A Bayesian Approach for Sample Size Determination in Method Comparison Studies

Kunshan Yin[a], Pankaj K. Choudhary[a,1], Diana Varghese[b] and Steven R. Goodman[b]

[a]Department of Mathematical Sciences

[b]Department of Molecular and Cell Biology

University of Texas at Dallas

Abstract

Studies involving two methods for measuring a continuous response are regularly conducted in health sciences to evaluate agreement of a method with itself and agreement between methods. Notwithstanding their wide usage, the design of such studies, in particular, the sample size determination, has not been addressed in the literature when the goal is simultaneous evaluation of intra- and inter-method agreement. We fill this need by developing a simulation-based Bayesian methodology for determining sample sizes in a hierarchical model framework. Unlike a frequentist approach, it takes into account of uncertainty in parameter estimates. This methodology can be used with any scalar measure of agreement available in the literature. We demonstrate this for four currently used measures. The proposed method is applied to an ongoing proteomics project, where we use pilot data to determine the number of individuals and the number of replications needed to evaluate agreement between two methods for measuring protein ratios. We also apply our method to determine sample size for an experiment involving measurement of blood pressure.

Key Words: Agreement; Concordance correlation; Fitting and sampling priors; Hierarchical model; Sample size determination; Tolerance interval; Total deviation index.

[1] Corresponding author. Address: EC 35, PO Box 830688; University of Texas at Dallas; Richardson, TX 75083-0688; USA. Email: pankaj@utdallas.edu; Tel: (972) 883-4436; Fax: (972) 883-6622

1 Introduction

Comparison of two methods for measuring a continuous response variable is a topic of considerable interest in health sciences research. In practice, a method may be an assay, a medical device, a measurement technique or a clinical observer. The variable of interest, e.g., blood pressure, heart rate, cardiac stroke volume, etc., is typically an indicator of the health of the individual being measured. Hundreds of method comparison studies are published each year in the biomedical literature. Such a study has two possible goals. The first is to determine the extent of agreement between two methods. If this inter-method agreement is sufficiently good, the methods may be used interchangeably or the use of the cheaper or the more convenient method may be preferred. The second goal is to evaluate the agreement of a method with itself. The extent of this intra-method agreement gives an idea of the repeatability of a method and serves as a baseline for evaluation of the inter-method agreement.

The topic of how to assess agreement in two methods has received quite a bit of attention in the statistical literature. See [1] for a recent review of the literature on this topic. There are several measures of agreement — such as limits of agreement [2], concordance correlation [3], mean squared deviation [4], coverage probability [5] and total deviation index or tolerance interval [4, 6], among others. On the other hand, the topic of how to plan a method comparison study — in particular, the sample size determination (SSD), has not received the same level of attention. Although several authors [5, 6, 7, 8] provide frequentist sample size formulas associated with agreement measures of their interest, they are restricted to only the evaluation of inter-method agreement. However, a simultaneous evaluation of both intra- and inter-method agreement is crucial because the amount of agreement possible between two methods is limited by how much the methods agree with themselves. A new method, which is perfect and worthy of adoption, will not agree well with an established standard if the latter has poor agreement with itself [9]. But we are not aware of any literature on SSD when the interest lies in evaluating both intra- and inter-method agreement. This involves determining the number of replicate measurements in addition to the number of individuals for the study.

Our aim in this article is to fill this gap in the literature. We develop a methodology for determining the number of individuals and the number of replicate measurements needed for a method comparison study to achieve a given level of precision of simultaneous inference regarding intra- and inter-method agreement. We take a Bayesian route to this SSD problem as it allows us to overcome limitations of frequentist procedures. In the frequentist paradigm, one generally casts the problem of agreement evaluation in a hypothesis testing framework, and then derives an approximate sample size formula by performing an asymptotic power analysis along the lines of [10, sec 3.3]. But this formula involves the asymptotic standard error of the estimated agreement measure, which is a function of the unknown parameters in the model. So typically one obtains their estimates, either from the literature or through a pilot study, and substitutes them into the SSD formula. However, by merely inserting their values in the formula no account is taken of the variability in the estimates. This issue can be addressed in a Bayesian paradigm. See, e.g., the review articles [11, 12] and the references therein.

Our approach is to use a suitable feature of the posterior distribution of the agreement measure of interest as the measure of precision of inference, which is a monotonic function of number of individuals and number of replications. This property is used to derive an SSD procedure. We do not have explicit sample size formulas. Instead, the procedure has to be implemented via simulation. It can be used in conjunction with any scalar measure of agreement currently available in the literature.

We now introduce two examples that we use to illustrate the application of the proposed methodology. In both cases, we treat the available data as pilot data and use the estimates obtained from them to determine sample sizes for future studies. These examples represent two qualitatively different method comparison studies. As shown in Section 3, in the first example, the measurements on the same individual, whether from the same method or two different methods, appear only weakly correlated. This scenario is somewhat unusual. In contrast, the measurements are highly correlated in the second example, which is the more typical scenario. These distinct scenarios allow us to get a better insight into the workings of the proposed methodology.

Example 1 (Protein ratios): In this ongoing project, we are interested in SSD for a study to compare two softwares — XPRESS [13] and ASAPRatio [14] for computing abundance ratios of proteins in a pair of blood samples. The former is labor intensive and requires much user input, whereas the latter is automated to a large extent. These softwares are used in proteomics for comparing protein profiles under two different conditions to discover proteins that may be differentially expressed. See, e.g., [15] for a general introduction to proteomics. The response of specific interest to us is the protein abundance ratios in blood samples of two healthy individuals. These ratios are expected to be near one since both the samples come from healthy people. We have ratios of 8 proteins in blood samples of 5 pairs of individuals computed using the two softwares. However, only a total of 60 ratios are available as not all proteins are observed in every sample. These data were collected in the laboratory of the last author. □

Example 2 (Blood pressure): In this case, we consider the blood pressure data of [16]. Here systolic blood pressure (in mmHg) is measured twice, each using two methods — a manual mercury sphygmomanometer and an automatic device, OMRON 711, on a sample of 384 individuals. The manual device serves as the reference method and the automatic device is the test method. □

This article is organized as follows. In Section 2, we describe the proposed SSD methodology. Its properties are investigated in Section 3, and its application to the two examples introduced above is illustrated in Section 4. Section 5 concludes with a discussion.

All the computations reported in this article were performed using the statistical package `R` [17]. We also used the `WinBUGS` [18] package for fitting models to the pilot data sets by calling it from `R` through the `R2WinBUGS` [19] package.

2 SSD for method comparison studies

Let y_{ijk}, $k = 1, \ldots, n$, $j = 1, 2$, $i = 1, \ldots, m$, represent the k-th replicate measurement from the j-th method on the i-th individual. Here m is the number of individuals and n is the common number of replicate measurements from each method on every individual. A common number of replicates is assumed to simplify the SSD problem. Replicate measurements refer to multiple measurements taken under identical conditions. In particular, the underlying true measurement does not change during the replication.

Any SSD procedure depends on the model assumed for the data, the inference of interest, and the measure of precision of inference. So we first describe them in Sections 2.1, 2.2 and 2.3, respectively. Section 2.4 explains how to compute the precision measure. Finally, Section 2.5 describes the SSD procedure.

