Journal of Clinical Monitoring and Computing (2009) 23:311–314
DOI: 10.1007/s10877-009-9198-4

USING A PREDICTION APPROACH TO ASSESS AGREEMENT BETWEEN TWO CONTINUOUS MEASUREMENTS

Cody Hamilton, PhD[1] *and James D. Stamey, PhD*[2]

Hamilton C, Stamey JD. Using a prediction approach to assess agreement between two continuous measurements.
J Clin Monit Comput 2009; 23:311–314

ABSTRACT. The problem of assessing agreement between two devices occurs with great frequency in the medical literature. If it can be demonstrated that a new device agrees sufficiently with a device currently in use, then the new device can be approved for general use. This work discusses how a prediction interval can be used to estimate the whether a future difference between two devices will be within acceptable limits with reasonable confidence. The method is illustrated with an example involving measurements of peak expiratory flow.

KEY WORDS. measure of agreement, prediction interval.

INTRODUCTION

The problem of assessing statistical agreement occurs in many fields of science, but is especially prevalent in the area of clinical research. If one can demonstrate that a new device provides measurements that closely agree with those from a device already in general clinical use, then one can substitute the new device for the old one. The deployment of this new device may present a variety of advantages: it may be less invasive, less expensive, more transportable, or easier to use than the current device. Recent clinical examples involving the assessment of statistical agreement between continuous measurements can be found in Neder and Stein [1] and McGaughran et al. [2].

There is a rich body of research with regards to this problem. Typically, the methods assume that the differences recorded between the two devices are independent and normally distributed with common mean and variance (see Choudhary and Nagaraja [3] and Barnhart et al. [4]. for a good overview). Several variations of the intraclass correlation coefficient have been presented: see for example Fleiss [5] and St. Laurent [6]. This metric measures the proportion of the total variation (sum of variability between devices, between patients, and within patients) explained by the device-to-device variability. The concordance correlation coefficient of Lin [7] provides a coefficient that can be divided into two pieces, one describing the linear relationship between the two devices and the other investigating how close that linear relationship is to the 45 degree line denoting perfect agreement. Each of these correlation coefficients presents a statistic that can range between zero and one: the closer the coefficient is to one, the stronger the evidence supporting agreement. Muller and Buttner [8] point out that

From the [1]Department of Clinical Operations, Edwards Lifesciences, One Edwards Way, Irvine, CA 92614, USA; [2]Department of Statistical Science, Baylor University, Waco, TX 76798-7140, USA.

Received 8 June 2009. Accepted for publication 10 August 2009.

Address correspondence to C. Hamilton, Department of Clinical Operations, Edwards Lifesciences, One Edwards Way, Irvine CA 92614, USA.
E-mail: Cody_Hamilton@edwards.com

correlation coefficients are all sensitive with regards to patient-to-patient variability: a set of differences between the devices recorded on patients with wide variation in terms of the underlying measurement being assessed will demonstrate a larger coefficient than an equivalent set of differences recorded on patients with less variation. In addition, it is difficult to determine a clinical decision rule for interpreting the coefficient—how close does the coefficient have to be to one in order to conclude agreement? The total deviation index of Lin [9] estimates the limits that contain a given proportion of the distribution of differences with a specified level of confidence. As these limits are presented on the original measurement scale (rather than on a zero-to-one scale), they can be readily assessed by a clinical audience with regards to clinical acceptability. Lin et al. [10]. points out that the total deviation index also provides greater statistical power than the concordance correlation coefficient. The method does assume, however, that the standardized bias (the bias divided by the standard deviation of the differences) between the two measurements is small (Lin [9]). Bland and Altman [11] estimate the 2.5th and 97.5th percentiles of the underlying distribution from which the differences between the two devices are ostensibly drawn—these percentile estimates are known as the 'limits of agreement.' Like the TDI, these limits are presented on the original measurement scale. Bland and Altman also derive asymptotically appropriate confidence intervals for these estimates. Choudhary and Nagaraja [3] point out that these intervals provide a large sample tolerance interval for the differences. Exact confidence intervals are presented by Liu and Chow [12]. Hamilton and Stamey [13] have pointed out that the 'raw' limits of agreement (i.e. without the accompanying confidence intervals) provide a reference interval only—they do not provide statistical bounds on the differences between the two measurement devices (except for large sample sizes), nor can they be used directly for statistical inference.

