University of York Department of Health Sciences

Measurement in Health and Disease

Assessing Agreement Between Methods of Clinical Measurement

Martin Bland

http://martinbland.co.uk/

This talk is based on:

Bland JM, Altman DG. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307-310.

http://www-users.york.ac.uk/~mb55/meas/ba.htm

Often wish to measure variables where direct measurement without adverse effects is difficult or impossible.

E.g. cardiac stroke volume, blood pressure.

True values remain unknown.

Indirect methods are used.

A new method has to be evaluated by comparison with an established technique rather than with the true quantity.

If the new method agrees sufficiently well with the old, the old may be replaced.

Neither method provides an unequivocally correct measurement, so we try to assess the degree of agreement.

PEFR MEASURED WITH WRIGHT PEAK FLOW AND MINI WRIGHT PEAK FLOW METER

Subject	Wright peak flow meter (l/min)	Mini Wright peak flow meter (l/min)
1	494	512
2	395	430
3	516	520
4	434	428
5	476	500
6	557	600
7	413	364
8	442	380
9	650	658
10	433	445
11	417	432
12	656	626
13	267	260
14	478	477
15	178	259
16	423	350
17	427	451

Non-representative sample of colleagues, family, and friends.

We want to know by how much the new method is likely to differ from the old.

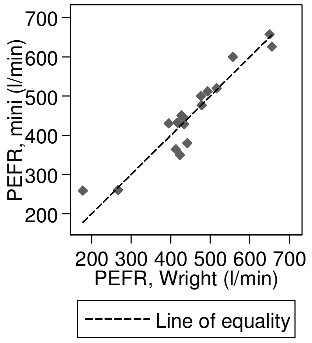
If this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably.

How far apart measurements can be without causing difficulties will be a question of judgment.

Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size.

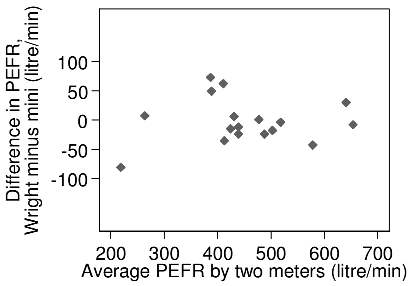
Helps to plot the data.

Scatter plot with line of equality:



Note equality of scales.

Plot of the difference between the methods against their mean:



No obvious relation between the difference and the mean.

Note equality of scales.

No obvious relation between the difference and the mean.

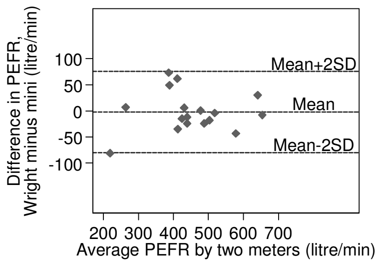
Under these circumstances we can summarise the lack of agreement by calculating the bias, estimated by the mean difference $\bar{d}$, and the standard deviation of the differences, s.

If there is a consistent bias we can adjust for it by subtracting $\bar{d}$ from the new method.

For the PEFR data the mean difference (large meter minus small meter) is –2.1 l/min and s is 38.8 l/min.

We would expect most of the differences to lie between $\bar{d} - 2s$ and $\bar{d} + 2s$.

We would expect most of the differences to lie between $\bar{d} - 2s$ and $\bar{d} + 2s$.



“Bland Altman plot”.

Limits of agreement: “Method of Bland and Altman”.

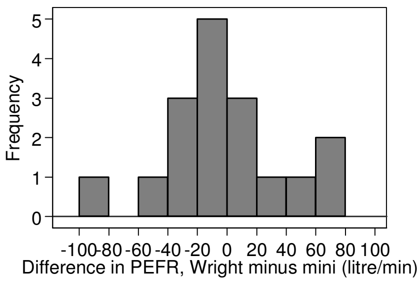
If the differences are Normally distributed (Gaussian), 95% of differences will lie between these limits.

More precisely, between $\bar{d} - 1.96s$ and $\bar{d} + 1.96s$.

Such differences are likely to follow a Normal distribution because we have removed a lot of the variation between subjects and are left with the measurement error.

The measurements themselves do not have to follow a Normal distribution, and often they will not.

We can check the distribution of the differences by drawing a histogram.



If this is skewed or has very long tails the assumption of Normality may not be valid.

