Post-referee version, 22 December 2006

Agreement between methods of measurement with multiple observations per individual

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

[1]Professor of Health Statistics

Dept. of Health Sciences

University of York

York, UK

[2]Professor of Statistics in Medicine

Centre for Statistics in Medicine

Wolfson College Annexe

Linton Road

Oxford, UK

[*] Correspondence to J M Bland

Professor of Health Statistics

Dept. of Health Sciences

University of York

York YO10 5DD

email: mb55@york.ac.uk

Abstract

Limits of agreement provide a straightforward and intuitive approach to agreement between different methods for measuring the same quantity. When pairs of observations using the two methods are independent, i.e. on different subjects, the calculations are very simple and straightforward. Some authors collect repeated data, either as repeated pairs of measurements on the same subject, whose true value of the measured quantity may be changing, or more than one measurement by one or both methods of an unchanging underlying quantity. In this paper we describe methods for analysing such clustered observations, both when the underlying quantity is assumed to be changing and when it is not.

Introduction

The limits of agreement (LoA) method (Altman and Bland 1983, Bland and Altman 1986) for assessing the agreement between two methods of medical measurement is widely used. (Bland and Altman 1993, Ryan and Woodall 2005). We obtain the differences between measurements by the two methods for each individual and calculate the mean and standard deviation. We then estimate the 95% limits of agreement as the two values mean minus 1.96 standard deviations and mean plus 1.96 standard deviations. These limits are expected to contain the difference between measurements by the two methods for 95% of pairs of future measurements on similar individuals.

The motivating scenario for the LoA method is the case where each individual has one measurement made by each of the methods X and Y. It is valuable, however, to obtain replicate measurements by each method on each individual so that the repeatability of the two methods can be compared (Bland and Altman 1999). Such data comprise a mixture of between and within-individual information on the differences between methods. We did not state in early publications that the LoA method assumes independent observations [Altman

and Bland (1983) Bland and Altman (1986)], as this important requirement is not specific to the LoA approach but rather applies to all types of statistical analyses. If each pair of X and Y measurements is treated as if from a different individual the structure of the data is ignored and incorrect estimates are likely; specifically, the interval between the limits of agreement may be too narrow.

In this paper we look at how to apply the LoA method when we have repeated measurements on each of a group of subjects. We consider separately two somewhat different situations.

Concepts

The key principle of the LoA method is to examine the average difference between the methods, and also to consider the variability in those differences across individuals. It is an implicit assumption that the difference between the two methods is reasonably stable across the range of measurements, and we will assume this condition holds for the purpose of this paper. We have discussed elsewhere possible strategies when this condition is not met, including transformation of the data (Bland and Altman 1999).

Table 1 and Figure 1 show some typical data in which pairs of measurements were made sequentially on each of a group of subjects. Here 60 pairs of measurements of cardiac ejection fraction by two methods were made on 12 individuals, with 3-7 replicates per individual. First, we might ignore the replication and treat these as 60 independent pairs of measurements and calculate the mean and standard deviation of their differences. As noted above, these limits of agreement could be too narrow. An alternative would be to average all the observations on the same subject. The limits of agreement calculated in this way would be for the average of several measurements and would be too narrow for a single measurement. This approach is appropriate only when the usual clinical measurement is the average of that number of observations.

As an extreme example, Barry *et al.* (1997) reported the comparison of bioimpedance and continuous thermodilution two methods of cardiac output using 2390 observations from just 7 patients.

A somewhat different problem is shown in a study by Almén *et al.* (1991) who reported the glomerular filtration rate (GFR) in the left and right kidneys of 20 patients using both a gamma camera and computed tomography (CT). They presented the GFR of each kidney as a percentage of the total GFR for that patient. Unfortunately, they use data from both kidneys in their comparison of the two methods, but they have effectively analysed all the data twice for each patient, as the difference between methods with the left kidney is minus that for the right kidney. Their plot displays point symmetry as a consequence of plotting each point as both (X,Y) and (100-X, 100-Y). Had they calculated limits of agreement they would have found that the mean difference was exactly zero.

There are two different situations to consider for replicated data. We can think of the observations for the same subject as a series of measurements of a quantity that does not vary over the period of observation. An example is measurements of carotid artery stenosis taken on the same day. Or we can think of them as pairs of measurements by two methods of a changing quantity, where it is the instantaneous measurement for the subject which we want to capture. This second situation could arise either when the quantity being measured is unstable, such as blood pressure or daily excretion of some chemical, or when observations are made under different conditions – e.g. before and after exercise. The distinction is important, as it determines whether we need to consider pairing of observations by the two methods. Indeed, for the first (constant) case we do not require equal replication of each method for each individual, whereas this is a requirement for the second (non-constant) case.