2.1 Modelling the data

We assume that the data follow the model

$$y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk},\ k = 1, \ldots, n,\ j = 1, 2,\ i = 1, \ldots, m, \tag{1}$$

where β_j is the effect of the j-th method; b_i is the effect of the i-th individual; and b_{ij} is the individual × method interaction. Oftentimes, there is no need for this interaction term. Moreover, when $n = 1$, this term is not identifiable and we drop it from the model. We assume that $b_i|\psi^2 \sim$ independent $\mathcal{N}(0, \psi^2)$, $b_{ij}|\phi^2 \sim$ independent $\mathcal{N}(0, \phi^2)$ and $\epsilon_{ijk}|\sigma_j^2 \sim$ independent $\mathcal{N}(0, \sigma_j^2)$. In addition, these random variables are assumed to be mutually independent. In what follows, we use $\boldsymbol{\gamma}$ as a generic notation for the vector of all relevant parameters without identifying them explicitly. They will be clear from the context.

Let the random vector (y_1, y_2) denote the bivariate population of measurements from two methods on the same individual. Also, let (y_{j1}, y_{j2}) denote the bivariate population of two replicate measurements from the j-th method on the same individual. The model (1) postulates that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} |\boldsymbol{\gamma} \sim \mathcal{N}_2\left(\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} \psi^2 + \phi^2 + \sigma_1^2 & \psi^2 \\ \psi^2 & \psi^2 + \phi^2 + \sigma_2^2 \end{bmatrix}\right),$$

$$\begin{bmatrix} y_{j1} \\ y_{j2} \end{bmatrix} |\boldsymbol{\gamma} \sim \mathcal{N}_2\left(\begin{bmatrix} \beta_j \\ \beta_j \end{bmatrix}, \begin{bmatrix} \psi^2 + \phi^2 + \sigma_j^2 & \psi^2 + \phi^2 \\ \psi^2 + \phi^2 & \psi^2 + \phi^2 + \sigma_j^2 \end{bmatrix}\right),\ j = 1, 2. \tag{2}$$

Next, let $d_{12} = y_1 - y_2$ and $d_{jj} = y_{j1} - y_{j2}$ respectively denote the population of inter-method differences and the population of intra-method differences for the j-th method. It follows that

$$d_{12}|\boldsymbol{\gamma} \sim \mathcal{N}(\mu_{12} = \beta_1 - \beta_2, \tau_{12}^2 = 2\phi^2 + \sigma_1^2 + \sigma_2^2), \quad d_{jj}|\boldsymbol{\gamma} \sim \mathcal{N}(0, \tau_{jj}^2 = 2\sigma_j^2),\ j = 1, 2. \tag{3}$$

To complete the Bayesian specification of model (1), we assume the following mutually independent prior distributions:

$$\beta_j \sim \mathcal{N}(0, V_j^2),\ \psi^2 \sim IG(A_\psi, B_\psi),\ \phi^2 \sim IG(A_\phi, B_\phi),\ \sigma_j^2 \sim IG(A_j, B_j),\ j = 1, 2, \tag{4}$$

where $IG(A, B)$ represents an inverse gamma distribution — its reciprocal follows a gamma distribution with mean A/B and variance A/B^2. All the hyperparameters in (4) are positive and need to be specified. Choosing large values for V_j^2 and small values for A's and B's typically lead to noninformative priors. For β_j, an alternative noninformative choice for prior is an improper uniform distribution over the real line. It can also be thought of as taking $V_j^2 = \infty$ in (4). The resulting posterior is proper for fixed values of the remaining hyperparameters. These prior choices are quite standard in the Bayesian literature [20, ch 15]. See [21, 22] for some recent suggestions for priors on variance parameters. Under the prior (4), the joint posterior of the parameters is not available in a closed-form. So we use the Gibbs sampler [20, ch 11] — a Markov chain Monte Carlo (MCMC) algorithm, described in the Appendix, to simulate draws from the posterior distribution.

2.2 Evaluation of agreement

Let θ_{12} be a measure of agreement between two methods and θ_{jj} be the corresponding measure of intra-method agreement for the j-th method, $j = 1, 2$. The former is a function of parameters of the distribution of (y_1, y_2), whereas the latter is the same function of parameters of the distribution of (y_{j1}, y_{j2}). These distributions are given in (2). We assume that θ_{12} and θ_{jj} are scalar and non-negative, possibly after a simple monotonic transformation, and their low values indicate good agreement. All the agreement measures mentioned in the introduction satisfy these criteria except the limits of agreement, which quantifies agreement using a lower limit and an upper limit.

Let U_{12} and U_{jj} be upper credible bounds for θ_{12} and θ_{jj}, respectively, each with $(1-\alpha)$ posterior probability. The agreement is considered adequate when these bounds are small. To compute them, we first obtain a large number of draws from the joint posterior of the model parameter vector $\boldsymbol{\gamma}$ by fitting the model (1) to the data and using (4) as the prior distribution. Then, noting that θ_{12} and θ_{jj} are functions of $\boldsymbol{\gamma}$, we use the draws of $\boldsymbol{\gamma}$ to get draws from the posteriors of θ_{12} and θ_{jj}. Their $(1-\alpha)$-th quantiles are taken as the bounds U_{12} and U_{jj}, respectively.

2.3 A measure of precision of inference

Consider for now only the evaluation of the inter-method agreement using the credible bound U_{12} for θ_{12}. It satisfies $F_{12|\mathbf{y}}(U_{12}) = 1-\alpha$, where $\mathbf{y}$ represents the observed data, and $F_{12|\mathbf{y}}$ represents the posterior cumulative distribution function (cdf) of θ_{12}. Although this condition ensures $\{\theta_{12} \leq U_{12}\}$ with a high posterior probability, it says nothing about how the posterior density of θ_{12} is distributed near U_{12}. In particular, if a large portion of this density is concentrated in a region far below U_{12}, then the inference is not very precise.

Various suggestions have been made in the literature for measuring the precision of a credible interval. See [12] for a summary. But they are relevant only for a two-sided interval, and not an one-sided bound such as U_{12}, which is of interest in this article. So below we develop a new measure of precision that is appropriate for upper bounds. Similar arguments can be used to develop a precision measure for lower bounds.

It is well-known that, under certain regularity conditions, the posterior distribution of a population parameter converges to a point mass as m increases to infinity [20, ch 4]. This suggests that one can focus on the posterior probability $Pr(\delta U_{12} \leq \theta_{12} \leq U_{12}|\mathbf{y})$, for a specified $\delta \in (0,1)$, as a measure of precision. Further, considering that SSD takes place prior to data collection, our proposal is to take $E\{Pr(\delta U_{12} \leq \theta_{12} \leq U_{12}|\mathbf{y})\} = 1 - \alpha - E\{F_{12|\mathbf{y}}(\delta U_{12})\}$, or equivalently

$$T_{12}(m, n, \delta) = E\{F_{12|\mathbf{y}}(\delta U_{12})\}, \tag{5}$$

as the measure of precision of U_{12}. The expectation here is taken over the marginal distribution of $\mathbf{y}$. A small value for T_{12} indicates high precision of inference since it implies a small probability mass below δU_{12}. In other words, a large probability mass is concentrated between δU_{12} and U_{12}. Also, for a fixed (m, n), T_{12} is an increasing function of δ — increasing from zero when $\delta = 0$ to $(1 - \alpha)$ when $\delta = 1$. For the bound U_{jj}, one can similarly take

$$T_{jj}(m, n, \delta) = E\{F_{jj|\mathbf{y}}(\delta U_{jj})\},\ j = 1, 2, \tag{6}$$

as the measure of precision. We have $T_{11} = T_{22}$ when identical prior distributions are used for the parameters of the two methods.

