Published in final edited form as:

Stat Med. 2017 October 15; 36(23): 3621–3635. doi:10.1002/sim.7352.

A Framework for the Meta-Analysis of Bland-Altman Studies Based on a Limits of Agreement Approach

Elizabeth Tipton and

Teachers College, Columbia University

Jonathan Shuster

University of Florida

Abstract

Bland-Altman method comparison studies are common in the medical sciences and are used to compare a new measure to a gold-standard (often costlier or more invasive) measure. The distribution of these differences is summarized by two statistics, the ‘bias’ and standard deviation, and these measures are combined to provide estimates of the Limits of Agreement (LoA). When these LoA are within the bounds of clinically insignificant differences, the new non-invasive measure is preferred. Very often multiple Bland-Altman studies have been conducted comparing the same two measures, and random-effects meta-analysis provides a means to pool these estimates. We provide a framework for the meta-analysis of Bland-Altman studies, including methods for estimating the LoA and measures of uncertainty (i.e., confidence intervals). Importantly, these LoA are likely to be wider than those typically reported in Bland-Altman meta-analyses. Frequently Bland-Altman studies report results based on repeated measures designs, but do not properly adjust for this design in the analysis. Meta-analyses of Bland-Altman studies frequently exclude these studies for this reason. We provide a meta-analytic approach that allows inclusion of estimates from these studies. This includes adjustments to the estimate of the standard deviation and a method for pooling the estimates based upon robust variance estimation. An example is included based on a previously published meta-analysis.

Keywords

meta-analysis; Bland-Altman; method comparison study; random effects; robust variance estimation

1. Introduction

An important area for meta-analysis that has not received much attention is method comparison studies conducted using the Bland-Altman approach [1 – 2]. Bland-Altman studies are widely used in medicine to compare two methods of measurement: one that is new to one that is considered the “gold standard” but is often invasive or costly. The goal of a method comparison study is to determine if the new measure can be used in practice, and

when used, the degree of similarity or difference that can be expected compared to the standard measure.

The popularity of Bland-Altman methods means that there are often multiple studies conducted comparing the same two measures. Meta-analyses of method comparison studies are common and highly cited, particularly in the fields of anesthesia [3 – 4] and cardiology [5 – 7]. For example, a meta-analysis of cardiac output measurement techniques combined results across 25 studies [8] (776 citations), while a meta-analysis of pulse oximetry studies included 27 Bland-Altman studies [5] (117 citations). Importantly, even a single Bland-Altman study with large numbers of repeated measures per person can be thought of as a meta-analysis, with the role of “study” assumed by the individual subject.

Despite this interest and use, methods for meta-analyzing Bland-Altman studies are underdeveloped. Bland-Altman studies typically report estimates of two measures – the *bias* of a test (i.e., the average difference between the new measure and the standard measure) and the *standard deviation* of these differences. These two measures are typically combined into the *limits of agreement*, roughly a 95% prediction interval for these differences. To date, meta-analyses of Bland-Altman studies have focused primarily on providing pooled estimates of the bias and standard deviation across studies; for example, [9] provide an approach to pooling estimates of and testing hypotheses regarding heterogeneity of bias and standard deviation across studies, while [10] provides a method for meta-analyzing standard deviations and mean-squared-error across studies. In comparison, estimates of the limits of agreement – while the primary focus of individual Bland-Altman studies – are rarely the focus of Bland-Altman meta-analyses. To date, no meta-analytic papers have focused on methods for estimating these limits or measures of their uncertainty.

The goal of this paper is to provide a unified framework for the meta-analysis of Bland-Altman studies. After reviewing Bland-Altman measures for primary studies, we frame the discussion of meta-analyses by identifying the parameters of interest. We then provide an approach for estimating the limits of agreement and their associated confidence intervals. This estimation strategy assumes that the bias and standard deviation are appropriately estimated in primary studies, which is not always the case. Instead, it is routine for studies to treat repeated measures data as independent. After outlining this problem, we then provide an adjustment to the estimate of the standard deviation, and provide a robust estimation strategy that allows for appropriate confidence intervals for the limits of agreement to be calculated. In order to illustrate the approach, we provide an example briefly re-analyzing data from a previously conducted meta-analysis. We conclude with a discussion of future directions for research in this area.

2. Bland-Altman Studies

In a single Bland-Altman method comparison study, the focus is on summarizing the difference between two tests – a new test (Y) and a standard test (X). To be more formal, for study j assume there are individuals $i = 1 \dots n_j$, and that each individual is measured once with the new test (Y_{ij}) and once with the standard test (X_{ij}). Assume that in the population, the distribution of Y_{ij} has mean $E(Y_{ij}|j) = \mu_{yj}$ and variance $V(Y_{ij}|j) = \sigma_{yj}^2$ and similarly, the

distribution of X_{ij} has mean $E(X_{ij}|j) = \mu_{xj}$ and variance $V(X_{ij}|j) = \sigma_{xj}^2$. The estimated difference is modeled as,

$$D_{ij} = Y_{ij} - X_{ij} = (\mu_{yj} - \mu_{xj}) + (\varepsilon_{yij} - \varepsilon_{xij}) = \delta_j + \varepsilon_{ij}, \quad (2.1)$$

where $\delta_j = E(D_{ij}|j)$ is referred to in the literature as the *bias* of the new test Y (compared to the standard measure X) in study j , and $\sigma_j = \sqrt{V(D_{ij}|j)}$ is the *standard deviation* of the differences between the new test Y and standard test X in study j . The parameters δ_j and σ_j are estimated with their sample quantities,

$$\hat{\delta} = D_j = \sum D_{ij} / n_j \quad (2.2)$$

$$\hat{\sigma} = S_j = \sqrt{[\sum (D_{ij} - D_j)^2 / (n_j - 1)]}. \quad (2.3)$$

In study j if we assume that $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, then a 95% prediction interval for the test comparison can be estimated using $D_j \pm 1.96S_j$. These limits – or the simplified $D_j \pm 2S_j$ – are referred to as the *limits of agreement* (LoA). If the LoA are within the bounds of clinically insignificant differences, then the new test – often cheaper or less invasive – can be substituted in practice for the standard test.

Bland and Altman [2] suggest that confidence intervals also be reported for the limits of agreement. In particular, they focus on the *outer* limits – the lower 95-percentile limit for the lower-value of the LoA and the upper 95-percentile limit for the upper-value of the LoA. In order to construct these, they note that

$$V(D_j \pm 2S_j) \approx 3S_j^2 / n_j \quad (2.4)$$

and they argue that the 95% confidence intervals can then be calculated using the t-distribution as $(CI-LoA_L, CI-LoA_U)$, where

$$CI-LoA_L = (D_j - 2S_j) - t(n_j - 1, .025) \sqrt{(3S_j^2 / n_j)} \quad (2.5)$$

$$CI-LoA_U = (D_j + 2S_j) + t(n_j - 1, .025) \sqrt{(3S_j^2 / n_j)}. \quad (2.6)$$

Unfortunately, these confidence intervals are too often not reported in primary studies.

