Chapter 1

Statistics Used to Assess Monitors and Monitoring Applications

Lester A. H. Critchley

Introduction

An evidence-based approach now prevails when recommending medical treatments. This applies as much to the latest therapies as to appropriate methods to monitor patients and their response to treatment. For an evidence-based approach to be successful, however, it must be based on good-quality clinical data from well-conducted research. The quality of clinical studies and their data is now graded according to the level of evidence they provide,[1] and guidelines exist on how to properly conduct clinical research. Cochrane reviews have set standards for best evidence. Working groups such as the National Institute for Clinical Excellence (NICE) and Resuscitation Council (UK) demonstrate how such an approach can be transformed into up-to-date guidelines and courses. When assessing the value of emerging clinical monitoring technologies for perioperative, emergency room, and critical care use, researchers should be aware that clinical validation studies must be of a sufficient standard to be of use in evidence-based reviews. This perspective drives the approach of this chapter, with a focus on cardiac output (CO) monitoring, since most of the literature on these statistical methods has arisen from analysis of this variable.

Cardiac Output Measurement

Cardiac output is the sum of stroke volumes expelled from the heart over one minute; it can be measured from either the pulmonary or the systemic circulations. As the arterial system leaving the heart branches, it is not possible to measure total CO at a distal point such as the arm or descending aorta, and corrections are needed (e.g., arterial pulse contour analysis and esophageal Doppler).[2,3] Measurement of CO at its source, the heart, is also difficult to achieve in the clinical setting because of restricted access, unless one is performing open-heart surgery. Instead, at-a-distance (e.g., transthoracic Doppler) or surrogate (e.g., bioimpedance) methods are utilized, which result in lack of precision.[4,5] Compared to measuring other more accessible hemodynamic variables such as blood pressure or heart rate, lack of accuracy and precision has hampered the development of routine CO monitoring in the clinical setting.[5]

Cardiac output can be measured accurately using techniques such as the Fick method and radionuclide imaging studies. These methods, however, have several limitations. They are only applicable in settings such as the physiology laboratory or radiology department; they are inapplicable at the point of care and therefore cannot be used in operating room, emergency medicine, or critical care settings. Furthermore, Fick and radionuclide studies only provide single readings, and there is need for technologies that measure CO on a frequent or continuous basis. The clinical significance of being able to assess changes or trends in CO is only now being recognized, and this is highlighted by the designs of recently marketed CO devices and the statistical approaches to their validation.

All validation studies require a reliable reference method against which comparisons are made. For CO monitoring, the accepted reference method has been and remains single bolus thermodilution using a pulmonary artery Swan–Ganz catheter. The pulmonary artery catheter, however, is now seldom used in clinical practice, and its use is associated with significant risk to patients.[6,7] Clinical validation studies incorporating pulmonary artery catheter measurements are mostly restricted to cardiac surgery and liver transplant. Some recent research studies have used the less invasive transpulmonary thermodilution method, which is employed in the PiCCO (Pulsion, Munich, Germany) and VolumeView (Edwards Lifesciences, Irvine, CA, USA) systems. Errors arise in thermodilution measurement because of injectate

and dead space issues,[8] and the degree of inaccuracy varies between clinical settings and different manufactured devices.[9] The precision of the thermodilution method is generally accepted to be ± 20%,[10,11] and this margin of error has played a significant role in the ongoing development of validation statistics.

Cardiac output is not a static variable; its value constantly changes. Achieving a steady state in which simultaneous comparative readings can be taken often proves difficult, and this hampers the collection of good-quality validation data.

Protocol Design and Data Collection

The need for ethical approval and patient consent is an obvious prerequisite for publication. Poorly planned data collection and inadequate sample size will limit the usefulness of collected data and thus the ability to publish the study findings. Common mistakes are (i) failure to blind investigators to comparative readings, (ii) failure to achieve simultaneous readings during steady-state hemodynamics, (iii) failure to have sufficient range of readings, (iv) failure to collect sufficient data resulting in inconclusive results, (v) inconsistent number and timing of repeated measurements from individuals (i.e., irregular data collection), and (vi) failure to collect serial data pairs that show adequate changes and hence fail to facilitate trend analysis. A well-designed study has clearly defined times of data collection, which are of sufficient number to allow comprehensive analysis.[12]

Sample size is difficult to calculate in this type of research, even if a pilot study is performed, because of the range of different variables and outcomes involved. A more pragmatic approach may be based on reviewing the sample sizes used in previous studies that were successful in detecting effects. Comparative studies with cohorts of over 30 patients and 6 or more serial data pairs are recommended.[12]

Background to Validation

Thirty years ago scatter plots and regression and correlation analyses were the principal analytical methods used to show how reliably a new measurement method compared to a reference standard.[13] Regression and correlation, however, only evaluate the degree of association between two measurement methods; they do not quantify accuracy. Quoting correlation coefficients and p values confirms little.

