Assessing agreement between methods of measurement

Douglas G. Altman[1*] and J. Martin Bland[2]

[1] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom.

[2] Department of Health Sciences, University of York, York, United Kingdom

* Address correspondence to this author at: Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom. E-mail doug.altman:csm.ox.ac.uk.

[3] This article has been cited more than 27000 times since publication.

Featured Article: Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-10.[3]

We began our careers as medical statisticians in the 1970s, working in a department of epidemiology. The quality of measurements was well-recognized in epidemiology, as illustrated by the development of the "random-zero" sphygmomanometer to enable unbiased measurement of blood pressure to be made. But it was after we had moved on to separate clinical research environments that we independently faced many scenarios relating to how to measure biological quantities, initially in the fields of rheumatology and cardiology. We discovered that the standard way that measurement studies were then analyzed was by correlation coefficients, whether authors were comparing measurements obtained by different observers using the same method, or measurements obtained using different methods.

It was obvious to us that correlation did not provide a useful answer to the agreement between two methods of measurement. For example, we could double the size of all the observations made using one of the methods without altering the correlation coefficient. The question was not whether the two methods agreed, but how closely they agreed, for which the differences between the pairs of measurements would be much more informative. Accordingly, we developed some very simple ideas based around the mean and standard deviation (SD) of the differences, from which we could estimate a range within which we would expect a high percentage (usually 95%) of observed differences to lie – we called these "limits of agreement". These limits depend on assumptions that the mean and SD of the differences are unrelated to the magnitude of the measurement and that they follow an approximately normal distribution. We thought that a plot of the data would be really helpful to check these assumptions. Specifically, we suggested a histogram of the differences between the paired measurements and a plot of the differences versus their mean. We saw this plot, now often called a "Bland-Altman plot", as giving support for the numerical summary, not as the method itself *(1).*

We first presented our thoughts at a conference of statisticians in Cambridge in 1981, and our paper based on the talk appeared in 1983 *(2).* This paper was inaccessible to our clinical colleagues so we thought it would be useful also to publish the ideas in a medical journal. And so, in January 1986, our paper was published in *The Lancet*. It is astonishing that it has become one of the most cited statistical papers ever published.

Our *Lancet* paper included a data set we collected especially for the study, consisting of replicate measurements of peak expiratory flow using two different instruments –the sample was colleagues, family and ourselves (we were the two people with the best lung function). It was obvious to us that it would be good to include a table of the raw data – sadly this practice remains rare despite the current focus on transparency.

We certainly were not the first people to recognize that correlation coefficients did not assess agreement, but rather association. Notably, we quoted Westgard and Hunt, who wrote wisely in *Clinical Chemistry* in 1973, "The correlation coefficient ... is of no practical use in the statistical analysis of comparison data" *(3).* Their preferred approach used prediction intervals based on least squares regression. That the topic is of critical importance in clinical chemistry is illustrated by the fact that this was one of five references in our 1983 paper to articles in chemistry journals.

At the time we thought our simpler approach was obvious and so our ideas were unlikely to be original, but we couldn't find a similar proposal. We since found that Donald Mainland had made similar comments about correlation 40 years earlier *(4)* but still have not found the same suggestion about how to compare methods of measurement.

Since the first paper we have extended the approach in various ways, in particular to allow analysis of replicated measurements *(5,6).*

Method comparison studies are clearly a topic of permanent relevance. Our 1986 paper is the most cited paper in *The Lancet* and is still being cited several times a day, but our other related papers are the most cited papers in several other journals *(2,5,6).* Measurement is fundamental to clinical practice and research. After more than 40 years as medical statisticians we remain disappointed that too few people realize that.

[703 words]

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

References

1. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995;24(Suppl. 1):S7-S14.

2. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307-17.

3. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. Clin Chem 1973;19:49-57.

4. Mainland D. An experimental statistician looks at anthropometry. Ann NY Acad Sci 1955;63:474-83.

5. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Meth Med Res 1999;8:135-60.

6. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharmaceut Stat 2007;17:571-82.