BJA

British Journal of Anaesthesia, 116 (6): 750–8 (2016)

doi: 10.1093/bja/aew094
Review Articles

REVIEW ARTICLES

Methodology of method comparison studies evaluating the validity of cardiac output monitors: a stepwise approach and checklist†

L. J. Montenij[1], W. F. Buhre[2], J. R. Jansen[3], C. L. Kruitwagen[4] and E. E. de Waal[1,*]

[1]Department of Anaesthesiology, Intensive Care and Emergency Medicine, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands, [2]Department of Anaesthesiology and Pain therapy, Maastricht University Medical Centre, P. Debeyelaan 25, 6229 HX, Maastricht, The Netherlands, [3]Department of Intensive Care, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands, and [4]Department of Biostatistics and Research Support, Julius Centre, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

*Corresponding author. E-mail: e.e.c.dewaal@umcutrecht.nl

Abstract

The validity of each new cardiac output (CO) monitor should be established before implementation in clinical practice. For this purpose, method comparison studies investigate the accuracy and precision against a reference technique. With the emergence of continuous CO monitors, the ability to detect changes in CO, in addition to its absolute value, has gained interest. Therefore, method comparison studies increasingly include assessment of trending ability in the data analysis. A number of methodological challenges arise in method comparison research with respect to the application of Bland–Altman and trending analysis. Failure to face these methodological challenges will lead to misinterpretation and erroneous conclusions. We therefore review the basic principles and pitfalls of Bland–Altman analysis in method comparison studies concerning new CO monitors. In addition, the concept of clinical concordance is introduced to evaluate trending ability from a clinical perspective. The primary scope of this review is to provide a complete overview of the pitfalls in CO method comparison research, whereas other publications focused on a single aspect of the study design or data analysis. This leads to a stepwise approach and checklist for a complete data analysis and data representation.

Key words: cardiac output; trends; validation studies

Method comparison research aims to evaluate the validity of a new monitor against an established reference technique, and is of specific interest in cardiac output (CO) monitoring.[1 2] After establishing validity, other types of research are needed to evaluate the extent to which new monitors alter haemodynamic management, effects on patient outcome, and cost-effectiveness. Method comparison studies face a number of methodological challenges. A number of reviews have been published, most of them discussing a component of the application of Bland–Altman analysis in this setting.[3–7] Despite these reviews, many studies do not meet a number of fundamental principles.[3 8] This may lead to incorrect conclusions and erroneous applications in clinical practice. This review therefore aims to provide a complete overview of the methodological considerations in method comparison studies concerning new CO monitors. Each component of the study design, data analysis, or data interpretation is followed by a recommendation. In addition, we focus on evaluation of trending ability, which has become increasingly important with the emergence of continuous systems.[9 10] Current methods to analyse trending ability have a number of

† This Article is accompanied by Editorial Aew110.

limitations.[11] As an alternative, the concept of 'clinical concordance' and a corresponding error grid method for evaluation of trending ability from a clinical perspective is introduced. Finally, the methodological issues are summarized, resulting in a stepwise approach and checklist for CO method comparison research. The use of this checklist could lead to a more complete and homogeneous presentation of data, which may facilitate systematic reviews and meta-analyses in the future.

Bland–Altman analysis: concept

Each new CO monitor should be evaluated for its accuracy and precision; accuracy refers to the ability to measure CO close to its true value, whereas precision refers to the spread of repeated measurements (Fig. 1A). Measurement of the 'true' CO is extremely difficult in clinical practice, and reference techniques can provide only an approximation.[1 2] This problem can be handled in part using Bland–Altman analysis.[12–14] This method evaluates agreement between two measurement techniques, rather than validating the experimental technique against a perfect reference. As a result, only conclusions about interchangeability between the experimental and reference technique can be drawn. Bland–Altman analysis determines the bias, or mean difference between the experimental and reference technique, as a measure of accuracy.[12–14] As a measure of precision, the 95% limits of agreement (LOA) are used (Fig. 1B). The LOA are generally

