SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–22

DOI: 10.1177/0962280215601133
smm.sagepub.com
SAGE

Assessing agreement between two measurement systems: An alternative to the limits of agreement approach

Nathaniel T Stevens, Stefan H Steiner and R Jock MacKay

Abstract

The comparison of two measurement systems is important in medical and other contexts. A common goal is to decide if a new measurement system agrees suitably with an existing one, and hence whether the two can be used interchangeably. Various methods for assessing interchangeability are available, the most popular being the limits of agreement approach due to Bland and Altman. In this article, we review the challenges of this technique and propose a model-based framework for comparing measurement systems that overcomes those challenges. The proposal is based on a simple metric, the probability of agreement, and a corresponding plot which can be used to summarize the agreement between two measurement systems. We also make recommendations for a study design that facilitates accurate and precise estimation of the probability of agreement.

Keywords

Bias, homoscedasticity, interchangeability, measurement error, probability of agreement, repeatability

1 Introduction

Accurate and precise measurements in medical and other contexts are of paramount importance. However, accuracy and precision may come at a cost; an accurate and precise measurement system—defined here to be the devices, people, and protocol used to make a measurement—may be costly in terms of time, money, resources, or may be invasive. In this case, new measurement systems that are less expensive, less time-consuming, less labor-intensive, or less-invasive may be developed. Interest often lies in comparing a new measurement system to an existing one. To do so, we perform a measurement system comparison (MSC) study.

The goal of this comparison may vary in emphasis by context. Dunn[1] highlights four possible goals that can be described as follows: (i) calibration problems, which deal with establishing a

Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Science, University of Waterloo, Canada

Corresponding author:
Nathaniel T Stevens, Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W, M3 4123 Waterloo, ON N2L 3G1, Canada.
Email: nstevens@uwaterloo.ca

relationship between a new system and an existing one that can be used to appropriately adjust the new system's measurements; (ii) comparison problems, which deal with assessing the level of agreement between two measurement systems whose measurements are on the same scale; (iii) conversion problems, which deal with the comparison of two systems whose measurements are on different scales; and (iv) gold-standard comparison problems, which deal with the comparison of a new measurement system with a system that is known to make measurements without error. In addition to these, a fifth goal, which we term superiority, may be to determine whether the new system is in fact better, in terms of accuracy and/or precision, than the existing one.

Each of these goals is important; they are appropriate in different contexts, and at times appropriate within the same context. In the present article, we focus mainly on problem (ii). That is, we wish to quantify the level of agreement between two measurement systems and hence determine whether the agreement is sufficient for the two systems to be used interchangeably. Thus, for this article "comparing measurement systems" is synonymous with "assessing agreement". Secondary to this, we will also demonstrate how the proposed methodology may be used to address calibration, conversion, gold-standard comparison problems, and superiority.

This choice of emphasis is driven largely by the literature. Bland and Altman[2,3] provide a method of assessing agreement between measurement systems, called the limits of agreement approach, that has been cited over 30,000 times. This citation record suggests that the statistical evaluation of agreement is a common goal. The importance of assessing agreement is also evident in various regulations set forth by the US Food and Drug Administration. For example, the FDA[4] mandates that agreement be formally assessed in the context of bioequivalence studies. As well, the FDA[5] commands the use of the limits of agreement approach when demonstrating substantial equivalence between a premarket measurement device and an existing one. Similarly, the assessment of agreement is recommended by the Mayo Clinic[6] when validating assays and by the Clinical and Laboratory Standards Institute[7] when comparing measurement procedures. Furthermore, when reporting the results of a method comparison study some academic journals, for example the *Annals of Clinical Biochemistry*,[8] require that a limits of agreement analysis be included. Other journals, *Clinical Chemistry*[9] for example, strongly recommend its inclusion. What is proposed here is a more transparent and informative alternative to this approach.

In a typical MSC study, some characteristic—the measurand—of a number of subjects is measured one or more times by both measurement systems. We denote the number of subjects by n, and we use r to denote the number of measurements on each subject by each system. For now we assume that r is the same for both systems and all subjects, but we discuss relaxing this assumption in section 5.

We adopt the common convention of describing data of this form with the following linear mixed effects structural model that relates the measurements by two systems[10]

$$\begin{aligned} Y_{i1k} &= S_i + M_{i1k} \\ Y_{i2k} &= \alpha + \beta S_i + M_{i2k} \end{aligned} \tag{1}$$

where $i = 1, 2, \ldots, n$ indexes the subjects, $j = 1$ corresponds to the reference measurement system, $j = 2$ the new measurement system and $k = 1, 2, \ldots, r$ indexes the replicate measurements. Thus Y_{ijk} is a random variable which represents the value observed on system j's k^{th} measurement of subject i. S_i is a random variable that represents the unknown true value of the measurand for subject i. In model (1) we assume that $S_i \sim N(\mu, \sigma_s^2)$ and that subjects are sampled randomly from the target population, but we discuss relaxing these assumptions in section 5. M_{ijk} is a random variable which represents the measurement error of system $j = 1, 2$. We further assume that the M_{ijk} are independent of each other and independent of S_i, and that they are distributed $N\left(0, \sigma_j^2\right)$ where σ_j quantifies the measurement variation, or repeatability, of system j. In this article we assume that σ_j is constant across true values and hence each measurement system is homoscedastic. We briefly discuss the heteroscedastic case in section 5.

The parameters $\alpha,$ and β quantify the bias of the new measurement system relative to the reference system. We refer to α as the fixed bias since it increases or decreases the average measurement of the second system by a fixed amount and we call β the proportional bias because it biases the second system's measurements by an amount that is proportional to the true value.[11] It would of practical interest to estimate the absolute bias of each system; i.e. include an α_j and β_j for both systems. However, because both measurement systems are prone to error, the true values of the measurand are unknown and so we cannot estimate the absolute bias of the measurement systems; we can only estimate relative bias. If, however, the reference measurement system is a gold-standard, the relative biases α and β can be interpreted as the absolute bias of the new system. We discuss this point further in section 3.4.

Based on equation (1), we say that the two measurement systems are identical if $\alpha = 0$, $\beta = 1$ and $\sigma_1 = \sigma_2$. However, the two systems do not need to be identical to be used interchangeably. Informally we say that two systems can be used interchangeably if, most of the time, their measurements are similar. In other words, two measurement systems agree and could be used interchangeably if $Y_{i2} - Y_{i1}$, the difference between single measurements on a given subject by each system, is small. Typically this happens when $\alpha \approx 0$, $\beta \approx 1$, and both σ_1 and σ_2 are small, relative to σ_s. We formalize the notion of interchangeability in sections 2 and 3.

Note that this formulation of the problem assumes that the measurements by both systems are on the same scale. If, however, the measurements by the two systems are on different scales, i.e. degrees Celsius versus degrees Fahrenheit, then relative bias is confounded with the conversion between scales. In this situation, two measurement systems may be interchangeable even if $\alpha \neq 0$ and $\beta \neq 1$. This is discussed further in section 3.4.

A variety of statistical techniques exist for assessing agreement between two measurement systems. Excellent reviews of existing techniques are provided by Choudhary and Nagaraja,[12] Barnhart et al.,[13] Lin,[14] and Carstensen.[4] As mentioned previously, the most widely cited technique is the limits of agreement approach due to Bland and Altman.[2,3] In section 2, we describe this approach and identify several problems associated with it that can lead to misinformed judgments of interchangeability. In section 3, we introduce a novel analysis method which facilitates a better understanding of the relationship between the two measurement systems. We illustrate this new method with two examples from the literature. In section 4 we provide recommendations for the design of an MSC study, and we end with a summary and discussion in section 5.

2 The limits of agreement technique

2.1 Description

The "limits of agreement" approach is the most widely used technique for assessing interchangeability of measurement systems. It was first introduced by Bland and Altman in 1983,[2] but the wide uptake did not begin until the publication of Bland and Altman's second paper on the topic which appeared in the *Lancet* in 1986.[3] This latter article has been cited over 30,000 times and is one of the ten most frequently cited statistical articles ever.[15]

To describe this technique, suppose we have one measurement by each system on each of n subjects. The limits of agreement approach characterizes the agreement between two measurement systems by evaluating the difference between measurements made on the same subject. Using a scatter plot, known as a "difference plot", the observed differences for subject i, $d_i = y_{i2} - y_{i1}$, are plotted against the observed averages: $a_i = (y_{i1} + y_{i2})/2$.

