BJA

British Journal of Anaesthesia, 117 (5): 569–75 (2016)

doi: 10.1093/bja/aew320
Review Article

Reporting of method comparison studies: a review of advice, an assessment of current practice, and specific suggestions for future reports

A. Abu-Arafeh[1], H. Jordan[1] and G. Drummond[2,*]

[1]Department of Anaesthesia Critical Care and Pain Medicine, Royal Infirmary of Edinburgh, 51 Little France Crescent, Old Dalkeith Road, Edinburgh EH16 4SA, UK and [2]Anaesthesia Critical Care and Pain Medicine, Division of Health Sciences, The University of Edinburgh, Edinburgh Medical School Deanery of Clinical Sciences, 49 Little France Crescent, Edinburgh EH16 4SB, UK

*Corresponding author. E-mail: g.b.drummond@ed.ac.uk

Abstract

Background. Anaesthetic journals frequently publish studies comparing measurement methods. A common method of analysis is the Bland and Altman plot, which relates the difference between paired measurements to the mean of the pair. Previous reviews have shown that key data are often omitted from reports using this method of analysis, and the analysis of more complex data is frequently insufficient.

Methods. We identified articles by searching reports, and subsequent citations, considering use of the method. We assembled a list of frequent and important criteria from these articles. These key features were tested by assessing articles in the yr 2013 and 2014, in five anaesthetic journals: Anaesthesia, Anesthesiology, Anesthesia and Analgesia, The British Journal of Anaesthesia, and The Canadian Journal of Anaesthesia.

Results. We found 29 features suggested for reporting such studies. Eight of these were frequently found. We chose 13 key features. In the journal articles reviewed to test these features, three features were almost always reported: the data structure, a plot of the bias, and the limits of agreement of the differences. Often, features required for adequate interpretation of the studies were absent, notably an *a priori* decision of acceptable limits of agreement, and an estimate of the precision of the limits of agreement.

Conclusions. Bland and Altman analysis remains poorly reported. Our formal list of key criteria will assist authors in providing all the relevant features of a study. We explain errors that may be made in reporting, and suggest methods for analysis, including easily available software.

Key words: Accepted for publication; Editor's key points; Research design; Standards; Software

To compare different methods of measurement, study results are often presented and evaluated using the general method popularised by Bland and Altman, whose paper in 1986 became one of the most commonly cited in statistics.[1] This process compares measurements made by two different methods, and has been widely adopted by anaesthetists. An example is shown in Figure 1. This plot displays the difference between a pair of measurements made with the two methods, in relation to the

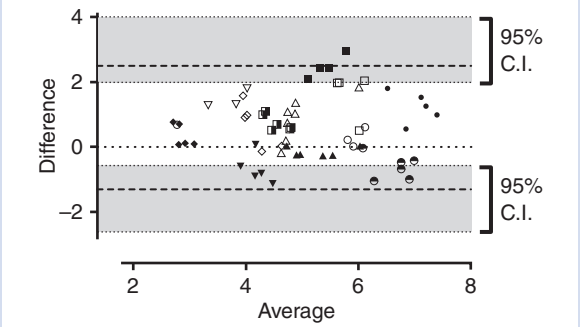

Fig 1 Influence of multiple measurements in several subjects (indicated by the different symbols) on the precision of the limits of agreement. Data replotted from Bland and Altman (2007) J Biopharm Stat 17 571. Grey areas represent the 95% CI for the Limits of Agreement.

mean of this pair of measurements. The values fall within the "limits of agreement" which summarize the overall matching between the two methods of measurement. If measurements with the two methods are similar, then the differences between them will be small, with an average near zero, they will be consistent over the range of measurement values, and the limits of agreement will be narrow. Statistical treatment of the data is simple, if only one pair of measurements is taken in each of a number of separate subjects. However if several measurements are made in multiple subjects (as in the example in Figure 1), the limits of agreement are less easily calculated, and are not exactly known. In the example shown in Figure 1, we present the limits of agreement with their 95% confidence intervals. Answering the final question "do these methods yield results that are in agreement?" depends in large part on the overall range of the limits of agreement and also their confidence intervals. A misinformed answer to this question could mean that a new, unreliable measurement device is inappropriately used to guide clinical practice.

