David S. Warner, M.D., Editor

Agreed Statistics

Measurement Method Comparison

J. Martin Bland, M.Sc., Ph.D.,* Douglas G. Altman, D.Sc.†

Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. By J. Martin Bland, Douglas G. Altman. Lancet 1986; 1(8476):307–10. Abstract reprinted with permission of Elsevier, copyright 1986.

ABSTRACT:

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analyzed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described, together with the relation between this analysis and the assessment of repeatability.

WE first met in 1972, when J. Martin Bland (J.M.B.), M.Sc., Ph.D., joined the Department of Clinical Epidemiology and Social Medicine, St. Thomas's Hospital Medical School, University of London. Douglas G. Altman (D.G.A.), D.Sc., had been working there, in his first post, since late 1970. J.M.B. came from 3 yr in the agrochemical industry. We had a lot in common and soon became friends, but we did not work directly together and did not publish together until after we both left St. Thomas's in 1976, D.G.A. for the Medical Research Council at Northwick Park and J.M.B. for St. George's Hospital Medical School, London. Our first joint publication was a letter in the *Lancet* in 1977, and we have now published approximately 90 articles and letters, including our long-running series Statistics Notes in the *British Medical Journal*.‡ Our most frequently cited publication is our 1986 *Lancet* article "Statistical methods for assessing agreement between two methods of clinical measurement,"[1] which by August 2011, 25 yr after publication, has been cited more than 18,000 times. Nearly 1,000 of these citations are in the anesthesiology literature.

Identifying a Problem

The work described in this article began around 1978, when we each independently came across the problem of agreement be-

* Professor, Department of Health Sciences, University of York, Heslington, York, United Kingdom. † Professor of Statistics in Medicine and Director, Centre for Statistics in Medicine, University of Oxford, Oxford, United Kingdom.

Received from the Department of Health Sciences, University of York, Heslington, York, United Kingdom. Submitted for publication September 14, 2011. Accepted for publication October 5, 2011. Dr. Bland's travel was supported by a Senior Investigator Award from the National Institute for Health Research, London, United Kingdom. All other support was provided from institutional and/or departmental sources. Copyright on the title page figure is held by J. Martin Bland and Douglas G. Altman.

Address correspondence to Dr. Bland: Department of Health Sciences, University of York, Heslington, York YO10 5DD, United Kingdom. martin.bland@york.ac.uk.

‡ martinbland.co.uk/pubs/pbstnote.htm. Accessed September 15, 2011.

 Anesthesiology 2012; 116:182–5

tween different methods of clinical measurement. A cardiologist colleague brought J.M.B. a paper and said "There's something wrong with this, but I don't know what it is." It was a paper comparing two methods of measuring cardiac stroke volume.[2] A group of patients had been measured by the standard dye dilution method and by an electrical impedance method. There was a significant correlation between these measurements. The authors had also made several pairs of measurements on each of 20 patients, and of course 20 is a magic number to a statistician. They found that only 1 of the 20 sets of measurements on a single person gave a statistically significant correlation, and they concluded from this that the two methods did not agree. It occurred to J.M.B. that if an individual's stroke volume was constant, we would be correlating only the measurement errors of the two methods. We would thus expect the correlation to be 0, so we would expect 1 of 20 tests to be significant, exactly what they found. So the result is what would be expected whatever the agreement was like, and their conclusion didn't follow from the design and analysis.

D.G.A. had come across a similar problem in a study of between-observer variation in leg and knee circumference measurements. The publication about that study included a brief footnote about the issue: "It is incorrect to use the correlation coefficient to compare sets of measurements of the same variable. In such circumstances the correlation largely reflects the variability of the subjects being measured. For example, for our least reliable measurement at 15 cm above the patella the correlation between the measurements taken by two observers was 0.99. It is the differences between the measurements that should be investigated."[3] Shortly afterward D.G.A. discussed method comparison studies in an article in the *British Medical Journal* and emphasized the importance of looking at between-method differences, rather than correlation.[4]

Limits of Agreement

We were intrigued that we had both stumbled across this question, and we agreed that, in addition to the problem of being dependent on the range of true values being measured, correlation measures relationship, not agreement. If one measurement is always twice as big as the other, they are highly correlated, but they do not agree.