One purpose of this plot is to evaluate whether the differences are related to the averages, a surrogate for the unknown true values. If no relationship appears to exist, the distribution of the differences is summarized by the limits of agreement, defined as

$$\bar{d} \pm 1.96 s_d \tag{2}$$

where, for a sample of $i = 1, 2, \ldots, n$ subjects, $\bar{d}$ and s_d are respectively the sample average and standard deviation of the observed differences. Assuming the differences roughly follow a normal distribution, these limits represent the interval within which we expect 95% of the differences to lie. Horizontal reference lines corresponding to the upper and lower limits of agreement and the average difference $\bar{d}$, are added to the plot.

To decide whether two measurement systems agree sufficiently to be used interchangeably, one must compare the limits of agreement to the clinically acceptable difference (CAD). Bland and Altman[2,3] define the CAD to be the maximum allowable difference between two measurements that would not adversely affect clinical decisions. How far apart two measurements can be before it causes difficulties is not a statistical question; instead the answer must be based on clinical judgment. In many situations the CAD is defined as an interval around zero: $(-c, c)$.

In what follows, we refer to the upper and lower limits of agreement as ULA and LLA, respectively. If the limits of agreement are contained within the CAD, i.e. $-c \leq LLA < 0 < ULA \leq c$, one concludes that the differences will be clinically acceptable at least 95% of the time, and the measurement systems are deemed interchangeable. Otherwise, if the limits of agreement fall outside the CAD, it is likely that measurements by the two systems will too often differ by more than the allowable amount. In this situation, one concludes that the two measurement systems do not agree sufficiently and should not be used interchangeably.

Since the introduction of the limits of agreement approach, Bland and Altman have authored many articles which clarify the method, and guide its use in nonstandard situations. For example, they suggest alternate methods of calculating limits of agreement if the differences appear to depend in some way on the true value,[16] or if replicate measurements are made.[17] However, whether in the simple or more complex cases, problems can still arise and investigators can be misled. We describe these problems in the next section.

2.2 Problems with the limits of agreement approach

Though the limits of agreement method is simple to implement, its simplicity can also be its downfall. Because no model is assumed the relationship between measurement systems is oversimplified which inhibits informed comparisons. In this section we demonstrate problems that are inherent to the approach or that arise as a result of misuse.

A serious problem associated with misuse is that although Bland and Altman recommend measuring each subject two or more times by each measurement system, replicate measurements are rarely made in practice.[1,18] This could, in part, be because the example presented in their landmark *Lancet* paper ignores the fact that each system made two measurements on each subject, and uses only the first measurement on each subject to compare the two systems. Replicate measurements are ignored in examples in a subsequent paper as well.[16]

To fully understand the relationship between the two measurement systems, and hence to decide if they are interchangeable, it is important to model their relationship as in equation (1) and estimate all of the corresponding parameters. Without replicate measurements we cannot separately estimate all of the parameters in equation (1), a limitation which Barnett[13] refers to as the *problem of identifiability* and that Voelkel and Siskowski[18] refer to as the *problem of indeterminacy*. This issue arises because there are six parameters to estimate, but without replicate measurements the data provide only five minimally sufficient statistics. A consequence is that without separate estimates of α and β, we cannot distinguish between fixed and proportional bias, and so the biases become confounded. As well, without separate estimates of the two repeatabilities, σ_1 and σ_2, we cannot determine which system is more precise, and we risk rejecting a new measurement system which is more precise than the existing one.

Bland and Altman[2,16] oppose the use of such structural models and instead use the difference plot, as described above, to visualize the relationship between two measurement systems. However, this plot cannot disentangle confounding biases, it does not indicate which system is more precise and hence it does not provide adequate information about this relationship. Without the additional information gained by replicate measurements, the difference plot can be misleading.

To illustrate the effect of not explicitly estimating and comparing σ_1 and σ_2, consider the comparison of two measurement systems when the new system is unbiased ($\alpha = 0$ and $\beta = 1$). In this situation the standard deviation of the differences is $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2}$, which is estimated by s_d, defined in equation (2). Because $\alpha = 0$ and $\beta = 1$, these systems should agree on average, but acceptable agreement may be hindered by large variability in one or both systems. For example, when σ_1 is large but σ_2 is small, σ_d and hence s_d might still be large enough to push the limits of agreement outside the CAD, leading one to reject interchangeability. It is true that agreement should be small in this case, but when this happens an unsuspecting practitioner unknowingly rejects a new measurement system which is more precise than the existing one, even though both are unbiased. Bland and Altman[3,16] acknowledge that in using their technique this problem is a possibility. However, we feel that this is a potentially serious problem that practitioners should avoid.

Another problem inherent to the technique is one which we call *false correlation*. As stated by Bland and Altman[2,3] one purpose of the difference plot is to detect whether there is a relationship between the differences and the averages (which are a surrogate for the unknown true values of the measurand). By using the averages on the horizontal axis, we are supposedly protected against the appearance of a pattern when no real relationship between differences and true values exists, i.e. when there is no proportional bias ($\beta = 1$). To investigate this issue we consider the correlation between differences $D = Y_2 - Y_1$ and averages $A = (Y_1 + Y_2)/2$ for a particular subject when $\beta = 1$

$$Corr(D, A) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)\left(4\sigma_s^2 + \sigma_1^2 + \sigma_2^2\right)}} \tag{3}$$

If $\sigma_1 = \sigma_2$ then D and A are uncorrelated. But if the repeatabilities are not equal, a more realistic assumption, the differences and averages are correlated. It is interesting to point out that Bland and Altman[19] initially acknowledge that σ_1 and σ_2 may not be equivalent, and hence this correlation may be nonzero, but they suggest in a subsequent paper that the correlation in equation (3) should be zero because the variability of each measurement system should be the same: "as they should if they are measurements of the same thing" (p. 91).[20] However, just because both systems are measuring the same thing, does not imply that the repeatabilities should be the same.

A serious issue arises here. In the absence of an actual relationship between differences and true values, the Bland and Altman difference plot can suggest a significant relationship exists. As well, the presence of a false negative correlation could mask the existence of a true positive relationship, and vice versa. Thus the existence of a false correlation can confuse the relationship between two measurement systems and may lead to misinformed judgments of interchangeability. That said, false correlation can be identified and accounted for if replicate measurements are taken and the individual variance components in equation (1) are separately estimated.

Because the limits of agreement approach can be misleading in the absence of replicate measurements, we do not recommend its use in this case. In fact, we do not recommend the comparison of measurement systems at all, if replicate measurements are not available.

Bland and Altman[5,6] describe extensions to the limits of agreement technique when replicate measurements are available. They recommend averaging the replicate measurements on a single subject by a particular measurement system, and constructing the difference plot using the differences and averages of the averaged measurements on each subject. By doing this the limits of agreement as defined in equation (2) are too narrow and so the calculation of s_d is adjusted to account for the reduction in measurement variation that results from working with the average of replicate measurements instead of individual measurements. Although this results in limits that more accurately reflect the distribution of differences in single measurements, the approach is not without difficulties.

First, by plotting averages of the replicate measurements, a transparent display of the raw data is unavailable. A plot of the averages can mask large differences in the replicate measurements on the same subject by each system, and can make the level of agreement between the two measurement systems appear stronger than it truly is. A second issue is that Bland and Altman's method of calculating the limits of agreement in this situation assumes that "the difference between the two methods is reasonably stable across the range of measurements" (p. 572).[17] In other words, this technique assumes there is no proportional bias ($\beta = 1$), and so its applicability is limited. A third problem is that although replicate measurements are made, there is no explicit comparison of repeatabilities, i.e. σ_1 and σ_2 in equation (1), and so it is still possible to reject interchangeability with a more precise measurement system if the measurement variation in the reference system is large.

Another issue that exists, that is not a fault of the limits of agreement approach, is that in general the technique is widely misused. In fact, Bland and Altman acknowledge the misuse of their technique when they say "the 95% limits of agreement method has been widely cited and widely used, though many who cite it do not appear to have read the paper" (p. 91).[20] To investigate this, Mantha et al.[21] and Dewitte et al.[22] undertook large-scale literature reviews of MSC studies analyzed by the limits of agreement technique, and found a variety of problems. The most pervasive and alarming was that in more than 90% of the articles examined the authors did not define a clinically acceptable difference. These authors were unaware that the crux of the limits of agreement approach, and the basis upon which interchangeability is determined, is the comparison of the limits of agreement to the clinically acceptable difference. Without this comparison, an assessment of interchangeability is ill-informed.