The whole approach to validation statistics changed in the 1980s when J. M. Bland and D. G. Altman introduced a new method of comparing measurements based on bias, the difference between pairs of comparative readings.[14] Bias was plotted against the average of each pair, and the standard deviation of the bias provided a statistic called *limits of agreement* (i.e., 95% confidence intervals for the bias). Bland and Altman, however, never provided guidance as to how the limits of agreement should be used to confirm clinical utility, leaving this to the discretion of the user. This was particularly unsatisfactory when Bland–Altman analysis was applied to CO studies where the reference method, usually thermodilution, was imprecise. Limits of agreement of less than 1 liter/min were considered to be acceptable,[15,16] but no provision for (i) variations in baseline CO or (ii) imprecision of the reference method was made.

To enable outcomes from Bland–Altman style CO studies to be compared in 1999, Critchley proposed the use of percentage error (PE), a statistic calculated from the limits of agreement (i.e., 95% confidence interval of the bias) divided by the baseline CO for the study.[10] A benchmark for acceptance of a new technique of less than 28.4% was set, which was rounded up to less than 30%. This benchmark was based on a reference method's precision of 20% and acceptance of the test method also being set at 20%. Although PE has been criticized over the years for being too strict,[11,17,18] its simplicity and robustness as an analytical tool have withstood the test of time.

In more recent decades, following advances in clinical medicine and monitoring technology, it has become increasingly important to have bedside monitors that accurately follow the vital signs of hospitalized patients. Unfortunately, Bland–Altman analysis does not assess the ability of devices to detect changes; it is limited to assessing accuracy of readings and agreement between methods.[19,20] Thus, new statistical approaches were developed, referred to as *trend analysis*.[21] Many researchers new to clinical monitoring, however, fail to recognize the need to show trending and restrict data collection to that suitable for Bland–Altman analysis.

How to effectively address the issue of trending capability has not been fully resolved in the literature. In a recent review of CO studies, Critchley and colleagues reported that only 20% of the studies performed some form of trend analysis; the analytical methods employed were (i) Bland–Altman analysis

of tables and histograms, (ii) regression analysis of scatter plots, and (iii) analysis of direction of change.[21]

When analyzing CO data from hospital patients, commonly used trend analysis methods are (i) concordance on a four-quadrant plot and (ii) polar plot analysis.[22,23] Both these analyses rely on comparing serial data from reference and test methods, calculating the serial change in consecutive readings (ΔCO), and excluding data where the change is small (i.e., < 10–15% change). The polar method involves transforming the data from a simple (x, y) Cartesian format to a radial format (radius, angle). Polar plots provide greater information about the agreement between two methods that is lost when just direction of change is used. Criteria for acceptable trending have been proposed for CO monitoring.[21,23] A more detailed description of these methods follows.

Bland–Altman Analysis

Practically all CO validation studies published today use Bland–Altman analysis and provide a Bland–Altman plot (Figure 1.1). The plot shows bias collected from the whole or subgroups of the study. Each plot should display horizontal lines indicating mean bias and the 95% confidence intervals or limits of agreement. Inspection of the plot allows one to assess (i) the distribution or spread of data, (ii) the degree of agreement between methods (i.e., size of the limits of agreement), and (iii) any systematic changes in bias as CO increases (i.e., offsets in calibration). One common problem with presenting Bland–Altman plots is using inappropriate scales, especially when more than one plot is shown. Rather than choosing scales that fill the page with data points, the axis of each plot should have similar scales and ranges. Otherwise visual comparisons between plots are difficult to perform. Very often the Bland–Altman plot is accompanied by an (x, y) scatter plot that shows the raw data (Figure 1.1), but regression lines and correlation coefficients are often omitted.

