the appropriate construction and use of the Bland–Altman plot are the following. The $x$ axis should be constructed by the mean of the methods and the $y$ axis in a way that is most sensible to the concentration range of the $x$ data (absolute: small range; percentage: medium range; log-scale: large range). The 95% limits of agreement should reflect the actually observed nature of the differences (whether or not there is a relationship between difference and magnitude) (9). Most important, interpretation of the data should be done by comparison of the observed limits of agreement with a priori ones.

As a final note, we want to remark that in this journal, already in 1981, a similar plot (with the $y$ axis constructed as a ratio) was proposed for the evaluation of method-comparison data (12). Strange to say, this report has been overlooked.

This work was supported by the Research Fund of the University Ghent (Grant BOF 011109000).

References


Katy Dewitte
Colette Fierens
Dietmar Stöckl
Linda M. Thienpont*

Laboratorium voor Analytische Chemie
Universiteit Gent
Harelbekestraat 72
9000 Gent, België

*Author for correspondence. Fax 32-9-264-81-98; e-mail Linda.thienpont@rug.ac.be.

In the following commentary, Drs. Altman and Bland elaborate on the issues raised in the above letter:

Commentary on Quantifying Agreement between Two Methods of Measurement

Our interest in the analysis of method comparison studies stemmed from discussions about consulting problems we were independently working on in the late 1970s. Examination of published papers showed that, at that time, most authors were using the Pearson correlation coefficient. It was obvious to us that this method did not assess agreement, but association, and that a high correlation was no guarantee of good agreement.

We felt that a comparison of two methods of measurement, such as different assays, should attempt to quantify the differences and that $P$ values were largely irrelevant. The question is not whether the two methods agree, but how closely they agree. Our statistical approach was based on investigation of the distribution of the between-method differences. We suggested summarizing the data by the mean and 95% range of the differences, which we called the 95% limits of agreement. The graph, which many think is the whole of our method, was intended as a visual check that the approach was reasonable and that the data were “well-behaved”. Thus the graph shows whether the variability of differences between methods is roughly constant across the range of measurement, but the key element of the approach is to examine and summarize the individual differences between the two methods. Indeed, in our original paper we included histograms of these differences. This distribution should be approximately normal, and (apart from occasional outliers) this is usually what we see.

Our first two papers outlined the basic ideas (1, 2), but a recent report contained the fullest exposition of our method, including various extensions to deal with replicated observations and complex relationships between the between-method difference and the magnitude of the measurement (3).

Our original work related to clinical rather than laboratory measurements, but it was soon obvious that, broadly speaking, the same issues arose. Concerns about the use of the Pearson correlation coefficient had been expressed in this Journal as long ago as 1973 (4), but the method remained widespread for decades (and it still is in the wider medical literature).

The idea of plotting difference vs mean was not new (5), but as far as we knew its use had not been proposed in this context. The same type of plot was suggested as a general purpose approach for method comparison studies at around the same time (6), although without any suggestion for quantifying the differences between the methods.

A particular issue that we were aware of from the start is that there are some measurements where the between-method (and within-method) variability increases as the measurement increases. We found that the SD of the differences tended to be proportional to the size of the measurement, so that log transformation of the original data led to
differences in the log scale that were unrelated to the size of the measurement. This situation arises commonly in clinical chemistry. Our original suggestion was to take logarithms of the original data, the natural approach to a mathematician, but working with the ratio of the two methods (3) or the percentage difference (7) gives almost the same answers and is more transparent to the laboratory scientist.

Another important issue is that the full comparison of the performance of two methods of measurement ought to include repeated measurements. Such repeat data can be used to compare observers or instruments, or simply to assess random error.

The very wide uptake of the limits of agreement approach has naturally been very pleasing. We have been aware, however, that sometimes the method has not been adopted with full understanding. For example, we have seen it suggested that two methods agree well because most of the observations lie within the 95% limits of agreement. The limits are calculated so that this will always be the case. We welcome the review by Dewitte et al. of method comparison studies in this Journal. They found that some authors are not making the best use of the method. The most common problem was that investigators plotted the differences against the values obtained by one method. We note, however, that this error will matter much less in situations where there is relatively little measurement error vs between-individual variation, as is the case for many analyses presented in this Journal.

Other aspects of methodology that could benefit from more thought include sample size and how the participants are selected.

Lastly we want to comment on interpretation. We agree that the acceptability of a new method should be based on clinical rather than statistical criteria (2) and that the criteria should ideally be prespecified. Relating the criterion of acceptability to goals for analytical quality may be a helpful suggestion (8).

References

Douglas G. Altman* J. Martin Bland1

1 Cancer Research UK Medical Statistics Group Centre for Statistics in Medicine Institute of Health Sciences Old Road, Headington Oxford OX3 7LF, United Kingdom

2 Department of Public Health Sciences St. George’s Hospital Medical School London SW17 0RE, United Kingdom

*Author for correspondence. E-mail doug.altman@cancer.org.uk.