In Bland and Altman (1986) we described how to deal with the constant case, where the true value of the quantity is not changing, but only for the simple case when the number of

observations for each subject is the same. In Bland and Altman (1999) we discussed the more general case where the number of observations varies, and also the non-constant situation where we are trying to capture the instantaneous value of a changing quantity. That paper is quite technical and not easily accessible to many researchers, so in this paper we describe these methods more simply and provide a worked example.

In both cases, the key to the analysis of such data is that the repeated observations by a method on an individual will be scattered around the mean value of all the possible observations by that method, which we might consider to be that person's true value. There may be both a bias, where one method tends to give consistently higher measurements than the other, and heterogeneity, when the between method-differences vary across individuals more than expected simply by chance. This phenomenon, which we also call a subject by method interaction, is seen clearly in Figure 2.

In each case, we shall estimate limits for the difference between measurements by the two different methods on the same subject. We shall begin with the non-constant case where the true value varies, because it is rather simpler.

For both approaches we want the agreement to be the same or at least similar over the range of measurement. We can check this assumption by plotting the difference against the average of the two methods (Figure 2). We have included a zero line in Figure 2. It is clear that there is a bias, the RV cardiac ejection fraction tending to be larger than the IC, but no obvious variation in agreement across the range of measurements.

Method where the true value varies

Calculations for the "non-constant" situation are relatively straightforward, using the difference between methods for each pair. We want to estimate the mean difference and the standard deviation of differences about the mean. To do this we must estimate two different variances: that for repeated differences between the two methods on the same subject and that

for the differences between the averages of the two methods across subjects. The model is that the observed difference is the sum of the mean difference (bias), a random between subjects effect (heterogeneity) and a random error within the subject. The within-subject variance is assumed constant and observations within the subject are independent. The variance for single differences between pairs of measurements on different subject is found by summing the between subjects and within subjects variances (Bland and Altman 1999).

The first variance, that within subjects, can be estimated very easily using one way analysis of variance, using the difference in matched pairs as a response. We must assume that this within-subject variance is the same for all subjects. We check that it is unrelated to the subject mean as the best estimate of the magnitude of the measurement for that subject. Figure 3 shows the standard deviation of the differences for the subject against the average measurement for that subject. There is no suggestion that there is a relationship between the variability of the differences and the magnitude of the ejection fraction.

The one-way analysis of variance is shown in Table 2. The estimated variance of multiple between-method differences for the same subject is the residual mean square or mean square error, 0.170714026. (We will retain all the decimal places until the end of the calculation, and then round to more practical numbers.) The other component of the variance, for differences between the average difference across subjects, can also be found from this table, using the difference between the mean squares for subjects and the residual mean square, 4.2090856 – 0.170714026 = 4.0383716. We must divide this by a value which depends on the numbers of observation on each subject. If the number of observations on subject *i* is m_i, this divisor is

$$\frac{\left(\sum m_i\right)^2 - \sum m_i^2}{(n-1)\sum m_i}$$

where *n* is the number of subjects. If all the subjects have the same number of observations, *m*, this factor reduces to *m*. For the ejection fraction data, n = 12, $\sum m_i$ = 60 (the total number of observations), and $\sum m_i^2$ =312. Hence

$$\frac{\left(\sum m_i\right)^2 - \sum m_i^2}{(n-1)\sum m_i} = \frac{60^2 - 312}{(12-1)\times 60} = 4.9818182$$

The estimated component of variance which represents the heterogeneity is 4.0383716 divided by 4.9818182 = 0.81062203. (In our earlier paper (Bland and Altman 1999), we incorrectly stated that this number should be used to multiply the difference in the sum of squares, rather than divide it.) The total variance for single differences on different subjects is estimated by the sum of these two components: 0.170714026 + 0.81062203 = 0.98133606. The standard deviation is the square root of this, which is 0.99062408.

The estimated bias, the mean difference, can be estimated simply from the mean of the individual differences. This method automatically weights the observations correctly. The average is 0.6021667. Hence the 95% limits of agreement are estimated to be 0.6021667 – 1.96 × 0.99062408 to 0.6021667 + 1.96 × 0.99062408. This gives –1.3394565 to +2.5437899 which we can round to –1.3 to +2.5. These are the 95% limits for RV minus IC, so we estimate the ejection fraction measured by the RV to be between 1.3 units less than IC and 2.5 units greater. We can add these limits to the difference against average plot, as shown in Figure 4. The limits appear to fit the data well.