Bland–Altman analysis requires each data pair to be independent of all other pairs and ideally from separate subjects.[14] If data pairs are related (i.e., they come from the same subject), the size of 95% confidence intervals and limits of agreement for the analysis will be reduced. Use of repeated measures (i.e., data pairs from the same subject) is common in CO studies; thus, the data analysis should correct for

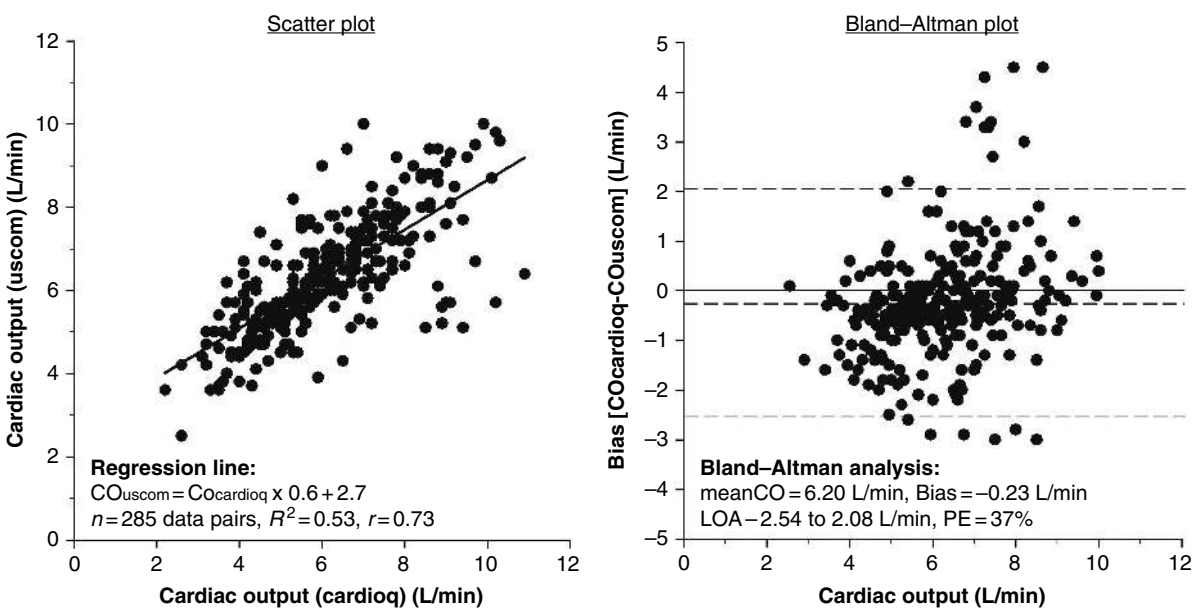

Figure 1.1 Scatter plot with regression line and accompanying Bland–Altman plot. Statistical analysis data are added to each plot. The Bland–Altman plot also displays the mean bias and limits of agreement of the analysis (dashed horizontal lines). Data are from a study that compared two Doppler CO measurement methods, transthoracic (USCOM) and esophageal (CardioQ).

Source: Huang L, Critchley LA. An assessment of two Doppler-based monitors to track cardiac output changes in anaesthetized patients undergoing major surgery. *Anaesth Intens Care* 2014;42:631–9. LOA: limits of agreement, PE: percentage error.

repeated measures by either (i) Bland and Altman or (ii) Myles and Cui methods, which differ slightly in complexity.[24,25] Statistical software programs that perform Bland–Altman analysis should also adjust for repeated measures; journal editors and reviewers expect that authors will employ such corrections and describe them in their manuscripts.

Percentage error is a key outcome statistic arising from CO studies that perform a Bland–Altman analysis.[10] It is used to compare findings of CO studies with findings of other published studies. It also allows criteria to be set for acceptance of a new CO monitor prior to starting a study. Most authors will use the less than 30% benchmark, from Critchley's 1999 paper that based the criteria on a 20% precision for thermodilution CO measurement and the need for less than 20% measurement error (i.e., 95% confidence intervals or precision).[11] A 20% error represented up to a 1 liter/min variation in CO if the mean CO was 5 liter/min.