In this section, we have described the limits of agreement technique for comparing measurement systems, and although it is widely used we have demonstrated some of the challenges associated with the approach. Given that it is so widely used, it is clear that there is need for an analysis method that more accurately quantifies the agreement between two measurement systems and that is better safeguarded against misuse.

3 The alternative: Probability of agreement

In this section we propose a new method of analysis as an alternative to the limits of agreement approach. We propose a simple metric, the probability of agreement, and an associated plot to quantify the agreement between two measurement systems and hence help to decide whether the two systems can be used interchangeably. This approach strives to overcome the deficiencies of the limits of agreement technique described in the previous section.

3.1 The probability of agreement

The limits of agreement technique seeks to assess agreement by comparing the distribution of observed differences to what is considered clinically acceptable. This is a sensible goal, but in practice this comparison seems to be misunderstood and often omitted. A more direct and intuitive method of achieving this goal is to quantify the probability that the observed differences are small enough to be considered clinically acceptable. Using the notation associated with model (1), and assuming a clinically acceptable difference has the form $CAD = (-c, c)$ we define $\theta(s)$, the *probability of agreement*

$$\theta(s) = P(|Y_{i2} - Y_{i1}| \leq c | S_i = s) \tag{4}$$

The probability of agreement is the probability that the difference between single measurements on the same subject by the two systems falls within the range that is deemed to be acceptable, conditional on the value of the measurand. Based on the distributional assumptions associated with (1), $\theta(s)$ can be written as

$$\theta(s) = \Phi\left(\frac{c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \tag{5}$$

where $\Phi(x)$ is the standard normal cumulative distribution function evaluated at x.

Using probabilities of this form, we construct the *probability of agreement plot* which graphically displays the estimated probability of agreement across a range of plausible values for s. On this plot we include approximate pointwise confidence intervals for each value of $\theta(s)$ which reflect the uncertainty associated with its estimation. Note we employ maximum likelihood estimation to obtain estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ from model (1) which are substituted into (5) to obtain $\hat{\theta}(s)$. We sketch the technical details associated with this estimation procedure, and describe how to obtain the standard errors necessary for calculating approximate confidence intervals in the Appendix.

This probability of agreement plot serves as a simple tool for displaying the results when comparing two measurement systems; it summarizes agreement transparently and directly while accounting for possibly complicated bias and variability structures. While the modeling and estimation of $\theta(s)$ is somewhat complicated, its interpretation is extremely simple and one that most nonstatisticians can understand.

Another benefit is that even if a more complicated model than equation (1) is assumed, the interpretation of the probability and the plot is unchanged. For example we may wish to relax the assumption that S_i is normally distributed or perhaps model heteroscedastic measurement variation. In both cases we might alter model (1), but our interpretation of the probability of agreement and of the probability of agreement plot remains the same. These generalizations are discussed in section 5.

With this method, the probability that is deemed to indicate acceptable agreement and hence interchangeability is context-specific and is not a statistical decision. Accordingly, in this article we demonstrate how to estimate and interpret $\theta(s)$, but how large it must be to indicate interchangeability must be decided by the user. One reasonable choice might be to require $\theta(s) \geq 0.95$, similar to the limits of agreement approach.

If the probability of agreement plot does not indicate acceptable agreement, (i.e. $\theta(s)$ is too low in the range of interest for s), then we recommend looking at the separate estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ to determine the source of disagreement. Although the probability of agreement plot is informative and simple to interpret, examining the individual parameter estimates is the most informative description of the relationship between the two system's measurements. We also discuss how their estimation can be used to address Dunn's[1] additional MSC goals in section 3.4.

When the value of $\theta(s)$ is largely unchanged across the possible values for s, or if we simply wish to focus on the most likely values of the measurand, we may summarize the probability of agreement with a single number. We define an unconditional version of the probability of agreement, denoted θ, which is, in a sense, the average value of $\theta(s)$ across the distribution of S_i. Using the components of model (1), the unconditional probability of agreement is

$$\theta = P(|Y_{i2} - Y_{i1}| \leq c) = \Phi\left(\frac{c - \alpha - (\beta - 1)\mu}{\sqrt{(\beta - 1)^2\sigma_s^2 + \sigma_1^2 + \sigma_2^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)\mu}{\sqrt{(\beta - 1)^2\sigma_s^2 + \sigma_1^2 + \sigma_2^2}}\right) \qquad (6)$$

Use of an estimate of this single-number summary is appropriate when the probability of agreement is similar for all values of s, or when the range of measurand values of interest is close to the mean, μ. Thus, we first recommend the use of equation (5) and the corresponding plot to assess agreement, and then if the plot suggests that it is appropriate one may choose to summarize agreement based on equation (6).

Bland and Altman[16] offer a nonparametric approach which calculates the proportion of observed differences that fall within an acceptable range. Their method, although similar in spirit to the probability of agreement, does not address modeling the underlying relationship between measurement systems and, consequently, does not provide enough information to make an informed judgment regarding the interchangeability of two measurement systems.

The probability of agreement as defined in equation (4) may be viewed as a generalization of what Lin et al.[23] refer to as coverage probability. Here, we extend this idea to model (1) which is more general than what Lin et al.[23] consider in that proportional bias, replicate measurements and between-subject variation are considered. Another key difference between the proposed method and Lin's coverage probability is the manner in which it is used. For a fixed values of s, Lin et al.[23] consider testing hypotheses of the form $H_0 : \theta(s) \geq \theta_0$ versus $H_A : \theta(s) < \theta_0$, where agreement is rejected if H_0 is rejected. The emphasis of the probability of agreement approach on the other hand, is estimation as opposed to hypothesis testing; here we are interested in estimating and visualizing the agreement between two systems across a typical range of values for the measurand. Furthermore, by explicitly modeling the relationship between the two systems, we are able to identify the source of disagreement, should disagreement be indicated

3.2 Model checking

The first step in this procedure is to look at the data and decide whether the intended analysis is appropriate. In this context we suggest checking two assumptions of model (1). Specifically, we should check whether (i) the unknown true values of the measurand are normally distributed, and (ii) the repeatability is constant across the range of true values. We can assess each of these assumptions respectively with a *modified QQ-plot* and a *repeatability plot*. The latter plot also has the benefit of allowing us to check for outliers in the individual measured values.

To assess whether the measurand values are normally distributed, for each measurement system separately we average the replicate measurements on a particular subject and create a QQ-plots of these n averages. By working with the averages we reduce the effect of the measurement variation, allowing us to better examine the between-subject variation and the distribution of S_i. If the normality assumption holds both of these plots should yield a relatively straight line. To aid in their interpretation, we suggest overlaying the quantiles of 50 simulated normal datasets with mean and variance equal to the sample mean and sample variance of the n averages as suggested by Oldford.[24] Doing so depicts a region that we could expect the observed points to lie, if they came from a normal distribution. If this modified QQ-plot suggests that the normal distribution is a reasonable assumption for S_i then model (1) is applicable. However, if it does not, then an alternative to the maximum likelihood approach should be used. In section 5, we discuss a moment-based estimation procedure which does not require the normality assumption.

To decide whether the measurement variation for each system is constant across true values of the measurand we suggest constructing a repeatability plot for each measurement system. The plot is an individual values plot of the residuals of the replicate measurements on each subject versus the average of those replicate measurements, ordered by size. If the residuals seem unrelated to the averages this suggests that the measurement variation is homoscedastic. If however there appears to be a dependency between the residuals and averages, for example if variability in the residuals increases as the average increases, we conclude the measurement variation is heteroscedastic. The exact structure of heteroscedasticity will depend on the nature of the relationship between the residuals and averages. If the repeatability plots suggest heteroscedasticity of any kind in one or both measurement systems then model (1) is no longer appropriate and another approach must be taken. We discuss this issue in section 5.

In the next section, we present an example in which we illustrate this model checking approach.

3.3 Blood pressure example

To illustrate how to determine whether two measurement systems are interchangeable using the probability of agreement and the associated plot, we use systolic blood pressure data from an example published by Bland and Altman.[16] In this example, 85 subjects are measured three times by each of two observers, labeled “J” and “R”, both using a sphygmomanometer. While this is technically a comparison of two observers using the same measurement system, it is statistically equivalent to the comparison of two measurement systems; we can think of observer J as measurement system 1 (MS1), and observer R as measurement system 2 (MS2).

Before proceeding with the analysis, we check the model assumptions in accordance with the previous section. Figure 1 depicts the modified QQ-plots (upper panels) and repeatability plots (lower panels) by measurement system for this example. Examining the modified QQ-plots, we see that the blood pressure values in this particular sample are somewhat right-skewed, but we also see that the observed points fall within the grey region, indicating there is no evidence against the normal assumption. In examining the repeatability plots, we see that the points are randomly scattered with no clear trends indicating that the repeatability of each measurement system is homoscedastic. Thus we conclude that model (1) is appropriate.