Repeated reviews since 1999[2–8] have shown that the Bland and Altman method is inconsistently used and inadequately reported. Several of these reviews specified features that if reported, would allow proper evaluation of a published study. Unfortunately these helpful suggestions were often not laid out systematically as an explicit list of specific requirements. Because previous reviews of the Bland and Altman method did not formally list the key features required, valuable suggestions and recommendations for adequate reporting appear to be rarely followed.

The most important element of inadequate reporting relates to the "limits of agreement". This is a critical feature of the method. These limits are an estimate, based on the experimental sample provided by the study, and represent the likely scatter of the average differences. The limits of agreement can only be used properly if the confidence intervals of these limits are known.[6,9] These confidence intervals are affected, often substantially, by the structure of the data, particularly when several measurements are made in each of a number of subjects (Fig. 1). This form of data is frequent in clinical studies. Measurements recorded on the same patient could be expected to vary less than measurements recorded from separate patients.[10] Confidence intervals for the limits of agreement are rarely presented in medical studies.[4,11] This may in part be because methods to calculate these values easily are not readily available,[10,12] compared with software that is used to carry out other frequently used statistical tests.

Reporting guidelines are now common for many types of scientific study.[13] Using guidelines should improve the standard of published research, and allow more effective pooling of study data.[14] In an attempt to improve the standard of reporting comparisons of methods of measurement, we reviewed all available material on Bland and Altman analysis. We collected published material which suggested features that should be reported when a comparison of methods was conducted, and drew up a practical summary. We tested this list to assess the use of reporting standards, by examining comparison studies recently published in major journals of anaesthesia. Our findings suggest that journals should provide explicit guidance for the reporting of comparison studies.

Methods

Establishing criteria

We assembled all the papers (original articles, editorials, and letters, in the English language) we could find that discussed, criticised, or recommended how comparison studies using continuous data should be reported. Having identified six obvious source articles[2,3,5,6,8,15] published between 1990 and 2007, we then used links from these articles to "related papers" or similar facilities in PubMed, the ISI citation index, and Google Scholar. A conventional search for original articles in PubMed failed to return many of these articles, possibly because relevant publications were not original papers, but were editorials or even letters, or because search terms are too literal.

We searched PubMed using a broad strategy using MeSH keywords that were associated with our source articles, (i.e. "[Data Interpretation, Statistical] AND Monitoring, Physiologic/ statistics & numerical data*"). This yielded many more general articles on standards or guidance on reporting.

We consulted "instructions to authors" provided by the following journals: Anesthesiology, Anesthesia and Analgesia, The British Journal of Anaesthesia, Anaesthesia, and The Canadian Journal of Anaesthesia. We wrote to the Editor-in-Chief of each of these journals asking if their journal provided any specific guidance for editors, assessors or referees, relating to statistical matters in general and specifically in regard to comparison studies. In several cases, a repeat request was required. Two journals failed to respond.

We found 111 papers that were potentially useful in providing guideline material (Supplementary material A). These papers included not only those relevant to the method of Bland and Altman, but also more general articles on reporting comparison studies, and more general guidelines on reporting. We reviewed each of these articles and noted all the suggestions made concerning reporting criteria which were directly relevant to the Bland and Altman method. The papers that were used to provide these suggestions are indicated in the supplement. From these suggestions, we assembled the most frequent and pertinent criteria (Supplementary material B). The results section below (Results –Setting Criteria) reports how these were assembled into a list of 13 key items, that would allow a practical measure of the completeness of presentation of Bland and Altman comparisons (Table 1).