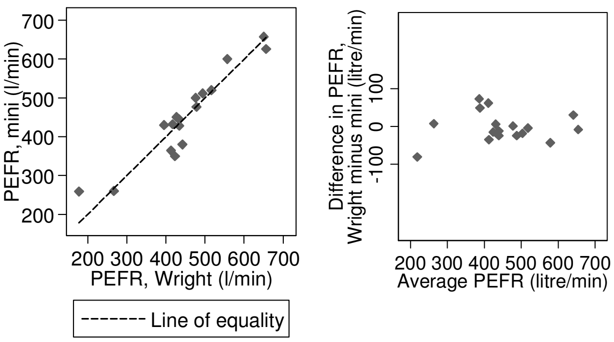
Provided differences within the l would not be clinically important, we could use the two measurement methods interchangeably. We shall refer to these as the “95% limits of agreement". For the PEFR data we get:

$\bar{d} - 2s$ = −2.1 − (2x38.8) = −79.7 l/min

$\bar{d} + 2s$ = −2.1 + (2x38.8) = 75.5 l/min

Thus, the mini meter may be 80 l/min below or 76 l/min above the large meter, which would be unacceptable for clinical purposes.

This lack of agreement is by no means obvious in the scatter diagram.



95% confidence intervals can be found for the 95% for the limits of agreements.

For the PEFR meters:

lower limit of agreement, 95% CI is −114.3 to −45.1 l/min.

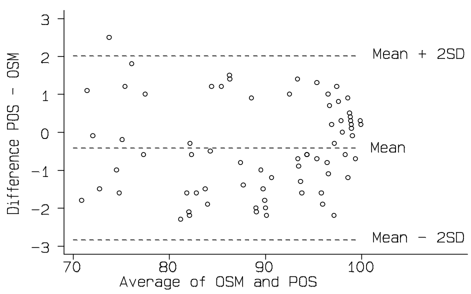
upper limit of agreement 95% CI is 40.9 to 110.1 l/min.

These intervals are wide, reflecting the small sample size and the great variation of the differences.

They show, however, that even on the most optimistic interpretation there can be considerable discrepancies between the two meters and that the degree of agreement is not acceptable.

The first real application:

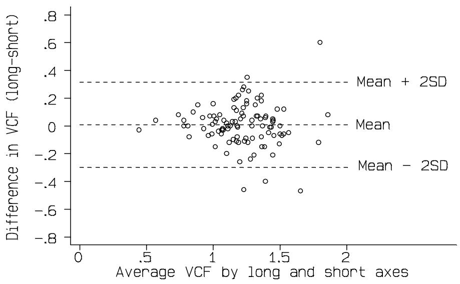
oxygen saturation monitor and pulsed saturation oximeter.



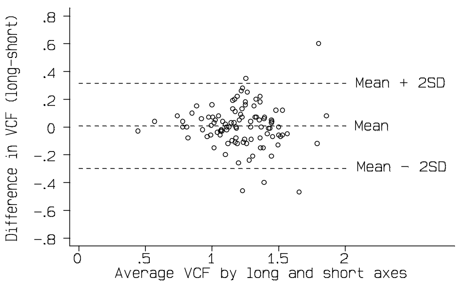
Note that I got the scales wrong and use 2*s*, not 1.96*s* !

Tytler JA, Seeley HF. The Nellcor N-101 pulse oximeter - a clinical-evaluation in anesthesia and intensive-care. *Anaesthesia* 1986; **41**: 302-305.

Relation between difference and mean

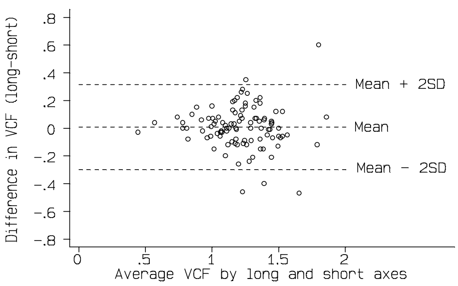


D'Arbela PG, Silayan ZM, Bland JM. Comparability of M-mode echocardiographic long axis and short axis left ventricular function derivatives. *British Heart Journal* 1986; **56**: 445-9.



The scatter of the differences increases as the VCF increases.

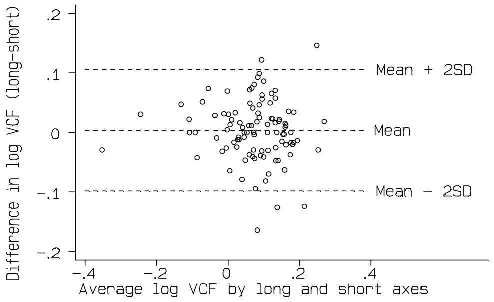
We could ignore this, but the limits of agreement would be wider apart than necessary for small VCF and narrower than they should be for large VCF.



