaluation of the agreement, albeit the inadequacy of this approach [3,15]. That the paired *t*-test with a nonsignificant result does not indicate agreement is another limitation of this approach.

In view of these problems, agreement analysis centered in synthetic measures or in statistical tests could lead to scarcely useful results, from the applied standpoint.

Altman and Bland [11], in 1983, proposed to quantify agreement by construction of limits of agreement. These statistical limits were calculated by using the mean and the standard deviation of the differences. To check the assumptions of normality of differences and other characteristics, they used a graphical approach. The graphic is a scatter plot XY, in which the Y axis shows the difference between the two measurements (A−B) and the X axis presents the average of these measures ((A+B)/2). An evaluation of the correlation between these two new figures can complement the analysis. With this graphic, the evaluation of the magnitude of the disagreement, the identification of outliers, and the observation of any bias are easily performed.

Notwithstanding the opposition of certain authors [13], we consider the Altman-Bland approach the preferred one to evaluate agreement between two measurements. We think so in view of its simplicity, and, even more, of its potential for the identification of pairs of observations whose differences reach beyond clinical tolerances. In addition, if one counts the frequency within certain limits, studying agreement is made easier, an idea that has already been pointed out earlier [2,16,17]. This idea is central to the graphical approach to agreement that we present next.

Another graphical approach to assess agreement—rather similar to the approach proposed here but yet been cited only rarely—is the mountain plot [18,19]. This graphic, also called the folded empirical cumulative distribution plot, is prepared by computing a percentile for each ranked difference between two measurements. To get a folded plot, one performs the subtraction 100 percentile for all percentiles over 50. The authors recommend it as a complement of the Altman and Bland plot [19].

3. The new approach

The Altman-Bland approach permits an evaluation of the agreement and incorporates some limits of tolerance that have clinical relevance. It is possible, however, to extend the agreement evaluation through a graphic capable of expressing the degree of agreement (or disagreement) as a function of several limits of tolerance. We can, for example, construct a graphic, such as the Kaplan-Meier, used in the analysis of survival data [20]. The "failure" would happen exactly at absolute values of the observed differences between the methods. Thus, if in the X axis we have the module of the observed differences, and in the Y axis we have the proportion of cases with differences that are at least the observed difference (x_i), then we have a step function typical of a survival analysis, without censored data, with the Y axis representing the proportion of discordant cases. A possible name for this new approach could be "survival-agreement plot."

As an example, let us consider Table 1, which registers the inferior pelvic infundibular angle (IPIA) for 52 kidneys, evaluated by means of computerized tomography and urography. Due to the financial costs of a tomography, obtaining reliable results through urography would be convenient for the diagnoses and treatment of renal lithiasis. It is possible to detect a disagreement between these two methods. This disagreement should be evaluated through the incorporation of some "clinical information," to answer

Table 1
Inferior pelvic infundibular angle (IPIA), in degrees, by urography and tomography (n = 52 kidneys)

	Method			Method	
Kidney	Urography	Tomography	Kidney	Urography	Tomography
1	100º	97º	27	40º	45º
2	58º	77º	28	70º	60º
3	95º	74º	29	63º	50º
4	55º	59º	30	103º	94º
5	79º	79º	31	95º	91º
6	95º	85º	32	80º	66º
7	60º	78º	33	72º	63º
8	88º	78º	34	68º	65º
9	68º	68º	35	48º	58º
10	94º	96º	36	70º	75º
11	60º	74º	37	90º	105º
12	64º	64º	38	60º	65º
13	88º	76º	39	80º	80º
14	57º	60º	40	96º	90º
15	66º	78º	41	54º	58º
16	67º	71º	42	80º	75º
17	76º	67º	43	88º	83º
18	95º	103º	44	70º	78º
19	85º	95º	45	90º	85º
20	105º	78º	46	79º	65º
21	80º	70º	47	100º	90º
22	85º	80º	48	85º	76º
23	82º	78º	49	108º	100º
24	102º	102º	50	53º	65º
25	100º	102º	51	58º	40º
26	75º	77º	52	49º	53º

if the difference between the methods actually does or does not have any relevance, from the clinical standpoint.

