BJA

"I would have everie man write what he knowes and no more."–Montaigne

BRITISH JOURNAL OF ANAESTHESIA

Volume 99, Number 3, September 2007

British Journal of Anaesthesia 309–11 (2007)
doi:10.1093/bja/aem214

Editorial I

Using the Bland–Altman method to measure agreement with repeated measures

Medical researchers often need to compare two methods of measurement, or a new method with an established one, to determine whether these two methods can be used interchangeably or the new method can replace the established one.[1–6] In most of these situations, the 'true' value of the measured quantity is unknown.

In a series of articles, Bland and Altman[7–9] advocated the use of a graphical method to plot the difference scores of two measurements against the mean for each subject and argued that if the new method agrees sufficiently well with the old, the old may be replaced. Here the idea of *agreement* plays a crucial role in method comparison studies. There are numerous published clinical and laboratory studies evaluating agreement between two measurement methods using Bland–Altman analysis. The original Bland–Altman publication[7] has been cited on more than 11 500 occasions—compelling evidence of its importance in medical research.

The Bland–Altman method calculates the mean difference between two methods of measurement (the 'bias'), and 95% limits of agreement as the mean difference (2 SD) [or more precisely (1.96 SD)]. It is expected that the 95% limits include 95% of differences between the two measurement methods. The plot is commonly called a Bland–Altman plot and the associated method is usually called the Bland–Altman method. The Bland–Altman method can even include estimation of confidence intervals for the bias and limits of agreement, but these are often omitted in research papers.[8]

The presentation of the 95% limits of agreement is for visual judgement of how well two methods of measurement agree. The smaller the range between these two limits the better the agreement is. The question of how small is small depends on the clinical context: would a difference between measurement methods as extreme as that described by the 95% limits of agreement meaningfully affect the interpretation of the results?

Repeated measurements for each subject are often used in clinical research. Two recent articles in the *British Journal of Anaesthesia* use such a design.[5 6] When repeated measures data are available, it is desirable to use all the data to compare the two methods. However, the original Bland–Altman method[7] was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for repeated measures data. However, as a naïve analysis, it may be used to explore the data because of the simplicity of the method.

Examples of the misuse of agreement estimation for repeated measures data can be found readily in the anaesthetic literature: Opdam and colleagues[3] did repeated measurements of cardiac output in six subjects, but incorrectly analysed and plotted 251 paired data sets using the standard Bland–Altman technique. Niedhart and colleagues[4] compared a processed EEG device's electrode placement on each side of the head in 12 subjects, but analysed and plotted 22 860 paired data sets. Such examples of incorrect use are widespread in the anaesthetic and critical care literature. Bland and Altman have provided a modification for analysing repeated measures under stable or changing conditions, where repeated data were collected over a period of time.[9] As an alternative, we propose using random effects models for this purpose.

Random effects model for repeated measures data

With repeated measures data, we can calculate the mean of the repeated measurements by each method on each individual. The pairs of means can then be used to compare the two methods based on the 95% limits of agreement for the difference of the means. The bias between these two methods will not be affected by averaging the repeated measurements. However, the variation of the differences of the original measurement will be underestimated by this practice because the measurement error is, to some extent, removed. Therefore, some advanced statistical calculation is needed to take into account these measurement errors.

Random effects models can be used to estimate the within-subject variation after accounting for other observed and unobserved variations, in which each subject has a different intercept and slope over the observation period.[10] On the basis of the within-subject variance estimated by the random effects model, we can then create an appropriate Bland–Altman plot.[9] The sequence or the time of the measurement over the observation period can be taken as the random effect.

Following Bland and Altman,[9] the SD of the difference between the *means* of the repeated measurements can be calculated based on the within-subject SD estimates. However, the purpose of drawing the Bland–Altman plot is not for showing the difference between the means against the average of the means, but for a single measurement. Therefore, we need to further calculate the SD of the difference of a single measurement between the two methods according to a formula provided by Bland and Altman using standard statistical software.[9]

To illustrate this approach, we have re-analysed an existing data set comparing two methods of measuring oxygen consumption before, during, and after cardiac surgery:[1] inspired gas analysis (GVO_2) and the reverse Fick method (FVO_2) based on arterial and mixed venous blood gas analysis. In the original study, 20 subjects were studied on about seven occasions, with bias and limits of agreement calculated separately for each of these seven time points.[1] An analysis based on pooling the 144 paired measurements of GVO_2 and FVO_2 ignores the repeated nature of the data, but if we apply the original Bland–Altman method,[7] we obtain the following agreement plot (Fig. 1). The 95% limits of agreement (−128, 88) contain 95% (137/144) of the difference scores. The mean difference (bias) of the measurements between FVO_2 and GVO_2 methods is −20 ml min^{-1}. The SD of the difference is 50 and the width of the 95% limits of agreements is 216. But this approach is invalid, as it assumes each of the 144 data pairs are independent of each other. This cannot be accepted because oxygen consumption in any of the subjects will be correlated with subsequent measurements in that individual.