We decided to write an article about measurement studies, and D.G.A. found two other methods of analyzing agreement, testing the null hypothesis that the regression slope is equal to 1 and testing the difference between means, which were also deeply flawed. When we were preparing our article, we thought that if we were to say that everybody is doing things in the wrong way and then stopped, it would fall a little flat. We should say what we thought was the right analysis. We agreed that we should start with the difference between measurements by the two methods, one minus the other. Having obtained a set of numbers, as any statistician would, we found the mean and SD. Then 95% of differences would be between the mean minus 1.96 standard deviations and the mean plus 1.96 standard deviations. We called these the 95% limits of agreement and suggested this analysis as a possible approach. (We have not been entirely consistent about this and have sometimes used 2 standard deviations as an approximation to 1.96, all the fault of J.M.B.).

Fig. 1. Drs. Bland and Altman in 1981 on the occasion of the first public presentation of their new method. Copyright J. Martin Bland and Douglas G. Altman.

We presented our ideas at a statistical conference, which was a first for both of us; we had spoken only at medical meetings before this (fig. 1). We did not claim any great originality for the limits of agreement idea but said that it was the obvious statistical approach. Indeed, we were sure that someone was going to stand up and say "of course Fisher did this in 1932" (we have seen it happen), but nobody did and nobody ever has. To us it was a very simple idea that any statistician would suggest. That it hadn't happened may reflect the fact that few statisticians had been actively involved in analyzing that type of data. Recently, D.G.A. discovered that a broadly similar approach (but without the idea of limits) had been described in 1955 by the great pioneer of statistics in medicine, Donald Mainland (1902–1985).[5] He also criticized correlation in this context: "Even when the coefficient is + 0.95 or higher, it does not tell us whether, for the purpose in hand, the differences between the duplicate readings are trivial or serious."[6] Notably, Mainland was an anatomist, and his research had involved a lot of measurement. However, he did not succeed in popularizing the approach.

Like most statistical analyses, the limits of agreement method requires some assumptions. The mean and the SD of the differences are assumed to be the same for everybody. They should be the same for subjects with a large value of the quantity being measured and for subjects with a small value, for example. We suggest checking this by plotting difference against the average of the two methods, using average as the best estimate of the magnitude that we have. We suggested adding the mean and limits of agreement as horizontal lines in the difference *versus*

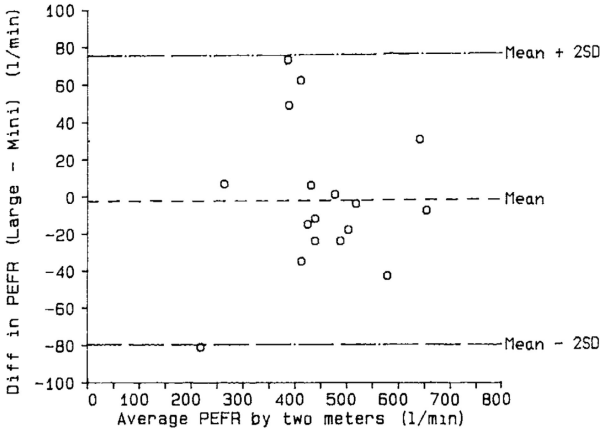

Fig. 2. Difference *versus* mean plot, with mean difference and 95% limits of agreement. (Reprinted, with permission, from Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1(8476):307–10. Copyright 1986 Elsevier.)

mean plot, which should then include approximately 95% of the observations (fig. 2).

If the variability of differences is not constant, the most frequent pattern is that the SD of differences increases with the magnitude of the quantity being measured, often with the SD being proportional to the mean. We can deal with this by a log transformation of the observations. We can then antilog the limits of agreement to give 95% limits for the ratio of one method to the other, rather than for the difference.

Another assumption of the limits of agreement method is that the differences should have an approximately normal distribution. This is necessary for the 1.96 multiplier, but it doesn't have to be met very closely, and it is unlikely to be a problem if the first assumption is met. We can check it by a histogram or a normal quantile plot of the differences.