For this analysis of variance, we must assume that the repeated differences for a single subject are independent. This might be a rather strong assumption. For the ejection fraction data, for example, subjects were in the operating theatre undergoing surgery and there may be changes over time in the ejection fraction. Hence there would be autocorrelation in the ejection fraction, which might produce autocorrelation in the differences. One visual check on the assumption of independence would be to plot observed differences against order.

Figure 5 shows such a plot, assuming that the data were supplied to us in temporal order. There appears to be autocorrelation for some subjects and, indeed, the order by subject interaction is highly significant. Whether this influences the estimate of the variance within subjects is unclear, but this is certainly an area where further work is needed.

Would it matter if we ignored the subject and treated the 60 observations as if they were from 60 different subjects? Not much in this case. The mean difference would be unchanged and the standard deviation would be 0.9610571 compared to 0.99062408. The limits of agreement would be –1.2815052 to +2.4858386, so to one decimal place they would also be –1.3 to 2.5. This similarity is because the number of pairs per subject is quite small and less than the number of subjects. However, the limits are slightly narrower than they should be. As the number of pairs per subject rises, the limits will become narrower.

Method where the true value is constant

In the "constant" case where the true value does not change any pairing of measurements made by the two methods simultaneously will not be informative. We do not keep the link between them and may have different numbers of measurements on a subject by the two methods. Indeed, there may in fact not be any pairing in the first place.

For each method separately, the variability will be made up of three components: the variability across individuals of the true quantity being measured, the variability of each individual's average values about overall average for that method, which we call heterogeneity, and the variability of repeated measurements about the average for an individual. We assume that these are independent, that observations within a subject are independent, and that the error variances within the subject are constant.

As described in section 5 in Bland and Altman (1999), we derive the estimated standard deviation for individual differences from the variance of the subject mean differences and an extra term derived from the separate measurement error of each method.

We will use the same data to illustrate the method. We expect to get wider limits of agreement than before because the true value may not be constant in this situation.

We first find the two measurement errors using one-way analyses of variance for each method separately (Table 3). The within subjects variances are obtained from the mean square for the residual, 0.107227795 for the RV method and 0.137874069 for the IC method. We next find the mean RV and IC for each subject and the differences between them. The mean of these average differences across the 12 subjects, RV minus IC, is 0.7092361, and their variance is 0.91269114, which corresponds to a standard deviation of 0.9553487. For these analyses to be valid, the within-subject standard deviation of each measurement must be constant and unrelated to the magnitude. We can check this by plotting the individual subject standard deviation against the individual mean, for each method separately (Figure 6). There is no problem for IC, but there may be for RV. However, with a small sample it is hard to judge whether there may be a problem and for this illustration we decided to accept the standard deviation as constant.

We then increase this variance by a term that allows for it being derived from the average of several observations. We multiply each of the separate within-subject variances found earlier by

$$\left(1-\frac{1}{n}\sum\frac{1}{m_i}\right)$$

The m_i may be different for the two methods, as it is possible to have different numbers of observations by the two methods on a subject. If the numbers of observations on each subjects are the same, m, this expression reduces to

$$\left(1-\frac{1}{m}\right)$$

For the ejection fraction data, n = 12 and $\sum 1/m_i$ = 2.5166664, so the multiplier is

$$1-\frac{1}{12}\times 2.5166664 = 0.7902778$$

Hence we estimate the variance for individual differences by the variance of differences between subject means plus the multiplier times the sum of the measurement error variances for each of the methods:

$$0.91269114 + 0.7902778\times 0.107227795 + 0.7902778\times 0.137874069 = 1.1063897$$

The standard deviation is the square root of this value, i.e. 1.0518506.

The limits of agreement can now be found. As before, the weighted mean difference is 0.6021667, and the 95% limits of agreement are 0.6021667 – 1.96 × 1.0518506 to 0.6021667 + 1.96 × 1.0518506, which gives –1.4594605 to 2.6637939. We could round these values to –1.5 to +2.7.

These limits are slightly wider than those where we retained the pairing information, because they are for agreement in measuring the average ejection fraction over a period rather than at an instant. The variability of ejection fraction over the measurement period has been included in the random variation. The small difference between the two sets of limits is because in these data the ejection fraction varied only slightly over the measurement period. These were the limits which were originally provided for Bowling *et al*. (1993).

As with the case of paired data, there may be correlation within the subject. This would be for the separate methods rather than the differences, as the method assumes no pairing. We plot each variable separately against order (Figure 7). Although we might observe some subjects where there are apparent trends, in fact the order by subject interactions are not significant for either method of measurement. However, as for the paired case, this topic would be worth further investigation.