The typical published Bland-Altman study thus provides two parameter estimates: the estimated bias (D_j) and the estimated standard deviation (S_j). These are typically reported

both separately and in the combined limits of agreement ($D_j \pm 2S_j$). In the next section, we outline the overall goals of a meta-analysis of these results.

3. Meta-Analysis Parameters

In a meta-analysis, the goal is to pool results from $j = 1 \dots m$ independent studies, each reporting estimates of bias, standard deviation, and limits of agreement. If we had access to the full data from all m studies, we might imagine writing the model combining all of the data as,

$$D_{ij} = \delta_j + \varepsilon_{ij} = \delta + \eta_j + \varepsilon_{ij} \quad (3.1)$$

which disaggregates the observed difference into three parts: the common bias (δ), the study-specific bias residual (η_j) and the patient-specific residual (ε_{ij}). Assume that $E[\eta_j] = E[\varepsilon_{ij}|j] = 0$, $\text{Var}[\eta_j] = \tau^2$, $\text{Var}[\varepsilon_{ij}|j] = \sigma_j^2$ and that $\text{Cov}[\eta_j, \varepsilon_{ij}|j] = 0$. Furthermore, assume that the variation in measurement differences is not necessarily constant across studies, that is, that $E[\sigma_j^2] = \sigma^2$ and $\text{Var}[\sigma_j^2] > 0$. Note that this assumption is weaker than requiring that each study has the same within-study variation (i.e., $\sigma_j^2 = \sigma^2$). This is important since different inclusion/exclusion criteria, study samples, or measurement processes might result in a larger or smaller variation in differences in a particular study.

Looking across the population of patients in many studies, differences between a new test Y and the standard test X can be summarized in terms of a distribution with a mean (the population bias) and standard deviation (population standard deviation). Using the law of iterated expectations we can see this *population bias* is

$$E(D_{ij}) = E[E(D_{ij}|j)] = E[\mu_{yj} - \mu_{xj}] = \delta. \quad (3.2)$$

Furthermore, applying the law of total variance, the *population standard deviation* can be written,

$$\begin{aligned} \sqrt{V(D_{ij})} &= \sqrt{\{E[\text{Var}(D_{ij}|j)] + \text{Var}[E(D_{ij}|j)]\}} \\ &= \sqrt{\{E[\sigma_j^2] + \text{Var}[\eta_j]\}} = \sqrt{(\sigma^2 + \tau^2)}. \end{aligned} \quad (3.3)$$

Importantly, this standard deviation takes into account two sources of variation – the average within-study variation (pooled across the study-specific variances, σ_j^2) and the between-study variation (in study-specific biases, τ^2). Thus the limits of agreement for the total population of patients studied can be written

$$\delta \pm 2\sqrt{(\sigma^2 + \tau^2)}. \quad (3.4)$$

In the next section, we will provide estimators for each parameter (δ , σ^2 , and τ^2) as well as standard errors.

Importantly, the standard deviation just defined differs from the standard deviation currently estimated in meta-analyses of Bland-Altman studies. In these studies, it is typical instead to estimate the limits of agreement

$$\delta \pm 2\sigma \quad (3.5)$$

(e.g., [4, 6, 7, 8]). These limits of agreement, however, indicate the range of differences for the *average study* not for the *total population* of individuals studied.

In order to understand how these might differ, Figure 1 compares the parameters in an extreme case. In the figure, the within-study precisions are very small (i.e., for all studies, $\sigma_j^2 = .25^2$), indicating that the tests are perhaps similar, though the degree of bias varies strongly from study-to-study (i.e., $\delta = 0$, $\tau^2 = .5^2$). Focusing on the average study (the black solid line) obscures this between-study heterogeneity, which the population distribution (the black dashed line) makes obvious.

In the remainder of this paper, we focus on the estimation of the limits of agreement based upon the population standard deviation ($\sigma^2 + \tau^2$). We argue that this standard deviation more accurately reflects the true heterogeneity found in practice, particularly for a physician using the procedure for the first time. In the Example, we will compare these approaches using real data and discuss the implications for practice.

4. Pooling Estimates from Independent Data Designs

In a meta-analysis of Bland-Altman studies, the goal is to estimate the population limits of agreement, $\delta \pm 2(\sigma^2 + \tau^2)$. These limits of agreement are a function of three parameters: the average difference between test Y and X in the population (the population bias, δ), the average within-study variation in differences between test Y and X (σ^2), and the variation in bias across studies (τ^2). In this section, we focus on the case in which within each study, each patient is measured a single time. In Section 5 we relax this assumption, allowing multiple measurements.

Study specific estimators

If we assume that in study j , the individual differences D_{ij} are conditionally normally distributed (which Bland & Altman [2] argue is a reasonable assumption, even when the outcomes themselves are not normally distributed), then we can show that the sampling distribution of the study specific bias estimate D_j is itself normal with

$$E(D_j|j) = \delta_j \quad V(D_j|j) = \sigma_j^2/n_j \quad (4.1)$$

where the latter can be estimated consistently by substituting S_j^2 for σ_j^2 , and where n_j is the total number of patients in the study. Importantly, the estimated sampling variance is independent of the bias estimate, thus allowing for the use of inverse-variance weights (which are standard in meta-analysis).

Similarly, under normality we have

$$(n_j - 1)S_j^2 / \sigma_j^2 \sim \chi^2(n_j - 1) \quad (4.2)$$

which implies that

$$E(S_j^2) = \sigma_j^2 \quad V(S_j^2 | j) = 2\sigma_j^4 / (n_j - 1). \quad (4.3)$$

In this case, there are two problems for pooling these S_j^2 across studies. First, the sampling distribution of S_j^2 is not normal, particularly when the number of individuals n_j is small and second the sampling variance of the estimate is a function of the parameter.

A solution to these two problems is the log-transform, which both normalizes the distribution and stabilizes its variance. Following the delta-method¹ and results from [10], we can show that the distribution of $\log(S_j^2)$ is approximately normal with

$$\begin{aligned} E\{\log(S_j^2)\} &\approx \log(\sigma_j^2) - 1/(n_j - 1) \\ \text{Var}\{\log(S_j^2)\} &\approx 2/(n_j - 1). \end{aligned} \quad (4.4)$$

Thus an approximately unbiased estimator of $\log(\sigma_j^2)$ is

$$\log(S_j^2) + 1/(n_j - 1) \quad (4.5)$$

and the sampling variance of this estimator is no longer a function of the parameter itself.