Cecconi and colleagues have questioned the logic of assuming a 20% error in the reference method.[26] They recommended measuring its precision and using the error to set new acceptance criteria a priori. Their rational was that (i) the error in thermodilution or other reference method is very variable and 20% is just an approximation, and (ii) any significant variation from 20% would result in lesser or greater errors in the test method to be accepted, if the acceptance criteria are set at the standard 30%. Their approach to measuring the reference method's precision was to perform serial steady-state measurements from which the coefficient of variation was calculated and precision derived.[26]

Trend Analysis

Trending capability, the ability to follow changes in CO, can be assessed either by (i) multiple paired comparisons in a small number of subjects (i.e., $n = 6–10$ laboratory animals) or (ii) as part of a larger scale clinical trial with up to 8–10 comparative measurements in 20 or more patients. Statistical approaches are different for the two settings. Small cohort studies are dealt with later in the section Time Plots and Regression Analysis.

Concordance Analysis

For larger cohort clinical trials the current approach is concordance analysis using direction of change.[21,22] This analysis is based on serial data, and ΔCO is the study variable calculated from the difference between consecutive readings. Direction of change in CO can either be increased (i.e., positive direction change) or decreased (i.e., negative direction change); the magnitude of change is not included in the analysis. In the trial a test method is compared to a reference method, which provides pairs of directions of change of readings that can either agree (i.e., concord) or disagree. Concordance is measured as the proportion of readings that agree.

To make concordance analysis easier to visualize, a four-quadrant plot is drawn of ΔCO reference against ΔCO test (Figure 1.2). Data where directions of change agree fall into the right upper and left lower quadrants. The ratio of the number of data pairs where directions of change agree over the total number of data pairs for the study provides the concordance presented as a percentage.

Data pairs where the serial change in CO is small, however, can often have directions of change that disagree due to random errors in measurement; this is referred to as *statistical noise*. To eliminate statistical noise from the concordance analysis an exclusion zone is used that removes data where the change in CO is less than 10–15% of the mean CO for the study. The setting of limits for the exclusion zones is based on a receiver operator characteristic (ROC) curve analysis.[22]

Current advice for acceptable trending ability in CO studies is greater than 92%.[21] Ideally, confidence limits should be calculated for the concordance, which is based on sample size. The ΔCO data is treated as a binomial (i.e., direction of change either agrees or disagrees), and the standard deviation of the concordance ratio (p) is $\sqrt{[np(1-p)]}$, where n is the number of data points. A good example of how this statistic is generated and used is found in Axiak-Flammer and colleagues.[27]

Polar Plots

The introduction of polar plots (Figure 1.2) was to address the problems that (i) the four-quadrant plot method did not include magnitude of change and (ii) all data pairs were treated equally despite size.[14,21,23] By converting the data to (i) a radial distance that represented the size of the combined changes in CO from the two paired readings (i.e., average absolute change in ΔCO) and (ii) an angle that represented the degree of agreement (i.e., the greater the degree of disagreement the larger the angle), more information about the comparison between the two measurement

