International Biometrics Conference: July 2014

Limits of Agreement: Birth of a Simple Idea

J Martin Bland

Professor of Health Statistics
University of York

Douglas G Altman

Centre for Statistics in Medicine
University of Oxford

http://martinbland.co.uk

The paper

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

Reported in top 10 most highly cited statistical papers up to 2003, with 8,151 citations, sixth most highly cited.

Now: 23,630 citations on Web of Science (30th June 2014).

Also most highly cited in *Lancet*, next has 8,592 citations.

Ryan TP, Woodall WH. The most-cited statistical papers. *Journal of Applied Statistics* 2005; **32**: 461-74.

The paper

What won?

Kaplan EL & Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**, 457-481.

25,869 citations (now 38,529)

Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 187-220.

18,193 citations (now 28,358)

Where it began

DGA and JMB first met in 1972.

JMB joined the Department of Clinical Epidemiology and Social Medicine at St Thomas's Hospital Medical School, University of London, after three years in the agricultural chemical industry.

DGA had been working there, in his first post, since late 1970.

Did not publish together until after we both left St. Thomas's in 1976, DGA for the Medical Research Council at Northwick Park and JMB for St. George's Hospital Medical School, London.

Where it began

First joint publication was a letter in the *Lancet*:

Bland JM, Altman DG. Enteric disease in San-Francisco. *Lancet* 1977; **ii**: 306-306.

(0 citations)

More than 90 articles and letters, including our long-running series *Statistics Notes* in the *British Medical Journal.*

A simple problem

Around 1978, a cardiologist colleague brought JMB 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:

Keim HJ, Wallace JM, Thurston H, Case DB, Drayer JIM, Laragh JH. Impedance cardiography for the determination of stroke index. *Journal of Applied Physiology* 1976; **41**: 797-9.

A simple problem

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.

They found that only one of the 20 sets of measurements on a single person gave a statistically significant correlation.

Concluded from this that the two methods did not agree.

A simple problem

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 zero and so we would expect one out of 20 tests to be significant, exactly what they found.

So the result is what would be expected ***whatever the agreement was like.***

Their conclusion did not follow from the design and analysis.

A simple problem

DGA 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 ... It is the differences between the measurements that should be investigated.”

Kirwan JR, Byron MA, Winfield J, Altman DG, Gumpel JM. Circumferential measurements in the assessment of synovitis of the knee. *Rheumatology and Rehabilitation* 1979; **19**: 78-84.

A simple problem

We were intrigued that we had both stumbled across this question.

Agreed:

- correlation depends 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

A simple problem

Decided to write an article about measurement studies.

DGA found two other methods of analyzing agreement:

- testing the null hypothesis that the regression slope is equal to one.
- testing the difference between means.

Also deeply flawed.

So what was the right analysis?

A simple solution

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 standard deviation.

Then 95% of differences would be between the mean minus 1.96 standard deviations and the mean plus 1.96 standard deviations, assuming a Normal distribution and constant mean and standard deviation.

We called these the 95% limits of agreement and suggested this analysis as a possible approach.

(Sometimes used 2 standard deviations as an approximation to 1.96, all the fault of JMB.)

A simple solution

We presented at a statistical conference, the Institute of Statisticians, a first for both of us.

Altman and Bland at the Institute of Statisticians Conference, Cambridge, 1981.

A simple solution

Did not claim any originality for the limits of agreement idea. It is the obvious statistical approach.

Sure that someone was going to stand up and say “of course Fisher did this in 1932”.

Nobody did and nobody ever has.

To us it was a very simple idea that any statistician would suggest.

Perhaps few statisticians had been actively involved in analysing that type of data?

A simple solution

Broadly similar approach (but without the idea of limits) had been described in 1955 by the great pioneer of statistics in medicine, Donald Mainland.

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.”

Mainland D: An experimental statistician looks at anthropometry. *Annals of the New York Academy of Sciences* 1955; **63**: 474-83.

Checking the assumptions

Limits of agreement method requires some assumptions.

The mean and the standard deviation of the differences are assumed to be the same for everybody.

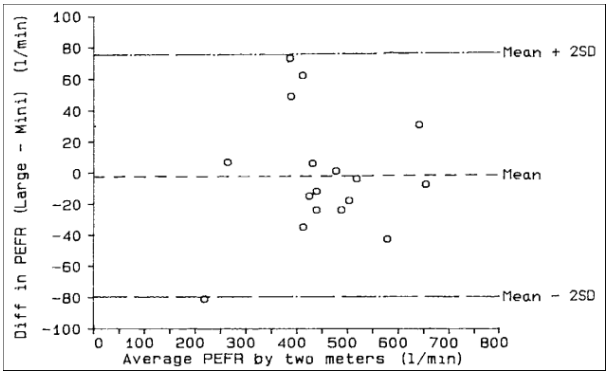
They should be independent of the quantity being measured, for example.

Can check by plotting difference against the average of the two methods, using average as the best estimate of the magnitude that we have.

A standard statistical procedure.

Checking the assumptions

We suggested adding the mean and limits of agreement as horizontal lines in the difference vs. mean plot, which should then include about 95% of the observations.



Called the “Bland Altman plot” — but not by us!

Checking the assumptions

Another assumption: the differences should have an approximately Normal distribution.

Necessary for the 1.96 multiplier, but doesn’t have to be met very closely.

Unlikely to be a problem if the first assumption is met.

Check by a histogram or a Normal quantile plot of the differences.

Publications

We sent our paper to *The Statistician*, which was the journal of the Institute of Statisticians.

Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307-17. (1,510 citations)

Waited for things to change, but measurement researchers just carried on correlating.

Urged by colleagues to produce a version for a medical audience with a worked example, so we did.

The paper appeared in the *Lancet* in 1986.

Publications

Others followed:

Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085-7. (1,138 citations)

Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine* 1990; **20**: 337-340. (276 citations)

Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; **8**: 135-160. (2,518 citations)

Publications

Others followed:

Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology* 2003; **22**: 85-93. (471 citations)

Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* 2007; **17**: 571 – 582. (329 citations)

Publications

Others have not only used this very simple idea, but have developed it in many ways.

In Web of Science:

Bland AND Altman in Topic: 19,626 publications.

Bland AND Altman in Title: 79 publications.

Publications

Reprinting:

Bland JM, Altman DG. Statistical methods for assessing agreement between measurements. *Biochimica Clinica* 1987; **11**: 399-404. (Reprinted from the *Lancet*.)

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies* 2010; **47**: 931-936. (Reprinted from the *Lancet*.)
(96 citations!)

Publications

All about us:

Bland JM, Altman DG. This week's citation classic: Comparing methods of clinical measurement. *Current Contents*, 1992; CM20(40) Oct 5, 8.

Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. *International Journal of Epidemiology* 1995; **24** (Suppl), S7-S14.

Bland JM, Altman DG. Agreed Statistics: Measurement Method Comparison. *Anesthesiology* 2012; **116**: 182–185.

The keys:

- be in the right place at the right time,
- have a prepared mind.

" … high impact is … achieved … by the presence of a unique skill set that is used by researchers to identify strategic niches and gaps which ultimately results in higher impact."

Zelko H, Zammar GR, Ferreira APB, Phadtare A, Shah J, Pietrobon R. Selection mechanisms underlying high impact biomedical research - a qualitative analysis and causal model. *PLoS ONE* 2010; **5**: e10535.
(1 citation)