BJA

British Journal of Anaesthesia, 2015, 1–3

doi: 10.1093/bja/aev356
Editorial

EDITORIAL

Disagreement between cardiac output measurement devices: which device is the gold standard?

Y. Le Manach[1,*] and G. S. Collins[2]

[1]Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada, and
[2]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, UK

*Corresponding author. E-mail: gary.collins@csm.ox.ac.uk

A common research question in perioperative haemodynamics research concerns the assessment of whether a new measurement device can replace an existing device (often referred to as method comparison studies). Typically, a new measurement method is being compared with an established reference method (unfortunately often referred to as the '*gold standard*'). In a recent issue of the journal, Biais and colleagues[1] reported the comparison of two cardiac output measurement devices, one based on pulse wave transit time (i.e. the new devices) and the other one based on transthoracic echocardiography (i.e. the reference method '*gold standard*'). The study concluded that devices were not interchangeable and that the new device cannot guide haemodynamic interventions in critically ill patients. Their conclusion was based on observing percentage errors exceeding the limits of 30%, suggested by Critchley and Critchley.[2]

Disagreement is not necessarily an error

In order to understand method comparison studies, several terms need to be clearly defined to avoid misinterpretation when evaluating and comparing a new device to the reference device. Accuracy refers to the closeness of a measurement from the new device to the reference device (often referred to as bias), whereas precision refers to the reproducibility or repeatability of a set of measurements. A perfect device would produce both accurate and precise measurements. However as true cardiac output is unknown, evaluating accuracy tends to reflect the accuracy of a new device against the reference method, which may or may not be accurate. In contrast, evaluating precision does not require the true cardiac output value and can be adequately evaluated in the majority of the studies. Percentage error refers to the observed differences in measurements obtained from the two devices being compared, but under the assumption of a well-calibrated reference device. However, an important issue for cardiac output measurement is that observed differences can be attributed to either the new device, or the reference device or indeed both, as some degree of measurement error of cardiac output is likely in either device. Consequently, the term 'agreement' is probably more appropriate as the intention is to compare two devices both affected by errors in their measurement of the true cardiac output.

In an attempt to move away from the misleading correlation coefficient, Bland and Altman highlighted the importance of correctly assessing measurement agreement and introduced their popular graphical method (known as the Bland-Altman plot).[3] The Bland-Altman method involves plotting the difference (*x*-axis) between measurements from the two devices for each individual against their average (*y*-axis). Any relationship between the measurement error and an estimate of the true value (average of the two measurements) can therefore be examined. If the difference between the two devices is sufficiently small not to cause problems in clinical interpretation (within the 95% limits of agreement), then the new device can be considered either as a candidate to replace the current device or be used interchangeably. However, the main limitation of this method is that the observed disagreement (i.e. the difference between the two measurements) is assumed to be in the new device. The consequence is the method could be conservative and likely to reject any new device when the reference device is not precisely and accurately measuring the true value. An alternative to the Bland-Altman method that attempted to quantify acceptable limits of agreement between two measurement devices has been proposed.[2] Using an error-gram, Critchley and colleagues derived the relationship between the accuracy of reference device and

the limits of agreement between the two devices. They recommended that limits of agreement of up to 30% be accepted, and Biais and colleagues referred to this approach to help in the interpretation of their results. However, the 30% limits of agreement are only appropriate if the precision of the new device is within 20% of the reference device. As highlighted by Critchley, the choice of 30% was based on clinical judgement from more than 15 yr ago,[4] and it would be interesting to re-examine this, and observe whether this remains today, or whether more (or less) precise measurements can be used. In most of the studies using averaged thermodilutions as the reference device, there is a greater risk of rejecting the new device because the '*reference*' method was not as accurate as it should have been (e.g. not averaged thermodilution, alternative method used as reference). Consequently, a fair evaluation of a new device should also include an evaluation of the reference device, particularly when the true value the device is attempting to capture cannot be measured, such as cardiac output. The evaluation should include repeated measurements on individuals, as this enables an assessment of agreement between the two devices and an evaluation of each device has with itself to assess repeatability.[5]

In order to illustrate the impact of the reference device characteristics on the evaluation of a new device, we carried out a set of simulations, in which we compared an almost perfect device (precision 4%, perfect accuracy), using a reference device with different levels of precision in 150 pairs of measures (Fig. 1). These simulations demonstrate that according to the level of precision of the reference device, a perfect device can be rejected according to the 30% limits of agreement method proposed by Critchley.[2] Consequently, the choice of the reference method plays a major role in the evaluation of a new device. To retain an imprecise reference device would result in rejecting any precise new devices, and would eventually accept similarly imprecise devices.

The limit of acceptable agreement

Studies comparing cardiac output devices generally do not prespecify a clinically meaningful limit of acceptable agreement before any analyses are conducted.[6] Clinical interpretation of device comparison studies is therefore often done *post hoc*, based on the observed results, with no reference as to any acceptable limit of agreement, and therefore inconsistently. Clearly this approach is flawed, because the width of the 95% limit of agreements confidence interval is influenced by the sample size. Many studies evaluating cardiac output measurement devices are conducted on small sample sizes (e.g. less than 100 pairs of independent measurements).[6] Small sample sizes naturally have a considerable impact on the robustness of the conclusions about the possible interchangeability of the devices. Estimating