Again, this independence between the $\log(S_j^2)$ and its sampling variance (i.e., $2/(n_j - 1)$) will be important for the implementation of inverse variance weighting in meta-analysis. Without the transformation, when σ_j^2 is overestimated (underestimated) the weights will be understated (overstated) respectively when non-transformed inverse variance estimates are used.

¹Note that we use a second order delta-method approximation to reduce the bias of the estimator, but we use a first-order approximation for the variance. Simulation results suggest that the more complex second-order variance approximation did not improve later performance of the combined LoA, so we do not include this here.

Estimating the population bias (δ)

The parameter δ can be estimated by pooling the estimates D_j across studies using random-effects meta-analysis [12]. We focus here on a random-effects approach since there is no a priori reason to believe that the true bias (i.e., average differences between tests Y and X) is identical across studies. The weighted least squares estimator of δ can be written,

$$D = \sum w_{1j} D_j / \sum w_{1j}, \quad (4.6)$$

where D_j is the estimated bias in study j . When the weights w_{1j} are defined to be inverse-variance, i.e., $w_{1j} = 1/\{S_j^2/n_j + \tau^2\}$, [13] shows that D is the minimum-variance unbiased estimator of δ , with

$$\text{Var}(D) = 1 / \sum w_{1j} \quad (4.7)$$

Furthermore, since the study specific bias estimator D_j is normally distributed, the pooled estimator D is also normally distributed and confidence intervals and hypothesis tests can be constructed using critical values from the $t(m-1)$ distribution, where m is the total number of studies.

Importantly, in the above formulation of weights, the between-study heterogeneity in bias (τ^2) must be estimated. To do so, the method-of-moments estimator is common [14] (since it provides a closed form solution and is robust to departures from normality) and is defined as,

$$T^2 = [Q_1 - (m-1)] / (R_{11} - R_{12}/R_{11}) \quad (4.8)$$

where $R_{1q} = \sum w_{1j}^q$ and $Q_1 = \sum w_{1j} (D_j - D)^2$ is Cochran's Q-statistic used for testing homogeneity.

Estimating the average study-variation (σ^2)

Recall that in each study j , the true within-study across-patient variation in test differences is σ_j^2 . Across studies, the average of this within-study variation is $E(\sigma_j^2) = \sigma^2$. Like δ , the parameter σ^2 can be estimated using random-effects meta-analysis. Since the study specific estimators S_j^2 are not normally distributed, however, this pooling is conducted on the log-transformed values instead. We can thus estimate $\log(\sigma^2)$ using,

$$\log(S^2) = \sum w_{2j} \log(S_j^2) / \sum w_{2j}, \quad (4.9)$$

where the weights w_{2j} are defined to be inverse-variance, i.e., $w_{2j} = 1/\{2/(n_j - 1) + \omega^2\}$, where $\omega^2 = \text{Var}[\log(\sigma_j^2)]$. This heterogeneity parameter ω^2 can be estimated using

$$G^2 = [Q_2 - (m - 1)] / (R_{21} - R_{22}/R_{21}) \quad (4.10)$$

where $R_{2q} = \sum w_{2j}^q$ and $Q_2 = \sum w_{2j} (\log(S_j^2) - \log(S^2))^2$.

Following the same approach as for the bias, the sampling variance of the estimate can be written,

$$\text{Var}[\log(S^2)] = 1 / \sum w_{2j} \quad (4.11)$$

and confidence intervals and hypothesis tests for $\log(\sigma^2)$ can again be constructed using critical values from the $t(m - 1)$ distribution, where m is the total number of studies. These lower and upper bounds can be back-transformed (using exponentiation) if intervals for σ^2 are desired.

Sampling distribution of between-study heterogeneity estimator (T^2)

The estimator T^2 of τ^2 is required not only indirectly for estimating the overall bias δ , but also directly as a part of the limits of agreement. In order to calculate confidence intervals for the limits of agreement, the sampling variance of T^2 is thus required. Several possible approaches are available [15 – 16]; here we investigate two options.

The first variance estimator can be written,

$$\text{Var}_1(T^2) = A_0 + A_1 \tau^2 + A_2 \tau^4 \quad (4.12)$$

where $A_0 = 2(m - 1)/(R_{11} - R_{12}/R_{11})^2$, $A_1 = 4/(R_{11} - R_{12}/R_{11})$, $A_2 = 2(R_{12} - 2R_{13}/R_{11} + R_{12}^2/R_{11}^2)/(R_{11} - R_{12}/R_{11})^2$. This estimator is derived by [15] based on the method-of-moments estimator of τ^2 and holds when $E(\eta_j) = 0$, $V(\eta_j) = \tau^2$ and the η_j 's are normally distributed.

An alternative, approach is to use the variance estimator,

$$\text{Var}_2(T^2) = 2 / \sum (V(D_j | j) + \tau^2)^{-2} = 2 / \sum (S_j^2 / n_j + \tau^2)^{-2} \quad (4.13)$$

This variance estimator is also provided by [15] and is derived in relation to the maximum likelihood estimator of τ^2 . Here we investigate this as an alternative variance estimator for the method-of-moments estimator T^2 , since it is simpler to calculate and is not as strongly a function of the parameter τ^2 as the estimator $\text{Var}_1(T^2)$.

Importantly, the sampling distribution of the estimator T^2 – unlike that of either D or $\log(S^2)$ – is not generally normally distributed. [15] shows that when $\tau^2 = 0$, $Q_1 \sim \chi^2_{(m-1)}$ and thus the distribution of T^2 follows a shifted, scaled $\chi^2(m - 1)$ distribution. When $\tau^2 \neq 0$, they

show that T^2 follows a gamma distribution, the parameters of which can be estimated with a Satterthwaite approximation and a numeric algorithm. If samples are large enough, however, [15] shows that the maximum likelihood estimator follows a normal distribution. In practice, the fact that T^2 is not normally distributed means the estimated population limits of agreement are also not exactly normally distributed, particularly when the number of studies included in the meta-analysis is small. Later we will compare the performance of these two variance estimators when creating confidence intervals for the population limits of agreement using a simulation study.

Sampling distribution of the estimate of the population limits of agreement

In a Bland-Altman meta-analysis, the goal is to estimate the population limits of agreement (LoA), defined as the interval (LoA_L , LoA_U). These can be estimated using the plug-in-estimator for both the lower- and upper-bounds as,

$$\begin{aligned} LoA_L &= D - 2\sqrt{\{exp[\log(S^2)] + T^2\}} \\ LoA_U &= D + 2\sqrt{\{exp[\log(S^2)] + T^2\}}, \end{aligned} \quad (4.14)$$

where D is estimated using Eqn 4.6, $\log(S^2)$ is estimated from Eqn 4.9, and T^2 is based on Eqn 4.8. Just as Bland and Altman [2] argued that the estimated limits should be accompanied by confidence intervals, we argue that these pooled limits of agreement should also be reported with confidence intervals.