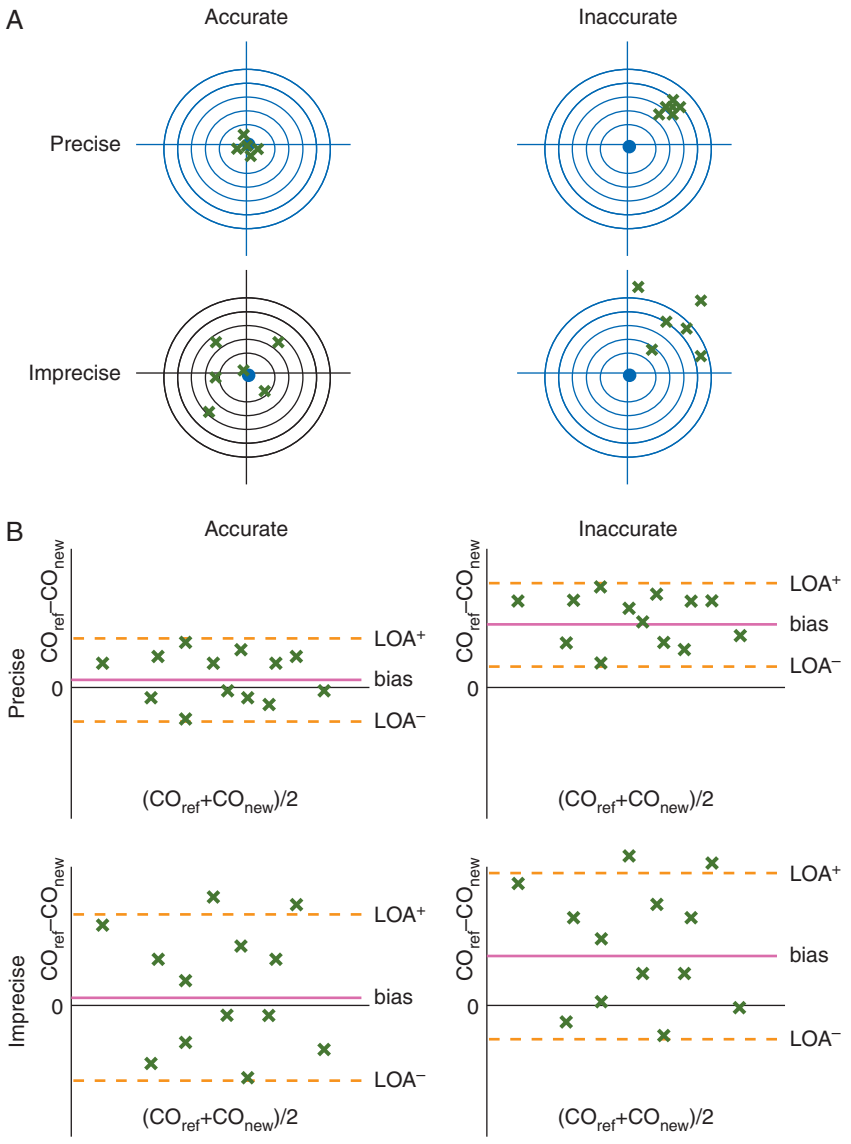

Fig 1 Accuracy and precision and the relation with bias and the limits of agreement as determined with Bland–Altman analysis. (A) Accurate measurements are close to the true value, irrespective of the spread of the measurements; in contrast, precise measurements are close to each other, irrespective of their deviation from the true value. Valid cardiac output monitors are both accurate and precise. (B) In Bland–Altman plots, accurate cardiac output monitors show a bias (continuous line) close to the line 'x=0', whereas precise monitors show limits of agreement close to the bias (dotted lines). CO_{exp}, cardiac output of the experimental technique; CO_{ref}, reference cardiac output; LOA^+, upper limit of agreement; LOA^-, lower limit of agreement.

determined as:

$$\text{LOA} = (\text{bias}) \pm t_{\alpha,n-1} * (\text{SD})$$

in which SD is the standard deviation of the differences, n the sample size, and $t_{\alpha,n-1}$ the t-value corresponding to the degrees of freedom (n–1) and a type I error (α) of 0.05. The LOA therefore represent the limits enclosing 95% of the differences. The bias and LOA can be depicted in a Bland–Altman plot (Fig. 1B). The mean error or percentage error is calculated as follows:

$$\text{Mean error } (\%) = 100\% * t_{\alpha,n-1} * \frac{(\text{SD})}{(\text{mean CO})}$$

Consequently, the mean error is a measure of interchangeability relative to the underlying CO and therefore a more appropriate parameter to compare the results from different studies. For calculation of the LOA and mean error, a t-value of 1.96 is often used. Strictly speaking, this value holds true only in infinitely large sample sizes. It is advisable to use correct t-values in small studies (e.g. <20 subjects), as a value of 1.96 will underestimate the real LOA and mean error.

Pitfalls in the application of Bland–Altman analysis

Bland–Altman analysis has a number of important pitfalls, which are discussed in the next sections, each followed by a recommendation. These recommendations are summarized in a checklist (Table 1).

Normal distribution

The differences between the experimental and reference technique should be normally distributed. Usually, this will be the case, even if the individual CO measurements with the experimental or reference technique do not follow a normal distribution.[14] If not, a straightforward non-parametric approach is available.[14 15] Normal quantile–quantile (QQ) plots or histograms of the differences provide a visual check of normality.[16] In addition, the Kolmogorov–Smirnov or Shapiro–Wilk test can be applied. Nonetheless, small studies may pass these tests because of insufficient statistical power to demonstrate non-normality. In contrast, large studies tend to be tested non-normal even if the deviation from a normal distribution is small.[16]

Recommendation

Check the differences between the experimental and reference technique for normality by combining a visual check and statistical test.