If the differences are proportional to the mean, a logarithmic transformation should remove the relationship.

We can then apply the limits of agreement analysis described to the transformed data.

After logarithmic transformation :



Still shows some relation between the difference and the mean VCF, but there is some improvement.

The mean difference is 0.003 on the log scale, s = 0.051, and the limits of agreement are –0.098 and 0.106.

The limits of agreement have somehow to be related to the original scale of measurement.

If we take the antilogs of these limits, we get 0.80 and 1.27.

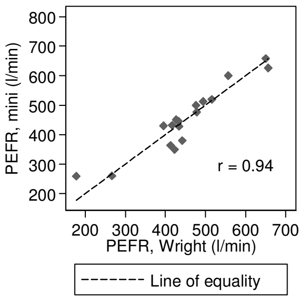
However, the antilog of the difference between two values on a log scale is a dimensionless ratio.

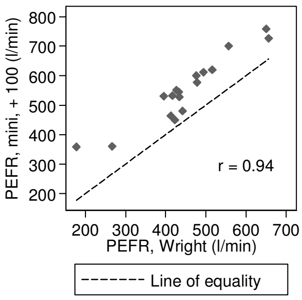
The limits tell us that for about 95% of cases the short axis measurement of VCF will be between 0.80 and 1.27 times the long axis VCF.

The log transformation is the only transformation giving back-transformed differences which are easy to interpret, and we do not recommend the use of any other in this context.

Correlation coefficients do not measure agreement

They ignore bias. If we add 100 to one of the measurements, the correlation is unchanged.





The relationship to repeatability

Repeatability is relevant to the study of method comparison because the repeatabilities of the two methods of measurement limit the amount of agreement which is possible.

If one method has poor repeatability the agreement between the two methods is bound to be poor too.

When the old method is the more variable one, even a new method which is perfect will not agree with it.

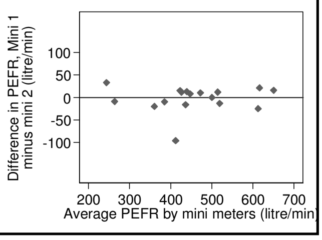
The best way to examine repeatability is to take repeated measurements on a series of subjects.

The relationship to repeatability

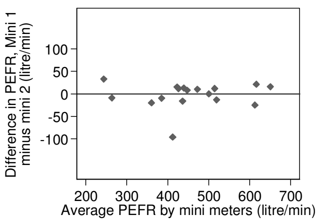
Mini Wright peak flow meter

Subject	First PEFR (l/min)	Second PEFR (l/min)
1	512	525
2	430	415
3	520	508
4	428	444
5	500	500
6	600	625
7	364	460
8	380	390
9	658	642
10	445	432
11	432	420
12	626	605
13	260	227
14	477	467
15	259	268
16	350	370
17	451	443

Plot differences against mean for each subject:



The relationship to repeatability



No relation between the difference and the size of the PEFR.

A clear outlier.

Retained this measurement for the analysis, although I suspect that it was technically unsatisfactory.

The relationship to repeatability

Calculate the mean and standard deviation of the differences.

Mean difference should here be zero since the same method was used.

(If the mean difference is significantly different from zero, we will not be able to use the data to assess repeatability because either knowledge of the first measurement is affecting the second or the process of measurement is altering the quantity.)

For the PEFR by the mini meter, the standard deviation of differences between the 17 pairs of repeated measurements is 28.2 l/min.

The relationship to repeatability

Expect 95% of differences to be less than two standard deviations.

This is the standard deviation of the difference between two measurements on the same person.

It is equal to √2 times the within subjects standard deviation, s_w.

Two standard deviations of differences = $2\sqrt{2}\, s_w$.

This is the repeatability coefficient.

For the mini meter, the coefficient of repeatability
= 2× 28.2 = 56.4 l/min.

For the large meter the coefficient is 43.2 l/min.

The relationship to repeatability

For the mini meter, the coefficient of repeatability = 56.4 l/min.

For the large meter the coefficient is 43.2 l/min.

Compare these repeatability coefficients to the limits of agreement, –80 l/min to +76 l/min.

We estimate that the mini meter will be within 56 l/min of another measurement by itself, but only within 80 l/min of a measurement by the Wright peak flow meter.

We can conclude that not all the variation between the two instruments is because of their measurement error, but there is some other source of variation.