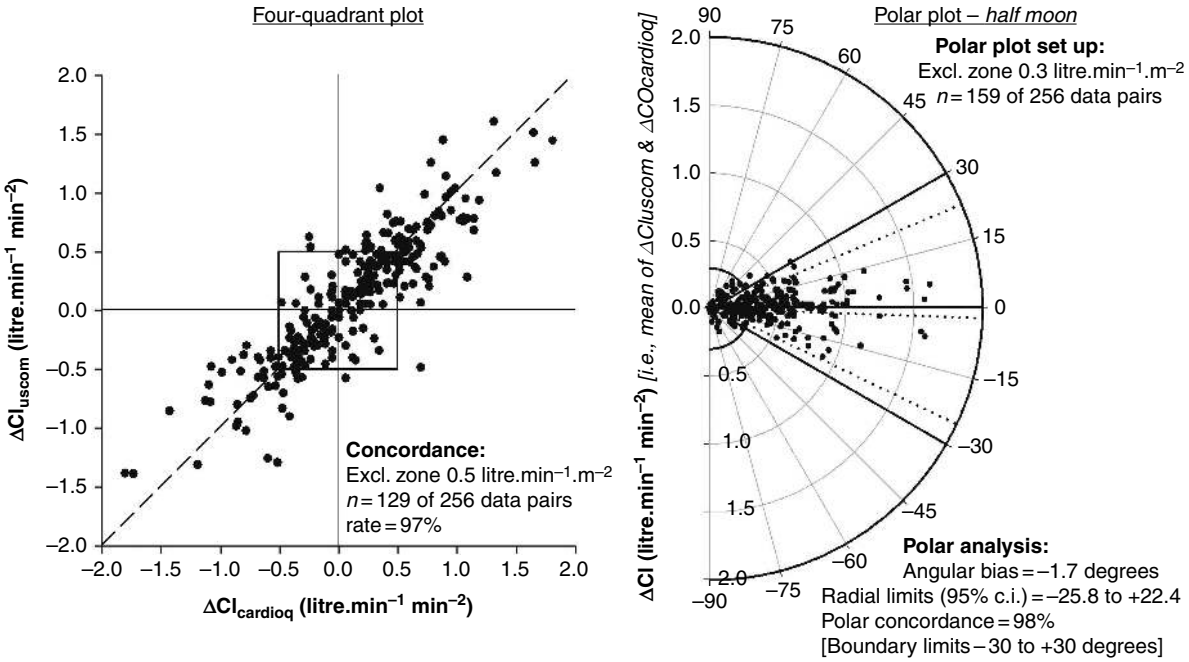

Figure 1.2 Four-quadrant and polar plots showing changes (ΔCO). Four-quadrant plot has zero axes crossing at its center, creating four zones. A central exclusion zone (square) is shown: Data lying within this zone are excluded because they contain a high level of random variation compared to changes in CO (i.e., statistical noise). The line of identity $y = x$ (dashed line) also is shown. Ideally, all data points should lie along this line. Data that lie within the upper right and lower left quadrants agree (i.e., direction of change agrees). Results of concordance analysis are printed in the plot. Polar plot is of a semicircle, or half-moon, design in which both positive and negative changes in CO are shown together. A central exclusion zone also is shown (half circle). Zero- and 30-degree axes are highlighted (solid lines). Mean polar angle and 95% radial limits of agreement for the polar analysis also are shown (dotted lines). Polar concordance rate is based on the proportion of data points that lie within 30 degrees of the polar axis (zero degrees). Results of the polar analysis are shown.

Source: Huang L, Critchley LA. An assessment of two Doppler-based monitors to track cardiac output changes in anaesthetized patients undergoing major surgery. *Anaesth Intens Care* 2014;42:631–9.

methods was retained. The concept of excluding data pairs where changes in CO were small and statistical noise may corrupt the analysis was also applied. However, the exclusion zone was reduced from 15% to 10% of the mean CO for the study because the combined change in ΔCO on the polar diagram (i.e., radial length) was derived from the average of the two ΔCO values, whereas in the four-quadrant plot the combined change was derived from the hypotenuse of a triangle produced by test and reference values and was √2 (or 1.42) times larger in size. The mean angle for all the data pairs provided a measure of misalignment in calibration or offset between methods. Empirically, a limit of ± 5% was set as the criterion for an acceptable offset. The radial limits of agreement were set at ± 30% and were based on a 2:1 ratio in size between ΔCO readings. These limits, however, were not based on sound statistical theory. To make the polar plot more visually friendly, one can rotate negative change data through 180 degrees to become a positive change, thus producing a half-moon rather than full-moon plot. Generating polar data from Cartesian (x, y) ΔCO data and drawing polar plots can be technically challenging. Some of the newer statistical programs now provide polar plot drawing and analysis software. Guidance can also be found in the original paper describing polar plots.[23] The polar method is probably best reserved for research groups performing high-quality validation studies. Mastering the technique of polar plots provides a greater appreciation of the data and trending ability.

Time Plots and Regression Analysis

Understanding the structure of one's data is the key to knowing which statistical methods are most appropriate. Data arising from validation studies can be considered as a two-dimensional matrix of paired readings representing subjects in one plane and serial

measurements from individual subjects in the other plane. Bland–Altman analysis is most appropriate when there are many subjects and few, if any, serial measurements, because the primary attribute being tested is the accuracy of a measurement technique as it is applied to a study sample. In studies where trending capability is being analyzed, multiple serial measurement pairs (n = 10 or more) are needed. For this type of study design, data can be analyzed on an individual subject basis (i.e., within subject) using regression analysis. Huang and colleagues performed a number of clinical studies comparing Doppler CO with bioimpedance CO methods during anesthesia for major surgery.[3,28] Their surgical model provided a range of ever-changing CO values. They plotted within subject serial changes in CO over time, for each monitoring modality and for each patient (n = 7 to 27 data points). They were able to visually identify divergences in the trend lines for CO between the different monitoring modalities and relate them to interventions during the surgery (Figure 1.3). They also used regression analysis as a method of quantifying the degree trending between the monitoring modalities for each subject. For CO studies using Doppler methods as the reference, they were able to set criteria when trending capability of the test bioimpedance method was considered acceptable. However, when regression analysis was applied to group data, the systematic differences in calibration between subjects introduced a second source of variation, and trending capability could no longer be easily evaluated using the correlation coefficients.