Proportionality

The bias and LOA are meaningful estimates only if they are uniform over the range of measurements.[14] If the difference between the techniques increases with an increase in CO, the bias will be overestimated in the low-CO range and underestimated in the high-CO range. This effect is called proportional bias and can be quantified by plotting a regression line in the Bland–Altman plot. If the slope of this line differs significantly from zero, proportional bias is present.[14 17] Nonetheless, in small studies, proportional bias cannot be ruled out because these studies may lack the statistical power to demonstrate this significant difference. Regression analysis should therefore be accompanied

Table 1 A stepwise approach and checklist to the design, data analysis, and data interpretation of cardiac output method comparison studies. CI, confidence interval; CO, cardiac output; LOA, limits of agreement; 4Q, four-quadrant; TDCO, thermodilution cardiac output; TPCO, transpulmonary thermodilution cardiac output

Study phase	Topic	Checklist item
Design	Measurement protocol	Create a protocol for the timing and recording of CO measurements, considering haemodynamic fluctuations, dependence of paired measurements, and the response time of (continuous) systems
	Criteria for agreement	Define criteria for acceptable bias and LOA or mean error, depending on the clinical context
	Sample size	Consider a sample size calculation (suggested method in Supplementary Appendix B or method by Bland),[21] or assess the appropriate sample size based on historical data
	Reference technique	Choose a highly precise reference technique (e.g. TDCO or TPCO)
Data analysis	Normal distribution	Check whether the differences are normally distributed by combining a visual check and statistical test
	Bland–Altman analysis	Calculate the bias, LOA, mean error, and their corresponding 95% CIs, using correct t-values
		A correction for the use of paired measurements should be applied unless both autocorrelation and clinical circumstances indicate that the measurements are independent
		Check the presence of proportional bias, spread, or both, visually in the Bland–Altman plot and with regression analysis. If present, consider regression analysis to display the bias or LOA as a function of the underlying CO, or data transformation
Interpretation	Reference precision	Determine the repeatability of the experimental and reference technique for correct interpretation of the LOA and mean error
	Response time	Consider changes in CO and differences in response time if one or more continuous techniques disagree; if necessary and appropriate, measurements can be postponed
Data analysis	Trending ability	If applicable, consider the clinical concordance method as an alternative or addition to 4Q and polar analysis

by a visual check of the Bland–Altman plot. The spread of the differences may also be non-uniform over the range of CO measurements. This proportional spread can be identified visually in the Bland–Altman plot as a change in the scatter of the differences. In addition, the absolute values of the residuals as obtained with linear regression can be plotted against the mean CO.[14] If the bias or LOA are non-uniform, transformation of the data or regression analysis can be applied to prevent under- and overestimation in specific measurement ranges;[14 17] however, this limits the interpretation of the study results.

If the bias or LOA are uniform over the range of measurements, the difference between two systems is relatively larger in the lower range in comparison with the higher range. A uniform bias of 0.6 litre min^{-1} represents a 20% mean deviation if CO is 3.0 litre min^{-1}, but a 10% deviation if CO is 6.0 litre min^{-1}. In contrast, if the bias or LOA are non-uniform, this percentage deviation may be constant. A non-uniform bias of 0.3 litre min^{-1} at 3.0 litre min^{-1} and of 0.6 litre min^{-1} at 6.0 litre min^{-1} represents a constant 10% deviation. Measurement error may therefore be constant in an absolute (e.g. 0.3 litre min^{-1}) or relative (e.g. 10%) sense.

Recommendation

Check the presence of proportional bias or spread visually in the Bland–Altman plot and with regression analysis. If present, consider regression analysis to display the bias or LOA as a function of the underlying CO, or data transformation.

Paired measurements

Many studies use multiple measurements in the same subject. Bland–Altman analysis without correction for paired measurements may underestimate the SD of the differences, leading to falsely narrow LOA and confidence intervals (CIs).[5 6 12 14] As illustrated by Hamilton and Lewis,[5] this effect increases with a small number of subjects, large number of measurements per subject, and little within-subject variance in comparison to between-subject variance. This emphasizes the need for correction for paired measurements in studies investigating continuous CO monitoring devices in the absence of major haemodynamic changes. Consecutive measurements will tend to correlate, reducing the within-subject variance. In contrast, major haemodynamic changes may increase the within-subject variance to an extent that measurements become independent.[18] We therefore suggest determining the autocorrelation of repeated measurements first. If this autocorrelation is not negligible, a correction for the use of paired measurements should be applied. Two methods are available for this purpose.[6 14] Bland and Altman[14] provide a method to determine the LOA from the within-subject variances of the experimental and reference methods and the variance of the differences between the within-subject means. Alternatively, Myles and Cui[6] use the average of repeated measurements and use a random effects model to correct for the reduction in variation that occurs by using this average. In addition to these statistical approaches, it is advisable to separate consecutive measurements in time, especially in the absence of major haemodynamic fluctuations. In this way, substantial correlation between consecutive measurements can be prevented.

