Veterinary
Anaesthesia and Analgesia

Formerly the Journal of Veterinary Anaesthesia

Veterinary Anaesthesia and Analgesia, 2014, **41**, 227–232 doi:10.1111/vaa.12143

PERSPECTIVE

Issues and pitfalls in method comparison studies

Sonja Hartnack

Section of Epidemiology, Vetsuisse Faculty, University of Zurich, Zurich, Switzerland

Correspondence: Sonja Hartnack, Section of Epidemiology, Vetsuisse Faculty, University of Zurich, Winterthurerstr. 270. 8057 Zurich, Switzerland. E-mail: Sonja.Hartnack@access.uzh.ch

Abstract

Method comparison studies are needed for validation of new methods of measurements, for example, non-invasive blood pressure measurements against standard reference methods. After a brief introduction into method comparison studies, this paper is organized in three sections. The first section deals with the widely, though not always appropriately, used classical Bland-Altman plot with the limits of agreement and its extensions with non-constant bias and multiple observations. The second section comments on other statistical approaches including correlation coefficients, linear regressions and sensitivities and specificities which are sometimes seen in method comparison studies. The third section proposes the usage of linear mixed effects models as a flexible way to deal with questions associated with method comparison studies.

Keywords agreement between two measurements, Bland-Altman plot, measure agreement with repeated measures, method comparison study.

Introduction

Accurate, precise and reliable measurements are indispensable in veterinary medicine. Although these three terms are not always used consistently, many authors agree that ‘accurate’ refers to measurements being close to the true value of interest – an assessment which might be difficult to make in the absence of a perfect gold standard. ‘Precise’ means how close or similar repeated measurements of the same entity are, thus indicating the degree of variability of a measuring method. Closely related terms to the concept of precision are repeatability and reproducibility. The term ‘reliable’ is often used to describe the capacity of a method to differentiate between healthy and affected individuals.

New methods of measurements are being continuously developed and need to be validated in method comparison studies. In the last years numerous studies validating indirect blood pressure measurements in veterinary anaesthesia have appeared. Issues related to validating new methods of measurements had already been dealt with in the nineteenth century by pioneers in statistics like Pearson and Fisher. The Pearson correlation coefficient is still, albeit inappropriately, used in method comparison studies. Typical questions in method comparison studies are related to the assessment of (dis-)agreement between two measuring methods, and determining an acceptable level of disagreement. The latter is clearly a clinical and not a statistical question and should be agreed upon before analysing the data of a method comparison study. Complete agreement between two different methods measuring on a continuous scale is not possible as errors are inherent to each measurement.

The results of two methods of continuous measurements might differ because one method might in general always yield lower or higher values. This is called bias or (mean) difference. This bias might be constant over the whole range of values measured or it might vary. For example it could increase with higher measured values. Results from different methods of measurements might also differ because one method is more variable or less precise than the other one. Additionally, one method may be influenced more than the other by variations in another factor (a covariate). For example, one method might be more sensitive to variations in temperature or different cuff positions for indirect blood pressure measurements.

Classically, in method comparison studies, one part of the study has been focussed on the assessment of agreement by using samples that are measured once individually by each method, resulting in an assessment of bias. Another part of the study – most often independently analysed from the first part- focussed on reproducibility/repeatability or precision of parameters by using replicates measured by the same method. Coefficients of variation (CV) have been used for this purpose.

Depending on the aim of the method comparison study different descriptive and analytical tools for method comparison studies are available. Different aims could be: describing a bias, deciding if two methods are interchangeable and one method could replace the other, providing formulae to convert the measurement results of one method into the other method, and assessing the importance of (external) factors affecting bias and/or variability.

This paper is organized in three sections. The first deals with the widely, though not always appropriately used, classical Bland-Altman plot with the limits of agreement and its extensions with non-constant bias and multiple observations. The second section comments on other statistical approaches which are sometimes seen in method comparison studies. The third section proposes the usage of linear and linear mixed effects models as an elegant state-of-the-art method.

Bland Altman plots

In the classical Bland Altman plot, the differences between single measurements of the same sample by two methods (y-axis) are plotted against the mean of the two measurements (x-axis) (Bland & Altman 1986, 1995). This graphical approach allows assessment of bias between the two methods of measurement, thus indicating a systematic difference between the two methods. Furthermore, Bland Altman plots might show if this bias is constant (parallel to the x-axis) or non-constant (a straight line drawn through the data points would have a slope different from 0). Bland Altman plots may also indicate if the variability of the differences is constant or varies across the range of values measured. A non-constant bias as well as a non-constant variability would both indicate that the two methods do not agree equally through the range of measurements (Bland & Altman 1995).

Bland and Altman also developed the so-called 'limits of agreement' (LOA). The upper and lower line of these limits are generated by the mean difference between the two methods of measurement (bias) and ± 1.96 * the standard deviation (SD). These limits will cover 95% of the differences between the measurements or for 95% of the samples measured by the two methods; the difference will be between these two lines. Since these limits are derived from the data at hand, considering that agreement was 'excellent, if 95% or more of the values were within the 95% LOA' (Williams et al. 2012) is a misunderstanding of this concept. The decision if the agreement between two methods is good, excellent or unsatisfactory is a clinical question and not a statistical one. The LOA could also be used to predict the results of one method by the results of the other by adding the bias and determining the 95% interval.

The application of the classical Bland Altman plot and LOA is only useful if the underlying assumptions are fulfilled. These include constant bias, constant variability, normality of the differences and independence of the values, thus using single pairwise measurements and not replicates. In the case of a non-constant bias, an extension to the classical Bland Altman plots and derived LOAs is available (Bland & Altman 1999; Carstensen 2010b). This extension, accommodating for a non-constant bias, consists of regressing the differences between both methods on the averages and use intercept, slope and residual standard deviation to determine a prediction interval (Carstensen 2010b). Log transformation of the measurements before determining LOAs and back-transformation of these to obtain limits for the ratio of the measurements has also been proposed (Bland & Altman 1986, 1999) in this case. Log transformations have also been proposed to deal appropriately with non-constant variability.

Nevertheless, some papers use the classical Bland Altman plot and derived LOAs even if a non-constant bias is evident from the Bland Altman plot or even a trend line is present (Seliskar et al. 2013). However, in this situation – due to their dependency on the averages between both methods – LOAs cannot be used directly for prediction intervals (Carstensen 2010b).

The normality assumption of the differences might in principle be alleviated by log transformations, using Bland Altman plots and assessing normality