For T_{12} and T_{jj} to be useful in SSD, they must be decreasing functions of m and n, and must tend to zero as m increases to infinity, keeping everything else fixed. Analytically verifying the first property is hard, but simulation can be used to clarify the approximate monotonicity (see Section 3). The second property follows from a straightforward application of the asymptotic properties of posterior distributions [12].

2.4 Computing the precision measure

The expectations T_{12} and T_{jj} are not available in closed-forms under the model (1). So we use the simulation-based approach of Wang and Gelfand [12] to compute them. Prior to their work, Bayesian SSD focussed mostly on normal and binomial one- and two-sample problems, partly because of a lack of a general approach for computing precision measures under hierarchical models that are typically not available in closed-forms. Their key idea is to distinguish between a *fitting prior* and a *sampling prior*. A prior that is used to fit a model is what they call a fitting prior. Frequently, this prior is noninformative and sometimes even improper, provided the resulting posterior is proper. On the other hand, for SSD in practice, we are usually interested in achieving a particular level of precision when $\boldsymbol{\gamma}$ is concentrated in a relatively small portion of the parameter space. This uncertainty in $\boldsymbol{\gamma}$ is captured using what they call a sampling prior. This prior is necessarily proper and informative. One choice for this is a uniform distribution over a bounded interval, but other distributions can also be used. In our case with model (1), the fitting prior is given by (4). Examples of sampling priors appear in Section 3.

Let $\boldsymbol{\gamma}^*$ denote a draw of $\boldsymbol{\gamma}$ from its sampling prior and $\mathbf{y}^*$ be a draw of $\mathbf{y}$ from $[\mathbf{y}|\boldsymbol{\gamma}^*]$, where "$[\cdot]$" denotes a probability distribution. This $\mathbf{y}^*$ alone is a draw from $[\mathbf{y}^*]$ — the marginal distribution of $\mathbf{y}$ under the sampling prior. Following [12], we compute the expectations $T_{12}(m, n, \delta)$ in (5) and $T_{jj}(m, n, \delta)$ in (6) with respect to $[\mathbf{y}^*]$, instead of $[\mathbf{y}]$. They are denoted as

$$T^*_{12}(m, n, \delta) = E\big\{F_{12|\mathbf{y}^*}(\delta U^*_{12})\big\}, \;\; T^*_{jj}(m, n, \delta) = E\big\{F_{jj|\mathbf{y}^*}(\delta U^*_{jj})\big\}, \; j = 1, 2, \tag{7}$$

where U^*_{12} and U^*_{jj} are counterparts of the credible bounds U_{12} and U_{jj} when $\mathbf{y}^*$ is used in place of $\mathbf{y}$. The expectations in (7) are approximated by Monte Carlo integration using the following steps:

(i) Draw $\boldsymbol{\gamma}^*$ from its sampling prior. Then use model (1) to draw $\mathbf{y}^*$ from $[\mathbf{y}|\boldsymbol{\gamma}^*]$.

(ii) Fit model (1) to $\mathbf{y}^*$ treating it as the observed data and $\boldsymbol{\gamma}$ as the parameter vector, and simulate draws from the posterior $[\boldsymbol{\gamma}|\mathbf{y}^*]$. The distributions in (4) are used as (fitting) priors for $\boldsymbol{\gamma}$ in this model fitting.

(iii) Use the draws of $\boldsymbol{\gamma}$ to get draws from the posteriors $[\theta_{12}|\mathbf{y}^*]$ and $[\theta_{jj}|\mathbf{y}^*]$.

(iv) Take the $(1-\alpha)$-th percentiles of draws of θ_{12} and θ_{jj} as their respective credible bounds U^*_{12} and U^*_{jj}.

(v) Find the proportion of draws of θ_{12} that are less than or equal to δU^*_{12}, and the proportion of draws of θ_{jj} that are less than or equal to δU^*_{jj}. These proportions respectively approximate the cdf's $F_{12|\mathbf{y}^*}(\delta U^*_{12})$ and $F_{jj|\mathbf{y}^*}(\delta U^*_{jj})$ in (7).

(vi) Repeat the steps (i)-(v) a large number of times, say, L, and take the averages of proportions in (v) as the approximated expected values in (7).

A URL for an R program for computing these averages is provided in the last section. As in the case of T_{12} and T_{jj}, the approximate monotonicity of T^*_{12} and T^*_{jj} with respect to the number of individuals m and the number of replicates n can be clarified using simulation. This issue is investigated in Section 3.

2.5 The SSD procedure

We would like to find sample sizes (m, n) to make each of the measures T^*_{12} and T^*_{jj}, $j = 1, 2$, defined in (7), sufficiently small, say, less than $(1-\beta) \in (0, 1-\alpha)$. There may be several combinations of (m, n) that give this precision of inference. So the cost of sampling must be taken into account to find the optimal combination. Let C_I be the cost associated with sampling an individual and C_R be the cost of taking one (replicate) measurement from both methods. Thus the total cost of sampling n replicates on each of m individuals is

$$C_T(m, n) = m(C_I + nC_R), \quad C_I, C_R > 0. \tag{8}$$

We now propose the following SSD procedure: Specify the sampling priors, a large $\delta \in (0, 1)$ and a small $(1-\beta) \in (0, 1-\alpha)$. Find (m, n) such that

$$\max\left\{T^*_{12}(m, n, \delta),\, T^*_{11}(m, n, \delta),\, T^*_{22}(m, n, \delta)\right\} \leq 1 - \beta, \tag{9}$$

and $C_T(m,n)$ is minimized. The condition in (9) ensures that the posterior probabilities of the intervals $[\delta U_{12}^*, U_{12}^*]$, $[\delta U_{jj}^*, U_{jj}^*]$, $j = 1, 2$, are all at least $(\beta - \alpha)$, providing sufficiently precise inference. In practice, one may take $\delta \geq 0.75$ and $(1-\beta) \leq 0.20$.

3 Simulation study

We now use simulation to verify monotonicity of the precision measures $T_{12}^*(m,n,\delta)$ and $T_{jj}^*(m,n,\delta)$, $j = 1, 2$, defined in (7), with respect to m and n. This investigation is done for four agreement measures — concordance correlation [3], coverage probability [5], mean squared deviation [4], and total deviation index [4, 6]. They are defined as follows for the evaluation of inter-method agreement.

concordance correlation: $\dfrac{2cov(y_1, y_2)}{\big(E(y_1) - E(y_2)\big)^2 + var(y_1) + var(y_2)}$.

coverage probability: $Pr(|y_1 - y_2| \leq q_0)$ for a specified small q_0.

mean squared deviation: $E\big((y_1 - y_2)^2\big)$.

total deviation index: p_0-th quantile of $|y_1 - y_2|$ for a specified large p_0.

The first one is based on the joint distribution of (y_1, y_2), whereas the others are based on the distribution of $y_1 - y_2$. For the evaluation of intra-method agreement of the j-th method, these measures are defined simply by substituting (y_{j1}, y_{j2}) in place of (y_1, y_2). Table 1 gives their expressions assuming distributions (2) and (3). It may be noted that the usual range of concordance correlation is $(-1, 1)$, but here they are restricted to be positive because $cov(y_1, y_2)$ and $cov(y_{j1}, y_{j2})$ under model (1) are positive. Moreover, the intra-method versions of the last three measures depend on the model parameters only through σ_j^2.