Following the advice of [2], we focus in particular on the *outer* confidence interval values – that is, the lower bound for the lower limit of agreement (LoA_L) and the upper bound for the upper limit of agreement (LoA_U). This means we are interested in the 95% confidence bounds,

$$\begin{aligned} CI-LoA_L &= LoA_L - t(m-1, 0.025)\sqrt{Var(LoA_L)}, \\ CI-LoA_U &= LoA_U + t(m-1, 0.025)\sqrt{Var(LoA_U)}. \end{aligned} \quad (4.15)$$

where $t(m-1, .025)$ is the critical value for the t-distribution with $m-1$ degrees of freedom. We use the t – distribution here to account for greater error when the number of studies (m) is small.

Using a delta-method approximation, we can show that $V(LoA_L) = V(LoA_U) = V(LoA)$, where

$$V(LoA) \approx Var(D) + B_1 Var[\log(S^2)] + B_2 Var(T^2) \quad (4.16)$$

Where $B_1 = S^4/(S^2 + T^2)$, $B_2 = 1/(S^2 + T^2)$, and $Var(D)$, $Var[\log(S^2)]$ and $Var[T^2]$ are defined according to Eqns 4.7, 4.11, and 4.12 or 4.13 respectively. Note that when S^2 is large relative to T^2 that B_1 dominates and the last term contributes little to the overall variation.

Simulation study

The approximation given above is based on asymptotic normality and an important question is how well this approach performs in smaller samples. In Appendix A we provide results from a comprehensive simulation study. We find three important trends. First, we found that in terms of 95% confidence interval coverage for the limits of agreement, the variance estimators for T^2 led to very similar result (see Figure 2 in Appendix A, with focus on the two plots on the diagonals); since Eqn 4.13 is simpler, we therefore suggest its use. Second, when the number of studies is very small ($m = 10$), the intervals given in Eqn 4.15 tend to over-cover for small values of τ^2 (empirical coverage 0.94 – 0.98) and tend to under-cover for very high values of τ^2 (empirical coverage 0.91 – 0.96). These extremes diminish, however, as the number of studies increases. Third, when there are a moderate to large number of studies ($m = 40$) this coverage greatly improves, particularly for large values of τ^2 .

5. Pooling Estimates from Repeated Measures Designs

In Section 4 we assumed that the study specific parameters (δ_j, σ_j^2) are estimated without bias. This requirement is met when the analyses reported are based upon independent measurements or, in repeated measures designs, when the correlation between measurements is accounted for in the analysis (e.g., using methods found in [11, 17]). Unfortunately, this is often not the case in Bland-Altman studies. For example, a study by [18] included 218 paired measures calculated on only 9 individuals, yet treated the resulting estimates of bias and standard deviation as if the measurements were independent. Not taking into account the correlation between the measurements, however, results in biased estimates of each of the within-study variances (σ_j^2), and thus a biased estimate of the population limits of agreement [17].

In order to address this concern, in this section, we provide two solutions. First, a bias-adjusted estimator of the study-specific standard deviation is provided. Second, since this approach requires assumptions about the correlation structure that are impossible to test, we then provide an alternative method for calculating confidence intervals for the limits of agreement that is robust to this possible misspecification.

Multiple measurements with equal replications within studies

We begin by focusing on a single study j , which now includes a total of N_j measurements. We now assume that the total number of measurements (N_j) is evenly divided across n_j individuals (i.e., $N_j = c_j n_j$) with c_j replications each. The total within-study variation (σ_j^2) is now a combination of the variation across measurements within patients (γ_j^2) and the variation across patients (ν_j^2). Importantly, we now assume that while these N_j measurements are not independent, the correlation between measures is *not* reported. In Appendix B details of the derivations are provided.

For a study j , we assume that the within-study variation in differences between measurements Y and X is estimated with

$$S_j^2 = \sum \sum (D_{ijk} - D_j)^2 / (N_j - 1). \quad (5.1)$$

By decomposing this into two parts,

$$(N_j - 1)S_j^2 = \sum \sum (D_{ijk} - \delta_j)^2 - \sum \sum (D_j - \delta_j)^2 \quad (5.2)$$

we can show that this estimated variance S_j^2 *underestimates* the true variance σ_j^2 , i.e.

$$E(S_j^2) = [N_j / (N_j - 1)] \{V(D_{ijk}) - V(D_j)\} \leq \sigma_j^2, \quad (5.3)$$

where the last relationship is only an equality when the measurements are independent. Furthermore, we can show that

$$E(S_j^2) > \sigma_j^2 [1 - (c_j - 1) / (N_j - 1)] \quad (5.4)$$

where this bound arises when the within-person measurements are *completely* correlated. This means that a *conservative* approach – in the sense that it is better clinically to underestimate rather than over-estimate the degree of similarity between tests Y and X – is to estimate the total variation (σ_j^2) using

$$S_j^{2*} = S_j^2 / [1 - (c_j - 1) / (N_j - 1)] = S_j^2 [(N_j - 1) / (N_j - c_j)]. \quad (5.5)$$

This increases the estimate, particularly when the number of patients (n_j) is small relative to the number of measurements (c_j). For example, if there are 10 patients, each with 20 measurements (see the Example), this correction increases the variance estimate by 11%.

Finally, note that we focus on this approach since it does not require information on the degree of correlation between measurements. In Appendix B, we also provide exact formulas for removing the bias when information on the correlation structure and degree of correlation is either provided or available from other sources.

Pooling estimates based on multiple measures

Recall that the population limits of agreement are a function of three parameters: the bias (δ), the average study-specific variation (σ^2), and the variation in bias between studies (τ^2). When each person contributes a single comparison between Y and X, in Section 4 we showed that δ and σ^2 can be estimated using weighted least-squares. However, when each person is measured multiple times, the resulting (typically unknown) correlation makes it difficult to specify exactly inverse variance weights, which then calls into question

properties of the random effects meta-analysis estimators. The result is confidence interval estimates for the population limits of agreement that may be too short.

An alternative estimation strategy is to use *robust variance estimation* (RVE) [19]. This approach does not require the weights to be exactly inverse-variance (see [20] for full discussion of weights). Additionally, this approach does not require the individual differences D_{ij} to be normally distributed, and allows for the inclusion of multiple effect sizes per study. The method has been widely studied, with several simulations indicating that it has accurate Type I error rates, even when the number of studies is small (see [19, 21, 22]). To date the method has been used in over 100 published meta-analyses in a wide range of fields, including several examples in medicine (e.g., [23 – 25]).