Using the data described above, we estimate $\theta(s)$ for s in the range $(\hat{\mu} - 3\hat{\sigma}_s, \hat{\mu} + 3\hat{\sigma}_s)$ and construct the probability of agreement plot given in Figure 2. Note that the calculation of these probabilities assumes a clinically acceptable difference with $c = 10$. This is somewhat arbitrarily

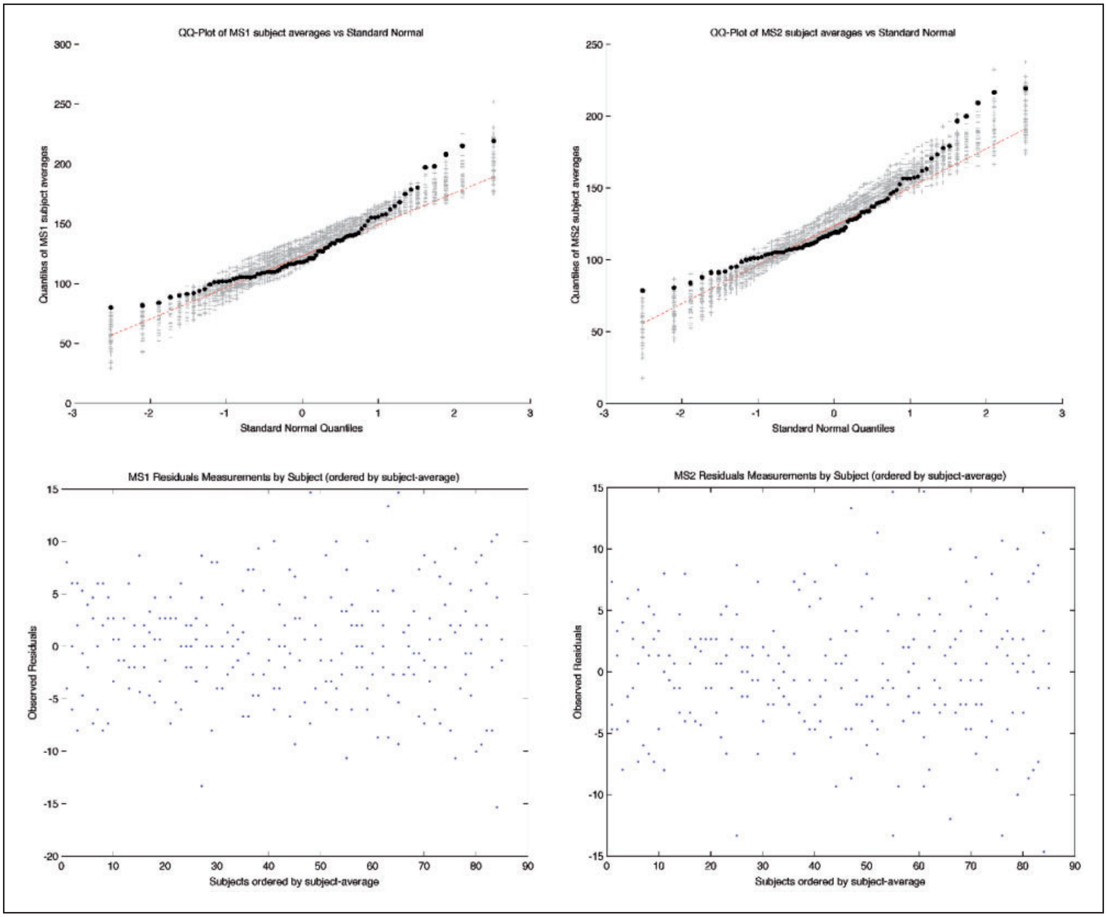

Figure 1. Modified QQ-plot and repeatability plot for observers "J" (MS1) and "R" (MS2) from the blood pressure data. Left panels correspond to observer "J" and right panels correspond to observer "R".

chosen since Bland and Altman[16] do not report a clinically acceptable difference for this example. To justify our assumed CAD we note that when assessing systolic blood pressure measuring devices, O'Brien et al.[25] provide criteria for grading such measurement systems. A blood pressure measurement device can be graded as A, B, C, or D depending on the proportion of differences that lie within ±5, ±10, and ±15 mmHg. These criteria are based on the difference between measurements by a new system and a sphygmomanometer, and are intended for assessing the adequacy of a new system relative to this standard. Our goal (assessing interchangeability) is different, but we assume this CAD is still relevant and use $c = 10$ for illustration. Note that the probably of agreement will increase for larger values of c and decrease for smaller values.

In Figure 2, we see that the probability of agreement is relatively constant (roughly 0.8) across the range of reasonable systolic blood pressures. It is not surprising then to find that the estimate of the unconditional probability of agreement θ is 0.799 with an approximate confidence interval given by (0.61, 0.98). Because there is little change in $\theta(s)$ across s, use of the unconditional probability seems reasonable for these data.

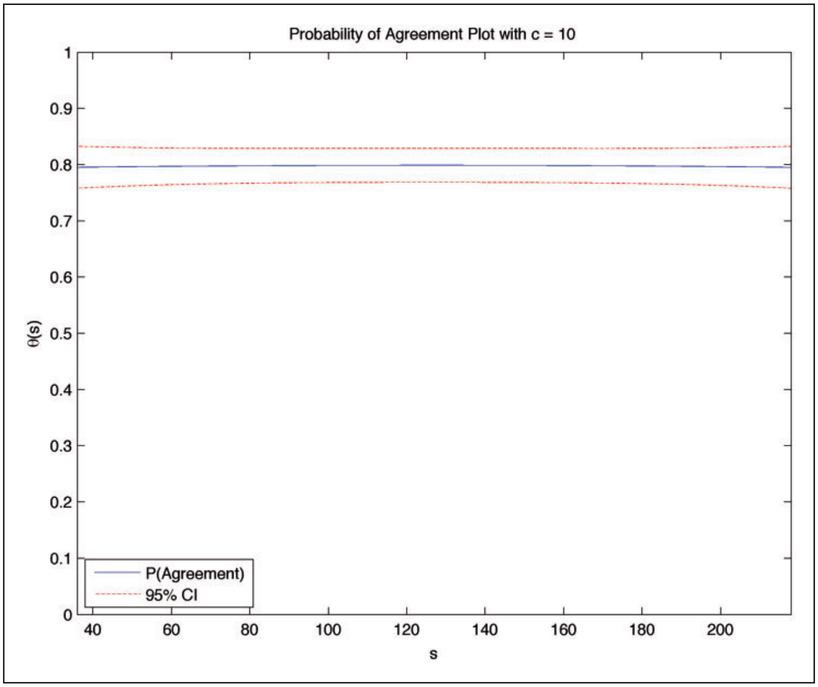

Figure 2. Probability of agreement plot comparing "J" and "R" for the blood pressure data.

Whether these results indicate agreement between the two measurement systems and that they could be used interchangeably depends on whether the investigators deem $\theta \approx 0.8$ to be sufficiently large. Suppose that $\theta \approx 0.8$ is not sufficiently large (perhaps $\theta \geq 0.95$ is necessary), leading us to conclude that the two measurement systems do not agree well enough to be used interchangeably. To identify the source of this disagreement we examine the individual parameter estimates and their asymptotic standard errors, which are shown in Table 1.

In light of the apparent disagreement, it is perhaps surprising to find that the fixed and proportional biases are negligible ($\alpha \approx 0$, $\beta \approx 1$) and the repeatabilities are very similar ($\sigma_1 \approx \sigma_2$), indicating that the distribution of the measurements made by each system are similar. The issue here is that although $\sigma_1 \approx \sigma_2$, both σ_1 and σ_2 are large relative to σ_s leading to large differences between individual measurements made by each system, causing the probability of agreement to be small.

In situations like this, when the reference system is highly variable, we may decide the new system is interchangeable with the reference even if the probability of agreement is small. For example, if the reference system is used routinely, perhaps a justification can be made for using a new system that is equally imprecise if it is, say, cheaper to operate.

Such a decision cannot be made by looking at the probability of agreement plot alone; although it accounts for complicated bias and repeatability structures, the probability of agreement masks the individual values of these parameters. Accordingly, we recommend that if the plot suggests disagreement between two systems, the individual parameter estimates be examined for guidance on a final decision.