The first two measures, namely the concordance correlation and the coverage probability, range between $(0, 1)$ and their large values indicate good agreement. For the purpose of SSD, we take their negative-log transformation to satisfy the assumptions that the agreement measure lies in $(0, \infty)$ and its small value indicates good agreement. On the other hand, no transformation is needed in the case of mean squared deviation and total deviation index as they range between $(0, \infty)$ and their small values indicate good agreement.

For the simulation study, we focus on the two examples introduced in Section 1. First, we need to choose sampling priors for the parameters in model (1). We now describe how we construct them by fitting appropriate models to the pilot data in the two examples.

In the first example, we take the ASAP software as method 1 and the XPRESS software as method 2. After a preliminary analysis, we model the available data as

$$y_{ijr} = \beta_j + b_i + \text{protein}_r + \epsilon_{ijr},$$

$$b_i|\psi^2 \sim \text{ independent } \mathcal{N}(0, \psi^2),\ \epsilon_{ijr}|\sigma_j^2 \sim \text{ independent } \mathcal{N}(0, \sigma_j^2),$$

where y_{ijr} represents the log-ratio of the r-th protein from the j-th method on the i-th blood sample pair, $r = 1, \ldots, 8,\ j = 1, 2\ i = 1, \ldots, m = 5$. The interaction term b_{ij} is not included here as the data do not have enough information to estimate it. To fit this model, we assume mutually independent (fitting) priors — uniform $(-10^3, 10^3)$ for β's and protein effects, and $IG(10^{-3}, 10^{-3})$ for variance parameters. The model is fitted using `WinBUGS` [18] by calling it from `R` [17] through the `R2WinBUGS` [19] package. Table 2 presents posterior summaries for $(\beta_1, \beta_2, \log\sigma_1^2, \log\sigma_2^2, \log\psi^2)$. Further, the central 95% credible intervals for $corr(y_1, y_2)$, $corr(y_{11}, y_{12})$ and $corr(y_{21}, y_{22})$, after adjusting for protein effects, are $(0.01, 0.44)$, $(0.01, 0.43)$ and $(0.01, 0.51)$, respectively. These correlations are computed using (2). They are relatively small for a method comparison study because the between-individual variation ψ^2 here is small in comparison with the within-individual variations σ_j^2, $j = 1, 2$. Under this model, the agreement measures listed in Table 1 do not depend on proteins. So for the purpose of SSD for a future study, we ignore their effects and assume model (1). Further, to examine the effect of sampling priors on SSD, we consider two sets of sampling priors for $(\beta_1, \beta_2, \log\sigma_1^2, \log\sigma_2^2, \log\psi^2)$ — independent normal distributions with means and variances equal to their posterior means and variances, and independent uniform distributions with ranges given by their central 95% credible intervals. The posterior summaries obtained using the pilot data are given in Table 2. Moreover, since the interaction term variance ϕ^2 could not be estimated in this small pilot study but might be present in a bigger study, we take the sampling prior for $\log\phi^2$ to be the same as that of $\log\psi^2$.

In the second example, we take the manual device as method 1 and the automatic device as method 2. These data are modelled as (1) with $(m, n) = (384, 2)$ but without the interaction term b_{ij}. This term is dropped from the model on the basis of an exploratory analysis of the data and the deviance information criterion [23]. This model is fitted along the lines of the previous model with the same (fitting) priors for β's and variances. Posterior summaries for $(\beta_1, \beta_2, \log\sigma_1^2, \log\sigma_2^2, \log\psi^2)$ are given in Table 2. The central 95% credible intervals for $corr(y_1, y_2)$, $corr(y_{11}, y_{12})$ and $corr(y_{21}, y_{22})$ are $(0.86, 0.90)$, $(0.86, 0.90)$ and $(0.85, 0.90)$, respectively. Such high correlations are typical in method comparison studies. In this case also, we consider two sampling priors for $(\beta_1, \beta_2, \log\sigma_1^2, \log\sigma_2^2, \log\psi^2)$ — they are constructed using posterior summaries in exactly the same way as the previous example except that b_{ij} term is not included in (1). This is because the present study, which itself is quite large, does not seem to suggest its need.

Our investigation of properties of precision measures $T^*_{12}(m, n, \delta)$ and $T^*_{jj}(m, n, \delta)$, $j = 1, 2$, given by (7), focuses on the following settings: $(\alpha, \delta) = (0.05, 0.80)$; fitting priors for β_j's in (1) as independent (improper) uniform distributions on real line, and as independent

$IG(10^{-3}, 10^{-3})$ distributions for variances; $m \in \{15, 20, \ldots, 150\}$; and $n \in \{1, 2, 3, 4\}$. We restrict attention to $n \leq 4$ as it is rare to find studies in the literature with more than four replicates. The Gibbs sampler described in Appendix is used for posterior simulation. The Markov chain is run for 2,500 iterations and the first 500 iterations are discarded as burn-in. These numbers were determined through a convergence analysis of the simulated chains. Finally, the expectations in (7) are approximated using $L = 2,000$ Monte Carlo repetitions. These calculations are performed entirely in R. It took about 45 minutes on a Linux computer with 4 GB RAM to compute one set of the three averages, $(T^*_{12}, T^*_{11}, T^*_{22})$.

The values of T^*_{11} and T^*_{12} in case of normal sampling priors are plotted in Figure 1 for concordance correlation (both examples), in Figure 2 for total deviation index with $p_0 = 0.8$ (both examples), and in Figure 3 for mean squared deviation and coverage probability with $q_0 = \log(1.5)$ (only example 1). The omitted scenarios have qualitatively similar results. The graphs confirm that the averages T^*_{11} and T^*_{12} can be considered decreasing functions of m and n. This is also true for T^*_{22} for which the results are not shown. Some of the curves do show minor departures from monotonicity, particularly in case of $n = 1$ when the curves drop rather slowly with m. But they are due to the Monte Carlo error involved in averaging and can be ignored. A slow decline also means that a precise evaluation of agreement is difficult even with a large m.

In both examples, we have $\max\{T^*_{12}, T^*_{11}, T^*_{22}\} = \max\{T^*_{11}, T^*_{22}\}$ for all (m, n), in case of all agreement measures except concordance correlation. It implies that to achieve the same precision of inference, as measured by (δ, β), these measures require a larger sample size for intra-method evaluation than for inter-method evaluation. This property also holds for concordance correlation in example 2, but it holds in example 1 only when $n = 1$. The Figures 1-3 also demonstrate that different sample sizes are needed for different agreement measures to attain the same precision of inference.

The above conclusions for normal sampling priors also hold for uniform sampling priors (results not shown). Upon further investigation, we find that the two priors lead to virtually identical average probabilities in case of example 2. Even in case of example 1, their differences are generally small — about 2-3% at most. Since these two priors represent different distributions over practically the same range of values, this suggests that for SSD, it does not matter much which distribution is used as both lead to similar results.

