Mantha et al.

1 Abstract

- In this era of medical technology assessment and evidence-based medicine, evaluating new methods to measure physiologic variables is facilitated by standardization of reporting results. It has been proposed that assessing repeatability be followed by assessing agreement with an established technique. If the limits of agreement (mean bias ± 2SD) are not clinically important, then one could use two measurements interchangeably. Generalizability to larger populations is facilitated by reporting confidence intervals. We identified 44 studies that compared methods of clinical measurement published during 1996 to 1998 in seven anesthesia journals.

- Although 42 of 44 (95.4%) used the limits of agreement methodology for analysis, several inadequacies and inconsistencies in reporting the results were noted. Limits of agreement were defined a priori in 7.1%, repeatability was evaluated in 21.4%, and relationship (pattern) between difference and average was evaluated in 7.1%.

- Only one of the articles reported confidence intervals. A computer macro for the Minitab statistical package (State College, PA) is described to facilitate reporting of Bland and Altman analysis with confidence intervals. We propose standardization of nomenclature in clinical measurement comparison studies.

Implications:

- A literature review of anesthesia journals revealed several inadequacies and inconsistencies in statistical reports of results of comparison studies with regard to interchangeability of measurement methods. We encourage journal editors to evaluate submissions on this subject carefully to ensure that their readers can draw valid conclusions about the value of new technologies.

- Validation of new technology for application to clinical medicine requires comparison with older techniques or assessment of outcomes. These processes, known as medical technology assessment and evidence-based medicine, have gained prominence through publication frequency (1,2). A standard nomenclature has evolved for reporting results after comparison of new methods to monitor physiologic variables with established ones. Thus, for example, the performance of a new monitor to measure cardiac output is compared with an established thermodilution technique.Statistical evaluations of such comparison studies are not simple.

- **Primary Aim** The primary aim of comparison studies is to determine whether the two methods agree sufficiently to be used interchangeability.

Because analysis with correlation and least squares linear regression (also known as calibration statistics) is fundamentally misleading, Bland and Altman favored a different statistical method for assessing agreement between two methods of measurement (35). Their analysis first calculates the difference in measurement values obtained by two methods on the same subject.

- The mean of such differences in a sample of subjects is the estimated bias (difference between methods), and the standard deviation (SD) of the differences measures random fluctuations around this mean. If the limits of agreement (mean difference ± 2SD) between two methods are not clinically important, one can use the two methods interchangeably. Another essential feature of the analysis is graphical representation of the data with between-method difference (y axis) plotted against the average (x axis). Such a graph allows one to evaluate any relationship between the measurement of error (difference) and the assumed true value (average).

- (**Use of CIs**) Because results obtained in a study furnish only the sample statistics, it is necessary for generalizability of results to other populations to report confidence intervals (CIs) (6,7). CIs show a range of values based on the observed data within which, with a specified probability, the population value lies.
  In Bland and Altman analysis (4), CIs for mean bias, mean bias 2SD, and mean bias + 2SD are of particular interest. We reviewed the statistical reporting of measurement comparison studies published in the anesthesia literature according to Bland and Altman analysis.

- We examined the table of contents of seven anesthesia journals (Anesthesiology, Anesthesia & Analgesia, Journal of Cardiothoracic and Vascular Anesthesia, Journal of Clinical Anesthesia, British Journal of Anesthesia, Anesthesia, and Canadian Journal of Anesthesia) published between January 1996 to December 1998. Articles with titles indicating evaluation of a new measurement technique were read. The primary goal was to identify comparison studies in which interchangeability of a new measurement technique with an established method. Animal studies were excluded.

- To ensure accurate data transcription, each eligible study was read at least twice by one author (SM) and graded by written criteria by using an extraction chart for each article. A second authors (JFF) opinion was taken in case of confusion regarding data transcription. From each study, data were retrieved based on written evaluation standards.

- Random audits to ensure accuracy of some data from each article were done by a third author (MFR).We evaluated the comparison studies according to Bland and Altman methodology (34) for the following five items:

  – repeatability,

- definition of limits of agreement,
- representation of x axis on Bland and Altman graph,
- evaluation of relationship (pattern) between difference (y axis data), and average (x axis data),
- report of CIs.

- For repeatability assessment of each study, we first determined whether repeatability is feasible (or practical), and then we determined whether repeatability was evaluated.

- Repeatability is determined by taking repeated measurements on a series of patients and calculating the mean and SD of differences. According to the definition of repeatability coefficient given by the British Standards Institute, the mean difference must not be significantly different from zero, and 95% of the differences are expected to lie within the range from 2SD to + 2SD of the mean (4).
  When reviewing a study for limits of agreement, two aspects were evaluated.

- We determined whether the authors correctly defined the limits as *mean bias ± 2SD.* In the methods section of each article, we looked for a statement defining maximum width for limits of agreement which would not impair medical care i.e., a priori definition of the limits.
  determined the x axis of a Bland and Altman graph for each study because of the potential for authors to erroneously use the x axis to represent the values of the established method rather than the average values of the two methods.