The limits of agreement enable us to estimate from an observed measurement by one method what the value of a measurement on the same person at the same time by the other method might be, as a range of possible values. If these limits are sufficiently narrow for us to draw the same conclusions about the quantity being measured, or the person being measured, we can conclude that the methods agree sufficiently well for the two methods to be used interchangeably. Interchangeability is an important property. For some measurements, methods are not interchangeable. For example, if we measure a patient's depression score as being 21, this can be interpreted only if we say "on the PHQ9 scale." On the Beck Depression Inventory or Hospital Anxiety and Depression Scale, it would be quite different.

Comparing methods of measurement should be a matter of estimating how closely two methods agree, not whether they agree or not. Most biologic measurements are made with error, and exact agreement will not happen, even when using the same method twice. We can decide how much disagreement might be acceptable before the study and that might well vary for different purposes.

Publications

We sent our article to *The Statistician*, which was the journal of the Institute of Statisticians (which since has merged with the Royal Statistical Society), at whose conference we had spoken, and it was published there in 1983.[7] We waited for things to change, but measurement researchers just carried on correlating. We were urged by colleagues to produce a version for a medical audience with a worked example, so we did. J.M.B. collected a set of lung function data from a convenience sample of colleagues, friends, and family (and ourselves!) for the purpose of illustration, and the article appeared in the *Lancet* in 1986. It was a great success. It is the most frequently cited article ever to appear in the *Lancet* by quite a long way and is 1 of the 10 most frequently cited statistical articles ever.[8] The previous article in *The Statistician* has now been cited more than 1,000 times and is the most frequently cited article to appear in that journal. The *Lancet* phoned J.M.B. to tell him about the acceptance just before Christmas 1985, a wonderful present.

The difference *versus* average plot was intended only as a check on assumptions, but this has produced more arguments than anything else we have ever written; people keep saying we should plot the standard method on the horizontal axis of the graph, not the average of the two measurements. However, it is well known that if we plot the differences against one of the measurements, we will get a relationship. For any variables X and Y, $X–Y$ will be positively related to X and negatively related to Y; it is just in the mathematics. (Here we knew we were following in the footsteps of others, such as Peter Oldham.[9]) In 1995, we felt compelled to write the article "Comparing methods of measurement, why plotting difference against standard method is misleading."[10] This also appeared in the *Lancet* and has had more than 800 citations.

The limits of agreement are sample estimates, and so are subject-to-sampling variation; they will vary from sample to sample. We can estimate confidence intervals for them, just as we would for a sample mean or a sample proportion. When the *Lancet* accepted the 1986 paper, David Sharp, the deputy editor, said that it should be shortened. Now this editorial position is familiar to most paper writers, but most unusually, he offered to do this for us. They improved the paper considerably in the process, but they cut out the paragraph describing the confidence interval. After J.M.B. told David Sharp how much better he thought the paper was after the editing, he asked for this one paragraph to be reinstated. He was asked "Is it really important?" J.M.B. said "Yes, it is," and David Sharp kindly agreed. We think it is a great pity that limits of agreement are often quoted without their confidence intervals. They are not hard to calculate.

The success of the *Lancet* article led to an increase in citations of the Statistician article, too, and these papers were declared a joint Citation Classic by the Institute of Scientific Information in 1992.[11] As the limits of agreement method became known and used, researchers began to approach us for help when things were a little more complicated than the simple examples in our 1986 article. We collected several of our solutions to these prob-

lems in an article published in *Statistical Methods in Medical Research* in 1999.[12] If we design our agreement study to have pairs of measurements by each method, we need to be able to estimate the limits of agreement between single observations from these data.