4 Illustration

The results in the previous section demonstrate, in particular, that the SSD procedure proposed in Section 2.5 can be used in conjunction with any of the four agreement measures listed in Table 1. In this section, we apply this procedure to determine sample sizes in case of the two examples introduced in Section 1. To avoid repeating the same ideas, we focus only on the total deviation index as the measure of agreement. We also investigate the robustness of various choices one has to make for SSD based upon this measure. Other agreement measures can be handled similarly. The method for computing T^*'s, and the sampling and fitting priors remain the same as in the previous section, with the exception that the uniform sampling priors will now be called "Uniform-I" priors. We also take $(\alpha, \beta, \delta) = (0.05, 0.85, 0.8)$. Our choice of $\delta = 0.80$ is motivated by bioequivalence studies [24, 25] where a difference of about 20% is considered a threshold for *practical equivalence.*

Table 3 presents values of the smallest m, separately for each $1 \leq n \leq 4$, such that $T^*_{12} \leq 1 - \beta$, $T^*_{jj} \leq 1 - \beta$, $j = 1, 2$, hold individually. In other words, for these (m, n) combinations we have at least $\beta - \alpha = 0.8$ individual probabilities in the intervals $[0.8U^*_{12}, U^*_{12}]$, $[0.8U^*_{jj}, U^*_{jj}]$,

$j = 1, 2$, where the probabilities below the upper bounds equal 0.95. Due to the discreteness of (m, n), the exact probabilities in these intervals may be slightly higher than 0.8. Their difference tends to increase with n and may be as much as 0.03 when $n = 4$. In Table 3, the sample sizes are provided for both examples by taking $p_0 \in \{0.80, 0.85, 0.90, 0.95\}$, and assuming normal and Uniform-I sampling priors.

As we expect from the simulation studies, larger sample sizes are needed for intra-method evaluation than for inter-method evaluation. Moreover, substantially large values of m are needed for intra-method evaluation with $n = 1$. But this scenario is not of much practical interest as precise estimation of within-individual variances σ_1^2 and σ_2^2 is difficult in this case. Furthermore, the sample sizes for intra-method evaluation do not depend on p_0 since the term involving p_0 in total deviation index appears as a multiplicative constant (see Table 1). On the other hand, in case of inter-method evaluation, sometimes the values of m do increase by one as p_0 increases to the next level, but their maximum difference over the four settings of p_0 is three and most of the differences are two or less. This demonstrates that the sample sizes for inter-method evaluation are also quite robust to the choice of p_0.

It is interesting to note that, with the exception of $n = 1$ case, the sample sizes for our two very different examples show remarkable similarity. In case of $n = 1$, the values of m for intra-method evaluation are substantially higher for the second example than the first example. This is because, as discussed in Section 3, the measurements on an individual in the second example are highly correlated, whereas they are only weakly correlated in the first example.

Suppose for the time being that SSD for *simultaneous* evaluation of intra- and inter-method agreement is to be performed assuming Uniform-I sampling priors. It follows from (9) and the results in Table 3 that the desired precision of inference can be achieved by any of the following (m, n) combinations: $(121, 1)$, $(58, 2)$, $(34, 3)$ and $(23, 4)$ for example 1; and $(655, 1)$, $(60, 2)$, $(33, 3)$ and $(23, 4)$ for example 2. These values do not depend on p_0. To find the optimal combination, recall that the cost of sampling is given by (8), where we expect that the relative cost $(C_R/C_I) \in (0, 1]$. Under this cost function, a combination (m_2, n_2) has lower total sampling cost than (m_1, n_1) if $(m_2 - m_1) < (C_R/C_I)(m_1 n_1 - m_2 n_2)$. In particular, if $(m_1 n_1 - m_2 n_2) > 0$ and $(m_2 - m_1) < 0$, then the (m_2, n_2) combination is better than (m_1, n_1) irrespective of the cost of sampling. A naive application of this criterion, without any subject-matter knowledge, suggests that a higher n is better. Hence $(23, 4)$ is the optimal sample size choice in both examples.

In case of example 1, the experimental protocol does allow us to draw enough blood from individuals to have four replications. So we decide to use $(m, n) = (23, 4)$ as the sample size for a future study. In case of example 2, however, it is unlikely that four replicate measurements of blood pressure can be obtained from each method without a change in the pressure. On the other hand, we do need replications, especially for evaluating intra-method agreement. So, as a compromise, we can take $(m, n) = (60, 2)$ for a future study.

We now investigate robustness of the sample sizes with respect to the specification of sampling priors and fitting priors. First consider the former. Results in Table 3 show that the values of m in case of Uniform-I sampling priors tend to be a bit higher than normal sampling priors. But they differ by at most two except in the case of intra-method evaluation with $n = 1$. Note, however, that these two sampling priors represent different distributions over essentially the same range of values.

Next, to specifically examine the impact of the range of sampling prior distributions, we determine sample sizes assuming independent uniform distributions over the central 50% posterior intervals of parameters as their sampling priors. These intervals based on the pilot data are reported in Table 2. We jointly refer to these priors as "Uniform-II". Their ranges are shorter than the ranges of Uniform-I priors, which represent uniform distributions over the central 95% posterior intervals. From Table 2, the differences between the two sets of ranges appear substantial in case of example 1, whereas they appear negligible in case of example 2. Upon comparing values of m in Table 3 for Uniform-I and Uniform-II priors, we find that there is virtually no difference between them in case of example 2. Even in case of example 1, their differences are considerable only for intra-method evaluation with $n = 1$. In all other cases, the differences mostly equal two or three but can be as much as five.

In a nutshell, these results suggest that except in the uninteresting case of intra-method evaluation with $n = 1$, the choice between a normal and a uniform sampling prior distribution over similar range of values is not important. More important is the choice of the range of parameter values, but even that affects the sample sizes only to a limited extent.

We now use normal and Uniform-II sampling priors to determine the optimal (m, n) combination that takes sampling cost and feasibility into account. This combination turns out to be the same as in the case of Uniform-I sampling priors — $(23, 4)$ for example 1 and $(60, 2)$ for example 2. This finding is somewhat odd as, in general, we do expect some difference in the optimal sample sizes for different sampling priors.

The above discussion assumed noninformative fitting priors for the parameters in model (1) — improper uniform distributions over real line for β_j's, and $IG(10^{-3}, 10^{-3})$ distributions for variance components. To examine sensitivity of the variance hyperparameter choice, we repeat determining the optimal (m, n) combination with two other noninformative choices, 10^{-1} and 10^{-2}, for the common hyperparameter value. Uniform-I sampling priors are assumed for this computation. We find that m changes by at most one — indicating that the sample sizes are not sensitive to the choice of hyperparameters, provided they are in the noninformative range. This conclusion is not surprising since the inference with a noninformative prior is dominated by the data.

5 Discussion

In this article, we described a Bayesian SSD approach for planning a method comparison study. It is conceptually straightforward and can be used with any scalar agreement measure. However, a disadvantage of this approach is that it is simulation-based and hence is computationally intensive. Nevertheless, the computations are easy to program in the popular software package `R`. An `R` program that can be used for SSD is publicly available on the website `http://www.utdallas.edu/~pankaj/Bayesian_ssd`.

This approach involves specifying sampling priors for the model parameters. Although the sample sizes seem somewhat sensitive to the ranges of parameter values, they seem quite robust to the shapes of their distributions. So for practical purposes, it suffices to specify a likely range of values for a parameter and use a uniform distribution over this interval as its sampling prior. Information about the likely parameter values can be obtained from literature or through a pilot study, as we do here. The sample sizes are also robust to the choice of fitting priors, provided they are noninformative. The proposed approach involves making rather subjective choices for (δ, β) that control how much precision of inference is desired. But, broadly speaking, their roles here are similar to the roles of effect size and desired power that also need to be specified in the usual frequentist SSD approach.