Assessing recent publications

Each author independently searched two calendar yr of issues for these journals (2013 and 2014, excluding supplements and special issues) to obtain a contemporary sample of comparison

Table 1 List of key features for adequate presentation of Bland and Altman analysis

1.	*A priori*, establish the acceptable limit of agreement. *A statement that the authors determined what was clinically relevant, and expected the observed LOA to fall within that range*
2.	Describe the data structure (i.e. single paired measurements, replicates, several measures in different subjects) *This feature may be implicit in the description of the study,(e.g. a single measure of NIBP with two devices in each patient). Replicates are immediately repeated measures(i.e. "true value constant") so that the innate variation of the measurements (one of the several sources of variation) can be determined. These are not frequent in anaesthetic literature but could be present for example replicate blood samples subjected to the same assay within a few s*
3.	If possible, estimate the repeatability of the measures. *(i.e. if replicates are available, estimate the differences between replicates and the SD of these values).*
4.	Plot the data and inspect for absence of trend and constant variance *For example, provide a distribution histogram of the differences which should be normally distributed. If the plot suggests that there may be a trend in the differences, see if there is a significant regression. See if the scatter of the differences depends on the mean of the observations (commonly, an increase in scatter with an increase in the mean)*
5.	If necessary, transform the data (e.g. ratio, log) to account for changes in variance in the differences *In other words, follow on from 4*
6.	Plot and report numerically the mean of the differences (Bias)
7.	Give an estimate of the precision (e.g. SD, or 95% CI) *This is usually the SD of the differences*
8.	Calculate and indicate the limits of agreement (LOA) of the differences
9.	Provide an estimate of precision (e.g. 95% CI) *The CI may be applied to the LOA to give the upper and lower bounds of the estimate of the confidence range.*
10.	Ensure that the range of the mean values is sufficient. A narrow range of original values will result in agreement being inevitable. *This can be verified by inspecting the results of randomly paired variables (see Preiss and Fisher)*
11.	Variance between and within subjects, or a statement that the CI of the LOA were calculated taking the data structure into account. *The CI for the limits of agreement can be substantially affected by the data structure*
12.	Software or computing processes used. *Readers may wish to check the results, validate the findings, or use the methods in a further study.*
13.	Statistical assumptions made, such as normality of the data. *If the normality of the data is tested, state the test used, and the result obtained*

studies in principal anaesthetic journals. After review and discussion, all authors agreed on the inclusion of the articles selected for scoring. These papers are listed in Supplement C. Each paper was assessed and scored by each author independently, using our list of criteria. A single mark was allocated for criterion present, and zero for not present. In many articles, the presentation of these criteria was not explicit: for example, the adequacy of limits of agreement was considered in the discussion section or by reference to previous publications, and not *a priori*. General difficulties with scoring were discussed by the authors, allowing refinement of the criteria, but the final scores allocated to each article remained individual.

The conduct of the study reported in each article, and the structure of the data, influenced the actual criteria that could be applied to each paper. One such criterion was reporting the repeatability of a measurement. This can only be assessed if several measurements have been made in conditions when the feature measured is not expected to vary. This condition, often expressed as "true value remains constant" is not frequent in anaesthesia: the alternative, "true value varies" is much more common. The other criterion that may not be always applicable is when several sets of measures are made in several subjects, a more complex structure compared with when only one pair of measures is made in each of several subjects. The former circumstance of multiple patients, each with multiple measurements, will be described as "repeated measures" as has been done previously.[6]

The adequacy of reporting of each criterion, when applicable, was summarized by the sum of the marks allocated by the authors, and expressed as a percentage of the occasions when this was an applicable criterion. We also noted the citations of "methods papers" made by each paper, to see if these were appropriate to the data considered.

Results

Results: setting criteria

Not all the journals surveyed provided guidelines to authors for reporting Bland and Altman comparisons. One journal suggested the "Bland and Altman method" but gave no further details. One journal provided more specific advice, by referring to two publications which we had included in our "source articles".[8,16] One journal referred assessors to the original paper by Bland and Altman[17] which is often insufficient, and gives no guidance on acceptable reporting of the variables.