Discussion

Most method comparison studies seem to use single observations by each method for each individual. There are, however, considerable advantages in collecting replicate observations so that the repeatability of the methods can be compared. The limits of agreement method is most easily applied to the simple, unreplicated case. In this paper we have illustrated two methods for analysing repeated measurements in the estimation of the agreement between two methods of measurement. Although these have been described previously (Bland and Altman 1999) no numerical example was given for the case when the true value of the quantity being measured varied and also there was an error in the mathematical description. Although a numerical example was given for the case when the true value did not vary, it was presented in very mathematical terms. We hope that this presentation will be more useful to researchers.

In these examples, the estimated limits are wider than those obtained if the data structure is ignored, as we would expect, but by only a small amount. It may be that this will be the case in many data sets. As noted above, a further possible approach is to average the repeated measurements for each subject and use only 12 pairs of means to calculate the 95% limits. That analysis would be expected to give limits that are too narrow. We again have the mean difference = 0.7092361 and the standard deviation of the differences = 0.9553487. The 95% limits of agreement become $0.7092361 - 1.96 \times 0.9553487$ to $0.7092361 + 1.96 \times 0.9553487$, which are –1.1632474 to +2.5817196, or –1.2 to +2.6. As expected, these limits are again narrower than the correct ones, though they are similar. There is little difference because there is much more variation between the subjects than for the repeated observations on a single subject. Averaging repeated observations for a subject removes only variation within the subject. There would be much greater narrowing of the limits if the variability between the differences within the same subject were similar to that for different subjects, i.e. if there

were less heterogeneity. On the other hand, ignoring the data structure altogether and treating the observations as independent would have less effect than it does in this example if there were less heterogeneity In the complete absence of heterogeneity the limits would be the same as for our analyses.

It must be better to have methods of analysis which do take the structure of the data into account and do not run the risk of producing limits of agreement which are too narrow. Incorrectly calculated limits would lead us to think that methods of measurement agreed more closely than they actually do, which could have adverse consequences.

Such analysis may be further improved by the development of methods to adjust for autocorrelation.

Acknowledgements

We thank Shaun Bowling for the data and the referees for helpful comments on the draft.

References

Almén T, Bergquist D, Frennby B, Hellsten S, Lilja B, Nyman U, Sterner G, Törnquist C. (1991). Use of urographic contrast media to determine glomerular filtration rate. Determining the glomerular filtration rate of each kidney with computed tomography and scintigraphy. *Invest Radiol* 26 Suppl 1:S72–S74.

Altman DG, Bland JM. (1983) Measurement in medicine: the analysis of method comparison studies. *Statistician* 32:307-317.

Barry BN, Mallick A, Bodenham AR, Vucevic M. (1997). Lack of agreement between bioimpedance and continuous thermodilution measurement of cardiac output in intensive care unit patients. *Crit Care* 1:71-74.

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

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

Bland JM, Altman DG. (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res* 8:135-160.

Bowling LS, Sageman WS, O'Connor SM, Cole R, Amundson DE. (1993) Lack of agreement between measurement of ejection fraction by impedance cardiography versus radionuclide ventriculography. *Crit Care Med* 21**:**1523-1527.

Ryan TP and Woodall WH (2005) The most-cited statistical papers. *J Appl Stat* 32:461-474.

Table 1. Cardiac ejection fraction (%) by two methods, radionuclide ventriculography (RV) and impedance cardiography (IC), for 12 subjects (Data provided by Dr LS Bowling REF)

Subj.	RV	IC		Subj.	RV	IC		Subj.	RV	IC
1	7.83	6.57		5	3.13	3.03		9	4.48	3.17
1	7.42	5.62		5	2.98	2.86		9	4.92	3.12
1	7.89	6.90		5	2.85	2.77		9	3.97	2.96
1	7.12	6.57		5	3.17	2.46		10	4.22	4.35
1	7.88	6.35		5	3.09	2.32		10	4.65	4.62
2	6.16	4.06		6	3.12	2.43		10	4.74	3.16
2	7.26	4.29		6	5.92	5.90		10	4.44	3.53
2	6.71	4.26		6	6.42	5.81		10	4.50	3.53
2	6.54	4.09		6	5.92	5.70		11	6.78	7.20
3	4.75	4.71		7	6.27	5.76		11	6.07	6.09
3	5.24	5.50		7	7.13	5.09		11	6.52	7.00
3	4.86	5.08		7	6.62	4.63		11	6.42	7.10
3	4.78	5.02		7	6.58	4.61		11	6.41	7.40
3	6.05	6.01		8	6.93	5.09		11	5.76	6.80
3	5.42	5.67		8	4.54	4.72		12	5.06	4.50
4	4.21	4.14		8	4.81	4.61		12	4.72	4.20
4	3.61	4.20		8	5.11	4.36		12	4.90	3.80
4	3.72	4.61		8	5.29	4.20		12	4.80	3.80
4	3.87	4.68		8	5.39	4.36		12	4.90	4.20
4	3.92	5.04		8	5.57	4.20		12	5.10	4.50