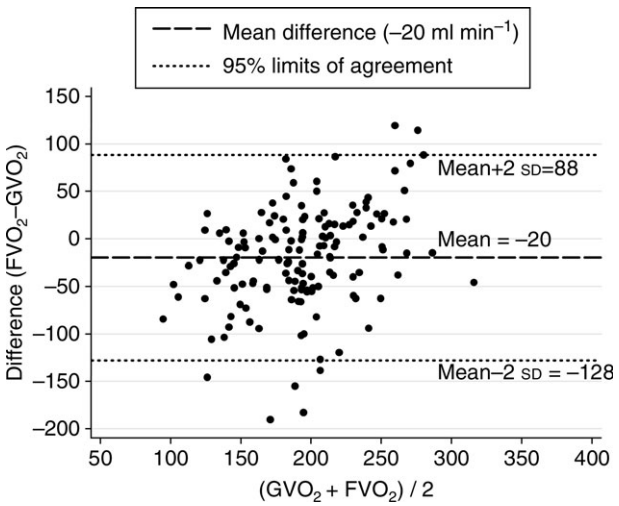

Fig 1 Bland–Altman plot ignoring the repeated nature in the data. The difference between Fick-derived oxygen consumption (FVO_2) and inspired/expired gas analysis-derived oxygen consumption (GVO_2) is drawn against the mean of GVO_2 and FVO_2 in the 144 paired measurements in the study.

As with the standard Bland–Altman method,[7] before the modified Bland and Altman method[9] can be applied for repeated measurement data, a check of the assumption that the variance of the repeated measurements for each subject by each method is independent of the mean of the repeated measures. This can be done by plotting the within-subject SD against the mean of each subject by each method (results not shown). If the assumption underpinning the modified Bland–Altman method is violated, then a log-transformation of the data may correct for this.[7 8]

In random effects modelling, a random effect is usually chosen to reflect the different intercept and slope for each individual with respect to their change of measurements over time. In this analysis, we use the time of the measurement as the random effect. As stated earlier, the main purpose of using the random effects model is to calculate the within-subject SD after the between-subject variation (agreement between methods) has been taken into account by this model. Furthermore, we can include known explanatory variables in the model to adjust for these covariates, in order to get a more precise estimate of the residual variation within a subject.

The difference between our proposed method and the Bland and Altman method[9] is that we used the random effects model to estimate the within-subject variance after adjusting for known and unknown variables. Bland and Altman[9] used one-way analysis of variance to estimate the within-subject variance. In general, the random effects model is an extension of the analysis of variance method and it can adjust for many more covariates than the analysis of variance method.

When using our data to fit a random effects model for GVO_2 and FVO_2 measurements separately, explanatory variables can include the baseline measurement (pre-induction) for each subject, mean measurement for each subject (over time), and the mean measurement between two methods for each measurement occasion.

Table 1 shows the within-subject SD after fitting the random effects model. When there is no covariate in the model (Model 1), the within-subject SD for GVO_2 34.1, which can be reduced to 19.8 when all the explanatory variables are included in the model. Similarly for FVO_2, the within-subject SD can be reduced to 20.5 when all explanatory variables are included in the model. We can create revised Bland–Altman plots by calculating the SD of the difference of a single measurement between the two methods. This will need the within-subject SD calculated earlier.

If we do not adjust for the mean of the two measurements (i.e. Model 4), then the 95% limits of agreement range from −154 to 95. The width of the interval is 249,

Table 1 Within-subject standard deviation (SD) and variables in the model to estimate agreement between Fick-derived oxygen consumption (FVO_2) and inspired/expired gas analysis-derived oxygen consumption (GVO_2)

Model and covariate	Within-subject SD	
	GVO_2	FVO_2
Model 1: no explanatory variables	34.1	47.6
Model 2: adjusting for baseline	34.0	47.2
Model 3: adjusting for mean value for the individual over time	32.6	46.3
Model 4: adjusting for baseline and mean value for the individual over time	32.5	46.2
Model 5: adjusting for baseline, mean value for the individual over time, and mean measurement between two methods	19.8	20.5