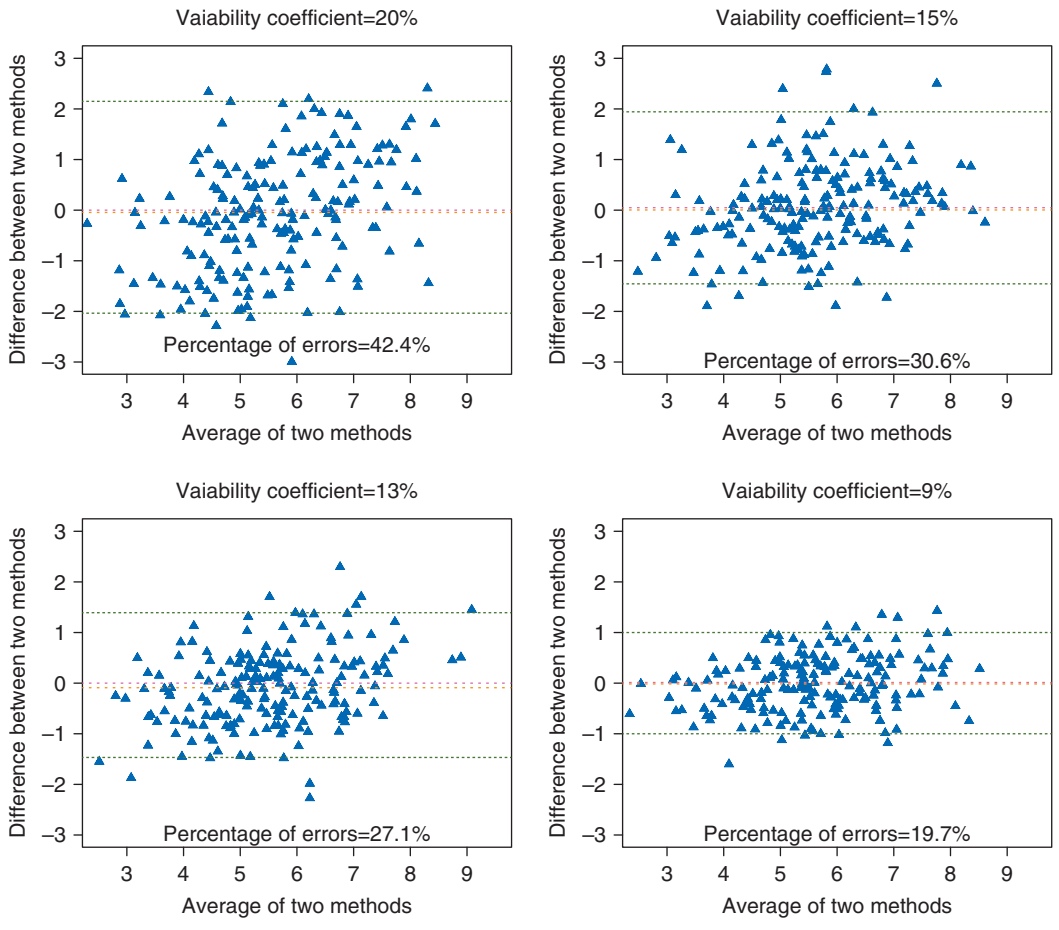

Fig 1 Impact of evaluating a perfect cardiac output device according to the precision of the reference method. Simulation conducted on 150 pairs of measurements.

the width of confidence interval for the 95% limit of agreement should be calculated, before collecting any data and used to determine the number of patients needed to get a clear and meaningful assessment of the device comparison.

Missing data

Missing data is ubiquitous in medical research. In cardiac output device evaluation, the existence of any missing values is rarely reported, including whether there are technical concerns regarding the measurement capability of the device. However, most of the missing values observed in cardiac output measurement device evaluation, relate to the inability of one (or both) device to produce a cardiac output value. The ability of the device to produce a measurement is crucial, if the device is to be considered interchangeable with the current measurement device. Devices able to produce unbiased measurement in 10% of the patients and no measurement in the 90% other patients, should clearly not be described as a perfect device.

The statistical methods used to describe the agreement between two devices do not account for this major consideration. Consequently, it is useful to report these unsuccessful measurements, possibly using a flow diagram (similar to the CONSORT flow diagram), describing which devices fails (and frequency) to estimate cardiac output and give reasons.

Measurement agreement studies are abundant during the perioperative period. The vast majority of the studies focus on the agreement between a new and reference device. Systematic reviews have shown these studies are frequently poorly conducted and often fail to report key information, to allow readers to adequately judge whether there is sufficient evidence, to suggest whether a new device can replace or be used interchangeably with the reference device.[6 7] The choice, and the description, of the reference devices, including an assessment of its repeatability is an important aspect to aid interpretation. An inappropriate reference device may falsely lead to rejecting a new device by rejecting a new device, which is easy to use, accurate and safe, based on the result of a flawed comparison with an inaccurate or invasive reference device. The consequence of this is the absence of any device in clinical practice, because the new device is regarded as imprecise and the reference one as too invasive. However, device measurement comparison studies are a necessary step in the evaluation of a new device; the goal of perioperative measurements remains the detection of a clinically abnormal condition requiring a treatment. We urge investigators to design their study by conducting appropriate sample size calculations, pre-specifying clinically important limits of agreements and conducting a detailed examination of both measurement devices, including an assessment of repeatability and reliability.

Declaration of interest

None declared.

References

1. Biais M, Berthezène R, Petit L, Cottenceau V, Sztark F. Ability of esCCO to track changes in cardiac output. *Br J Anaesth* 2015; **115**: 403–10
2. Critchley LAH, Critchley J. A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *J Clin Monit Comput* 1999; **15**: 85–91
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **327**: 307–10
4. Critchley LAH. Bias and Precision Statistics: Should We Still Adhere to the 30% Benchmark for Cardiac Output Monitor Validation Studies? *Anesthesiology* 2011; **114**: 1243–55
5. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–60
6. 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
7. Dewitte K, Fierens C, Stockly D, Thienpoint 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