Table 2. Analysis of variance table as produced by Stata

```
    Source |  Partial SS    df       MS           F     Prob > F
-----------+------------------------------------------------------
   Subject |  46.2999416    11   4.2090856      24.66     0.0000
  Residual |  8.19427323    48  .170714026
-----------+------------------------------------------------------
     Total |  54.4942149    59   .92363076
```

Table 3. One-way analyses of variance for RV and IC separately

RV method:

```
    Source |  Partial SS    df       MS           F     Prob > F
-----------+------------------------------------------------------
   Subject |  99.7289076    11  9.06626433      84.55     0.0000
  Residual |   5.14693415   48  0.107227795
-----------+------------------------------------------------------
     Total |  104.875842    59  1.77755664
```

IC method:

```
    Source |  Partial SS    df       MS           F     Prob > F
-----------+------------------------------------------------------
   Subject |  91.9533467    11  8.35939515      60.63     0.0000
  Residual |   6.61795533   48  0.137874069
-----------+------------------------------------------------------
     Total |  98.571302     59  1.67070003
```

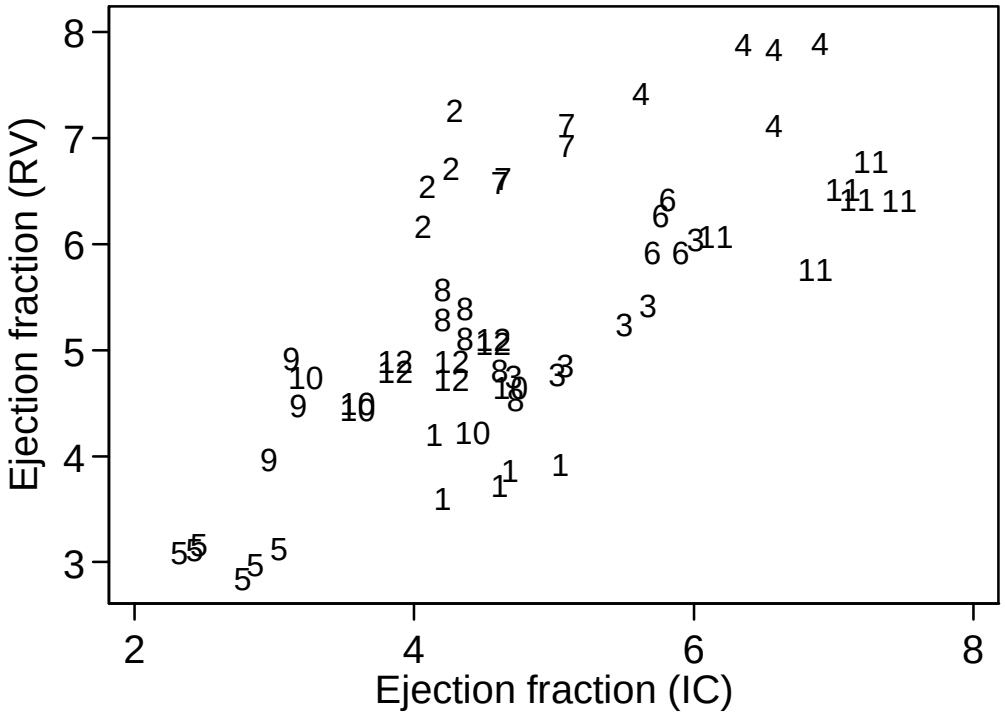

Figure 1. Scatter plot of the data of Table 1 (points are represented by the subject number)