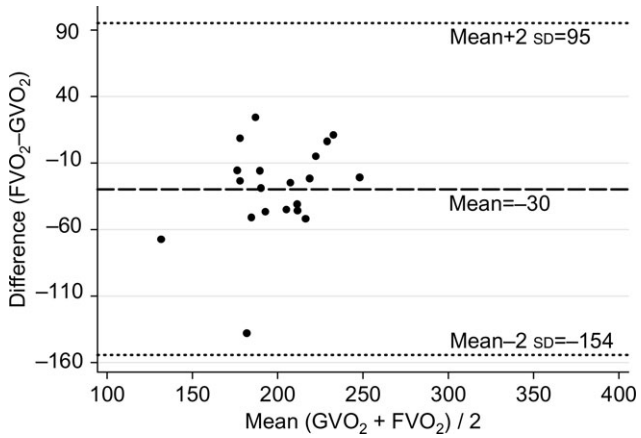

Fig 2 Revised Bland–Altman plot of the difference between inspired/expired gas analysis-derived oxygen consumption (GVO_2) and Fick-derived oxygen consumption (FVO_2) against the mean of the GVO_2 and FVO_2 in the 20 patients in the study. The within-subject variance is estimated by a random effects model which does not include the mean measurements of the two methods for each measurement occasion.

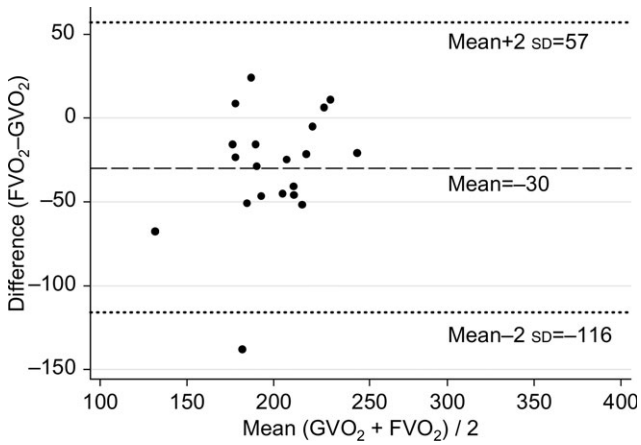

Fig 3 Bland–Altman plot of the difference between inspired/expired gas analysis-derived oxygen consumption (GVO_2) and Fick-derived oxygen consumption (FVO_2) against the mean of GVO_2 and FVO_2 in the 20 patients in the study. The within-subject variance is estimated by a random effects model which includes the mean measurements of the two methods for each measurement occasion.

suggesting unacceptable agreement (Fig. 2). However, if we use Model 5 (Table 1), which includes the mean measurements of the two methods for each measurement occasion, then the width of the 95% limits of agreement will be substantially reduced (Fig. 3). The 95% limits of agreement will be from −116 to 57, which include 95% (19/20) of all patients' difference data. This width of the interval is 173, which is narrower than that derived in Figure 2. It is also less than that derived from the standard Bland–Altman method.[7]

The standard Bland–Altman method[7] cannot be applied when estimating agreement between two measurement methods done on repeat occasions. However, a modification to this approach can be used.[9] In addition, we outline how our random effects models can account for the dependent nature of the data, and additional explanatory variables, to provide reliable estimates of agreement in this setting.

P. S. Myles*
Department of Anaesthesia and Perioperative Medicine, Alfred Hospital, Commercial Road, Melbourne, Victoria 3004, Australia
J. Cui
Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
**E-mail: p.myles@alfred.org.au*

References

1 Myles PS, McRae R, Ryder I, Hunt JO, Buckland MR. The association between oxygen delivery and consumption in patients undergoing cardiac surgery. Is there supply dependence? *Anaesth Intensive Care* 1996; **24**: 651–7
2 Myles PS, Story DA, Higgs MA, *et al.* Continuous measurement of arterial and end-tidal carbon dioxide during cardiac surgery: $P_{a\text{-}ET}CO_2$ gradient. *Anaesth Intensive Care* 1997; **25**: 459–63
3 Opdam H, Wan L, Bellomo R. A pilot assessment of the FloTrac(TM) cardiac output monitoring system. *Intensive Care Med* 2007; **33**: 344–9
4 Niedhart DJ, Kaiser HA, Jacobsohn E, Hantler CB, Evers AS, Avidan MS. Intrapatient reproducibility of the BISxp monitor. *Anesthesiology* 2006; **104**: 242–8
5 Anderson RE, Sartipy U, Jakobsson JG. Use of conventional ECG electrodes for depth of anaesthesia monitoring using the cerebral state index: a clinical study in day surgery. *Br J Anaesth* 2007; **98**: 645–8
6 Button D, Weibel L, Reuthebuch O, Genoni M, Zollinger A, Hofer CK. Clinical evaluation of the FloTrac/Vigileo^{TM} system and two established continuous cardiac output monitoring devices in patients undergoing cardiac surgery. *Br J Anaesth* 2007; **99**: 329–36
7 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; i: 307–10
8 Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–7
9 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–60
10 Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**: 963–74