Recommendation

A correction for the use of paired measurements should be applied unless both autocorrelation and clinical circumstances indicate that the measurements are independent. In the timing of consecutive measurements, the measurement protocol should consider the presence or absence of haemodynamic fluctuations.

Confidence intervals

Investigators should not forget to calculate 95% CIs for the bias, LOA and mean error, because they represent an estimation of their 'true' counterparts in a target population.[7 12] At first sight, bias and LOA in a study may seem clinically acceptable. If, however, the corresponding CIs are wide, considerable differences between two systems can still be present in the target population. To illustrate this, we reconstructed the CIs of the bias, upper and lower LOA, and mean error in a number of studies (Supplementary Appendix A). Considering the CIs in the data analysis would probably lead to different conclusions in some studies. The CI of the bias should not be confused with the LOA.[19] The CI of the bias indicates the limits for the bias in the target population, whereas the LOA refer to the spread of the differences in a specific study. The CI of the bias is calculated as bias$\pm t_{\alpha,n-1}$*SD/$\sqrt{n}$, and decreases with increasing sample size. Being a measure of spread, the LOA do not decrease by increasing the sample size.

Recommendation

The bias, LOA, and mean error should always be accompanied by their 95% CIs.

Agreement

Bland–Altman analysis does not provide definitive answers in terms of *P*-values. The acceptable level of agreement between a new and a reference CO technique is a matter of clinical judgment. A bias of 0.5 litre min^{-1} and LOA of ±1.0 litre min^{-1} may be acceptable for patients undergoing surgery with major haemodynamic disturbances, but not for patients with heart failure undergoing cardiac surgery. Clinically acceptable boundaries for bias and LOA or mean error should therefore always be defined in advance, depending on the target patients in which the new device is aimed to be used.[3] To a certain extent, the desirable level of agreement can be adjusted if the new device has clear advantages over the reference technique in terms of safety, handling in clinical practice, or costs.

Recommendation

Acceptable boundaries for the bias, LOA, and mean error should be defined in advance.

Sample size calculations

The use of predefined criteria for Bland–Altman variables facilitates the decision-making process of accepting or rejecting new CO monitors for clinical use. However, study results may have the tendency to end up close to the predefined criteria, as these criteria reflect the clinical context in which the study has been performed. If the 95% CIs are wide, there is a substantial risk that they include the predefined criteria, which hinders definite conclusions. It is therefore advisable to consider the appropriate sample size in advance. Sample size calculations for Bland–Altman analysis can be considered controversial, because the method is not a statistical test. Moreover, the variability of (repeated) measurements with the new technique is unknown. Despite this, we point to a number of methods to estimate the appropriate sample size. First, the use of a desired maximal width for the 95% CIs around the mean error enables sample size calculations. This method was applied in a previous study

by our group,[20] and is described in Supplementary Appendix B. Similar to this approach, the width of the CIs around the upper and lower LOA can be determined in terms of the SD, as described by Bland.[21] Third, sample sizes can be estimated based on historical data. We realize that these approaches can be debated, and researchers are free to consider their use; however, we advise reflection on this topic in the design phase of method comparison studies in order to reduce the risk of underpowering.

Recommendation

Sample size calculations may be considered to estimate the appropriate number of subjects.

Reference precision

The LOA and mean error are influenced by the precision of the reference technique.[4 22] This is reflected in the formula by Critchley and Critchley[22] to derive the mean error from the precision of the experimental and reference techniques, or:

$$\text{Mean error} = \sqrt{([\text{experimental precision}]^2 + [\text{reference precision}]^2)}$$

The use of imprecise reference techniques will therefore lead to wide LOA and high mean error, independent of the precision of the new device.[8 12] Intermittent thermodilution CO (TDCO) with a pulmonary artery catheter is frequently used as reference technique. In many studies, the precision of TDCO is assumed to be 20%, and experimental precision should not exceed this 20% to be interchangeable with TDCO. Consequently, the mean error should not exceed $\sqrt{(20^2+20^2)}$=28.3%, which is often rounded up to 30%.[22] The strict use of a 30% limit for the mean error will, however, lead to erroneous conclusions if reference precision is significantly smaller or larger than 20%. Precision of TDCO or alternative techniques, such as transpulmonary thermodilution (TPCO), may even be improved to 5%.[1 23–27] Both TDCO and TPCO can therefore be considered valuable as a reference technique, if properly performed. Moreover, this emphasizes the need for evaluation of reference precision in addition to experimental precision. The SD of repeated measurements or 'repeatability' can be used for this purpose.[3 8] Repeatability is defined as 2×SD of repeated measurements (SD_{rep}) divided by CO.[4] The squared values of experimental and reference repeatability can be added up as:

$$\text{Combined repeatability} = \sqrt{([\text{experimental repeatability}]^2 + [\text{reference repeatability}]^2)}$$

This 'combined repeatability' represents the maximal variation in repeated experimental and reference measurements that could explain the mean error. The mean error should therefore not exceed this value for the techniques to be interchangeable.[3 8]

Recommendation

The TDCO and TPCO may be precise reference techniques, if properly performed. Both experimental and reference repeatability should be determined for proper interpretation of the LOA and mean error.