- The relationship (correlation) between difference in measurement values and their average is evaluated to verify whether differences vary in any systematic manner over the range of measurement (3,4).

- (**Computation of CIs)** Bland and Altman (4) derived the following formulas for CIs needed in the analysis:For 95% CIs, t is the critical value for a 5% two-sided test drawn from tables of t distribution with n 1 degrees of freedom (df), where n is the sample size.
  The formula for calculating CI for mean bias (mean difference = $\bar{d}$) is:$\bar{d} \pm (t \times S.E.(d))$ 1 where SD = standard deviation of differences.

Image Tools The formula for calculating CI for limits of agreement (d- 2SD and d + 2SD) is MATH 2 MATH 3 For each study, we determined the following items: physiological variable assessed; the principle of the new monitoring method; the established method used for comparison; whether Bland and Altman analysis was used; whether repeatability was evaluated; whether definition of limits of agreement were made a priori (i.e., described in the methods); whether the x axis of the comparison graph represented the average values of two methods or the values of the established method; whether relationship (pattern) between

measurement error and the average value was evaluated; and whether CIs were reported. Equation 2 Equation 2 Image Tools Equation 3 Equation 3 Image Tools

- Finally, we also tried to infer the definitions of some terms peculiar to measurement, such as accuracy, precision, and parameter (810). However, we did not evaluate the studies based on the use of these terms.

- We identified 66 articles in which a new measurement method was evaluated. Three animal studies were excluded, as were 19 studies in which interchangeability was not the primary goal. In two other studies, conclusions were based on correlation regression analysis. These exclusions left 42 articles for further examination (1152). In all these studies, Bland and Altman analysis was used to project the results.

- Table 1 lists the statistical reporting of measurement comparison studies in these studies. We noted the use of Bland and Altman plot (difference versus average) in 38 articles (90.5%). Data transcription for evaluation and summarization was possible from all the but two studies (27,51). In these two studies, the opinion of one of the coauthors (JFF) was sought to solve the problem.

- Table 2 describes the methodology and reporting of the studies by physiological variables in chronological order of publication. Cardiac output was the most common physiological variable studied (12 of 42, 28.6%), and thermodilution technique was the most commonly used method for comparison.

- In 39 of 42 articles (92.9%), study subjects were patients (in intraoperative, postoperative, or critical care settings), whereas the rest were volunteers. According to our impressions, repeatability was feasible or practical in all but three studies (32,34,45).

- Irrespective of our impressions, repeatability was evaluated in only nine studies (21.4%). If the three studies are excluded, then repeatability reporting is 23.1%. In all but two studies (11,33), the limits of agreement were correctly represented as mean bias ± 2SD. But, the limits of agreement were defined a priori (described in methods) in only three studies: two studies measured blood pressure (29,30), and one measured cardiac output (20).

Two methods were judged to produce identical results in cardiac output measurement that varied substantially from the established thermodilution method. Finally, CIs for Bland and Altman statistics were reported in one study (38). At least three of five quality criteria set in our methods were satisfied in only three studies (20,29,38). Table 2A Table 2A Image Tools Table 2B Table 2B Image Tools Table 2C Table 2C Image Tools Table 2D Table 2D Image Tools Examination for definition of terms revealed that, in 23 studies (54.8Back to Top — Article Outline

Discussion

- Error quantification is an important component in the evaluation of new measurement techniques. Bland and Altman analysis is a statistical technique that quantifies error for repeatability and limits of agreement (34). Our study identified several inadequacies and inconsistencies in the statistical reporting of studies in which new measurement systems were evaluated, although 95% of the studies used Bland and Altman methodology for analysis.Repeatability is relevant in measurement comparison studies because poor repeatability (considerable variation in repeated measurements on the same subject) precludes the assessment of agreement between the two methods of measurement. Therefore, repeatability must be demonstrated before agreement between methods can be established.

- A conclusion about interchangeability should not be based on mean bias alone but also should consider limits of agreement. For example, if a new instrument for noninvasive blood pressure measurement records systolic pressure as 120, 140, 110, 120, and 130 mm Hg in a sample of five subjects and the corresponding values obtained by direct arterial monitoring are 140, 110, 110, 100, and 160 mm Hg, respectively, then mean bias $\pm$ 2SD is 0 ± 51.

- This example illustrates that one can be misled in agreement evaluation if the conclusion is based on the mean bias alone disregarding the limits of agreement. This survey identified one study with such an error (28).Ideally, the limits of agreement need to be defined a priori in the methods, and such a definition was given in only three studies (20,29,30). The American National Standards of the Association for the Advancement of Medical Instrumentation recommend that maximal bias of noninvasive arterial pressure, obtained from at least 85 patients, should not exceed 5 mm Hg $\pm$ 8 SD from a noninvasive reference method (53).