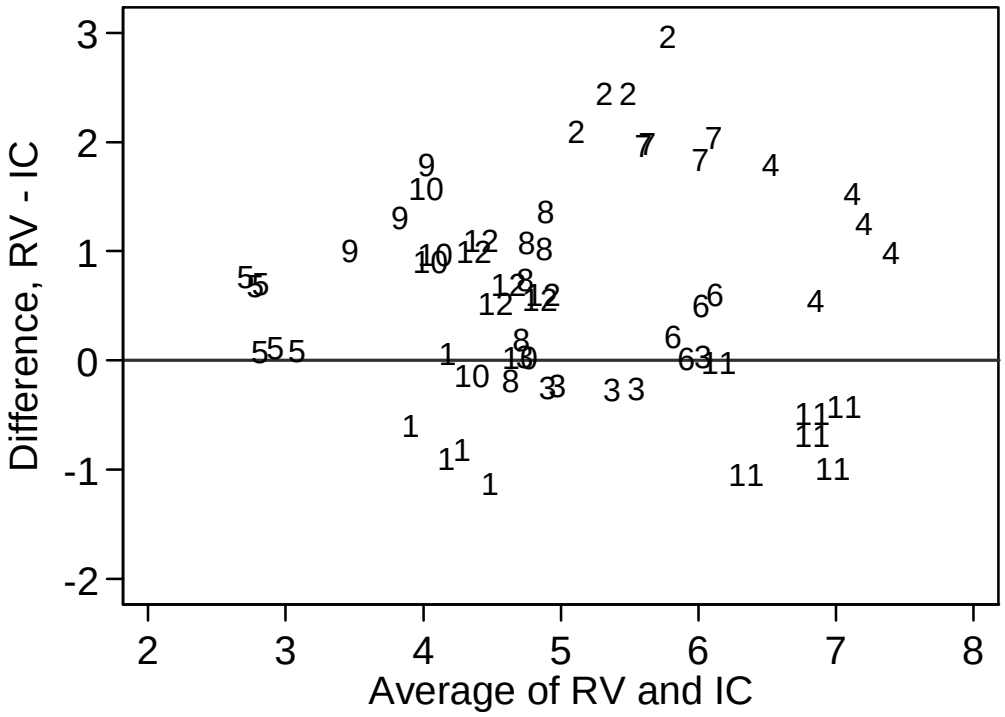

Figure 2. Scatter plot of difference between methods against the average of the two (points are represented by the subject number)

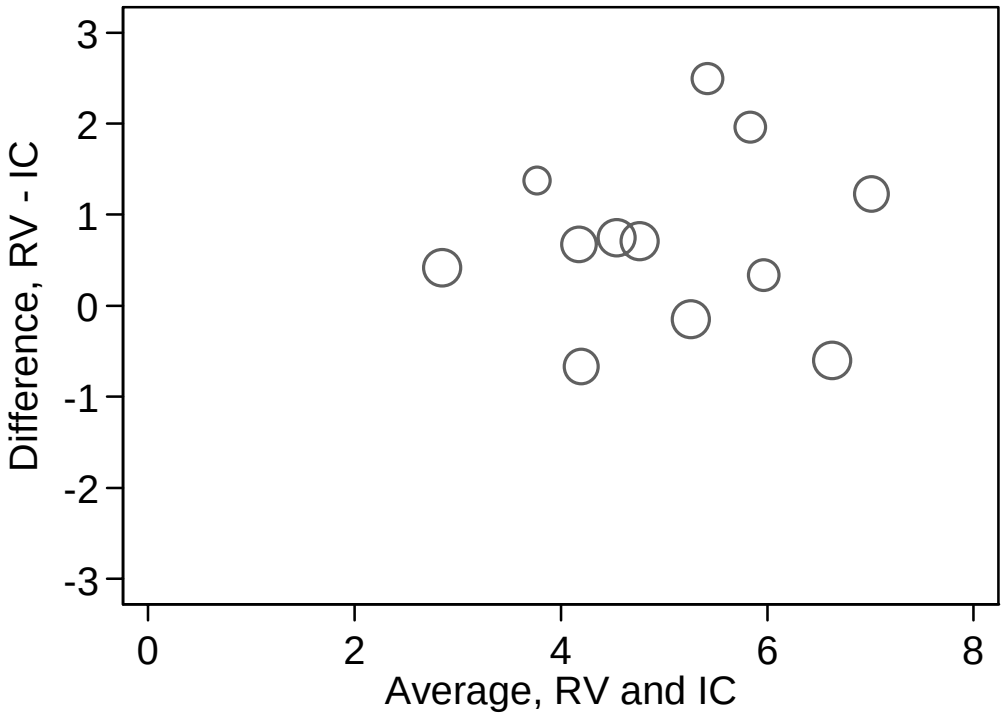

Figure 3. Scatter plots of standard deviation of measurement pair differences against subject mean for 12 subjects. (Area of circle is proportional to number of observations.)