We found 111 papers that could be relevant to our question "what do authors recommend should be reported when a Bland and Altman analysis is presented?" These papers are listed in supplementary material A. However, a substantial minority of these papers were suggesting alternative methods of analysis, and did not directly provide guidance about the Bland and Altman method *per se*. After these papers were excluded, we analysed the recommendations of 64 publications (indicated by * in the list) and noted the features that these publications suggested for adequate reporting. We noted a total of 29 unique

features. Eight of these features were found in five or more of these publications, and another 10 features were suggested by between two and five publications in the papers reviewed. These results are detailed in supplementary material B. The nine most popular features are listed below:

1. A *priori*, establish the acceptable limits of agreement
2. Describe the data structure (e.g. single paired measurements, replicates, several measures in different subjects)
3. If possible, estimate the repeatability of the measurement.
4. Plot the data and inspect for absence of trend and constant variance.
5. If necessary, transform the data (e.g. ratio, log) to account for changes in variance in the differences
6. Calculate and plot the mean of the differences (bias)
7. Give an estimate of the precision of the bias (e.g. SD, or 95% CI)
8. Calculate and indicate the limits of agreement (LOA) of the differences
9. Provide an estimate of precision of the LOA (e.g. 95% CI)

Figure 1 shows an example of the last feature. Surprisingly, none of the features in this table were mentioned in 29 out of the 65 publications that provided guidance, perhaps because some were "too obvious to mention". Such features are often presented in the plot itself (the plot of the difference between measures, against the mean of the two measures) and features of the population such as the total number of subjects, the numbers of measurements, and the sampling conditions, are usually expected to be provided in any research report.

We found one guidance paper which showed clearly that the usefulness of the limits of agreement may depend on the range of the values studied. This paper randomly reassigned data values between pairs of measurements. Analysis of these "shuffled pairs" showed that if the range of the measurements was small, the limits of agreement remained "acceptable".[18] Although this was the only paper we found that suggested a specific check for this feature, it is clearly an important factor. A simple example is shown in supplementary material D. As a result, we added the following requirement:

10. Ensure that the range of mean values in the data is sufficient.

Other features that we found in our survey were mentioned rarely, because they are only relevant in specific conditions. Thus, only four publications suggested that variance within and between subjects should be assessed. This knowledge is required to calculate the CI for the limits of agreement, when several measures are made in different subjects. The CI for the limits of agreement can be substantially affected by the data structure[19] (See Figure 1, and the discussion for further explanation) We therefore set a criterion:

11. Present Variance between and within subjects, or provide a statement that the CI of the LOA were calculated taking the data structure into account.

Finally, in our publication review, we found that only three papers discussed the computing methods used and only two provided sources for code. We therefore added these features as they are desirable in any report using statistical tests:

12. Software or computing processes used.
13. Statistical assumptions made, such as normality of the data.

These key criteria were assembled into a standard document with explanatory comments. (Table 1)

Results: assessing recent publications

The authors initially agreed that 44 papers from the sample of recent journal issues should be considered. (supplementary material C) After a first attempt at scoring, further discussion excluded two papers from further consideration. One (paper 1 in material C) had applied the comparison method to regression coefficients, which is certainly an unconventional and probably an inappropriate application of the method. The other (paper 41) was a comparison of methods of measurements of cardiac output, which referred to the Bland and Altman method in citations, but only applied polar plot analysis, and did not apply the Bland and Altman procedure. One paper was included which was unconventional (paper 22). Here, the Bland and Altman method was used to compare changes, not absolute values, of cardiac output measured by two different methods, before and after interventions. In all, 42 papers were scored.

