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

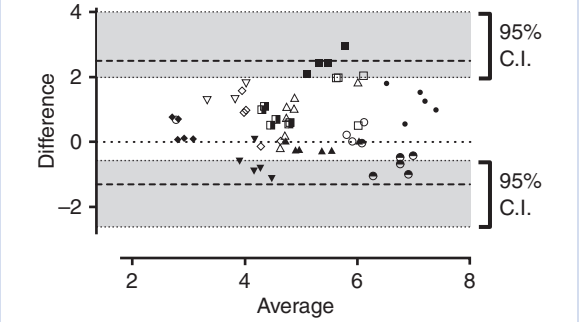

Fig 1 Influence of multiple measurements in several subjects (indicated by the different symbols) on the precision of the limits of agreement. Data replotted from Bland and Altman (2007) J Biopharm Stat 17 571. Grey areas represent the 95% CI for the Limits of Agreement.

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