The RVE approach can be used to estimate both δ and $\log(\sigma^2)$. To do so, the same weighted-least squares estimators found in Section 4 (Eqn 4.7 and Eqn 4.9) are used. Here any weights can be used, though following [19] approximately inverse variance weights will result in a more statistically efficient estimate. We thus suggest using the weights given in Section 4 for independent samples. Notably, these result in weights proportional to the number of patients n_j ; not the total number of measurements $(N_j = c_j^* n_j)$.

The main difference between RVE and random effects meta-analysis is in the formula for estimating the sampling variance of these estimators. For the general weights w_j and a generic pooled estimator Z , the RVE variance estimator can be written,

$$\text{Var}^R(Z) = [m/(m-1)] \sum w_j^2 (Z_j - Z)^2 / (\sum w_j)^2. \quad (5.6)$$

When estimating the bias δ , $Z_j = D_j$ and $Z = D$ while when estimating $\log(\sigma^2)$, $Z_j = \log(S_j^{2*})$ and $Z = \log(S^{2*})$. Note that if $(Z_j - Z)^2 = \text{Var}(Z_j) = 1/w_j$, then $\text{Var}^R(Z)$ would be equivalent to the model-based variance $\text{Var}(Z)$ found in Eqn 4.7 and 4.11. This specification is more general however and does not require this assumption. Using this approach, 95% confidence intervals δ and $\log(\sigma^2)$ can be calculated using the t – distribution with $m-1$ degrees of freedom [19] (though in more general meta-regression models, [21] shows that empirically estimated degrees of freedom perform better).

In this approach, the estimate of τ^2 follows from Eqn 4.8, and the estimate of the limits of agreement $\delta \pm 2 (\sigma^2 + \tau^2)$ and its confidence interval can be calculated according to Eqns 4.14 and 4.15, where now $\text{Var}^R[\log(S^{2*})]$ and $\text{Var}^R[D]$ are substituted instead. In Appendix A we provide results from a simulation study comparing the performance of this RVE approach in a wide range of conditions, comparing it head-to-head with the model-based estimator given in Eqn 4.7. In general, we find that the RVE based procedure leads to slightly lower coverage rates (see Figure 2, < 40 studies). We also find that as the number of studies increases, the procedure performs better, with only slight differences between the model-based and RVE approaches with as few as 20 studies (particularly for large values of τ^2). Like the model-based approach in the previous section, while coverage is not typically exactly 95% it is close (within the range of 0.90 – 0.98).

6. Example: Peyton & Chong (2010)

To illustrate this approach, we provide an example based on a previously conducted meta-analysis. This example provides a re-analysis of a meta-analysis of a new technology for cardiac output measurement [6]. This meta-analysis is highly cited (over 175 citations) and includes data from 52 studies. All analyses presented here were conducted in the statistical program **R** [26]. The **R** code is available as an online supplement.

This meta-analysis compared non-invasive and invasive techniques for cardiac output measurement [6]. For the sake of example, we focus only on studies using the pulse contour technique. In the original analysis, two measures were provided: a pooled estimate of the bias and a pooled estimate of the standard deviation. The pooled estimators, like ours provided in Eqn 4.7 and 4.9, were weighted averages; in their approach, however, the weights were based on sample size only (i.e., $w_j = n_j - 1$). The paper does not report estimates of the limits of agreement or measures of error (i.e., confidence intervals or standard errors) for any of the pooled estimators.

While the synthesis identified 52 studies reporting measures of bias and standard deviation (which they call “precision”), the original analysis only pooled data from 24 of these studies. The remaining studies – while reported in a table – were not included since the primary analyses were based on repeated measures designs that were improperly analyzed. Data from these 52 studies is included in Table 1.

In the analyses given here – and the data reported in Table 1 – we include a single set of measurements per study; when multiple were listed in the original meta-analysis, we include here the one that was included in the original meta-analysis (which privileged the first measure taken). Additionally, for the 28 studies originally excluded from the meta-analysis, we returned to the original articles and extracted information on the total number of patients (n_j).

We include three analyses here to illustrate our methods. First, we provide estimates of the limits of agreement and their lower- and upper- confidence bands for the 24 studies included in the original meta-analysis; these are based on the random effects meta-analysis methods found in Section 4 of this paper. Second, we provide estimates of the confidence bands for the LoA using the robust variance estimator, to compare these results to the model-based results. Third, we include estimates from the full set of 52 studies identified in Table 1. For this analysis, we use the adjusted estimator of σ_j^2 (i.e., S_j^{2*}). This is conservative in the sense that it may provide a slight over-estimate of the value of σ_j^2 when the correlation between measurements is not as strong. For this analysis, we calculate the confidence bands using the RVE method.

Results from the example are found in Table 2. These analyses show three trends. First, notice that summarizing the results of the Bland-Altman meta-analysis in terms of limits of agreement allows for easier interpretation and offers a direct parallel to the interpretation of findings from individual Bland-Altman studies. For example, the results indicate that with all data included, the two methods are estimated to differ from -2.22 to 2.18 units across *all*

patients studied. Importantly, this final interval is based on confidence intervals, which includes an accounting of sampling error.

Second, while some meta-analyses (e.g., [4]) do summarize the results in terms of limits of agreement, these studies do not include between-study variation in bias (i.e., τ^2) in these limits. In the example given here, this between-study variation in bias accounts for between 10 – 15% of the total variation, leading to total standard deviations (i.e., $(\sigma^2 + \tau^2)$) between 5 – 9% larger. Previous meta-analyses have rarely included measures of error in these limits, whereas the approach given here includes estimation of lower- and upper- confidence intervals for these limits. Both the inclusion of between-study heterogeneity and measures of sampling variation result in limits of agreement that are *wider*.

Third, many studies included in Bland-Altman meta-analyses include estimates based on repeated measures data that may not be properly accounted for in the analysis. Studies of this type accounted for 56% of the total number of studies in the example. As found in this example, it is common for primary studies to report all estimates of bias and standard deviation but to exclude those from repeated measures designs from the pooled estimators. As the analyses in Tables 2 indicate, excluding these studies can affect the conclusions. Moving from the second to third rows in the table indicates that including repeated measures designs can affect all parameters estimated, and not always in the same direction. Including more studies results in tighter confidence bands around the limits of agreement – because of the increase in the sample size – though the pay-off to additional studies is sometimes reduced by the increased measures of variation (σ^2 and/or τ^2).