In the 42 papers considered, we assessed 13 key features of the presentation of results, listed in Table 1. In some papers, not all features were relevant: for example, in 10 of the 42 studies analysed, single measurements were made in individuals. Those studies that involved "repeated measures" are indicated by * in the supplementary material.

Each author marked each paper independently. We found that some criteria, for example the exact data structure, were difficult to judge, requiring repeated careful reading to be certain of these features. The option of 'not applicable' was possible, and if all three markers considered a criterion 'not applicable' then that criterion score was removed from the possible maximum score for the paper, reducing possible score below 39. The score for each paper was expressed as a percentage of the possible score for that paper.

Thus the maximum possible score for each paper could vary from 27 to 39. The median of the scores was 59% (quartile values, 50 and 70%) (Fig. 2).

Some criteria, such as data structure, and the bias between the methods, were well reported: others such as the precision of the limits of agreement, were reported rarely (Fig. 3).

Considering the 32 studies in which multiple measurements were made in several individuals, 26 of these cited a reference

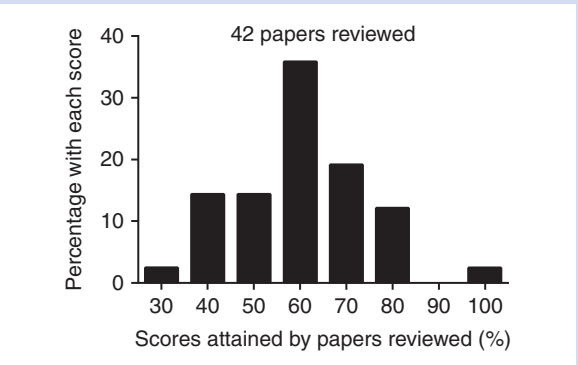

Fig 2 Distribution of scores achieved by the papers reviewed. The median score was 57%, (quartiles 50, 70).

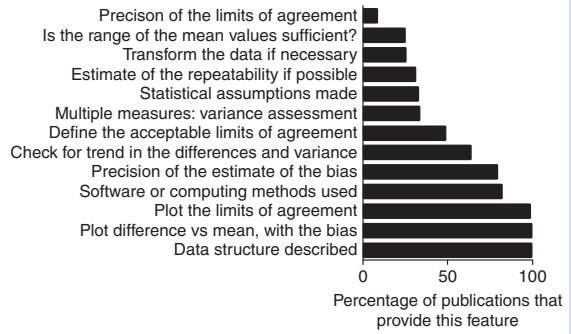

Fig 3 Frequency with which the criteria were found in the test survey.

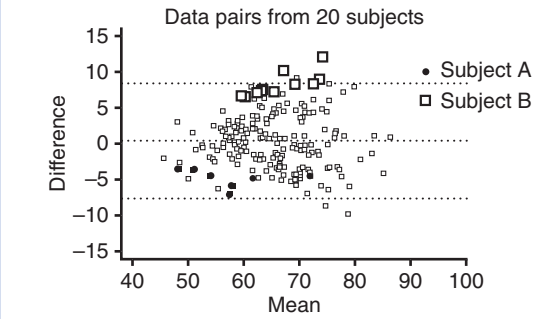

Fig 4 An example taken from one of the papers we surveyed, where a figure presented multiple measurements taken from 20 subjects. We emphasize two data sets from individual subjects with substantial variation in bias. (compare A with B). Other subjects in the data presented showed variable precision. These features were barely discernible in the original publication.

to an appropriate method of analysis, regarding management of variation between and within subjects. The other eight did not cite appropriate references for a method of analysis (These papers are indicated by † in supplementary material). Thirteen papers stated that the confidence intervals presented for the differences between the methods took account of the data structure, or presented information about the between- and within- subject variation, or both. However, only five of these papers presented 95% CI for the limits of agreement, although 30 papers provided a plot of the limits of agreement. In some papers, a different feature, the accuracy root mean square error, was used to express the variability of the differences.