Table 2 presents frequencies of differences, in modules, between the two graphs, as well as the accumulated percentage of cases, in consonance with these differences. To construct the graphic, we used column 3 of Table 2, which would be available, as well as its corresponding graph, from any standard statistical software used to perform survival analysis by the Kaplan-Meier method.

Table 2
Distribution of absolute difference between urography and tomography IPIA for 52 kidneys

Absolute difference (x_i)	Frequency	Cumulative percent ($>x_i$)
0º	5	90.4
2º	3	84.6
3º	3	78.8
4º	6	67.3
5º	7	53.8
6º	1	51.9
8º	3	46.2
9º	4	38.5
10º	7	25.0
12º	3	19.2
13º	1	17.3
14º	3	11.5
15º	1	9.6
18º	2	5.8
19º	1	3.8
21º	1	1.9
27º	1	0.0

Fig. 1 presents the results of the agreement analysis for the two methods used to measure the IPIA. We think that a useful interpretation of an analysis of agreement must lay explicit its dependence upon clinical limits of tolerance. The graphic, however, shows the discordance, to maintain the analogy with the survival analysis. Any measurement of agreement, thus would be calculated through the difference, and would be represented, in the graphic, by the distance between the curve and the superior limit of the Y-axis (100%). In Fig. 1, if we establish a tolerance limit of 5º, we will obtain an agreement of less than 50%. To obtain an agreement of 90%, a difference not inferior to 15º is needed. It is easy to visualize other estimates of agreement as a function of "clinical" limits (X-axis). The lack of precision of these estimates that are due to the sample size can be immediately obtained through any software that disposes of the Kaplan-Meier method.

This approach is also useful in the comparison of more than two measures. Bland and Altman [2] present the example of measurements of systolic blood pressure of 85 individuals, by two experienced observers (J and R) with a sphygmomanometer, and one other measurement, by a semiautomatic device. As there are three observations, three comparisons are possible. Fig. 2 shows the resulting graph. We can clearly observe a much greater agreement when the two observers are compared. For a difference of, at most, 2 mmHg—a very acceptable difference from the clinical standpoint—the agreement exceeds 90%. Comparing the measurements made with the semiautomatic blood pressure monitor to the measurements of any one of the two observers leads to worse, albeit similar, results. For instance, a degree of agreement of 90% occurs only when differences reach 50 mmHg, which is unacceptable from the clinical standpoint.

To verify if the agreement depends upon some categoric covariate, we can construct the agreement step curve for each level of the covariate and compare these curves. To deal with a continuous variable—including the magnitude of the measurement in itself—it would be necessary to subdivide the data, and then construct the agreement curves. When there is no dependence on the covariate, we would expect superposition of the agreement curves; if there is dependence, the category with the step curve nearest to the origin of the axes should exhibit the greatest agreement.

4. Discussion

The proposed approach presents advantages and disadvantages compared to the Altman-Bland proposal or to mountain plot. These two approaches, depending on the research interests, can complement or serve as alternative one to another. For small data sets, the Altman-Bland approach should be used, and the incorporation of any index would be of little value to the analysis.