- The **British Hypertension Society** considered the above criterion too liberal and proposed an alternative grading system according to the percentage of readings $\leq 5, \leq 10, \leq 15$ mm Hg from a noninvasive reference method (54). Unfortunately, both these criteria are not readily applicable in perioperative settings because these guidelines were planned for evaluating blood pressure instruments used in outpatient clinics. In *perioperative settings*, an invasive reference standard is usual. One cardiac output study defined the limits of agreement a priori as ± 1 L/min (20). Although not described in methods, two studies used valid criteria for limits of agreement while evaluating results (23,39). The intraarterial blood gas monitoring study (23) used published guidelines (55) to evaluate its results.

- The limits of agreement for blood gas measurements are as follows: PO2 range, 30.4 to 152 mm Hg; PCO2 range, 20.5 to 80.56 mm Hg; the limits must be ± 4.6 mm Hg of the reference. In another study in which an

intraoperative hemoglobin monitor was evaluated (39), the limits were empirically defined as ±1 g/dL from the laboratory reference method. Defining the limits of agreement for different physiologic variables may be a difficult aspect in designing the measurement comparison studies, especially in perioperative and critical care settings, because action limits (clinically important) depend upon the clinical scenario and the status of other related variables.

- **Importance of Design** Nevertheless, an attempt must be made to define such limits at a minimum after pooling data from other studies. Alternatively, a delphi survey (opinion from experts) may be used to design the study. Without a priori setting of limits, widely discrepant limits of agreement have been chosen (Table 2). Such varying limits seem too difficult to accept in practice and may mislead clinicians who are inexperienced in technology of evidence-based analysis.

- The x axis of the Bland and Altman analysis should ideally be represented by the average of measurement values obtained by two different methods because true value is unknown. Bland and Altman proved mathematically that the x axis must represent the average values of two methods (5). Three studies used values obtained by the established method alone on the x axis.

- The plot of difference against average in Bland and Altman analysis also allows us to investigate any possible relationship (correlation) between measurement error (difference between two methods) and the assumed true value (average value of two methods). Bland and Altmans suggestions are subject to the assumption that there is no pattern in the plot of difference versus average (3,4).

- (**Testing Correlation**)The correlation coefficient could be tested against the null hypothesis of $r = 0$ for a formal test of independence. Ideally, such independence should also be demonstrated during a repeatability experiment for each of the two methods. In other words, it is important to ensure that within-subject repeatability is not associated with the size of measurements. Otherwise, results of subsequent analysis might be misleading (3).

Computational Implementation

Although the computational scheme for CIs for Bland and Altman statistics is easy to comprehend, the algebraic calculations are tedious for repeated use. We devised a macro (see Appendix 1) for Minitab (Release 10 and above; Minitab Inc., State College, PA) to facilitate such computation and present it graphically. Minitab is statistical software that can be used for medical applications (56).

Accuracy and Precision

- Finally, standardization of nomenclature is an important issue in scientific writing. It is common to find the terms accuracy and precision in measurement comparison studies (810). Accuracy is defined as closeness of a measurement to its true value, and the term is used when a method is compared with an external standard. In practice, one is rarely comparing a measurement with the true value because a gold-standard method need not necessarily give the true value. Therefore it may be preferable to avoid the word accuracy in these contexts, and use of the term agreement may be preferable (D. G. Altman and M. J. Bland, written communication, 1999).

- **Precision** refers to closeness of values on repeated measurements obtained by the same method, i.e., a measure of repeatability. Confusion may arise with the use of the term precision because of another definition found in statistical literature. A statistical dictionary (57) defines it as follows: precision of an estimator is its tendency to have its values cluster closely about the mean of its sampling distribution. Thus, precision is related inversely to the variance of this sampling distributionthe smaller the variance, the greater is the precision.

- In fact, Bland and Altman used the term precision in the context of reporting CIs (4). In our survey of articles, precision was the most common incorrectly defined term and was used in contexts other than repeatability or reporting CIs. Therefore, in measurement comparison studies, avoiding the term precision and using the term repeatability may seem reasonable.

- If used, the term must clearly be defined (D. G. Altman, written communication, 1999). In medical literature, it is also common to find the word parameter used for variable, as in *We measured the following parameters: temperature, arterial blood pressure, pulse oximetry, end-tidal carbon dioxide and cardiac output.* In statistical literature, the term variable refers to quantities that vary from individual to individual. The term parameter refers to quantities defining a theoretical model (58) and is used to indicate numerical characteristics of a population that are analogous to the numerical characteristics of a sample (statistics).

- The unknown population parameter is estimated from a sample of values of a variable. Therefore substitution of the specific statistical term parameter for variable must be avoided.In this era of evidence-based medicine, standardization of statistical reporting of studies facilitates easy appraisal of published material.

- This survey has identified several inadequacies and inconsistencies in statistical reporting of measurement comparison studies. Such inadequacies render the validity of the conclusions in each of the articles in doubt. We encourage journal editors to evaluate submissions on this subject carefully

to ensure that their readers can draw valid conclusions about the value of new technologies.The authors thank Sally Kozlik for editorial assistance.