In the *Lancet* article, we described a method to do this. In the later article we extended the method to any number of repetitions by either method and we included unequal numbers of repetitions as well, where some people have more pairs of measurements than others. We also considered the case where the underlying quantity being measured is changing (such as blood pressure), so that we have several pairs of measurements on each person. We included a new approach to dealing with relationships between differences between methods and the magnitude of measurements, using regression. We also described a nonparametric approach for use when there were outliers. This article now has been cited more than 1,000 times and is the most frequently cited article in *Statistical Methods in Medical Research*. It was declared an ISI Current Classic in 2008.§ We also wrote a paper illustrating problems with analyses of measurement studies using examples from the radiology literature.[13] In 2007, we published an article about agreement between methods of measurement with multiple observations per individual,[14] which expanded suggestions in our 1999 article.[12] Both of these articles have become the most frequently cited articles in those journals.

Summing Up

We continue to be astonished by the impact of this work. Why have the 1986 *Lancet* article and the related articles been cited so frequently? These were important articles because researchers were carrying out inappropriate analyses and drawing potentially wrong conclusions from them. The limits of agreement method allows them to use a simple, intuitive analysis that can be done without special software and is easy to interpret. In the past, measurement studies were largely ignored by statisticians. When J.M.B. was first brought some measurement error data to analyze, he could not find the term "measurement error" in the index of any book on his shelves and had to proceed from first principles. Many of the previous generation of textbook writers in medical statistics spent their careers in universities or working for research councils and away from hospitals, where the measurers congregate. Donald Mainland is the great exception, but on the whole these writers did not say much about measurement studies. We were in close contact with many clinicians, and they often asked us about these studies. We were able to bring a detached statistical view to the design and analysis of measurement studies so that they could better meet the needs of these researchers. Other statisticians have tried different approaches, but if the design is simple and the data reasonably well behaved, our approach is so easy to apply and so intuitive that is has been the one adopted by researchers. They can carry it out without trying to find statistical support.

When we wrote our 1986 Lancet article, methodologic research was not the main professional role of either of us: D.G.A.'s was to carry out collaborative clinical research and J.M.B.'s to do collaborative epidemiologic research and teach medical students and doctors. We have now written approximately 90 joint papers, articles, and letters because we enjoy working together. We share a fascination with medical research and a desire to make it better and so to improve the medicine that depends upon it. We work well together. If one of us feels strongly about some point, the other knows when to give way. But most of all, we make one another laugh.

We are still working on measurement studies because we get many requests for help from researchers around the world. J.M.B. keeps a frequently asked questions list on martinbland.co.uk. We hope to find time to publish some of these (for example, on sample size estimation for measurement method agreement studies). We are continuing to develop these ideas and will do so as long as new questions arise.

§ http://sciencewatch.com/inter/pod/2008_2/. Accessed September 15, 2011.

References

1. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1(8476):307–10
2. Keim HJ, Wallace JM, Thurston H, Case DB, Drayer JI, Laragh JH: Impedance cardiography for determination of stroke index. J Appl Physiol 1976; 41:797–9
3. Kirwan JR, Byron MA, Winfield J, Altman DG, Gumpel JM: Circumferential measurements in the assessment of synovitis of the knee. Rheumatol Rehabil 1979; 18:78–84
4. Altman DG: Statistics and ethics in medical research. V: Analysing data. BMJ 1980; 281:1473–5
5. Mainland D: An experimental statistician looks at anthropometry. Ann NY Acad Sci 1955; 63:474–83
6. Mainland D: Elementary Medical Statistics, 2nd edition. Philadelphia, Saunders, 1963, pp 334
7. Altman DG, Bland JM: Measurement in medicine: The analysis of method comparison studies. Statistician 1983; 32: 307–17
8. Ryan TP, Woodall WH: The most-cited statistical papers. J Appl Stat 2005; 32:461–74
9. Oldham PD: Measurement in Medicine. The Interpretation of Numerical Data. London, English Universities Press, 1968
10. Bland JM, Altman DG: Comparing methods of measurement: Why plotting difference against standard method is misleading. Lancet 1995; 346:1085–7
11. Bland JM, Altman DG: This week's citation classic: Comparing methods of clinical measurement. Current Contents 1992; CM20(40):8
12. Bland JM, Altman DG: Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8:135–60
13. Bland JM, Altman DG: Applying the right statistics: Analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22:85–93
14. Bland JM, Altman DG: Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007; 17:571–82