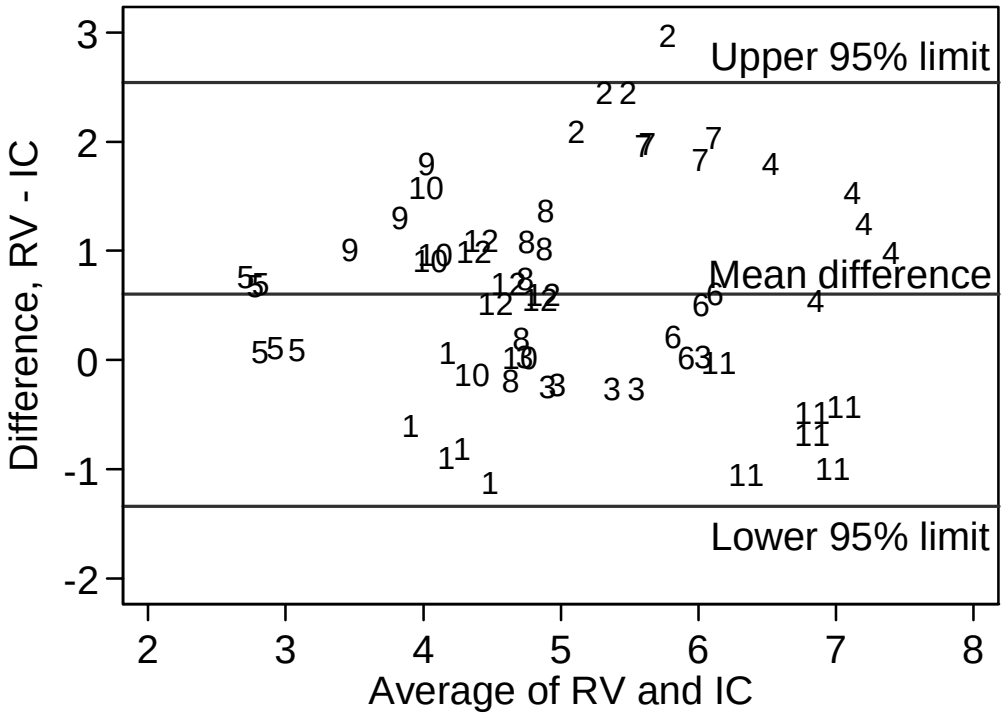

Figure 4. Scatter plot of difference between methods against the average of the two (points are represented by the subject number)

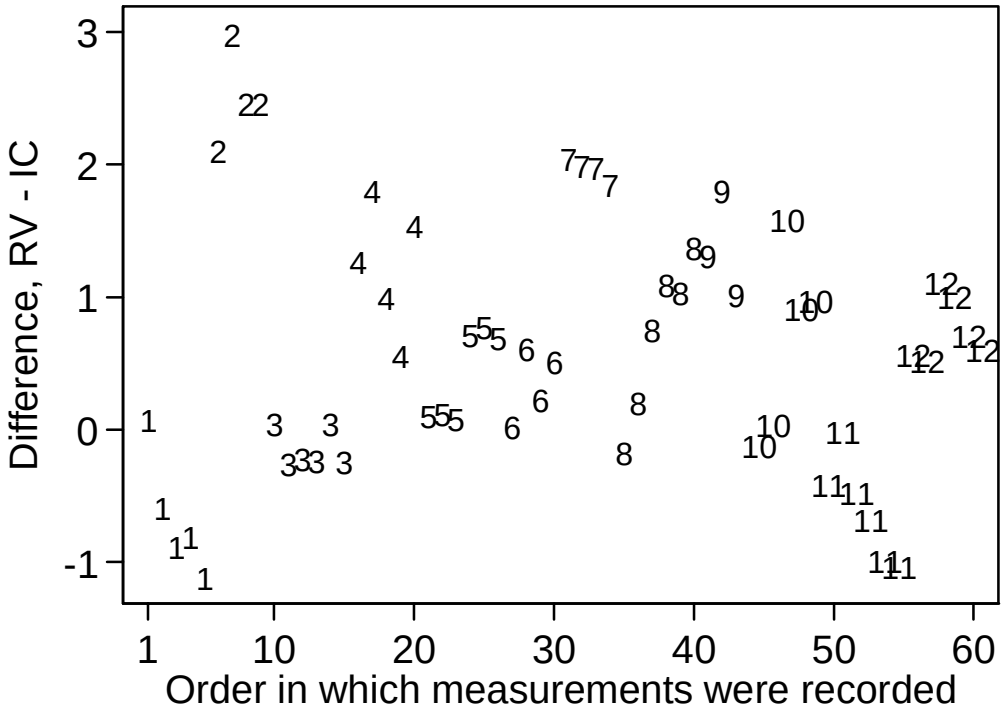

Figure 5. Scatter plot of difference between methods against the order in which the measurements were made (points are represented by the subject number)