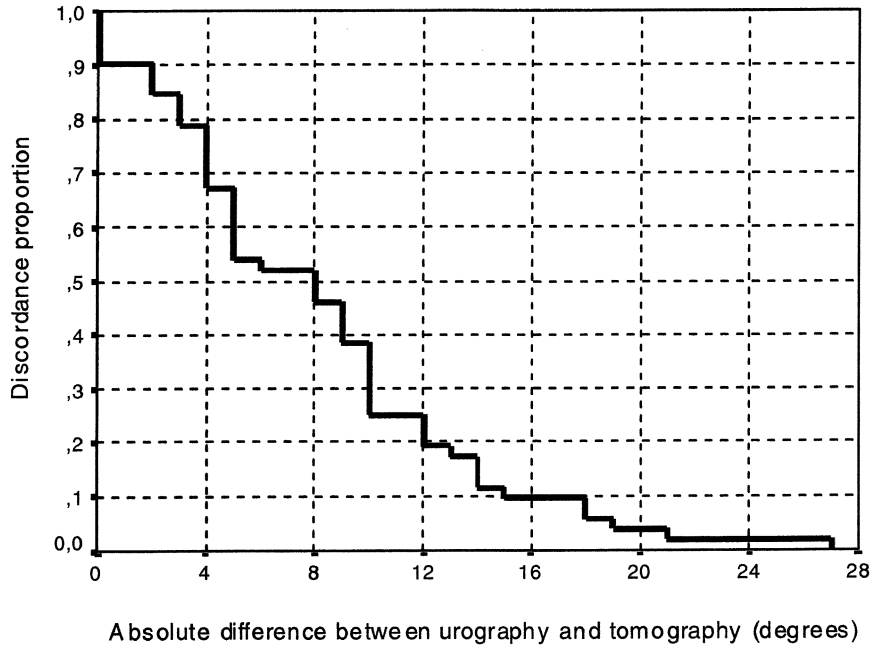

Fig. 1. Proportion of discordance between urography and tomography IPIA until "tolerance" limits.

In general, the use of measures of agreement—particularly in the case of a quantitative variable—is difficult, because these measures are calculated, and interpreted, almost strictly, from a statistical standpoint. The proposed approach has the advantage of expressing, by means of a graphic, an index that is easily interpreted, and depends upon the degree of relevance of the agreement, as judged by the researcher. Moreover, the form of the resulting step function can also yield much information: very high steps indicating that a better agreement will be reached more rapidly, that is, for smaller differences.

However, in choosing the module of the difference, for example, we loose sight of some characteristics of the differences, very much evident in the Altman-Bland approach or in the mountain plot. These characteristics can have a great impact in the study of agreement [21]. The average of the differences can serve as an estimate of the bias among the methods. Besides the bias, a tendency in the

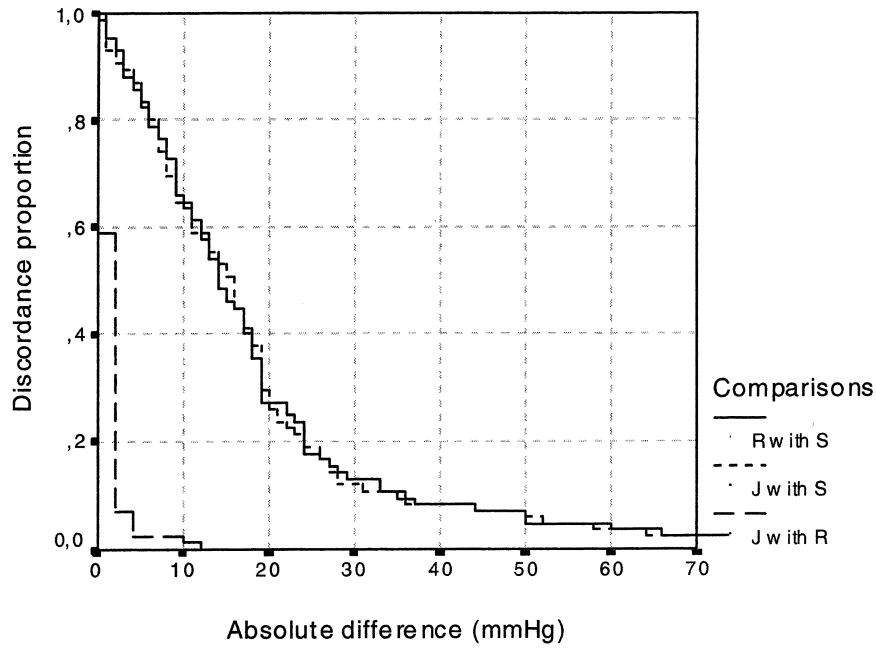

Fig. 2. Proportion of discordance between systolic blood pressure by two observrs (J and R) and a semiautomatic device (S) until "tolerance" limits.

differences as a function of the magnitude of the measurement or an increase of the differences indicating a greater error due to the measurement are not considered in the proposed approach, because it does not take into account the magnitude of the measurement. Nevertheless, if the magnitude of the bias does not have "clinical" relevance, it might be that the absence of this information may not impair the analysis.