For completeness we present the Bland and Altman replicate measures difference plot for these data in Figure 3. This plot also indicates disagreement, as the limits of agreement lie outside

Table 1. Maximum-likelihood estimates and asymptotic standard errors associated with the blood pressure data.

	Estimate	Asy. standard error
μ	127.3612	3.2937
α	−1.3623	2.1432
β	1.0108	0.016377
σ_s	30.1959	2.3421
σ_1	5.5655	0.28559
σ_2	5.4955	0.28347
θ	0.7985	0.09511

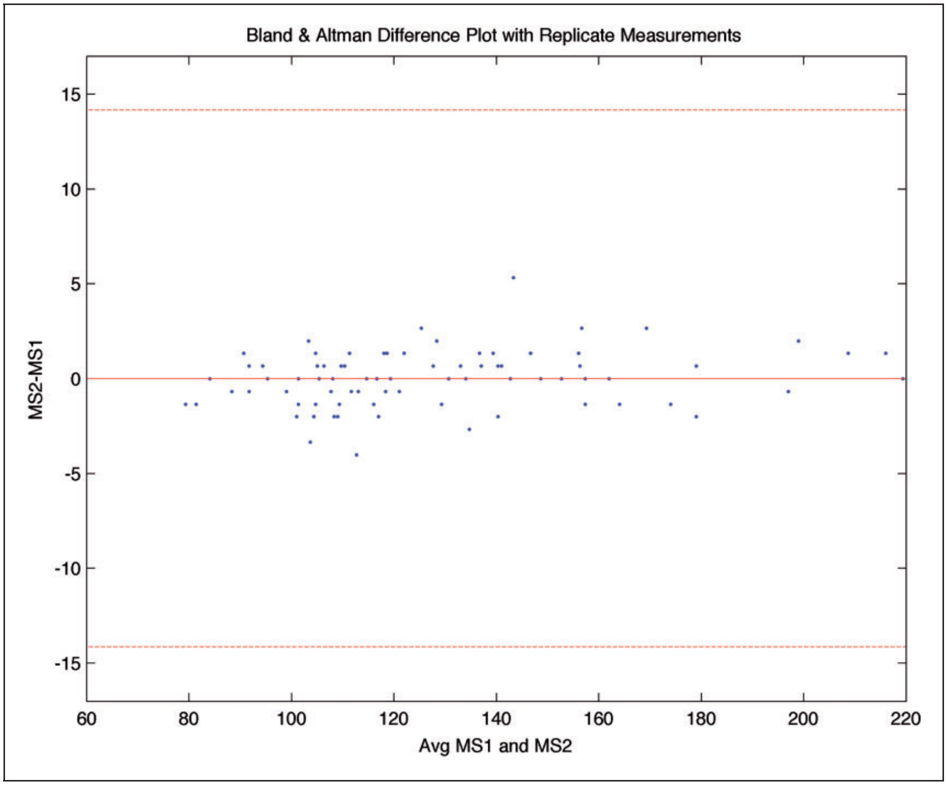

Figure 3. Replicate measures difference plot comparing "J" (MS1) and "R" (MS2) for the blood pressure data.

$CAD = \pm 10$. However, the difference plot does not quantify the disagreement as concisely as does the probability of agreement plot, nor does it offer any indication of the source of this disagreement.

This probability of agreement analysis technique and plot construction have been automated, and software is freely available at www.bisrg.uwaterloo.ca.

3.4 Addressing alternative goals of comparison

In section 1, we described several goals that may be considered important when comparing two measurement systems. The primary emphasis of the present article has been to quantify the agreement between two systems, with the goals of calibration, conversion, gold-standard comparison, and superiority having secondary importance. In this section we describe how the proposed methodology may be used to address these other goals.

If the probability of agreement plot suggests disagreement between two measurement systems, we have suggested the estimates of the parameters in model (1) be examined to identify the source of disagreement. Often disagreement will arise, in part, because of a systematic difference between the two systems, i.e. $\hat{\alpha} \neq 0$ and $\hat{\beta} \neq 1$. In this situation it may be of interest to calibrate the new system such that it agrees, on average, with the reference system. The corresponding adjustment to the new system's measurements is given by $Y_2^* = (Y_2 - \hat{\alpha})/\hat{\beta}$. When these adjusted measurements are compared to the reference system's measurements (i.e. Y_1), the probability of agreement plot should be constant across s, and any disagreement that remains is due to large variation in one or both systems. This plot may be referred to as a *potential agreement plot* as it displays the potential agreement between measurement systems after calibration. Note that this plot assumes $\hat{\alpha}$ and $\hat{\beta}$ are fixed values, and does not account for the uncertainty associated with their initial estimation.

If the measurements by both systems are on the same scale, then a systematic disagreement corresponds to the existence of relative bias. However, if the two systems measure on different scales then a systematic difference is due to a combination of the conversion between scales and relative bias. In this situation, Y_{i1} and Y_{i2} will not be similar and the probability of agreement plot will suggest disagreement even if relative bias is negligible. However, if the scale conversion is known it can be performed before analysis and agreement can then be quantified in the usual manner. Any remaining systematic difference (now just relative bias) can then be dealt with through calibration as described in the previous paragraph. Alternatively, if the scale conversion is unknown we can estimate α and β with the data on the original scales and perform a calibration which now simultaneously addresses the conversion between scales and relative bias.

The comparison to a gold-standard measurement system (one that measures without error) represents a somewhat different problem; it serves as an assessment of the accuracy and precision of the new measurement system. In this situation the parameters α and β represent the absolute bias of this system, and the probability of agreement becomes $\theta(s) = P(|Y_j - s| \leq c|S = s)$, which quantifies how closely the measurements by system j agree with the true value of the measurand. Estimation within this framework can be carried out with regression techniques and as we discussed earlier in this section, any bias (absolute bias in this case) can be addressed through calibration.

When deciding which of two measurement systems is superior (in terms of bias and precision), the probability of agreement is not overly useful; this decision is based solely on the parameter estimates. Fortunately, however, these estimates are obtained as a part of the probability of agreement analysis. If $\hat{\alpha} \approx 0$ and $\hat{\beta} \approx 1$, then the answer to which system is superior is based on a comparison of the repeatabilities, $\hat{\sigma}_1$ and $\hat{\sigma}_2$. If, however, a relative bias does exist we may perform a suitable calibration to eliminate this, meaning that the two systems will agree on average, in which case the decision again is still based on a comparison of repeatabilities.

3.5 Ventricle brain ratio example

To illustrate some of the ideas discussed in the previous section we introduce another example from the literature which compares two devices that use CAT scan images to measure ventricle brain ratios (VBR).[1] In the study, the VBR of $n = 50$ schizophrenic patients is measured $r = 2$ times by a hand-held planimeter on a projection of an X-ray image, and by an automated pixel count based on such images. Here, it is assumed that the planimeter (PLAN) can be regarded as the reference system and the pixel count (PIX) is assumed to be the new system. The raw data (which is on a log-scale) is presented in Dunn's book[1] along with a Bland and Altman difference plot.

Before the probability of agreement plot is constructed we assess the assumptions of model (1) in accordance with section 3.2. Though not shown here, the modified QQ-plots both suggest that a normality assumption for log(VBR) is reasonable, and the repeatability plots both suggest that the measurement variation of each system is homoscedastic on the log scale. We note that the model used by Dunn[1] to analyze these data accounts for a random subject-by-system interaction, which allows the effect of each system to differ from one subject to another. Though this model may be more appropriate, the parameters are not identifiable without further assumptions. Though the probability of agreement method may be carried out using such a model, for illustrative purposes we perform the analysis based on model (1).

The left panel of Figure 4 displays the probability of agreement plot for the VBR data. Without a clinically acceptable difference reported, we arbitrarily choose $c = 0.1$. As we can see, the level of agreement between the planimeter and pixel count depends highly on the true VBR, but is low for all values. Recall that the level of agreement will increase for larger c but its dependence on s will persist.