In both examples, among the (m, n) combinations that yielded the same precision, the combination with the higher n also happened to be more cost efficient one, at least in the range of n investigated. This was true irrespective of what the cost of taking a replicate was relative to the cost of sampling an individual. This conclusion is not true in general. It is easy to construct examples where a higher n is more cost efficient only when the cost of sampling a replicate is small relative to the cost of sampling an individual.

Acknowledgements

The authors thank Dr. Swati Biswas, the two reviewers and the Associate Editor for several helpful comments that greatly improved this article.

References

1. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* 2007; **17**:1–41.

2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**:307–310.

3. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268. Corrections: 2000; **56**:324–325.

4. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000; **19**:255–270.

5. Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* 2002; **97**:257–270.

6. Choudhary PK, Nagaraja HN. Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 2007; **137**:279–290.

7. Lin LI. Assay validation using the concordance correlation coefficient. *Biometrics* 1992; **48**:599–604.

8. Lin SC, Whipple DM, Ho CS. Evaluation of statistical equivalence using limits of agreement and associated sample size calculation. *Communications in Statistics, Part A–Theory and Methods* 1998; **27**:1419–1432.

9. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; **8**:135–160.

10. Lehmann EL. *Elements of Large Sample Theory.* Springer: New York, 1998.

11. Adcock CJ. Sample size determination: a review. *The Statistician* 1997; **46**:261–283.

12. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; **17**:193–208.

13. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology* 2001; **19**:946-951.

14. Li XJ, Zhang H, Ranish JR, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable isotope dilution and tandem mass spectrometry. *Analytical Chemistry* 2003; **75**:6648–6657.

15. Liebler DC. *Introduction to Proteomics: Tools for the New Biology.* Humana Press: Totowa, NJ, 2002.

16. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 2003; **59**:849–858.

17. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, Austria, 2006. `http://www.R-project.org`.

18. Spiegelhalter DJ, Thomas A, Best NG, Lunn D. *WinBUGS Version 1.4 User Manual*, 2003. `http://www.mrc-bsu.cam.ac.uk/bugs`.

19. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software* 2005; **12**:1–16.

20. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*, 2nd edition. Chapman & Hall/CRC: Boca Raton, 2003.

21. Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* 2006; **1**:473–514.

22. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**:515–534.

23. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 2003; **64**:583–616.

24. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283–319.

25. Wang W. On equivalence of two variances of a bivariate normal vector. *Journal of Statistical Planning and Inference* 1999; **81**:279–292.

26. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression.* Cambridge University Press: New York, 2003.

Appendix: Gibbs sampler for posterior simulation

The model (1) can be written in a hierarchical fashion as

$$\mathbf{y}|(\boldsymbol{\beta}, \mathbf{b}, R) \sim \mathcal{N}(X\boldsymbol{\beta} + Z\mathbf{b}, R),\ \mathbf{b}|G \sim \mathcal{N}(\mathbf{0}, G),$$

where X and Z are appropriately chosen design matrices consisting of zeros and ones; $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)'$ is the data vector; $\boldsymbol{\beta} = (\beta_1, \beta_2)'$; $\mathbf{b} = (b_1, \ldots, b_m, b_{11}, b_{12}, \ldots, b_{m1}, b_{m2})'$; $R = \text{diag}\{R_1, \ldots, R_m\}$ is the conditional covariance matrix of $\mathbf{y}$; and $G = \text{diag}\{\psi^2 I_m, \phi^2 I_{2m}\}$ is the covariance matrix of $\mathbf{b}$. Here I_m denotes an $m \times m$ identity matrix, and $\mathbf{y}_i$ and R_i are respectively a column vector and a covariance matrix of order $2n$, given as

$$\mathbf{y}_i = (y_{i11}, \ldots, y_{i1n}, y_{i21}, \ldots, y_{i2n})',\ R_i = \text{diag}\{\sigma_1^2 I_n, \sigma_2^2 I_n\}.$$

The parameters in this case are $(\boldsymbol{\beta}, \mathbf{b}, \psi^2, \phi^2, \sigma_1^2, \sigma_2^2)$. It is well-known that this model is *conditionally conjugate* under the priors (4) [26]. In particular, the *full conditional* distribution — the conditional distribution of a parameter given the remaining parameters and $\mathbf{y}$, of the column vector $(\boldsymbol{\beta}, \mathbf{b})$ is

$$\mathcal{N}\left(\{C'R^{-1}C + D\}^{-1}C'R^{-1}\mathbf{y},\ \{C'R^{-1}C + D\}^{-1}\right),$$

where $C = [X, Z]$, $D = \text{diag}\{B^{-1}, G^{-1}\}$, and $B = \text{diag}\{V_1^2, V_2^2\}$. The matrices involved in this distribution can be computed using QR decompositions. Further, the full conditionals of variance parameters are the following independent inverse gamma distributions:

$$\psi^2 \sim IG\Big(A_\psi + m/2, B_\psi + \sum_i b_i^2/2\Big); \quad \phi^2 \sim IG\Big(A_\phi + m, B_\phi + \sum_i \sum_j b_{ij}^2/2\Big);$$
$$\sigma_j^2 \sim IG\Big(A_j + (mn)/2, B_j + \sum_i \sum_k (y_{ijk} - \beta_j - b_i - b_{ij})^2/2\Big), \ j = 1, 2.$$

The Gibbs sampler algorithm simulates a Markov chain whose limiting distribution is the desired joint posterior distribution of the parameters [20, ch 11]. It iterates the following two steps until convergence: First, use the current draw of $(\psi^2, \phi^2, \sigma_1^2, \sigma_2^2)$ to sample from the normal full conditional of $(\boldsymbol{\beta}, \mathbf{b})$. Then, use the draw of (β, b) in the previous step to sample from the independent inverse gamma full conditionals of $(\psi^2, \phi^2, \sigma_1^2, \sigma_2^2)$. The estimates of variance parameters from a frequentist mixed model fit can be used as the starting points in this algorithm.

List of Tables

List of Figures

measure	inter-method	intra-method
concordance correlation	$\frac{2\psi^2}{\mu_{12}^2+2\psi^2+2\phi^2+\sigma_1^2+\sigma_2^2}$	$\frac{\psi^2+\phi^2}{\psi^2+\phi^2+\sigma_j^2}$
coverage probability	$\Phi(\frac{q_0-\mu_{12}}{\tau_{12}}) - \Phi(\frac{-q_0-\mu_{12}}{\tau_{12}})$	$\Phi(\frac{q_0}{\tau_{jj}}) - \Phi(\frac{-q_0}{\tau_{jj}})$
mean squared deviation	$\mu_{12}^2 + \tau_{12}^2$	τ_{jj}^2
total deviation index	$\tau_{12}\{\chi_1^2(p_0, \mu_{12}^2/\tau_{12}^2)\}^{1/2}$	$\tau_{jj}\{\chi_1^2(p_0, 0)\}^{1/2}$

Table 1: Expressions for various agreement measures for inter-method evaluation and intra-method evaluation of the j-th method, $j = 1, 2$. They are derived using (2) and (3) under model (1). Here $\Phi(\cdot)$ represents a $\mathcal{N}(0, 1)$ cdf; and $\chi_1^2(p_0, \Delta)$ represents the p_0-th quantile of a chi-squared distribution with one degree of freedom and non-centrality parameter Δ.

	mean	sd	2.5%	25%	75%	97.5%
Example 1 (Protein ratios)						
β_1	0.16	0.07	0.01	0.12	0.21	0.31
β_2	0.10	0.08	-0.07	0.05	0.15	0.25
$\log \sigma_1^2$	-3.48	0.30	-4.01	-3.69	-3.29	-2.84
$\log \sigma_2^2$	-3.26	0.30	-3.82	-3.45	-3.08	-2.62
$\log \psi^2$	-5.92	1.11	-7.96	-6.68	-5.27	-3.55
Example 2 (Blood pressure)						
β_1	133.40	0.98	131.36	132.76	134.09	135.29
β_2	131.24	0.98	129.36	130.55	131.89	133.09
$\log \sigma_1^2$	3.96	0.06	3.84	3.92	4.00	4.07
$\log \sigma_2^2$	3.98	0.06	3.86	3.94	4.02	4.10
$\log \psi^2$	5.94	0.07	5.81	5.90	5.99	6.08

Table 2: Posterior means, standard deviations and selected quantiles for various parameters. They are computed using the pilot data.