Changes in cardiac output and response time

Changes in CO introduce variability in repeated measurements, irrespective of precision (Fig. 2A). This does not affect the difference between experimental and reference CO if they are observed at exactly the same moment in time (Fig. 2B). In the case of differences in response time between experimental and reference CO, however, a difference between the techniques will appear. This has important consequences for studies evaluating continuous devices during haemodynamic changes. These devices need time to process changes in the underlying CO, in contrast to intermittent reference techniques without measurement delay. Discrepancy will therefore emerge during haemodynamic changes, which fade out in time.[20] The timing and recording of measurements is therefore important, and postponing measurements during acute haemodynamic changes should be considered.[20] In acute settings, however, observations are directly followed by therapeutic decisions. To be valid in this situation, monitoring systems should display short response times.

Recommendation

The response time of (continuous) monitoring systems should be taken into account, and the method of collecting and recording CO measurements should be defined clearly. If necessary and appropriate, measurements can be postponed.

Trending ability

An increasing number of studies focus on the ability to track changes in CO, in addition to determining its absolute value.[9 10] Evaluation of the trend in CO might be helpful to evaluate the effects of interventions and is intuitive, because CO is continuously changing as a result of a variety of influences, such as respiration, the autonomic nervous system, and changes in metabolic demand.[24 25 28] The absolute value of CO is useful to consider in the diagnostic work-up of critical care patients. A proper evaluation of trending ability requires that changes in CO are induced in a controlled set-up. Moreover, the timing of and recording of measurements should be described clearly. Differences in response time between the experimental and reference method should be taken into account, and reference CO should be precise, as described earlier.

Bland–Altman analysis, polar plot methodology, and four-quadrant concordance

Although Bland–Altman analysis evaluates the accuracy and precision of absolute CO readings, conclusions about trending ability may be drawn intuitively. If absolute CO measurements are precise, trending ability should be adequate, irrespective of accuracy. Accuracy refers to the mean deviation between a new CO monitor and true CO. This deviation will be fixed in highly precise monitors and therefore irrelevant in tracking CO changes. In imprecise monitors, the deviation from the underlying CO is variable, which makes trending impossible. Theoretically, precision can be used to determine which changes in CO will be followed reliably. Precision of ΔCO is defined as $\sqrt{2}$ times the precision of a single CO measurement.[4] As a result, a CO measurement device with a precision of 10% can reliably detect changes in CO of >14.1% ($\sqrt{2}$*10). On the contrary, precision of CO measurement needs to be <7.1% (10/$\sqrt{2}$) to detect a ΔCO of 10% reliably.

Two articles by Critchley and colleagues[9 10] review several methods to evaluate trending ability, including four-quadrant (4Q) concordance and polar plot methodology. The 4Q method plots the change in experimental CO (ΔCO_{exp}) against the change in reference CO (ΔCO_{ref}).[9] The percentage of data points in which ΔCO_{exp} and ΔCO_{ref} change in the same direction is called 4Q concordance. This represents a rather crude estimate of trending ability and does not consider the magnitude of ΔCO_{exp} and ΔCO_{ref}.