On the other hand, if considering the magnitude of the measurement is of importance, the proposed graphic could take it into account through the calculation of relative differences. For instance, in the X-axis of the graphic, instead of showing the modules of the differences, we would present the differences relative to the magnitude of the measurement, as follows:

$$|A-B| / [(A+B)/2]$$

Therefore, the graphic would be completely adimensional, what could be an advantage, for the sake of comparison. Although our proposal is descriptive in nature, it allows the use of inference resources, by means of tests associated to the Kaplan-Meier analysis. It is possible, for example, to use the log-rank test to evaluate whether the difference between two curves of agreement, for a certain categorical covariate is statistically significant. The sample size must be always taken into account, in particular, when the sample is subdivided.

5. Conclusion

We believe that the proposed graphical approach can serve as a complement—or in special situations as an alternative—to the Altman-Bland method for agreement analysis or to mountain plot also. It allows a simple interpretation of agreement that takes into account the "clinical" importance of the differences between observers or methods. In addition, it allows the analysis of reliability or agreement, by means of survival analysis techniques.

This graphical approach, furthermore, could be used in instances in which internal correlation of the analysis units is expected and in which there is interest in analyzing the difference between these units. For quantitative measurements, in which symmetric results are expected—measurements concerning, for example, pairs of eyes, ears, cerebral hemispheres, and kidneys—this technique can be very useful. Therefore, in matched studies of a natural origin, or in studies designed to improve efficiency, or the validity of the analysis, the researchers could use this new approach. Finally, this approach will have a practical advantage of leading the researcher to consider the magnitude of the differences observed, and, consequently, think about the practical importance of these differences.

Acknowledgments

We thank Daibes Rachid Filho. who provided the data set, and the reviewer, who made important comments and suggestions.

References

[1] Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg: An Aspen Publication; 2000.
[2] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60.
[3] Müller B, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465–76.
[4] De Vet H. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopaedia of biostatistics. Chichester: John Wiley & Sons; 1999. p. 3123–27.
[5] Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. Clin Chem 1973;19:49–57.
[6] Solberg HE, Gräsbeck R. Reference values. Adv Clin Chem 1989;27:1–79.
[7] Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Stat Methods Med Res 1999;8:93–112.
[8] King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Stat Med 2001; 20:2131–47.
[9] Byrt T, Bishop J, Carli JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.
[10] Shoukri MM. Measurement of agreement. In: Armitage P, Colton T, editors. Encyclopaedia of biostatistics. Chichester: John Wiley & Sons; 1999. p. 103–17.
[11] Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. J R Stat Soc D 1983;32:307–17.
[12] Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337–40.
[13] Streiner DL, Norman GR. Health measurements scales: a practical guide to their development and use. Oxford: Oxford University Press; 1995.
[14] Bartko JJ. General methodology II—measures of agreement: a single procedure. Stat Med 1994;13:737–45.
[15] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
[16] Latis GO, Simionato L, Ferraris G. Clinical assessment of gestational age in the newborn infant: comparison of two methods. Early Hum Dev 1981;5:29–37.
[17] O'Brien E, Petrie J, Littler W, de Swiet M, Padfield PL, Altman DG, Bland M, Coats A, Atkins N. The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. J Hypertens 1993;11(suppl 2):S43–S62.
[18] Monti KL. Folded empirical distribution function curves—mountain plots. Am Stat 1995;49:342–5.
[19] Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995;33:525–7.
[20] Borgan O. Kaplan-Meier estimator. In: Armitage P, Colton T, editors. Encyclopaedia of biostatistics. Chichester: John Wiley & Sons; 1999. p. 2154–60.
[21] Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491–4.