Flattering innovation: a popular method to test new devices is often misunderstood and misused

How do we test a new method to measure something? Is it as good as a tried and tested method? It may be quicker and easier to use, but is the result equivalent to the old, clunky method?

More than 30 years ago, two statisticians, Bland and Altman, described a better way to compare a new method with an old method (Fig. 1 panels A and B). This method became widely accepted. They asked the question “are the results from the two methods sufficiently similar, so that we can we can start to use new one?” Their example was a lung function test from number of people, measured by two different instruments. They showed convincingly that even if a simple graph of the results might suggest that the tests agreed, further analysis showed differences between the methods, and that the results couldn’t be relied on to be equivalent. They used the difference between each pair of measurements (old – new) and related those differences to the mean of the same pair of results ([old + new]/2). This graph shows the “limits of agreement” between the two methods (Fig. 1).

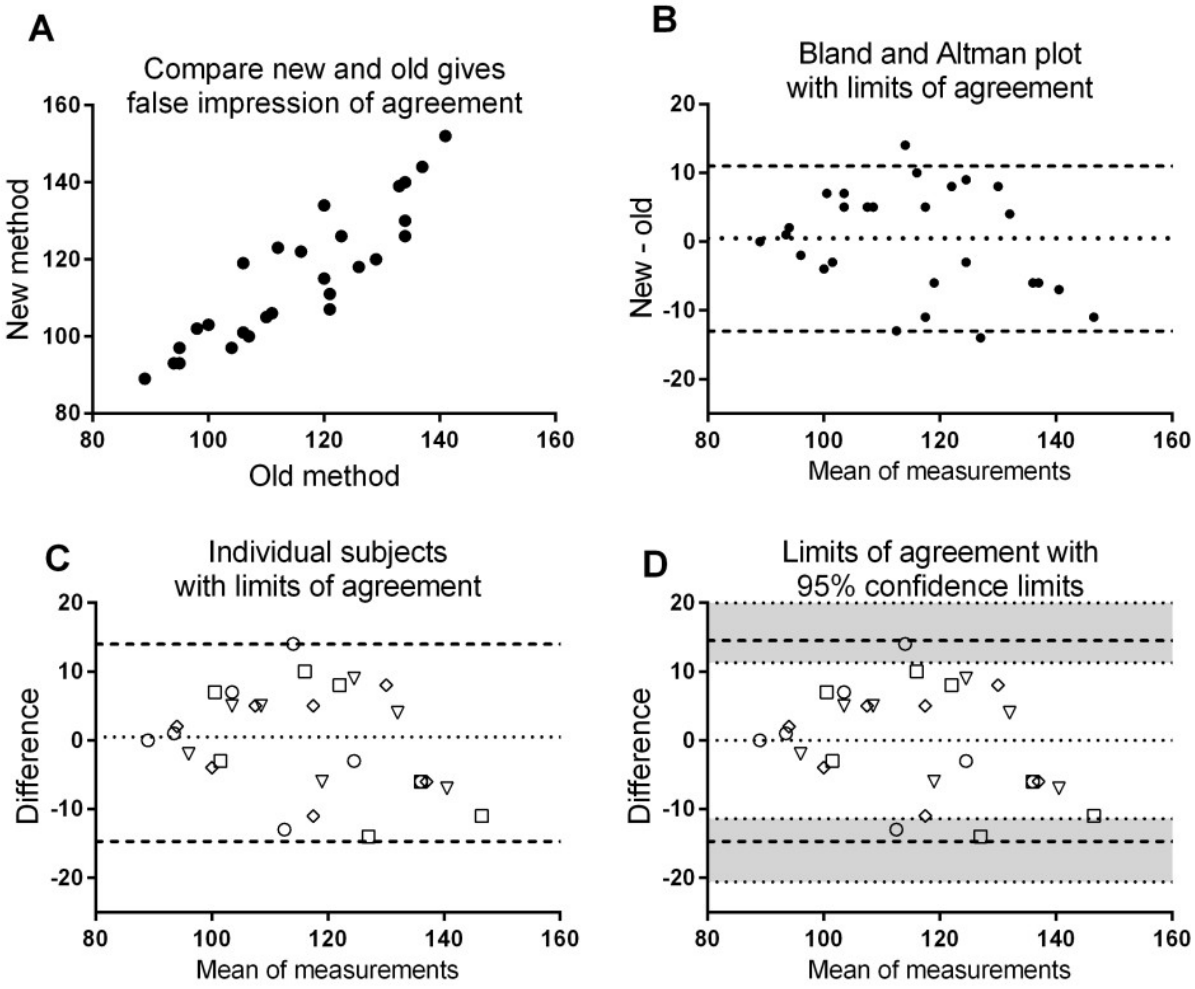

Fig. 1. Comparing blood pressure with two methods. Each panel is different displays from the same example data. A: An old method compared the actual measurements with the two methods: the relationship appears to be reasonable. B: Using the difference method, and assuming the paired measurements are from different people, there is a likely range of values of about ± 10 which might be acceptable for some uses. C: However, the data in panel B actually come from 4 individuals (different symbols) Each person was measured 7 times. The calculated limits of agreement are now wider and could be unacceptable. D These limits of agreement are not exactly known. When the 95% confidence bands are included, there is even more uncertainty. A measurement device that differed by 20 units would probably not be clinically useful!

However the original simple analysis assumes that each set of measures (old and new) come from a different person. Very often in later studies an instrument was used to make repeated measurements in the same person. Measurements such as blood pressure may not be the same

each time they are taken. If repeated measurements are taken, from several people, a new statistical complexity is introduced. The results vary because of variation between subjects, and because the values may also vary within a single subject. Using the original analysis suggested by Bland and Altman gives the wrong answer, exaggerating the agreement. (Fig. 1, panel C) Thus a new instrument could be wrongly judged to be equivalent to the old one.

A further problem is that even if the limits of agreement are calculated correctly, they are also less precisely known. The agreement limits come with an “uncertainty” factor because random sampling only gives an estimate of the truth. We should apply a “confidence interval” around the limits we have found. (Fig. 1, panel D) Calculating these intervals was complex and they were rarely reported by scientific papers, even though they could affect the conclusions.

As a result, new measurement devices should be more critically evaluated than they are at present.

Gordon Drummond
Anaesthesia Critical Care and Pain Medicine, University of Edinburgh

Publication

Limits of agreement may have large confidence intervals.
Drummond GB.
Br J Anaesth. 2016 Mar