Discussion

We tested our checklist of key features required for reporting Bland and Altman analysis by reviewing recent articles in major anaesthetic journals. We found that that current reporting of the Bland and Altman method is imperfect. Although data structure was described and bias was plotted reliably, the clinically acceptable limits of agreement were not established *a priori* in more than 50% of the publications reviewed. Bland and Altman consider this to be a key criterion.[10]

The journals we chose were exclusively anaesthetic, and those with the highest impact factor. It was used to test the list of criteria we devised, and not intended to investigate specifically the reasons for poor reporting. A possible further study would be the use of *a priori* limits in specific fields, such as cardiac output measurement, where some authors have already suggested limits.

Agreement between devices is commonly determined by the limits of agreement. The precision of this measure depends on the source of variance in the data, for example when multiple measurements are taken from each patient. We illustrate these different sources of variance, drawn from measurements taken from one of the papers we reviewed, in Figure 4. It shows how a compound figure of all the data could mislead, concealing differences in bias between subjects. Out of 32 papers in which multiple measurements were taken from each patient, only four received a score for providing an estimate of the precision of the limits of agreement. As we show in Figure 1, this precision may in some cases be limited, and proper knowledge of this feature may alter practical decisions about the value of a measurement device. This was the most frequent feature that was poorly reported (Fig. 3). Our limited observations provide little evidence for the reasons for this, which could include setting alternative limits such as "percentage error", the wish to conceal possibly poor agreement, and the lack of software available to derive these values.

Poor reporting of Bland and Altman analysis has been described before. In 1999, Critchley, Lee, and Ho[2] reviewed studies of cardiac output, searching for four features. In 2000, Mantha and colleagues[3] suggested five items that should be provided for adequate description of a comparison study.[3] Both groups reviewed samples of publications and found that few reported the features that they sought. A later study[5] had similar findings. Other authors have pointed out that other features are important: Myles and Cui[6] emphasised that the precision of limits of agreement depends on the structure of the data, and Preiss and Fisher[18] have shown that the range of the data affect the limits of agreement.[18]

In view of the past poor standard of reporting, and the evidence we present from recent publications that the present position is little better, we suggest that a practical, formal list of desirable criteria could be helpful for authors and reviewers. We derived our list based partly on a consensus of a large corpus of previous advice, and partly on more recent findings. If subsequent reports were required to use such a list, then study descriptions would be more complete, and perhaps more clear. We found that although we had a definition of the features we wished to find in our sample papers, and read the papers carefully, it was often hard to establish some important features. The most obvious of these was the data structure, which has a substantial influence on the method of analysis and interpretation of the results.[19,20] As with many topics, and has been shown when using other reporting criteria, omissions are more easily recognised when a checklist is used.

Poor reporting makes subsequent meta-analysis difficult. Important measures such as cardiac output, haemoglobin, and body temperature have been studied many times, but meta-analysis of such study reports is difficult because the data are presented in different ways and varying detail.[21–23] Others have noted this difficulty[24], which may result from poor instructions to authors on this topic,[25] and have suggested more formal rules for data presentation.[26,14]

Cardiac output measurement systems are often compared in anaesthesia journals. To set limits of agreement *a priori*, as suggested by most of the publications we surveyed, the investigator should define the agreement needed to allow a new measurement system to be substituted for a previously used device. This is fundamentally a clinical decision, affected by considerations such as how much a measurement needs to differ from a previous value to warrant intervention. Rather than setting an *a priori* value, Critchley and Critchley[2] considered the inherent repeatability of cardiac output measurements taken with different devices, derived a percentage measure using an estimate of the combined variances, and defined this as [plus or minus 2(SD) of the measurement differences/mean of the average measurements]. They assessed the repeatability of measurement methods, such as repeated thermodilution measures when the true value was constant, at about 20%. If the new method had a similar % error, then the theoretical minimum % error would be 30%, so they argued that such a value would indicate adequate agreement. Subsequently, this expression of % error has been used as a summary statistic to allow comparison of different measurement devices for cardiac output. It may easily be derived from the SD of the mean of the differences, if the mean of the measurement averages is also provided, and may allow different cardiac output studies to be compared. Thus, the mean of the average values should be provided when % error is not stated. This measure was advocated in only one of the publications we identified in our initial search. The concept of "acceptable" % agreement has been derived *a posteriori* from the variation of repeated measures obtained when the presumed true value is constant (which is not available in many clinical studies). More importantly, this measure does not incorporate any indication of the confidence limits of the estimated variance, which can vary between different studies, and could be considerable, so we suggest that the use of % error is inadvisable.