		Normal				Uniform-I				Uniform-II			
		n				n				n			
agreement	p_0	1	2	3	4	1	2	3	4	1	2	3	4
Example 1 (Protein ratios)													
1-1	n/a	120	57	33	23	121	58	34	23	107	54	32	23
2-2	n/a	103	54	32	23	108	56	33	23	96	51	31	23
1-2	0.80	61	29	21	16	62	30	22	17	60	28	19	15
1-2	0.85	60	29	20	16	61	30	21	17	58	27	19	15
1-2	0.90	59	28	20	16	59	29	21	17	58	27	18	14
1-2	0.95	58	27	19	15	58	28	20	16	57	26	18	14
Example 2 (Blood pressure)													
1-1	n/a	646	60	33	23	655	60	33	23	654	60	33	23
2-2	n/a	633	60	33	23	633	60	33	23	632	60	33	23
1-2	0.80	76	29	18	13	76	29	18	13	76	29	18	13
1-2	0.85	76	28	18	13	76	28	18	13	76	28	18	13
1-2	0.90	75	28	17	13	75	28	17	13	76	28	17	13
1-2	0.95	75	28	17	13	75	28	17	13	75	28	17	13

Table 3: Values of m for specified values of n that give the desired precision for agreement evaluation. Here "1-1", '2-2" and "1-2" respectively indicate intra-method agreement of method 1, method 2, and inter-method agreement. Results are presented assuming total deviation index is used as the measure of agreement with various common choices for p_0. The sample sizes for intra-method evaluation do not depend on p_0 — they are marked as "n/a".

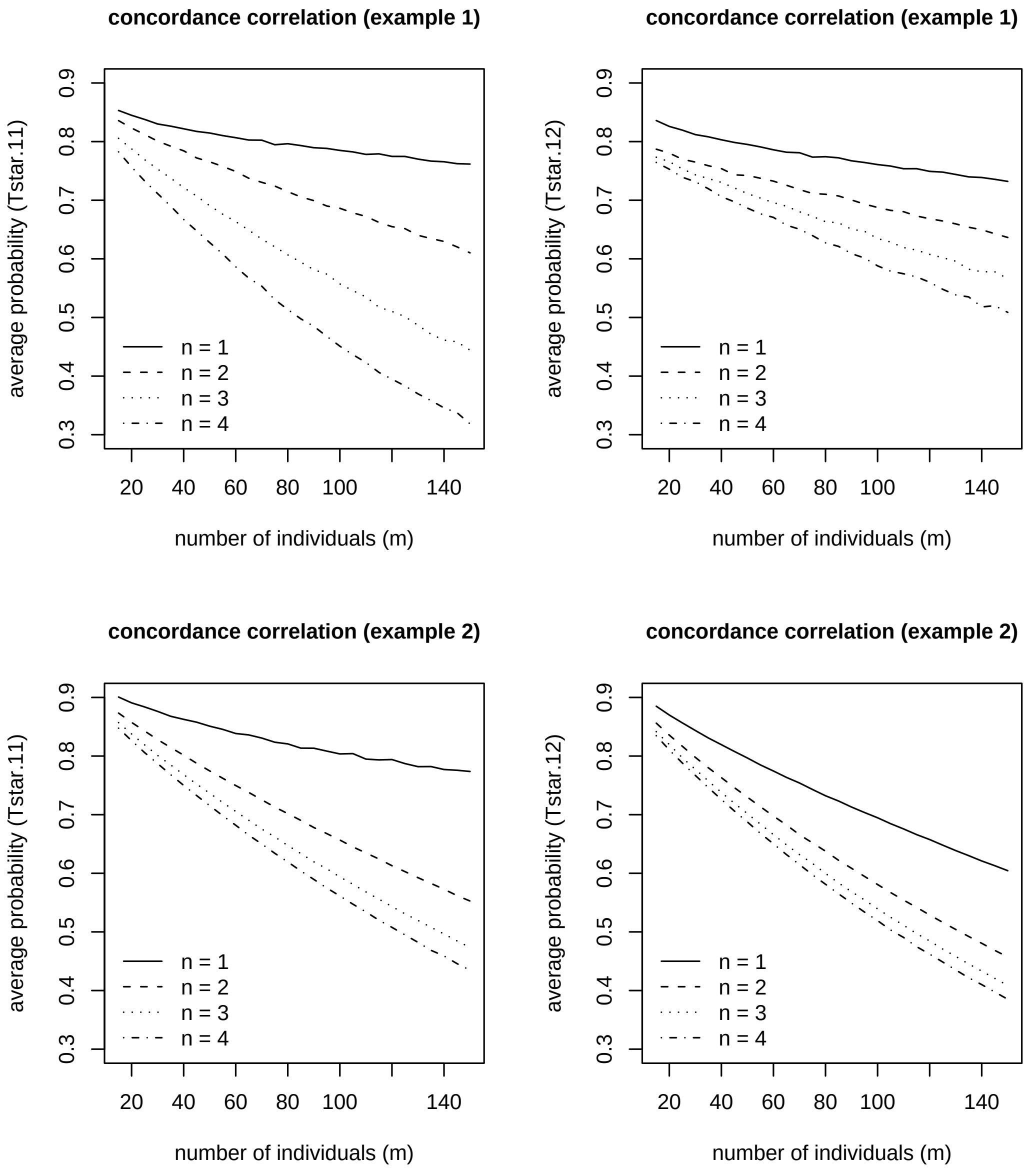

Figure 1: Average probabilities T_{11}^{*} (left panel) and T_{12}^{*} (right panel) when agreement is measured using concordance correlation. The top panel is for the protein ratios example and the bottom panel is for the blood pressure example. These probabilities measure precisions of inter-method agreement and intra-method agreement of method 1, respectively. Here n represents the number of replicates.