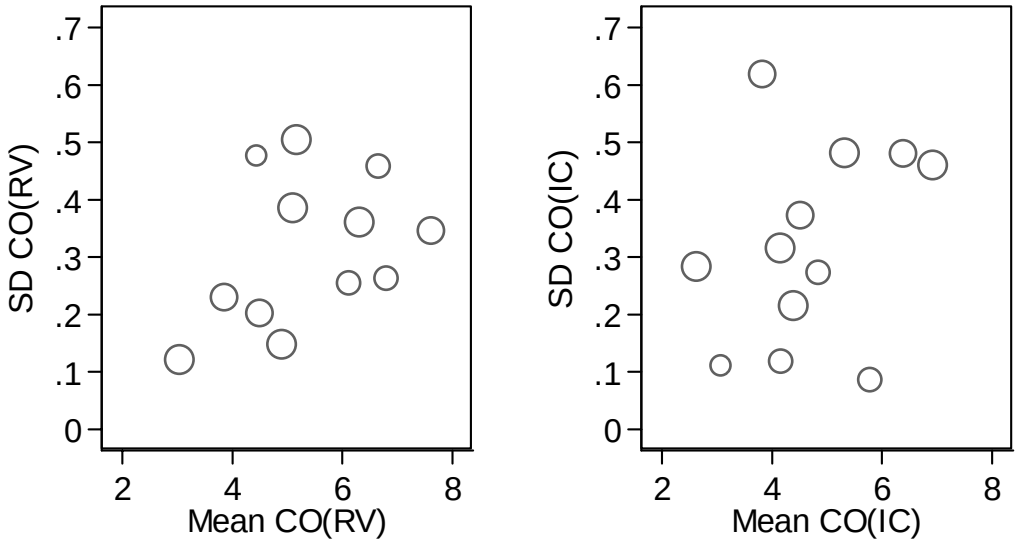

Figure 6. Scatter plots of standard deviation against mean for 12 subjects. (Area of circle is proportional to number of observations.)

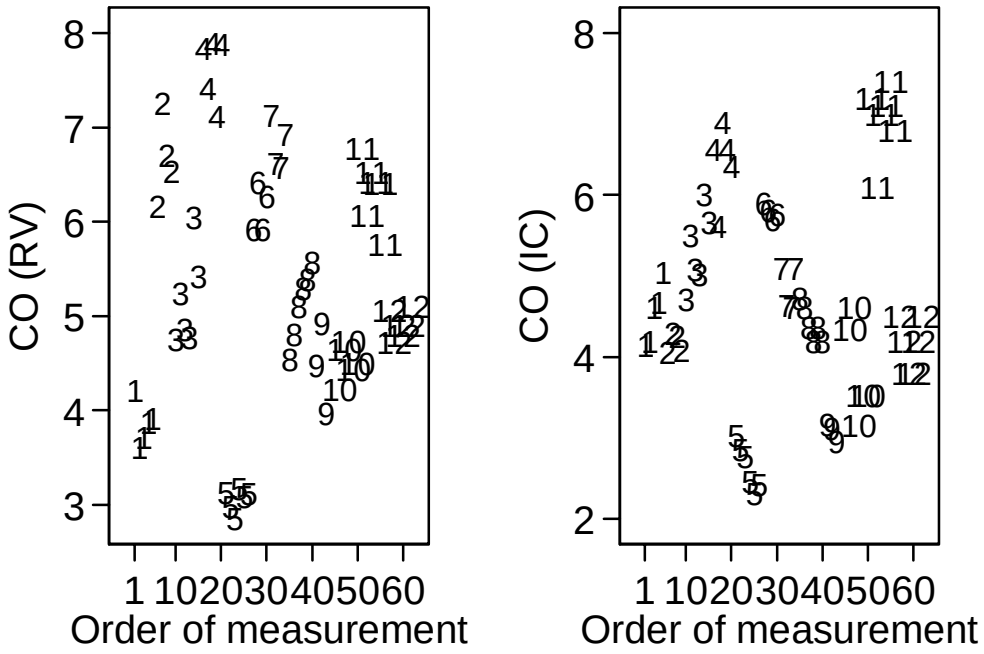

Figure 7. Scatter plot of cardiac output by each method against the order in which the measurements were made (points are represented by the subject number)