We could not devise an effective formal search process to find publications that would provide guidance. Permutations of standard MeSH or ISI terms were unhelpful, as were the index terms provided with articles that were clearly relevant. Generally, we found links such as "see related articles" and citations of the key articles that we had already identified, were more likely to provide helpful material. However our aim was not to collect all the information available on a topic, as is required for a systematic review of evidence. Rather, we wished to assemble sufficient material to develop a consensus view of frequent key features, and we are confident that we have assembled sufficient material (supplementary material B) to do this. What was remarkable was how some vital information was rarely detailed specifically (e.g. our key item 2, data structure). If an obvious feature such as this had not been explicitly identified as important, it could be that some other important features were also not included by these papers that we studied, and thus could not be chosen for inclusion in our criteria. Our method could not show up "holes" in current advice: this deficiency would only be met by a consensus of experts.

We tried to be objective when we drew up our list of appropriate criteria, and chose the most frequently suggested features. These were limited by pragmatic considerations. The selection of such features from the source articles was to some degree subjective, but refined by repeated reading of the chosen texts. However, the observations of Preiss and Fisher[17] that the data range affected the limits of agreement was judged important, even though this matter was not considered by any other authors. Unfortunately no software is readily available to carry out this check. However careful inspection of the plot will indicate the relative magnitude of the mean values and the differences. For guidance, we provide an example in the Supplementary material D. This shows a plot not dissimilar to many published, but easily generated by random association of samples from a single population. Apart from this feature, our list contains the predominant "top ranking" recommendations, plus obligatory features such as statistical methods. A consensus conference could have been an alternative way to generate a checklist, but this would require substantial resources that were not available to the authors.

Our test of these key features, conducted on a recent sample, had the advantage of being performed by three independent scorers. Each had been trained to search for the criteria, and each searched independently. Consensus was achieved for the "not relevant" judgement as that affected the denominator of the score of completeness of reporting, but judgement of present/not present remained an individual judgement.

We hope that if a checklist of explicit key features, such as we propose, or similar, is applied by both authors and editors, then the standard of presentation of comparison studies can be improved. In addition, analysis may be assisted by freely available software[27] which provides a simple guide to users to carry out and report these studies, based on recent publications on the topic.[10,12]

Authors' contributions

Study design/planning: A.A-A., H.J., G.D.
Study conduct: A.A-A., H.J., G.D.
Data analysis: A.A-A., H.J., G.D.
Writing paper: A.A-A., H.J., G.D. Revising paper: all authors

Supplementary material

Supplementary material is available at *British Journal of Anaesthesia* online.

Acknowledgements

We thank Erik Olofsen and David P Hall for their assistance and suggestions.

Declaration of interest

None declared.

References

1. Ryan TP, Woodall WH. The most-cited statistical papers. *J Appl Stat* 2005; **32**: 461–74
2. Critchley LAH, Critchley AJH. A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *J Clin Monit Comput* 1999; **15**: 85–91
3. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing methods of clinical measurement: reporting standards for bland and altman analysis. *Anesth Analg* 2000; **90**: 593–602
4. Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the bland-altman plot for interpretation of method-comparison studies: a critical investigation of its practice [4]. *Clin Chem* 2002; **48**: 799–801
5. Berthelsen PG, Nilsson LB. Researcher bias and generalization of results in bias and limits of agreement analyses: a commentary based on the review of 50 acta