This work discusses an alternative approach to the problem of assessing statistical agreement between two continuous measurements. This approach attempts to place limits on the deviation between the devices that may be expected for future patients; these limits are known as a prediction interval. We briefly derive the prediction interval for the case of independent, normally distributed differences with a common variance and then demonstrate how this interval provides better statistical properties than the Bland–Altman limits of agreement for smaller sample sizes.

METHODS

Let us suppose that two devices are used to record a measurement on each one of a set of n patients. Let d_i be the difference between the two measurements for the ith patient. Like previous authors (e.g. Bland and Altman [11]), we assume that the d_i are distributed normally with a common mean μ and variance σ^2. If we let $\tilde{d}$ denote a difference recorded between the two devices on some future patient then the expected value for $\tilde{d}$ is μ, which we may estimate with the mean of the observed differences $\bar{d}$. We seek to construct an interval that will contain $\tilde{d}$ with $100(1-\alpha)\%$ confidence, where α is the type I error. The prediction error for this future difference $\tilde{d}$ is $\mathrm{Var}\left(\tilde{d}-\bar{d}\right)=\sigma^2(1+1/n)$. By the Central Limit Theorem, $\tilde{d}-\bar{d}$ divided by the square root of its variance $\sigma^2(1+1/\mathrm{n})$ is normally distributed. In practice, however, the true variance σ^2 is unknown. We may approximate σ^2 with the sample estimator s^2 so that $\mathrm{Var}\left(\tilde{d}-\bar{d}\right)\approx s^2(1+1/n)$. It follows then that $\left(\tilde{d}-\bar{d}\right)/s\sqrt{1+1/n}$ follows a t distribution with $n-1$ degrees of freedom and that:

$$\Pr\left(-t_{n-1,1-a/2}\le\left(\tilde{d}-\bar{d}\right)\Big/s\sqrt{1+1/n}\right)\le t_{n-1,1-a/2})=1-\alpha$$

$$\Rightarrow\Pr\left(\bar{d}-t_{n-1,1-a/2}s\sqrt{(1+1/n)}\le\tilde{d}\le\bar{d}+t_{n-1,1-a/2}s\sqrt{(1+1/n)}\right)=1-\alpha$$

Thus a $100(1-\alpha)\%$ prediction interval for the future difference is:

$$\bar{d}\pm t_{n-1,1-a/2}s\sqrt{(1+1/n)}$$

This interval will contain the difference between the measurements recorded on a future patient with $100(1-\alpha)\%$ confidence. If this set is contained in a pre-specified, clinically acceptable range it can be concluded that a future difference would be within the acceptable clinical limits with at least $100(1-\alpha)\%$ confidence. Alternatively, one can simply conclude that a future difference between the two measurement devices will be within $\bar{d}\pm t_{n-1,1-\alpha/2}s\sqrt{(1+1/n)}$ with $100(1-\alpha)\%$ confidence.

The limits of agreement provided by Bland and Altman are

$$\bar{d} \pm z_{1-\alpha/2} s$$

where $\bar{d}$ and s are as defined above. These limits provide estimates of the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles[1] $\mu \pm z_{1-\alpha/2}\sigma$. While the limits for a reference interval useful for plotting against the raw differences, they do not guarantee any level of statistical coverage (see Hamilton and Stamey [13] for more discussion). It should be noted that the prediction interval is wider than the raw limits of agreement presented by Bland–Altman unless the sample size is large enough for $1/n$ to converge to zero and for s to converge to σ (i.e. for the true population variance to be considered known) in which case the intervals will be equivalent. Therefore, the limits of agreement cannot be used for statistical inference unless the sample size is large. Bland and Altman [15] have derived the following confidence interval for the limits of agreement