The parameter estimates displayed in Table 2 suggest that the source of disagreement is partly due to the fact that $\hat{\alpha} \neq 0$ and $\hat{\beta} \neq 1$, which indicate a systematic disagreement between systems. Using the calibration adjustment described in section 3.4, we calibrate the pixel count measurements to those made by the planimeter and redo the analysis. Table 3 displays the parameter estimates when the calibrated pixel measurements are used and the potential agreement plot (the probability of agreement plot for calibrated data) is shown in the right panel of Figure 4. As expected, the agreement between systems is constant across true VBR since $\hat{\alpha}^* \approx 0$ and $\hat{\beta}^* \approx 1$. However, the probability of agreement is still quite low (roughly 0.25), which results from a disparity

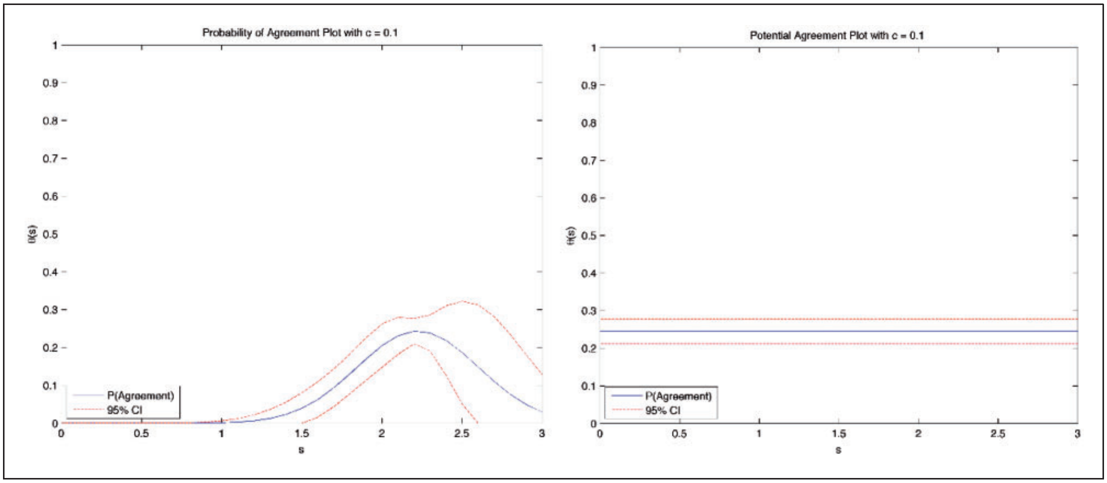

Figure 4. Probability of agreement plot (left panel) and potential agreement plot (right panel) comparing "PLAN" and "PIX" for the VBR data.

Table 2. Maximum-likelihood estimates and asymptotic standard errors associated with the VBR raw data.

	Estimate	Asy. standard error
μ	1.7861	0.0506
α	−1.9235	0.3905
β	1.8652	0.2160
σ_s	0.2771	0.0423
σ_1	0.3197	0.0227
σ_2	0.0401	0.0040
θ	0.1274	0.6427

Table 3. Maximum-likelihood estimates and asymptotic standard errors associated with the VBR data after calibration.

	Estimate	Asy. standard error
μ	1.7860	0.0506
α	0.0009	0.2091
β	0.9995	0.1157
σ_s	0.2772	0.0423
σ_1	0.3197	0.0227
σ_2	0.0215	0.0021
θ	0.2450	0.4905

between the repeatabilities of each system. Tables 2 and 3 indicate that the planimeter is extremely variable, and in fact $\hat{\sigma}_1 > \hat{\sigma}_s$. The pixel count, on the other hand, is much less variable on both the original and on the calibrated scales. Because $\hat{\sigma}_1$ is so large, we can conclude that agreement between the two systems is unlikely. We can further conclude that after suitable calibration, the pixel count appears to be the superior method of measuring VBR.

4 Planning MSC studies

When using the probability of agreement to decide whether two measurement systems are interchangeable, it is important to consider the design of the MSC study. The typical plan is to measure n subjects r times for a total of $N = nr$ measurements with each system. As has been stated several times, replicate measurements are necessary to ensure that the parameters in model (1), and hence the probability of agreement, can be estimated.

The emphasis of this article has been on estimating the agreement between two measurement systems. As such this investigation of study design is based on the assumption that precise estimation of the probability of agreement is of primary interest. If, however, agreement is based on a hypothesis test (such as the one discussed in section 3.1), then a power analysis approach would be more appropriate.

Here, we assume that N measurements can be made by each system, and the primary interest is to decide how to spend resources and hence decide how to allocate those measurements. As such we investigate the effect of the number of subjects n and the number of replicate measurements r on the precision with which θ can be estimated. Note that we base these comparisons on the unconditional probability of agreement, θ, instead of $\theta(s)$ because it is difficult to determine in general which values of s are relevant. As such, we investigate the effect of n and r on the estimate of θ, the probability of agreement for "typical" values of s. We compare designs using the asymptotic standard deviations of the estimator $\tilde{\theta}$, calculated from the Fisher information matrix, as described in the Appendix.

To ensure that the asymptotic results will allow us to appropriately rank the possible designs, we first conducted a simulation study to compare the asymptotic and simulated standard errors of $\hat{\theta}$ which we describe in section 4.1. This simulation confirmed that even for small sample sizes the simulated and asymptotic results agree, justifying the use of asymptotic results to investigate possible (n, r) combinations for a given value of N. In section 4.2, we make design recommendations for optimal estimation of the probability of agreement. In section 4.3, we investigate whether the manner in which subjects are selected effects the estimation of $\theta(s)$ and we report the results of a simulation study which compared three sampling protocols in their ability to accurately and precisely estimate $\theta(s)$.

4.1 Comparing simulated and asymptotic standard errors

In this simulation study, we compared the simulated and asymptotic standard errors of $\hat{\theta}$ for a variety of (n, r) combinations and parameter values. To cover a wide range of sample sizes, replicate measurements and parameter values, we considered:

- $n = 10$ to 200 in steps of 10 and $r = 2$ to 10 in steps of 1
- $\mu = 1,\ 10,\ 100$
- $\sigma_s = \mu/10,\ \mu/4$
- $\sigma_1 = \sigma_s/10,\ \sigma_s/4$
- $\sigma_2 = 3\sigma_1/4,\ \sigma_1,\ 5\sigma_1/4$
- $\alpha = 0,\ 0.5\mu$
- $\beta = 1,\ 1.1$

For each combination of n, r and the parameters, we generated 10,000 samples according to model (1) and for each sample determined the maximum likelihood estimate of θ and the asymptotic standard error associated with that estimate. We explain in the Appendix how we obtained the asymptotic standard error.

We then compared the simulated and asymptotic results by dividing the standard deviation of the 10,000 estimates of $\hat{\theta}$ by the average of the 10,000 asymptotic standard errors. Across all combinations of n, r and the parameters, the average of this ratio was 0.9915 and it ranged between 0.89 and 1.11 with the middle 50% lying between 0.97 and 1.02. Thus, overall the results suggest that the asymptotic standard deviation closely matches the simulated results for all designs. Accordingly we proceed to rank designs based on the asymptotic results.

4.2 Recommendations for MSC study design

For a particular combination of the parameter values and $N = 40,\ 60,\ 100,\ 120,\ 200$, we iterate through $2 \leq r \leq 10$ and take $n = N/r$. In the case that N/r is not an integer, we round this quantity down to the nearest integer to determine n, in which case $nr < N$. We then rank the designs according to the asymptotic standard deviation of θ, and consider the design associated with the smallest asymptotic standard deviation the 'best'. In doing this it became clear that the design in which each subject is measured twice, corresponding to $(n, r) = (N/2, 2)$, always has the smallest, or nearly the smallest, asymptotic standard deviation.

To investigate this further we compare the asymptotic standard deviations associated with the 'best' design and the design with two replicate measurements, i.e. $(n, r) = (N/2, 2)$. Specifically we divide the standard deviation corresponding to the best design by that of the $(n, r) = (N/2, 2)$ design. For $N = 40, 60, 100, 120, 200$, $2 \leq r \leq 10$, and the parameter values outlined in Section 4.1 we found the average of this ratio to be 1.01. Thus the asymptotic standard deviation associated with the $(n, r) = (N/2, 2)$ design is on average only 1% larger than the best design. We found the maximum of this ratio to be 1.065, which occurs when α is different from 0, β is different from 1 and when σ_1 and σ_2 are very different.

Software is available at www.bisrg.uwaterloo.ca which provides the best design for a particular combination of parameter values and maximum number of measurements N. However, because the best design depends on the values of the unknown parameters, and the $(n, r) = (N/2, 2)$ design is close to optimal across all parameter values we considered, we recommend its use. To select N, we can use the software described above to investigate the standard deviation of θ in the $(N/2, 2)$ design for various reasonable values of the unknown parameters.