Reporting Validation Study Data

Since 1999 there have been concerns in the literature regarding how validation study data have been reported, especially for studies using Bland–Altman analysis.[29–32] As recently as 2016, Abu-Arafeh and colleagues published a review of 111 papers from a two-year period, which concluded that Bland–Altman study data were poorly reported and of limited usefulness to evidence-based reviews.[33] Additionally, they proposed a list of 13 key issues to be included in reports and called for journals to provide more guidance on how Bland–Altman studies should be conducted and reported. In 2010, Critchley and colleagues reported similar findings in relation to reporting trend analysis data.[21] Based on the present author's experience as a researcher and journal

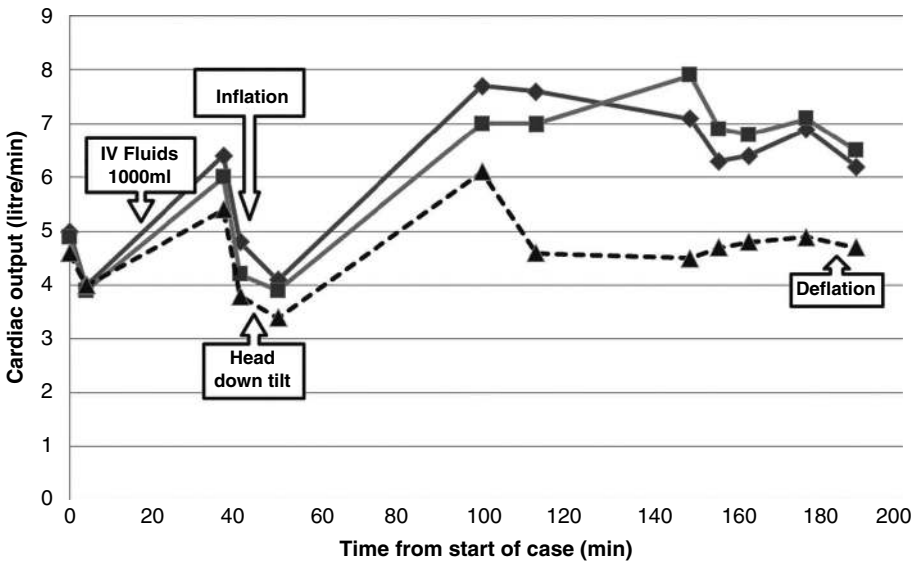

Figure 1.3 Time plot of comparative CO data collected during a laparoscopic surgical procedure. The two uppermost trend lines represent reference Doppler readings (diamond-USCOM and square-CardioQ). The lower dashed line represents a new bioimpedance device (triangles). Note how after 100 minutes there was a definite downward divergence of bioimpedance readings relative to the Doppler readings. Major interventions such as inflation and deflation of the pneumoperitoneum are shown. The precise cause of the downward divergence was unclear yet present in a number of other cases. Detection of this type of change in monitor readings is very important to developers yet does not show up in more classical group statistics using Bland–Altman and concordance analyses.

reviewer in the field of validation studies, the following recommendations are provided:

1. Provide a clear and thorough description of the study design, including (i) recruitment, (ii) number of subjects, (iii) how readings were taken, including blinding of investigators and steady-state synchronous readings, and (iv) timing of data collection points. Remember to mention ethical approval and consent.
2. Provide a well-described plan for analyzing the data in the methods section. Ideally, one should measure the precision of the reference method and use it when setting a priori criteria for acceptance of the test method.[26] A typical sequence for a simple test versus reference comparison study would be (i) results of any pilot studies such as reference method precision and power calculation (i.e., study size), (ii) inspection of study data using scatter plots, (iii) Bland–Altman analysis with details, and (iv) trend analysis using concordance and possibly polar plots. Acceptance criteria with references should be added to the relevant subsections.
3. The results section should start with the general demographics of the study population, including number of subjects and how many subjects were excluded and why. The power calculations justifying the size of the study, if performed, could be included at this point (see previous comments on study size).
4. Draw a scatter plot (optional) that shows the distribution of raw data (Figure 1.1). Multiple plots may be needed if subgroups of subjects have been included in the study design. Addition of a regression line and correlation coefficients is optional, as Bland–Altman recommended their exclusion.[14] Plots should contain, within the diagram or legend, essential information such as number of data points and relevant statistical outcomes, for example regression line equation and correlation coefficient (i.e., r or R^2).
5. Draw the Bland–Altman plot(s) (Figure 1.1). Make sure axes are appropriately scaled with sensible data ranges. If more than one plot is presented, the scales and ranges should be similar to facilitate visual comparison. Add horizontal lines for the mean bias and limits of agreement (i.e., 95% confidence intervals of the bias). Make sure the limits have been corrected for repeated measures, citing which methodology was used.[24,25] Some authorities are now asking for confidence intervals of the limits of agreement to also be included.[34] It is best to stick to simple numerical measurement units (i.e., CO in liter/min) rather than percentage changes; however, indexing variables to body surface area (BSA) (i.e., cardiac index = CO/BSA) is acceptable. Diagrams and legends should display essential numerical information about the plot(s).
6. Sufficient data to calculate the PE should be provided, including (i) the standard deviation of the bias or 95% confidence interval and (ii) mean CO for all the study data. Ideally, the PE should also be presented. The PE facilitates comparison of data with previous studies; one may wish to make such comparisons in the discussion section. In the methods section the criterion threshold should be set a priori that defines a PE that supports acceptance of the new technique. This requires some consideration regarding the precision of the reference method. Cecconi and colleagues recommend estimating the precision from coefficient of variation measurements for the reference method.[26] The current benchmark for PE is less than 30%, but this criterion should be set in the context of the precision of the reference method as a 20% error is presumed (see Axiak-Flammer et al. for guidance if an alternative reference method has been used[27]).
7. Depending on the study design and data structure, if a trend analysis is performed, then a four-quadrant plot should be drawn (Figure 1.2). Concordance analysis should be performed for studies with grouped data of sufficient numbers (e.g., $n > 20$ subjects) and serial data pairs (e.g., $n > 3$). An exclusion zone should be employed (i.e., 15% of mean CO for the study) to remove data where changes are small and data points lie close to zero. For CO studies the zone is set at 15% of the mean CO value. Remember that concordance is based on the variable ΔCO, *not* CO. Criteria for accepting a CO monitor as having good trending ability have been set at greater than 92%, where the reference method was single bolus thermodilution.[21] For studies with a small number of data pairs, the confidence intervals for the concordance also need to be calculated.[27]
8. A polar plot analysis may also be employed (Figure 1.2), following advice on generating the data from paired readings, creating the plots, and

interpreting the results.[11,21,23] Exclude central zone data that are less than 10% of the mean CO for the study. Key outcome data are the mean angle and 95% radial limits of agreement. They should be added to the polar plot as radial lines. The 30-degree radial axes should also be highlighted. Negative direction data points can be rotated through half a turn (i.e., 180 degrees), but not reflected, to provide a half-moon plot. Ideally, the main data outcomes, including number of data points, exclusion zone size, mean angle, and radial limits of agreement, should be added to the diagram or legend. Polar plots demonstrate (i) offsets in calibration between methods (i.e., mean angle of greater than 5 degrees) and (ii) the level of agreement between the methods (i.e., tightness of alignment of radial data points to the zero-degree axis or mean angle line). The 30-degree lines act as guides to good trending when 95% of data points fall within their boundaries.

9. For less commonly used methods of assessing trending, one should refer to the papers that describe them.

Noncardiac Output Studies

The application of validation statistics is not limited to just CO monitoring data. They can also be applied to blood pressure, oxygen saturation, and hemoglobin level monitoring. The main difference is the criteria used to determine acceptance thresholds and exclusion zones, because of their reliance on the precision of the reference method.