$$\bar{d} \pm z_{1-\alpha/2} s_d \pm t_{n-1,1-\gamma/2} s_d \sqrt{\left(1/n + z^2_{1-\alpha/2} \Big/ (2(n-1))\right)}$$

which for large sample sizes provides a tolerance interval containing $100(1-\alpha)\%$ of the population of differences with $100(1-\gamma)\%$ confidence ($1-\alpha$ determines the proportion of the differences that one wished to contain while $1-\gamma$ determines the level of confidence with which one wishes this proportion to be contained). While this interval will provide a level of coverage which the limits of agreement themselves do not, it may be too conservative for some applications. In that case, the prediction interval provides some guarantee of coverage (it will contain a future difference with $100(1-\alpha)\%$ confidence) which the raw limits of agreement do not.

RESULTS

As an example we consider data analyzed by Bland and Altman [11] involving 17 forced expiratory volume measurements recorded via a Wright peak flow meter and a mini Wright peak flow meter. These data are available from Martin Bland's webpage http://www-users.york.ac.uk/~mb55/datasets/pefr.dct. For the purposes of demonstrating the prediction interval, we consider only the first pair of measurement recorded one each patient. The average difference is −2.12, and the standard deviation of the differences is 38.77. For these data, the Bland and Altman's limits of agreement are (−78.1, 73.9). The 95% approximate tolerance interval computed via the formula provided in Bland and Altman [15] is (−112.2, 107.9). The 95% prediction interval is (−86.7, 82.4). Thus the prediction interval provides a nice balance between the narrow limits of agreement and the more conservative tolerance approach.

[1]The interested reader will note that s is not an unbiased estimator of σ, and hence the limits of agreement are not unbiased estimates of the population percentiles.

CONCLUSION

The prediction interval provides several benefits for small to moderate sample sizes. First, its interpretation is straightforward—the computed bounds indicate whether future observations taken by the new measurement will be within an acceptable range of the current standard with reasonable confidence. Second, these limits remain on the original measurement scale rather than on a 0 to 1 scale. Third, the method provides a balance between the raw Bland–Altman limits of agreement and the more conservative tolerance bounds.

As with any parametric test, the necessary assumptions behind the presented method should be checked. For example, like Bland and Altman [11] and Lin [7, 9], we assume that the variance of the differences does not vary over the range of underlying values measured. In addition, while the presented approach may detect disagreement between two devices, it will not indicate the nature of that disagreement. Bland and Altman [11] provide several excellent graphical checks to investigate such deviances.

REFERENCES

1. Neder JA, Stein R. A simplified strategy for the estimation of the exercise ventilatory thresholds. Med Sci Sports Exerc 2006; 38: 1007–1013.
2. McGaughran L, Voss LJ, Oliver R, Petcu M, Schaare P, Barnard JPM, Sleigh JW. Rapid measurement of blood propofol levels: a proof of concept study. J Clin Monit Comput 2006; 20: 109–115.
3. Choudhary PK, Nagaraja HN. Measuring agreement in method comparison studies—a review. In: Balakrishnan N, Kannan N, Najaraja HN, editors. Advances in ranking and section, multiple comparisons, and reliability. Boston: Birkhauser, pp. 215–44.
4. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat 2007; 17: 529–569.
5. Fleiss JL. The design and analysis of clinical experiments. New York, NY: Wiley.
6. St. Laurent RT. Evaluating agreement with a gold standard in method comparison studies. Biometrics 1998; 54: 537–545.
7. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255–268.
8. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994; 13: 2465–2476.

9. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med 2000; 30: 255–270.
10. Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. J Am Stat Assoc 2002; 97: 257–270.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307–310.
12. Liu J, Chow SC. A two-one-sided tests procedure for assessment of individual bioequivalence. J Biopharm Stat 1997; 7: 49–61.
13. Hamilton C, Stamey J. Using Bland–Altman to assess agreement between two medical devices—don't forget the confidence intervals. J Clin Monit Comput 2007; 21: 331–333.
14. Bland JM, Altman DG. Agreement between methods of measurement with multiple observation per individual. J Biopharm Stat 2007; 17: 571–582.
15. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8: 135–160.