4.3 Effect of subject sampling protocol

The design recommendation in the previous section assumes subjects are sampled randomly. To investigate the effect the sampling protocol has on the estimation of $\theta(s)$, we performed a small simulation study. For three different sampling protocols, we compared the true value of $\theta(s)$ to the simulated estimate, and compared the true asymptotic standard deviation to the simulated asymptotic standard error. These comparisons were made for $\theta(s)$ evaluated at small, medium and large values of s: $s = \mu - 2\sigma_s$, $s = \mu$, $s = \mu + 2\sigma_s$, respectively. The sampling protocols that we considered were random sampling, uniform sampling (equal number of subjects sampled between each decile of the distribution) and extreme sampling (subjects sampled equally from the upper and lower quarters of the distribution).

For every combination of the parameter values listed in section 4.1 and $n = 50, 100, 200$ and $r = 2, 3, 4, 5$ we simulated 100 datasets and estimated $\theta(\mu - 2\sigma_s)$, $\theta(\mu)$ and $\theta(\mu + 2\sigma_s)$ and their asymptotic standard errors. We then average these 100 estimates to obtain the simulated estimate and also determined the corresponding simulated asymptotic standard error. We then calculate the bias of the estimate as the true value of $\theta(s)$ minus the simulated estimate, and we examine the ratio of the asymptotic standard deviation to the simulated standard error.

The bias associated with estimating $\theta(s)$ for all sampling protocols, parameter values, sample sizes and values of s was on average 0.0001 or less. The only exception was estimating $\theta(\mu + 2\sigma_s)$ in the context of extreme sampling, in which case the average bias was 0.048, though it was 0.0001 for $\beta = 1$. Thus the manner in which subjects are sampled has little effect on the accuracy with which $\theta(s)$ is estimated.

Across all parameter values and sample sizes the average ratio comparing asymptotic and simulated standard errors for $\theta(\mu)$ was 1.00, 1.003, and 1.002 for random, uniform, and extreme sampling, respectively. The average ratios associated with $\theta(\mu - 2\sigma_s)$ and $\theta(\mu + 2\sigma_s)$ were also approximately 1 for random and uniform sampling, but significantly different from 1 in the case of extreme sampling. Thus we see general agreement between asymptotic and simulated precisions and so we conclude that approximate confidence intervals for $\theta(s)$ should be valid if subjects are sampled randomly or uniformly. In the case of extreme sampling such intervals should be valid for values of s close to μ.

5 Conclusions and discussion

In this article, we propose a new method of assessing agreement between two measurement systems: the probability of agreement. The probability of agreement is, for a particular value of the measurand, the probability that the difference between two measurements made by different systems falls within an interval that is deemed to be acceptable, defined by equation (4). This quantity can be translated into an informative plot which depicts the probability of agreement across a range of possible true values for the measurand. The result is a simple and intuitive summary of the agreement between two measurement systems. The benefit of this approach is that while the statistical modelling and estimation (which is handled by software) may be complicated for a nonstatistician, the interpretation is straight forward and intuitive, and the same regardless of which model is used and what assumptions are made. This ease of interpretation is of practical importance as it facilitates the wide-spread use of this technique, especially given that estimation and plot generation is automated with the software available at www.bisrg.uwaterloo.ca.

Here, we have assumed that the true values of the measurand follow a normal distribution. However, if normality does not hold we may apply a moment-based approach to estimating equation (4) that does not rely on this assumption.[26] We have also assumed that each system's repeatability is homoscedastic. However, if the measurement variation is heteroscedastic then we suggest using a model different from equation (1) that accounts for a dependence between the measurement variation and the unknown true value of the measurand.[26] Performing the present analysis on log-transformed data may also be effective. Even in the face of these adaptations, the probability of agreement and associated plot can still be interpreted and applied in the same way. We plan to explore this in future work.

We note that the level of agreement between two measurement systems depends critically on the value c, which defines the clinically acceptable difference; agreement will increase for larger values of c and decrease for smaller values. However, the choice of c is often a difficult decision to make in practice. In these situations we suggest repeating the analysis for different values of c to investigate the sensitivity of the conclusions to this value. For a particular choice of s we could summarize this analysis with a plot of $\theta(s)$ versus c to visualize the sensitivity of $\theta(s)$ to c. An example of such a plot is shown in Figure 5 for the blood pressure data when $s = \hat{\mu} = 127.3612$. An alternative approach might be to adapt Lin's total deviation index (TDI)[27] and invert the definition of equation (4) to calculate c for a value of $\theta(s)$ which is suitably large, then decide whether this value of c is practically acceptable.

In this article we assume that there are no operator effects; i.e. we implicitly assume that the measurement systems being compared have only a single operator, or if multiple operators exist, we assume that their effects are the same. One possible extension is to incorporate operator effects into the probability of agreement analysis. We have also assumed that each measurement system measures each subject r times. Another straight forward extension of this work would be to adapt the model and consider the case when the two systems make a different number of replicate measurements per subject, i.e. $r_1 \neq r_2$, or a different number of measurements on each subject.

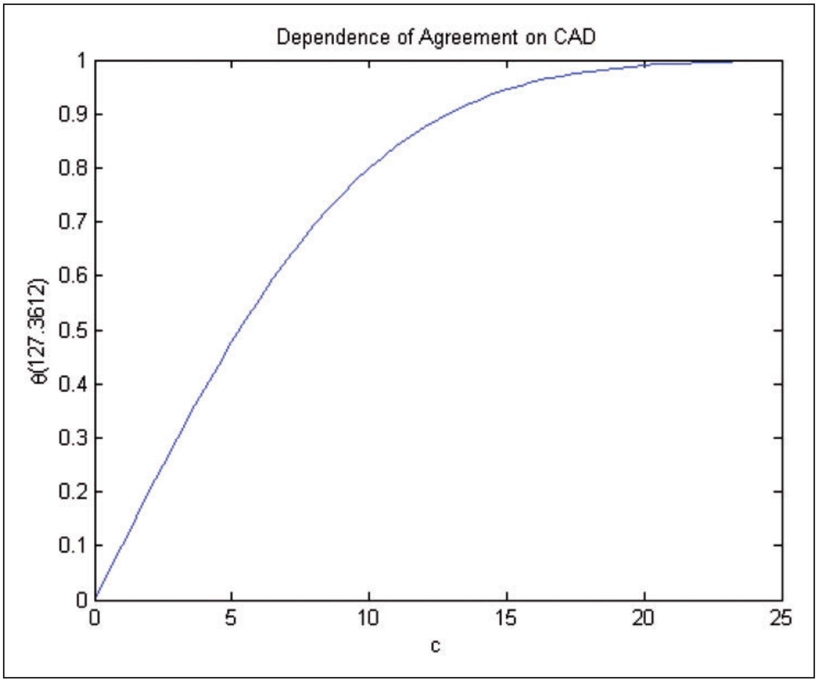

Figure 5. $\theta(s)$ plotted against c for blood pressure data when $s = \hat{\mu} = 127.3612$.

Funding

This research was supported by an NSERC PGS-D Scholarship and an NSERC Discovery Grant #102415.

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Dunn G. *Statistical evaluation of measurement errors: Design and analysis of reliability studies*, 2nd ed. London: Arnold, 2004.
2. Altman DG and Bland JM. Measurement in medicine: The analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.
3. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–310.
4. Carstensen B. *Comparing clinical measurement methods: A practical guide*. Chichester: Wiley, 2010.
5. U.S. Food and Drug Administration. Premarket notification. http://www.fda.gov/medicaldevices/deviceregulationandguidance/howtomarketyourdevice/premarketsubmissions/premarketnotification510k/default.htm#se (accessed 13 March 2015).
6. Mayo Clinic. Communique. Assay validation: What is it, why do we need it, and how do we do it? http://www.mayomedicallaboratories.com/articles/communique/2010/09.html (accessed 13 March 2015).
7. Clinical and Laboratory Standards Institute. Measurement procedure comparison and bias estimation using patient samples; approved guidelines. 3rd ed. Technical Report, EP09-A3, 2013.
8. Statistical guidelines for the annals of clinical biochemistry. http://acb.sagepub.com/site/includefiles/Statistical%20Guidelines.pdf (accessed 1 July 2015).
9. Clinical chemistry author instructions. http://www.clinchem.org/site/info_ar/info_authors.xhtml#standards (accessed 1 July 2015).
10. Barnett VD. Simultaneous pairwise linear structural relationships. *Biometrics* 1969; **25**: 129–142.
11. Ludbrook J. Confidence in Altman-Bland plots: A critical review of the method of differences. *Clin Exp Pharmacol Physiol* 2010; **37**: 143–149.