anaesthesiologica scandinavica papers using the altman-bland approach. *Acta Anaesthesiol Scand* 2006; **50**: 1111–3
6. Myles PS, Cui JI. Using the bland-altman method to measure agreement with repeated measures. *Br J Anaesth* 2007; **99**: 309–11
7. Cecconi M, Rhodes A, Poloniecki J, Della Rocca G, Grounds RM. Bench-to-bedside review: the importance of the precision of the reference technique in method comparison studies–with specific reference to the measurement of cardiac output. *Crit Care* 2009; **13**: 201
8. Critchley LA, Lee A, Ho AMH. A critical review of the ability of continuous cardiac output monitors to measure trends in cardiac output. *Anesth Analg* 2010; **111**: 1180–92
9. Stockl D, Rodriguez Cabaleiro D, Van Uytfanghe K, Thienpont LM. Interpreting method comparison studies by use of the bland–altman plot: reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic. *Clin Chem* 2004; **50**: 2216–8
10. Hamilton C, Lewis S. The importance of using the correct bounds on the bland-altman limits of agreement when multiple measurements are recorded per patient. *J Clin Monit Comput* 2010; **24**: 173–5
11. Chhapola V, Kanwal SK, Brar R. Reporting standards for bland-altman agreement analysis in laboratory research: a cross-sectional survey of current practice. *Ann Clin Biochem* 2015; **52**: 382–6
12. Zou GY. Confidence interval estimation for the bland-altman limits of agreement with multiple observations per individual. *Stat Methods Med Res* 2013; **22**: 630–42
13. Erb HN. Changing expectations: do journals drive methodological changes? Should they?. *Prev Vet Med* 2010; **97**: 165–74
14. Moher D, Simera I, Schulz KF, Hoey J, Altman DG. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Med* 2008; **6**: 13
15. LaMantia KR, O'connor T, Barash PG. Comparing methods of measurement: an alternative approach. *Anesthesiology* 1990; **72**: 781–3
16. Morey TE, Gravenstein N, Rice MJ. Let's think clinically instead of mathematically about device accuracy. *Anesth Analg* 2011; **113**: 89–91
17. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–7
18. Preiss D, Fisher J. A measure of confidence in bland - altman analysis for the interchangeability of two methods of measurement. *J Clin Monit Comput* 2008; **22**: 257–9
19. Olofsen E, Dahan A, Borsboom G, Drummond G. Improvements in the application and reporting of advanced bland-altman methods of comparison. *J Clin Monit Comput* 2015; **29**: 127–39
20. Hamilton C, Stamey J. Using bland-altman to assess agreement between two medical devices - don't forget the confidence intervals!. *J Clin Monit Comput* 2007; **21**: 331–3
21. Williamson PR, Lancaster GA, Craig JV, Smyth RL. Meta-analysis of method comparison studies. *Stat Med* 2002; **21**: 2013–25
22. Thiele RH, Bartels K, Gan TJ. Cardiac output monitoring: a contemporary assessment and review. *Crit Care Med* 2015; **43**: 177–85
23. Kim S-HH, Lilot M, Murphy LS-LL, *et al*. Accuracy of continuous noninvasive hemoglobin monitoring: a systematic review and meta-analysis. *Anesth Analg* 2014; **119**: 332–46
24. Schriger DL, Savage DF, Altman DG. Presentation of continuous outcomes in randomised trials: an observational study. *Br Med J* 2012; **345**: e8486
25. Schriger DL, Arora S, Altman DG. The content of medical journal Instructions for authors. *Ann Emerg Med* 2006; **48**: 743–9
26. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010; **7**: e1000217
27. Olofsen E. Bland-Altman analysis [Internet]. [cited 2016 Feb 13]. Available from https://sec.lumc.nl/method_agreement_analysis (accessed October 16 2016)

Handling editor: P. S. Myles