References

1. Oxford Centre for Evidence-based Medicine – Levels of Evidence (March 2009). www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/ (Accessed November 2017).
2. Sun JX, Reisner AT, Saeed M, Heldt T, Mark RG. The cardiac output from blood pressure algorithms trial. *Crit Care Med* 2009;**37**:72–80.
3. Huang L, Critchley LA. An assessment of two Doppler-based monitors to track cardiac output changes in anaesthetised patients undergoing major surgery. *Anaesth Intens Care* 2014;**42**:631–9.
4. Chong SW, Peyton PJ. A meta-analysis of the accuracy and precision of the ultrasonic cardiac output monitor (USCOM). *Anaesthesia* 2012;**67**:1266–71.
5. Wang DJ, Gottlieb SS. Impedance cardiography: more questions than answers. *Curr Cardiol Rep* 2006;**8**:180–6.
6. Koo KK, Sun JC, Zhou Q, et al. Pulmonary artery catheters: evolving rates and reasons for use. *Crit Care Med* 2011;**39**:1613–8.
7. Harvey S, Harrison DA, Singer M, et al. Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomized controlled trial. *Lancet* 2005;**366**:472–7.
8. Reuter DA, Huang C, Edrich T, Shernan SK, Eltzschig HK. Cardiac output monitoring using indicator dilution techniques: basics, limits, and perspectives. *Anesth Analg* 2010;**110**:799–811.
9. Yang XX, Critchley LA, Rowlands DK, Fang Z, Huang L. Systematic error of cardiac output measured by bolus thermodilution with a pulmonary artery catheter compared with that measured by an aortic flow probe in a pig model. *J Cardiothorac Vasc Anesth* 2013;**27**:1133–9.
10. 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.
11. Critchley LA. Bias and precision statistics: should we still adhere to the 30% benchmark for cardiac output monitor validation studies? *Anesthesiology* 2011;**114**:1245.
12. Biancofiore G, Critchley LA, Lee A, et al. Evaluation of an uncalibrated arterial pulse contour cardiac output monitoring system in cirrhotic patients undergoing liver surgery. *Brit J Anaesth* 2009;**102**:47–54.
13. Fuller HD. The validity of cardiac output measurement by thoracic impedance: a meta-analysis. *Clin Invest Med* 1992;**15**:103–12.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307–10.
15. LaMantia KR, O'Connor T, Barash PG. Comparing methods of measurement: an alternative approach. *Anesthesiology* 1990;**72**:781–3.
16. Wong DH, Tremper KK, Stemmer EA, et al. Noninvasive cardiac output: simultaneous comparison of two different methods with thermodilution. *Anesthesiology* 1990;**72**:784–92.
17. Michard F. Thinking outside the (cardiac output) box. *Crit Care Med* 2012;**40**:1361–2.
18. Peyton PJ, Chong SW. Minimally invasive measurement of cardiac output during surgery and critical care: a metaanalysis of accuracy and precision. *Anesthesiology* 2010;**113**:1220–35.
19. Critchley LA. Validation of the MostCare pulse contour cardiac output monitor: beyond the Bland and Altman plot. *Anesth Analg* 2011;**113**:1292–4.

20. Critchley LA. Meta-analyses of Bland-Altman-style cardiac output validation studies: good, but do they provide answers to all our questions? *Brit J Anaesth* 2017;**118**:296–7.
21. Critchley LA, Lee A, Ho AM. A critical review of the ability of continuous cardiac output monitors to measure trends in cardiac output. *Anesth Analg* 2010;**111**:1180–92.
22. Perrino AC, O'Connor T, Luther M. Transtracheal Doppler cardiac output monitoring: comparison to thermodilution during noncardiac surgery. *Anesth Analg* 1994;**78**:1060–6.
23. 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.
24. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat* 2007;**17**:571–82.
25. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth* 2007;**99**:309–11.
26. 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.
27. Axiak-Flammer SM, Critchley LA, Weber A, et al. Reliability of lithium dilution cardiac output in anaesthetized sheep. *Brit J Anaesth* 2013;**111**:833–9.
28. Huang L, Critchley LA, Zhang J. Major upper abdominal surgery alters the calibration of BioReactance cardiac output readings, the NICOM, when comparisons are made against suprasternal and esophageal Doppler intraoperatively. *Anesth Analg* 2015;**121**:936–45.
29. 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.
30. 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. *Clin Chem* 2002;**48**:799–801.
31. 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.
32. Bein B, Renner J, Scholz J, Tonner PH. Comparing different methods of cardiac output determination: a call for consensus. *Eur J Anaesthesiol* 2006;**23**:710.
33. Abu-Arafeh A, Jordan H, Drummond G. Reporting of method comparison studies: a review of advice, an assessment of current practice, and specific suggestions for future reports. *Br J Anaesth* 2016;**117**:569–75.
34. Drummond GB. Limits of agreement with confidence intervals are necessary to assess comparability of measurement devices. *Anesth Analg* 2017;**125**:1075.