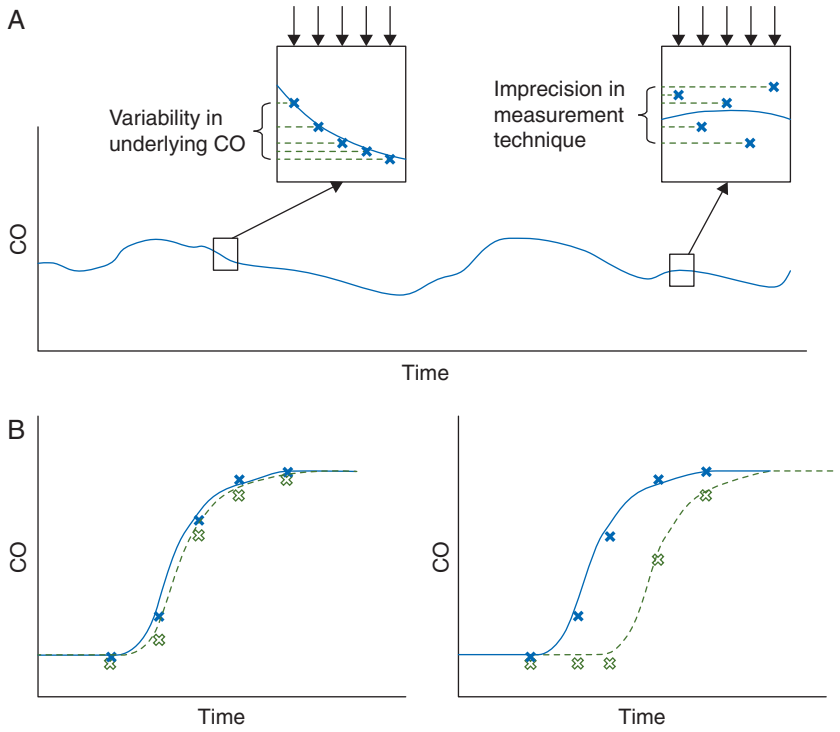

Fig 2 The influence of changes in the underlying CO (A) and difference in response time between the experimental and reference technique (B). (A) Both variability in the underlying CO and imprecision may lead to significant spread in repeated measurements. The left box indicates variability in five consecutive CO measurements by a decrease in underlying CO, whereas imprecision in the measurement technique itself explains the variability in the measurements in the right box. (B) Changes in underlying CO in time (continuous line) do not affect the differences between the experimental (open crosses) and reference (filled crosses) CO measurements if the response time is nearly the same (left panel). If the experimental CO device responds slowly (dotted line, right panel), the differences increase, resulting in wide LOA. CO, cardiac output; LOA, limits of agreement.

In contrast, the polar plot approach enables quantitative assessment of trending ability, which is a major advantage.[9 10] Nonetheless, a number of limitations need to be considered. First, interpretation of the polar variables is not straightforward. The translation of angular bias and radial LOA to clinical practice is not intuitive. Second, the criteria for good trending ability were validated, in a limited number of studies, against concordance and the opinion on trending ability by the authors. As a result, conclusions from polar plot analysis will have the tendency to agree with other statistical methods applied in the past, which limits the added value. Third, the criteria were determined with TDCO as the reference technique. In the case of another reference technique with different precision, the criteria should be adjusted.[1] Fourth, both polar plot and 4Q methods use exclusion zones to limit the influence of small changes in CO that may introduce random noise; however, this reduces statistical power and ignores potentially valuable information. The combination of small increases in ΔCO_{exp} (e.g. +1%) with small decreases in ΔCO_{ref} (e.g. −1%) or vice versa may be considered good trending, because these changes are both insignificant and unlikely to trigger therapeutic actions. In 4Q and polar analysis, these data pairs are excluded.

'Clinical concordance'

Alternatively, it is possible to pass a clinical judgment on each individual data pair. Each combination of ΔCO_{exp} and ΔCO_{ref} is designated as 'good' or 'poor' trending and depicted in a 'clinical concordance' plot (Fig. 3A). The designations are based on criteria from a clinical perspective. Changes in CO_{ref} in a patient are divided into the following categories:

non-significant change (ΔCO_{ref} ±5% or less);
moderate increase or decrease (ΔCO_{ref} ±5–15%); or
large increase or decrease (ΔCO_{ref} ±15% or more).

Each corresponding ΔCO_{exp} is assigned good trending if ΔCO_{exp} changes in the same direction and falls into the same category as ΔCO_{ref}. Depending on the clinical context, the number of categories and their criteria can be adjusted. 'Clinical concordance' can be defined simply as the percentage of 'good trending' assignments. The percentage 'poor trending' assignments directly informs the clinician about the risk for clinically relevant, erroneous trending information. Moreover, comparing the categories into which ΔCO_{exp} and ΔCO_{ref} fall provides insight into the extent to which ΔCO_{exp} and ΔCO_{ref} (dis)agree. In analogy with error grids used to validate new glucose measurement devices, this (dis)agreement can be further divided from the perspective of therapeutic consequences.[29] An error grid can be created to reflect the therapeutic consequences in specific zones in the concordance plot (Fig. 3B). The following zones can be distinguished.

(i) ΔCO_{exp} and ΔCO_{ref} change in the same direction and to the same extent, reflecting the following situations (in analogy with 'clinical concordance'): (a) CO_{exp} and CO_{ref} change