12. Choudhary PK and Nagaraja HN. Measuring agreement in method comparison studies - A review. In: Balakrishnan N, et al. (eds) *Advances in ranking and selection, multiple comparisons, and reliability: Methodology and applications*. Boston: Birkhauser, 2005, pp.215–244.
13. Barnhart HX, Haber MJ and Li L. An overview on assessing agreement with continuous measurements. *J Biopharm Stat* 2007; **17**: 529–569.
14. Lin L. Overview of agreement statistics for medical devices. *J Biopharm Stat* 2008; **18**: 126–144.
15. Ryan TP and Woodall WH. The most-cited statistical papers. *J Appl Stat* 2005; **32**: 461–474.
16. Bland JM and Altman DG. Measuring agreement in method comparison studies. *Stat Meth Med Res* 1999; **8**: 135–160.
17. Bland JM and Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharmaceut Stat* 2007; **17**: 571–582.
18. Voelkel JG and Siskowski BE. A study of the Bland-Altman plot and its associated methodology. Technical Report, Center for Quality and Applied Statistics, Rochester Institute of Technology, 2005.
19. Bland JM and Altman DG. Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet* 1995; **346**: 1085–1087.
20. Bland JM and Altman DG. Applying the right statistics: Analyses of measurement studies. *Ultrasound Obstet Gynecol* 2003; **22**: 85–93.
21. Mantha S, Roizen MF, Fleisher L, et al. Comparing methods of clinical measurement: Reporting standards for Bland and Altman analysis. *Anesth Analg* 2000; **90**: 593–602.
22. Dewitte K, Fierens C, Stöckl D, et al. Application of the Bland-Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clin Chem* 2002; **48**: 799–801.
23. Lin L, Hedayat AS, Sinha B, et al. Statistical methods in assessing agreement: Models, issues, and tools. *J Am Stat Assoc* 2002; **97**: 257–270.
24. Oldford RW. On modified QQ plots. Personal communication, 2014.
25. O'Brien, Petrie J, Littler W, et al. The British Hypertension Society protocol for the evaluation of measuring devices. *J Hypertension* 1993; **11**: S43–S62.
26. Stevens NT. *Assessment and comparison of continuous measurement systems*. PhD Thesis, University of Waterloo, Canada, 2014.
27. Lin L. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Stat Med* 2000; **19**: 255–270.
28. *Maple 18. Maplesoft*. Waterloo, ON: Waterloo Maple Inc. www.maplesoft.com.
29. *Matlab 8.2.0*. Natick, MA: The MathWorks Inc. www.mathworks.com.

Appendix

Here, we elaborate upon the maximum-likelihood procedure used to obtain point and interval estimates of $\theta(s)$. For a particular subject i, we order the random vector corresponding to its measurements by system and write $\boldsymbol{Y_i} = \left(\boldsymbol{Y_{i1}^T}, \boldsymbol{Y_{i2}^T}\right)^T$, where $\boldsymbol{Y_{ij}} = \left(Y_{ij1},\ Y_{ij2}, \ldots, Y_{ijr}\right)^T$ corresponds to the r measurements by system j on subject i. In what follows we let $\boldsymbol{J_a}$ be a column vector of a 1's, $\boldsymbol{J_{a\times b}}$ be an $a \times b$ matrix of 1's, and $\boldsymbol{I_a}$ be the $a \times a$ identity matrix. From model (1), we have $\boldsymbol{Y} \sim MVN\left(\boldsymbol{\mu}, \sum\right)$ with

$$\boldsymbol{\mu} = (\mu, \alpha + \beta\mu)^T \otimes \boldsymbol{J_r}$$

and

$$\sum = \sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & 1 \end{bmatrix} \otimes \boldsymbol{J_{r\times r}} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \otimes \boldsymbol{I_r}$$

where $\otimes$ denotes the Kronecker product.

In order to explicitly write down the log-likelihood function for subject i we must first obtain the inverse and determinant of the covariance matrix. Fortunately the form of $\sum$ allows us to write down $\sum^{-1}$ and $\left|\sum\right|$ explicitly

$$\sum^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes \boldsymbol{I_r} - \frac{\sigma_s^2}{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)} \begin{bmatrix} \frac{1}{\sigma_1^4} & \frac{\beta}{\sigma_1^2\sigma_2^2} \\ \frac{\beta}{\sigma_1^2\sigma_2^2} & \frac{\beta^2}{\sigma_2^4} \end{bmatrix} \otimes \boldsymbol{J_r}$$

$$\left|\sum\right| = \left(\sigma_1^2\sigma_2^2\right)^r \left\{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)\right\}$$

Denoting the observed data by y_{ijk} $i = 1, 2, \ldots, n$, $j = 1, 2$ and $k = 1, 2, \ldots, r$ (we distinguish the random variable Y_{ijk} by using a lower-case y_{ijk} to denote the observed data), the log-likelihood contribution from subject i with r replicate measurements by both systems is

$$-rln(2\pi) - \frac{1}{2} ln\left|\sum\right| - \frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{\mu})^{\boldsymbol{T}} \sum^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu})$$

since by model (1), $\boldsymbol{Y_i} \sim MVN(\boldsymbol{\mu}, \sum)$. We can explicitly write this as

$$\begin{aligned} l_i(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s) = & -rln(2\pi) - \frac{1}{2}\ln\left[1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)\right] - \frac{1}{2\sigma_1^2}\sum_{k=1}^{r}(y_{i1k} - \mu)^2 \\ & + \frac{b}{2\sigma_1^4}\left\{\sum_{k=1}^{r}(y_{i1k} - \mu)\right\}^2 - \frac{1}{2\sigma_2^2}\sum_{k=1}^{r}(y_{i2k} - \alpha - \beta\mu)^2 + \frac{b\beta^2}{2\sigma_2^4}\left\{\sum_{k=1}^{r}(y_{i2k} - \alpha - \beta\mu)\right\}^2 \\ & - \frac{b\beta}{\sigma_1^2\sigma_2^2}\sum_{k=1}^{r}(y_{i1k} - \mu)(y_{i2k} - \alpha - \beta\mu) \end{aligned}$$

where

$$b = \frac{\sigma_s^2}{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)}$$

Because we assume measurements made on different subjects are independent we obtain the full log-likelihood function by summing the log-likelihood contribution for each subject

$$l(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s) = \sum_{i=1}^{n} l_i$$

In order to calculate approximate confidence intervals for $\theta(s)$, we must obtain asymptotic standard deviations for $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ which are found using the expected Fisher information matrix. The expected Fisher information matrix is found by taking second partial derivatives of $l(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$, which are performed symbolically by Maple[28] to avoid errors, and by calculating the expected values of the necessary sums of squares. We do not give all of the formulas here, but note that we use the following results

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i1k} - \mu)^2\right] = nr(\sigma_s^2 + \sigma_1^2)$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i1k} - \mu)\right\}^2\right] = nr(r\sigma_s^2 + \sigma_1^2)$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i2k} - \alpha - \beta\mu)^2\right] = nr(\beta^2\sigma_s^2 + \sigma_2^2)$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right\}^{2}\right]=nr\left(r\beta^{2}\sigma_{s}^{2}+\sigma_{2}^{2}\right)$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i1k}-\mu)\right\}\left\{\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right\}\right]=nr^{2}\beta\sigma_{s}^{2}$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i1k}-\mu)\right]=0$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right]=0$$

We then invert the Fisher Information matrix numerically using Matlab.[29] This gives the asymptotic variances of $(\mu,\alpha,\beta,\sigma_1,\sigma_2,\sigma_s)$. But because we are interested in $\theta(s)$ and θ, we find their asymptotic variances by applying the delta method; we pre- and post-multiply the inverse of the Fisher information matrix by a suitable vector of partial derivatives: $\boldsymbol{D}_s$ for the asymptotic variance of $\theta(s)$, and $\boldsymbol{D}$ for θ.

$$\boldsymbol{D}_s=\frac{\partial\theta(s)}{\partial(\mu,\alpha,\beta,\sigma_1,\sigma_2,\sigma_s)}\qquad\qquad \boldsymbol{D}=\frac{\partial\theta}{\partial(\mu,\alpha,\beta,\sigma_1,\sigma_2,\sigma_s)}$$

Approximate confidence intervals for $\theta(s)$ and θ are calculated using asymptotic standard errors, which are obtained by evaluating their respective asymptotic standard deviations at the maximum likelihood estimates $(\hat{\mu},\hat{\alpha},\hat{\beta},\hat{\sigma}_1,\hat{\sigma}_2,\hat{\sigma}_s)$.