Finally, this increased heterogeneity is especially important for the interpretation of findings from a Bland-Altman meta-analysis. Following the approach given in [8], the original meta-analysis [6] noted that Thermodilution would only be acceptable if the percent error was less than 30% (where % error = LoA/mean cardiac output). Using the mean cardiac output reported in [6] of 5.78, this results in approximately 38% error for all three models estimated here, indicating that this test is not an acceptable alternative.

7. Discussion

This paper provides a framework for combining the results of multiple Bland-Altman studies on the same topic. Given the large number of Bland-Altman studies conducted in medicine, our goal has been to provide a method for making comparisons across all patients studied, thus providing practitioners with a better sense of the usefulness of new, non-invasive measurement methods. As the example indicates, meta-analyzing Bland-Altman studies can lead to important new conclusions, as well as questions, regarding the efficacy of a test in different circumstances.

This paper makes three important contributions. First, we show that the most relevant parameters in a Bland-Altman meta-analysis are those focused on understanding the distribution of comparisons across *patients*, not studies. In this framework, this goal is to estimate the combined limits of agreement and to provide a measure of precision for these estimated bounds. This contrasts with previous meta-analyses of Bland-Altman studies,

which to date have focused on estimating the bias and/or the average standard deviation across studies, often without efforts to quantify the degree of uncertainty in these estimates. Constructing these limits of agreement requires considering not only the average standard deviation, but also the variation in bias across studies. We also provide a method for calculating confidence intervals for these limits of agreement, thus taking into account sampling error. In combination, the interval that results is wider than that typically reported in current Bland-Altman meta-analyses.

Second, we provide an approach to include data from primary studies that are conducted using a repeated measures design, but that do not properly take this design into account in the analysis. When a patient is measured multiple times, the measures are correlated, though the degree of this correlation is not typically reported. We show that when measures are correlated, the sample standard deviation always under-estimates the true standard deviation and provide an adjustment that reduces this bias. This allows the population limits of agreement to be estimated based upon a larger set of data. The estimator we provide is conservative in that it tends to over-estimate (never under-estimating) the total standard deviation, thus resulting in slightly wider limits of agreement. We also provide an approach to estimating confidence bands for these limits based upon robust variance estimation (RVE), which does not require as strong of assumptions as standard random-effects meta-analysis.

Third, the approach we've developed here has applications beyond meta-analysis to the analysis of repeated measures data in Bland-Altman studies more generally. To see how, treat the individual patients as “studies”, now letting measurements act as “patients”. Since full information on the correlation structure is now available a covariance structure can be specified and separate estimates of each variance component (e.g., across measurements, across people) can be calculated. In the approach given in this paper, these measurements would be aggregated to a single measure per person before being combined to calculate the limits of agreement using the RVE approach. A benefit to this approach relative to other Bland-Altman repeated measures approaches is that it requires fewer assumptions.

There are many avenues for future research in this area. This paper has developed an approach that requires only a single set of measurements per study. In practice, however, Bland-Altman studies routinely report multiple bias and standard deviation estimates; for example, this might include measurements on different subgroups and/or at different time-points taken on the same patients. Additionally, we have focused here on models aimed at estimating the limits of agreement for the full population or subgroups, but not on methods for making comparisons across these groups. For this meta-regression models would also be needed. Both concerns can be addressed more generally using either multivariate meta-analysis or – extending the approach given here – through a more general implementation of robust variance estimation (see [19]). To do so will require the development of properties of the variance component estimates in these models, as well as a larger discussion on how to best report and interpret the effects of covariates and subgroups on not only *mean* differences (which is standard), but on *distributional* differences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was partly supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A: Simulation Study

Under normal errors the sampling distributions for both the pooled bias D and the pooled within-study variation $\log(S^2)$ are well established. However, the confidence interval procedures given Eqn 4.15 require three approximations. First, the sampling distribution of the between study heterogeneity estimator T^2 , while asymptotically normal, is not necessarily normal in small samples [15]; we have also presented two different options for estimating the variance of this statistic (Eqns 4.12 and 4.13). Second, the variance estimator for each of the LoA is a delta-method approximation. Third, we have provided two estimators for the variance of D and $\log(S^2)$, one based on known variances and weights (“model-based”) and one based on a robust alternative (RVE).

In order to investigate the performance of the confidence interval procedures proposed here, we conducted a simulation study varying five parameters: m (the number of studies), n_j (the sample size in each study), σ^2 , ω^2 (the variation in $\log(\sigma_j^2)$ across studies), and $\rho = \tau^2/(\sigma^2 + \tau^2)$. Note that we set $\delta = 0$ for all of these simulations, since results for the sampling distribution of D are well known. For each parameter, we focused on both small values as well as more moderate values and on values common in the literature (see the values given in Table 1, for example). Here these take values $m = 10, 20$, and 40 ; $n = 10, 20$, and 40 ; $\sigma^2 = 0.5, 1, 2$; and $\omega^2 = 0, 0.001^2, 0.1^2$, and 0.2^2 ; and $\rho = 0.01, 0.05, 0.10, 0.20$. This choice of values resulted in 12 different combinations of τ^2 values, which we categorized into 4 groups: less than $< .055$ (“very small”), $0.10 - 0.20$ (“small”), $0.20 - 0.30$ (“medium”), and 0.50 (“large”).

For each patient i in study j , based on the relationship,

$$\delta_{ij} = \delta + \eta_j + \varepsilon_{ij},$$

we simulated data such that $\eta_j \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, where

$$\log(\sigma_j^2) = \log(\sigma^2) + \nu_j$$

where $\nu_j \sim N(0, \omega^2)$. We focused our simulations on the normal data situation since the limits of agreement are defined based upon this assumption. In other situations, they act only as “reference” distributions, providing information on what the limits would be if the errors were actually normal. For each set of parameter values, we calculated 95% confidence

intervals for the LoA in four ways. These included two different estimators of $V(T^2)$ and two different estimators of $V(D)$ and $V(\log(S^2))$ – one model-based and one based on RVE.

We repeated this procedure for each combination 5,000 times and summarized the results in terms of the empirical coverage rate. The results are given in Figures 2 and 3. In both figures, horizontal gray lines indicate the true 0.95 coverage rate and the gray lines indicate simulation error. Figure 2 gives two results. First, the model-based variance estimator (of Eqn 4.7 and 4.11) tends to have higher coverage than the RVE based estimator (of Eqn 5.6), particularly in small samples (not shown here). Second, the two possible variance estimators for T^2 (Eqns 4.12, 4.13) give very similar results.