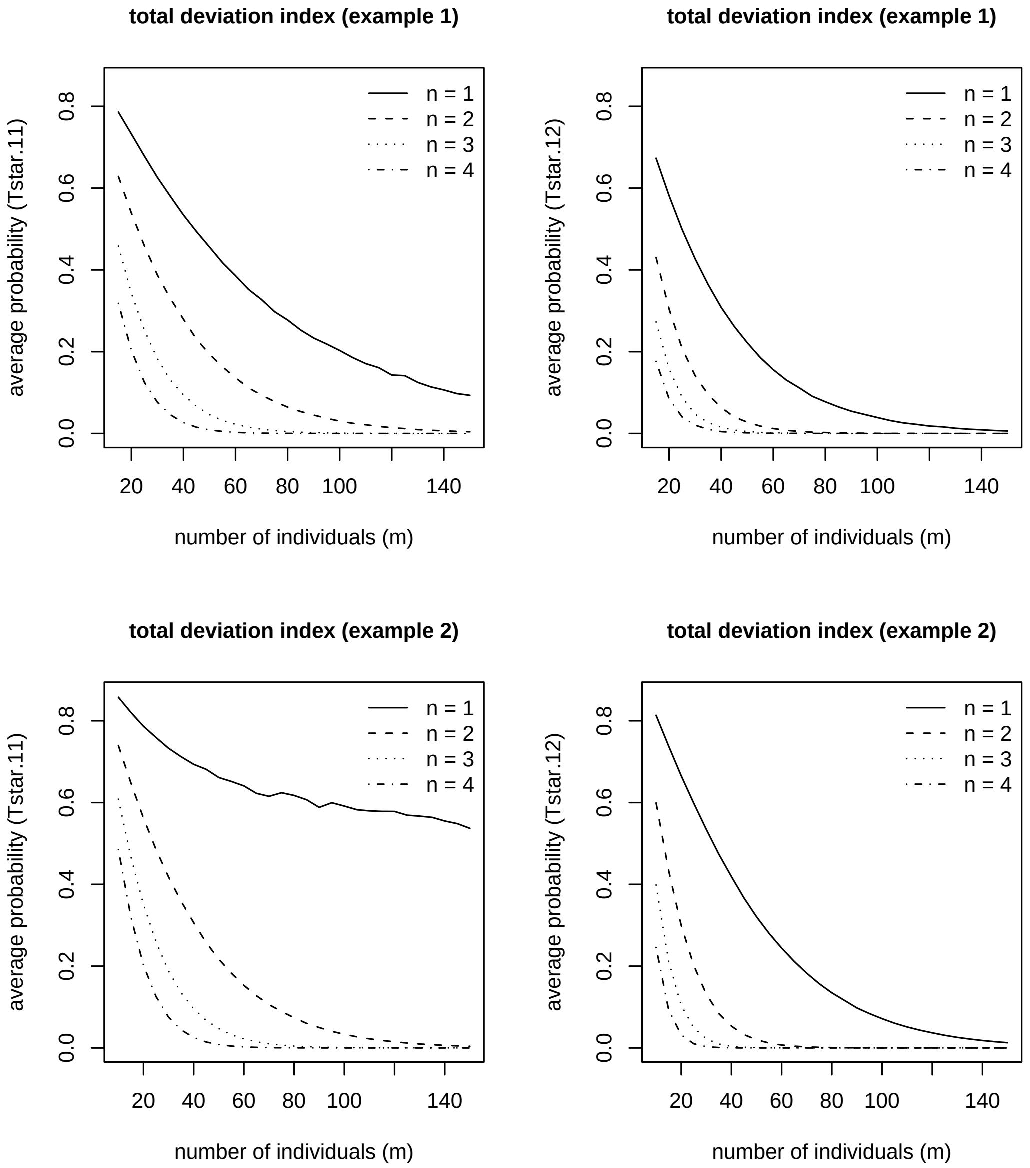

Figure 2: Average probabilities T^*_{11} (left panel) and T^*_{12} (right panel) when agreement is measured using total deviation index with $p_0 = 0.80$. The top panel is for the protein ratios example and the bottom panel is for the blood pressure example. Here n represents the number of replicates.

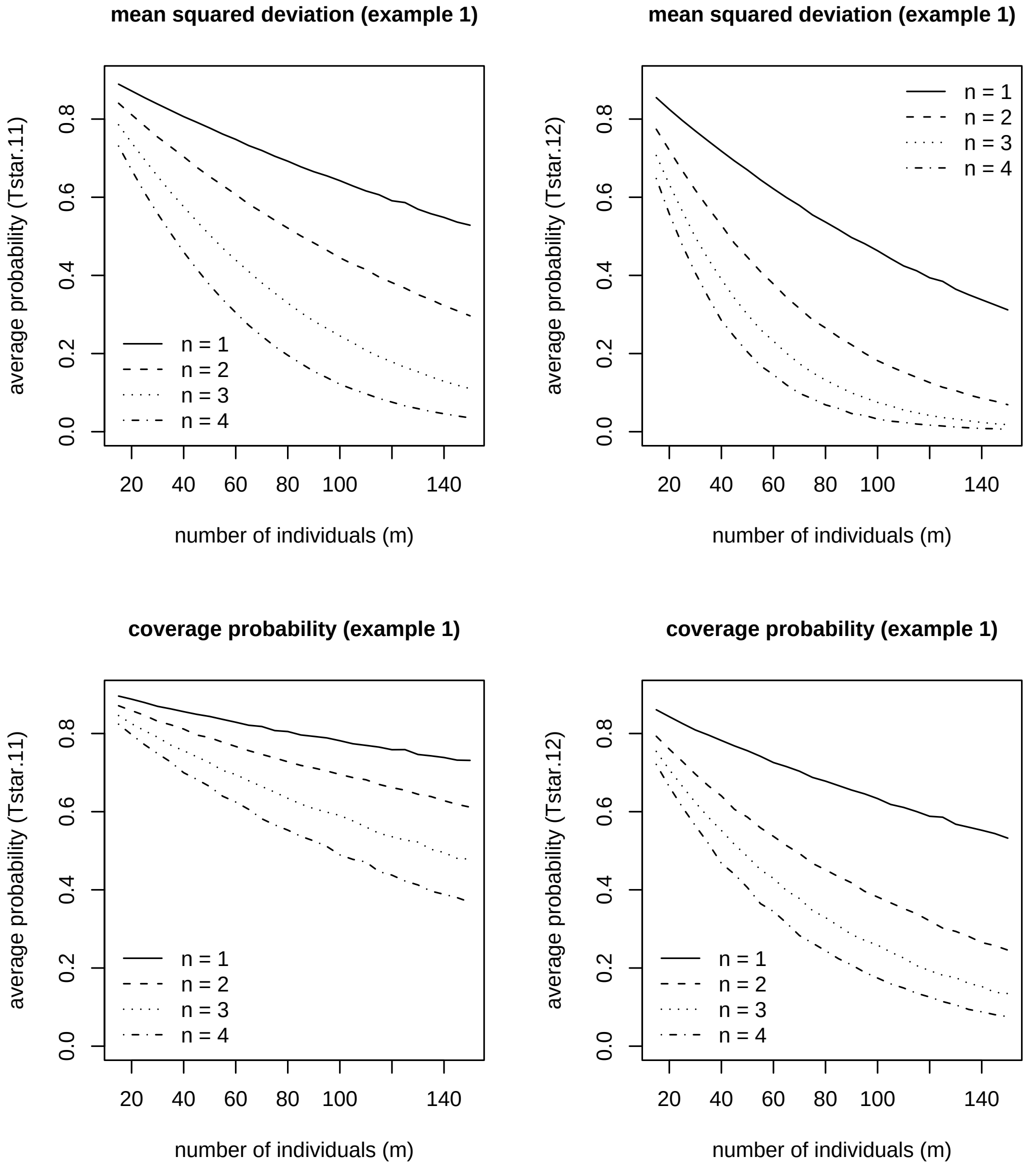

Figure 3: Average probabilities T^*_{11} (left panel) and T^*_{12} (right panel) when agreement is measured using mean squared deviation (top panel) and coverage probability (bottom panel) with $q_0 = \log(1.5)$. Here n represents the number of replicates and the results are presented only for the protein ratios example.