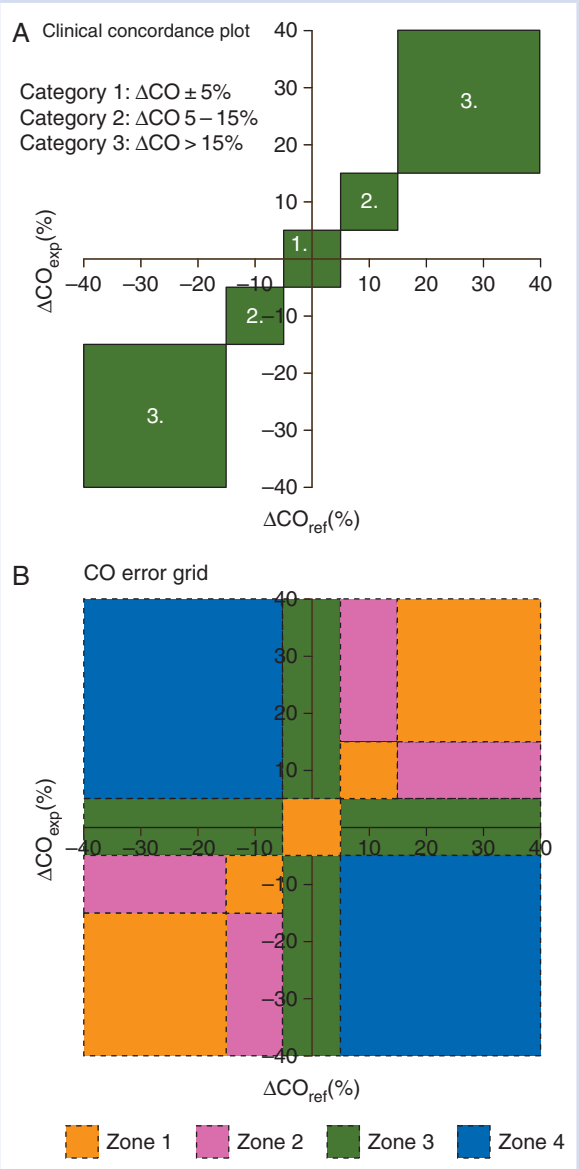

Fig 3 Clinical concordance and error grid plots. (A) Clinical concordance defines three categories (green squares), in which trending is 'good' from a clinical perspective. Clinical concordance represents the percentage of ΔCO_{exp} and ΔCO_{ref} data pairs falling into these categories. (B) The corresponding error grid uses multiple zones (rectangles in different shades of green) to define the level of agreement between ΔCO_{exp} and ΔCO_{ref} data pairs from the perspective of therapeutic consequences. The zones are based on the clinical concordance categories, and can be created by extending the lines that surround the clinical concordance squares. Zone 1 corresponds to the clinical concordance categories in which CO_{exp} and CO_{ref} change in the same direction and to the same extent. This results in correct treatment decisions. In Zone 2, CO_{exp} and CO_{ref} change in the same direction but not to the same extent, reflecting insufficient or exaggerated treatment. In Zone 3, CO_{exp} changes while CO_{ref} is constant or vice versa, reflecting unnecessary or withheld treatment. Zone 4 represents opposite changes in CO_{exp} and CO_{ref}, resulting in opposite treatment. ΔCO_{exp}, change in experimental cardiac output; CO_{ref}, change in reference cardiac output.

<5%; (b) CO_{exp} and CO_{ref} change between 5 and 15%; or (c) CO_{exp} and CO_{ref} change >15%. In this zone, correct treatment decisions are made with the new technique.

(ii) ΔCO_{exp} and ΔCO_{ref} change in the same direction but not to the same extent, reflecting the following situations: (a) CO_{exp} changes between 5 and 15% while CO_{ref} changes >15%; or (b) CO_{exp} changes >15% while CO_{ref} changes between 5 and 15%. In this zone, treatment may be insufficient (a) or exaggerated (b).

(iii) ΔCO_{exp} changes while ΔCO_{ref} is constant or vice versa, reflecting the following situations: (a) CO_{exp} changes >5% while CO_{ref} is constant; or (b) CO_{exp} is constant while CO_{ref} changes >5%. In this zone, unnecessary treatment may be initiated (a) or necessary treatment may be withheld (b).

(iv) ΔCO_{exp} and ΔCO_{ref} change in opposite directions, reflecting the following situations: (a) CO_{exp} increases >5% while CO_{ref} decreases >5%; or (b) CO_{exp} decreases >5% while CO_{ref} increases >5%. In this zone, opposite treatment may be initiated.

The clinical concordance method provides a crude measure of trending agreement (clinical concordance) in combination with the therapeutic consequences of trending disagreements (error grid). A worked example is provided in Supplementary Appendix C. The suggested method uses all data pairs in the data analysis, which is an important advantage. Moreover, the extent to which ΔCO_{exp} and ΔCO_{ref} agree is addressed from a clinical perspective, which enhances the interpretation and use in clinical decision-making. The definitions for the clinical concordance categories and zones in the error grid are, however, rather subjective, and the use of different definitions might hinder comparison between studies in the future. Additional research is therefore needed to validate this new approach against current methods for trending analysis. Clinical concordance and error grids are meant as an extension to current methods, such as 4Q concordance and polar plot methodology, not as a substitute.

Recommendation

The clinical concordance method should be considered as an alternative or addition to 4Q and polar analysis in the evaluation of trending ability.

Discussion

The present review article describes the methodological challenges with the application of Bland–Altman and trending analysis in CO method comparison research. Moreover, the concept of clinical concordance and a corresponding error grid method is introduced to evaluate trending ability from a clinical perspective. Based on the items discussed, a stepwise approach to the design and data analysis of CO method comparison research can be created (Table 1). This approach may serve as a checklist for new researchers in the field. In addition, it may help clinicians to interpret the results from these studies in their decisions to incorporate new CO monitoring techniques in daily practice.