Figure 3 compares results for both the model-based (in white) and RVE-based (in gray) estimators with $V(T^2)$ based on Eqn 4.9b. These are divided in panels in terms of number of studies and within panels by the value of τ^2 . There are three trends. First, differences between the model- and RVE-based estimators are larger in smaller sample sizes, and when they differ, the RVE-based estimator tends to have smaller coverage. Second, coverage varies highly in relation to the value of τ^2 (over 50% of the variation in coverage is explained by this). The worst case is when the number of studies is small and the variation τ^2 is large, in which case the confidence intervals tend to under-cover (but typically have coverage greater than 0.90). Third, when the number of studies is moderate to large ($m = 40$), coverage begins to stabilize and is typically (but not always) between 0.94 and 0.97.

APPENDIX B: Derivations and Proofs

We derive these results under a general covariance model that may arise in repeated measures data. This model includes both variation in patient-specific effects and auto-correlated AR(1) errors. For measurement $k = 1, \dots, c_j$ for patient $i = 1, \dots, n_j$ in a fixed study j , we have

$$D_{ijk} = \delta + \eta_{ij} + \phi(D_{ijk-1} - \delta) + \alpha_{ijk}$$

where $\eta_{ij} \sim N(0, \nu_j^2)$, $\alpha_{ijk} \sim WN(0, \sigma_w^2)$, $\text{Var}(D_{ijk}|i, j) = \sigma_w^2 / (1 - \phi)^2 = \gamma_j^2$, the total variation $\sigma_j^2 = \nu_j^2 + \gamma_j^2$, and $\rho_j = \nu_j^2 / \sigma_j^2$. Note here that $\rho_j \in (0, 1)$ is the correlation between patient true values of δ_{ij} across measurements and $|\phi_j| < 1$ is the correlation between errors one time-point apart.

This model reduces to the *compound symmetry* model when $\phi_j = 0$. In that case, the correlation between measurements is simply ρ_j . Conversely, when $\rho_j = 0$ the model reduces to the AR(1) auto-correlation model, where the correlation between measurements on the same person h time points apart is ϕ_j^h . When $\rho_j = 0$ and $\phi_j = 0$ the model includes two sources of variation and covariance.

Result

Let $D_j = \sum \sum D_{ijk} / N_j$, where $N_j = c_j n_j$, and $S_j^2 = \sum \sum (D_{ijk} - D_j)^2 / (N_j - 1)$. Then we can show that

- a. $E(D_j) = \delta_j$
- b. $E(S_j^2) \geq \sigma_j^2 [1 - (c_j - 1)/(N_j - 1)]$

when the covariance model is as defined above.

Proof

Result (a) is a straightforward application of the law of iterated expectations. For (b) note that since

$$(N_j - 1)S_j^2 = \sum \sum (D_{ijk} - D_j)^2 = \sum \sum (D_{ijk} - \delta_j)^2 - \sum \sum (D_j - \delta_j)^2,$$

then

$$(N_j - 1)E(S_j^2) = N_j [\text{Var}(D_{ijk}) - \text{Var}(D_j)],$$

and thus we need to know $\text{Var}(D_j)$. Note that we can write,

$$\text{Var}(D_j|j) = \sum \text{Var}(\sum D_{ijk}|j)/N_j^2 = \sum \{ \sum \text{Var}(D_{ijk}|j) + \sum \sum \text{Cov}(D_{ijk}, D_{ijk},) \} / N_j^2$$

Now note that $\text{Var}(D_{ijk}) = \sigma_j^2$ and $\text{Cov}(D_{ijk}, D_{ij(k-h)}) = \sigma_j^2 [(1 - \rho_j)\phi_j^h + \phi_j]$, and thus

$$\text{Var}(D_j|j) = (\sigma_j^2/N_j) [(2/c_j)(1 - \rho_j) \sum (c_j - h)\phi_j^h + (2\rho_j/c_j) \sum (c_j - h) - 1].$$

This can be further simplified by noting that,

$$\sum (c_j - h)\phi_j^h = c_j \sum \phi_j^h - \phi_j (d/d\phi_j) \sum \phi_j^h = (c_j - (c_j + 1)\phi_j + \phi_j^{c_j+1})/(1 - \phi_j)^2$$

which falls from properties of finite sums of geometric series, and since

$$\sum (c_j - h) = c_j(c_j + 1)/2.$$

This results in the variance,

$$\text{Var}(D_j|j) = \sigma_j^2 [(2A_j - 1) + \rho_j(c_j + 1 - 2A_j)]/N_j, \text{ where } A_j = [c_j - (c_j + 1)\phi_j + \phi_j^{c_j+1}]/[c_j(1 - \phi_j)^2].$$

Note that when $\phi_j = 0$,

$$\text{Var}(D_j|j) = \sigma_j^2 [1 + \rho_j(c_j - 1)]/N_j$$

and when $\rho_j = 0$,

$$\text{Var}(D_j|j)=\sigma_j^2[2A_j - 1]/N_j$$

Putting this together, we have

$$E(S_j^2)=\sigma_j^2[1 - \{2(A_j - 1)+\rho_j(c_j+1 - A_j)\}/(N_j - 1)].$$

Now the inequality is given by maximizing the portion in $\{.\}$, which occurs in limit as $\phi_j \rightarrow 1$ and $\rho_j \rightarrow 1$.

References

1. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *The Statistician*. 1983; 32:307–17.
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; i:307–310.
3. Chong SW, Peyton PJ. A meta-analysis of the accuracy and precision of the ultrasonic cardiac output monitor (USCOM). *Anaesthesia*. 2012; 67:1266–71. [PubMed: 22928650]
4. Kim S, Lilot M, Sidhu KS, Rinehart J, Yu Z, Canales C, Cannesson M. Accuracy and precision of continuous noninvasive arterial pressure monitoring compared with invasive arterial pressure: A systematic review and meta-analysis. *Anesthesiology*. 2014; 120:1080–97. [PubMed: 24637618]
5. Jensen LA, Onyskiw JE, Prasad NGN. Meta-analysis of arterial oxygen saturation monitoring by pulse oximetry in adults. *Heart and Lung*. 1998; 27:387–408. [PubMed: 9835670]
6. Peyton PJ, Chong SW. Minimally invasive measurement of cardiac output during surgery and critical care: A meta-analysis of accuracy and precision. *Anesthesiology*. 2010; 113:1220–35. [PubMed: 20881596]
7. Dorosz JL, Lezotte DC, Weitzkamp DA, Allen LA, Salcedo EE. Performance of 3-dimensional echocardiography in measuring left ventricular volumes and ejection fraction a systematic review and meta-analysis. *Journal of the American College of Cardiology*. 2012; 20:1799–808.
8. Critchley LAH, Critchley JAJH. A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *J Clin Monit*. 1999; 15:85–91.
9. Williamson PR, Lancaster GA, Craig JV, Smyth RL. Meta-analysis of method comparison studies. *Statistics in Medicine*. 2002; 21(14):2013–25. [PubMed: 12111884]
10. Rice MJ, Shuster JJ, Tipton E, Gravenstein N, Coursin DB, Wenzel V, Morey TE. Meta-analysis of Bland-Altman studies: noninvasive hemoglobin device accuracy. Under review.
11. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*. 1999; 8:135–160. [PubMed: 10501650]
12. Borenstein, M., Hedges, LV., Higgins, JPT., Rothstein, HR. *Introduction to Meta-Analysis*. Wiley; New York: 2009.
13. Hedges, LV., Olkin, I. *Statistical Methods for Meta-Analysis*. Academic Press; New York: p. 1985
14. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986; 7:177–188. [PubMed: 3802833]
15. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*. 1997; 16:753–768. [PubMed: 9131763]
16. Higgins J, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*. 2002; 21(11):1539–58. [PubMed: 12111919]
17. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*. 2007; 17:571–582. [PubMed: 17613642]
18. Opdam HI, Wan L, Bellomo R. A pilot assessment of theFloTrac cardiac output monitoring system. *Intensive CareMed*. 2007; 33:344–9 301.

19. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*. 2010; 1(1):39–65. [PubMed: 26056092]
20. Tanner-Smith EE, Tipton E. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research synthesis methods*. 2014; 5(1):13–30. [PubMed: 26054023]
21. Tipton E. Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*. 2013; 4(2):169–87. [PubMed: 26053656]
22. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*. 2015; 20(3):375. [PubMed: 24773356]
23. Sherman AC, Merluzzi TV, Pustejovsky JE, Park CL, Fitchett G, Jim HS, Munoz AR, Danhauer SC, Snyder MA, Salsman JM. A meta-analytic review of religious or spiritual involvement and social health among cancer patients. *Cancer*. 2015; 121(21):3779–88. [PubMed: 26258730]
24. Ding M, Bhupathiraju SN, Satija A, van Dam RM, Hu FB. Long-term coffee consumption and risk of cardiovascular disease: a systematic review and a dose-response meta-analysis of prospective cohort studies. *Circulation*. 2014; 129:643–59. [PubMed: 24201300]
25. Parkinson JR, Hyde MJ, Gale C, Santhakumaran S, Modi N. Preterm birth and the metabolic syndrome in adult life: a systematic review and meta-analysis. *Pediatrics*. 2013:e1240–e1263. [PubMed: 23509172]
26. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. URL <http://www.R-project.org/>

Figure 1.
Comparison of Pooled Distributions for All Patients versus Average Study

Figure 2.
Comparison of LOA Coverage (95% CI) by Method and Estimator

Figure 3.
Comparison of LOA Coverage (95% CI) by Estimator, Number of Studies, and τ^2

Table 1

Data for Example (from Peyton and Chong)

Sample	Study Name	Included in PC analysis	N	n	Bias	SD
1	MAYER_1	1	23	23	0.18	0.27
2	MAYER_2	1	15	15	0.13	0.30
3	SENN ET AL_1	1	25	25	-0.10	0.80
4	SENN ET AL_2	1	25	25	-0.30	0.60
5	BIANCOFIORE	0	290	29	-1.30	1.40
6	OSTERGAARD	1	25	25	0.51	0.93
7	MUTOH	0	179	16	-1.14	0.88
8	COMPTON	1	25	25	-1.90	1.94
9	DELLA ROCCA	0	126	18	-0.95	1.41
10	MAYER	1	40	40	0.04	0.29
11	MEHTA	1	12	12	-0.54	0.56
12	ZIMMERMAN	1	30	30	0.73	1.29
13	STAIER	1	30	30	0.16	0.70
14	MCGEE	0	561	84	0.20	1.28
15	CANNESON	0	166	11	0.26	0.87
16	SAKKA	1	24	24	-0.87	2.30
17	DE WAAL	1	22	22	-0.08	0.90
18	PRASSER	0	164	20	-0.02	1.46
19	MAYER	1	40	40	0.52	0.57
20	MANECKE	0	295	50	0.55	0.98
21	BREUKERS	0	56	20	0.14	1.00
22	BUTTON	1	31	31	0.60	0.9
23	CHAKRAVARTHY	0	438	15	0.15	0.33
24	OPDAM	0	218	9	0.01	0.60
25	SANDER	1	30	30	-0.2	1.40
26	BREUKERS	0	119	17	0.36	1.42
27	SENN	0	200	50	-0.20	0.70
28	COMPTON	0	102	49	-2.03	1.34
29	YAMASHITA	1	20	20	-0.31	0.46

Sample	Study Name	Included in PC analysis	N	n	Bias	SD
30	DE WILDE	0	199	24	-0.14	0.87
31	BUTTON	0	186	31	0.50	0.75
32	CHAKRAVARTHY	0	438	15	-0.13	1.12
33	OSTERGAARD	1	25	25	0.07	1.10
34	FELBINGER	0	53	17	0.44	0.52
35	SANDER	1	45	45	-1.40	1.70
36	WOUTERS	0	224	23	1.08	0.75
37	DE VAAL	1	24	24	0.37	1.58
38	DELLA ROCCA	1	58	58	0.26	0.57
39	TZENKOV	1	35	35	0.71	1.15
40	MIELCK	0	96	22	-0.40	1.39
41	DELLA ROCCA	1	62	62	0.02	0.74
42	ROMANO	1	18	18	-0.15	0.35
43	FELBINGER	0	360	20	0.62	0.91
44	RAUCH	0	380	25	-0.14	1.16
45	SEGAL	0	190	22	0.27	0.67
46	JANSEN	0	490	54	-0.10	0.50
47	HIRSCHL	1	29	29	0.65	1.20
48	ZOLLNER	0	228	19	0.31	1.25
49	MISSANT	0	149	20	-0.03	0.65
50	COSTA	0	151	23	-0.29	1.08
51	DE WILD	0	199	24	0.17	0.69
52	GARCIA RODRIGUEZ	0	402	24	-0.54	0.95

* Included in PC Analysis indicates if the study was included in pooled estimates of bias and standard deviation in the original meta-analysis.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Comparison of Meta-analysis Results for Example

Method	Studies	Bias	SD	τ^2 est.	LoA _L	LoA ^U	CI _L	CI ^U
Indpt + REMA	24	.03	.82	.09	-1.71	1.77	-2.13	2.19
Indpt + RVE	24	.03	.82	.09	-1.71	1.77	-2.16	2.22
All + RVE	52	-.02	.89	.15	-1.95	1.91	-2.25	2.20

Note: 'Indpt' means only primary studies reporting data from independent patients were included, while 'All' also includes primary studies with data from repeated measures designs.