Although this review focuses on Bland–Altman and trending analysis, the data analysis of method comparison studies should not be restricted to these statistical methods. As in any type of research, the data analysis should include a close look at the raw data, considering outliers, haemodynamic circumstances, and patient characteristics. The scatterplot depicting experimental against reference CO should be evaluated, together with the range of CO measurements, effects in regions with high- or low-CO states, and effects in subgroups of patients. This is important because the performance of CO monitors may differ considerably depending on (patho)physiological

conditions in the patient.[1 2] Moreover, method comparison research represents only the initial part of the validation process of new CO monitors.[30] Besides technical efficacy, the ultimate goal of any newly developed monitor is to improve patient outcome and to be cost-effective. Method comparison studies should therefore anticipate application in clinical practice. This was an important reason to point to the use of predefined criteria defined within a desired, future clinical context. In addition, clinical concordance was introduced as a clinically intuitive method for trending ability, in which the level of (dis) agreement is translated to therapeutic consequences. Researchers should challenge themselves to embed the clinical context into their studies, both for a better understanding of the results and in order to facilitate the implementation of new technology in daily care.

Authors' contributions

All authors have been involved in drafting the manuscript, have given final approval of the version to be published, and agree to be accountable for all aspects of the work.

Supplementary material

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

Declaration of interest

W.F.B. has received honoraria for lectures and was consultant for Edwards Lifesciences. W.F.B. has no non-financial competing interests. The other authors declare that they have no competing interests.

Funding

Departmental resources.

References

1. Peyton PJ, Chong SW. Minimally invasive measurement of cardiac output during surgery and critical care: a meta-analysis of accuracy and precision. *Anesthesiology* 2010; **113**: 1220–35
2. De Waal EE, Wappler F, Buhre WF. Cardiac output monitoring. *Curr Opin Anaesthesiol* 2009; **22**: 71–7
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. Cecconi M, Rhodes A, Poloniecki J, Della Rocca G. 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**: 201–6
5. 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
6. Myles PS, Cui J. Using the Bland–Altman method to measure agreement with repeated measures. *Br J Anaesth* 2007; **99**: 309–11
7. 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
8. 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
9. 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
10. Critchley LA, Yang XX, Lee A. Assessment of trending ability of cardiac output monitors by polar plot methodology. *J Cardiothorac Vasc Anesth* 2011; **25**: 536–46
11. Saugel B, Grothe O, Wagner JY. Tracking changes in cardiac output: statistical considerations on the 4-quadrant plot and the polar plot methodology. *Anesth Analg* 2015; **121**: 514–24
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–10
13. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–7
14. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–60
15. O' Brien E, Petrie J, Littler W, *et al.* The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *J Hypertens* 1993; **11**: S43–62
16. Razali N, Way YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Analyt* 2011; **2**: 21–33
17. Ludbrook J. Confidence in Altman–Bland plots: a critical review of the method of differences. *Clin Exp Pharmacol Physiol* 2010; **37**: 143–9
18. Jansen JR, Schreuder JJ, Settels JJ, *et al.* Single injection thermodilution. *Anesthesiology* 1996; **85**: 481–90
19. Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *Br J Anaesth* 2001; **90**: 514–6
20. Montenij LJ, Buhre WF, De Jong SA, *et al.* Arterial pressure waveform analysis versus thermodilution cardiac output measurement during open abdominal aortic aneurysm repair: a prospective, observational study. *Eur J Anaesthesiol* 2015; **32**: 13–9
21. Bland M. How can I decide the sample size for a study of agreement between two methods of measurement? Available from http://www-users.york.ac.uk/~mb55/meas/sizemeth.htm (accessed 13 June 2015)
22. Critchley LA, Critchley JA. A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *J Clin Monit Comput* 1999; **15**: 85–91
23. Jansen JR, Vesprille A. Improvement of cardiac output estimation by the thermodilution method during mechanical ventilation. *Intensive Care Med* 1986; **12**: 71–9
24. Nishikawa T, Dohi S. Errors in the measurement of cardiac output by thermodilution. *Can J Anaesth* 1993; **40**: 142–53
25. Jansen JR, Schreuder JJ, Settels JJ, Kloek JJ, Versprille A. An adequate strategy for the thermodilution technique in patients during mechanical ventilation. *Intensive Care Med* 1990; **16**: 422–5
26. Monnet X, Persichini R, Ktari M, *et al.* Precision of the transpulmonary thermodilution measurements. *Crit Care* 2011; **15**: R204

27. Gondos T, Marjanek Z, Kisvarga Z, *et al*. Precision of transpulmonary thermodilution: how many measurements are necessary? *Eur J Anaesthesiol* 2009; **26**: 508–12
28. Michard F, Teboul JL. Using heart-lung interactions to assess fluid responsiveness during mechanical ventilation. *Crit Care* 2000; **4**: 282–9
29. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 1987; **10**: 622–8
30. Pearl WS. A hierarchical outcome approach to test assessment. *Ann Emerg Med* 1999; **33**: 77–84

Handling editor: J. G. Hardman