Bayesian and Frequentist Methodologies for Analyzing Method Comparison Studies With Multiple Methods

Pankaj K. Choudhary[a,1] and Kunshan Yin[b]

[a]University of Texas at Dallas, Richardson, TX 75080, USA

[b]EQECAT, 475 14th Street, Oakland, CA 94612, USA

Abstract

Evaluation of agreement among multiple methods of clinical measurement is a topic of considerable interest in health sciences. As in an analysis of variance comparing more than two treatment means, when more than two measurement methods are compared, performing multiple comparisons and ranking pairs of methods on the basis of their extent of agreement are of primary concern. This article develops frequentist and Bayesian methodologies for this purpose. In particular, simultaneous confidence bounds and simultaneous credible bounds are developed for multiple comparisons. Moreover, two approaches are described for ranking method pairs — one based on simultaneous bounds and the other based on posterior probabilities of possible orderings. The proposed methodologies can be used with any scalar measure of agreement. Their small-sample performance is evaluated using simulation. Extension of the basic methodologies to incorporate covariates is illustrated using a blood pressure data set.

Key Words: Agreement; Concordance correlation; Multiple Comparisons; Ranking; Tolerance interval; Total deviation index.

[1]Corresponding author. Address: Department of Mathematical Sciences EC 35, 800 West Campbell Road, Richardson, TX 75080-3021, USA. Email: pankaj@utdallas.edu; Tel: (972) 883-4436; Fax: (972) 883-6622.

1 Introduction

Measurement of clinically important continuous variables, such as blood pressure, heart rate, cholesterol level, etc., is key to management of health. As the technology advances, new measurement methods become available that may be cheaper, easier to use or less invasive than the established ones. But before new methods are adopted for general use, they are validated by comparing them with established methods in method comparison studies. Such studies are quite common in health sciences research. This is evident from the fact that Bland and Altman (1986) who proposed the limits of agreement approach for evaluation of agreement between two methods currently has over 13,000 citations in the Web of Science — a citation database of the ISI Web of Knowledge.

In studies comparing two methods, generally the main objective is to quantify the extent of agreement between the methods and determine if they agree sufficiently well to justify their interchangeable use. In addition to the aforementioned limits of agreement approach, there are several other approaches for evaluating agreement between two methods, including concordance correlation coefficient (CCC, Lin, 1989), mean squared deviation (Lin, 2000), coverage probability and total deviation index (TDI) or tolerance interval (Lin, 2000, Lin et al., 2002, Choudhary and Nagaraja, 2007), and coefficient of individual agreement (Barnhart, Kosinski and Haber, 2007). See Barnhart, Haber and Lin (2007) for a recent comparative review of the statistical literature on this topic.

The focus of this article is on method comparison studies involving more than two methods. Such studies are also common in practice. Consider, e.g., Moura et al. (2009) who compare three methods — tissue Doppler imaging, pulsed-wave Doppler imaging and M-mode echocardiography, for measuring myocardial performance index. Another recent example is Welinder et al. (2009) who compare standard and EASI-derived 12-lead electrocardiogram with cardiac magnetic resonance imaging as gold standard for measuring Selvester score

for chronic myocardial infarction. An additional example involving measurement of blood pressure is provided later in this article.

As in an analysis of variance (ANOVA) comparing more than two treatments (Hsu, 1996), when more than two measurement methods are compared, the interest lies in performing *multiple comparisons*, i.e., *simultaneously* comparing the extent of agreement between all pairs of methods of concern, and ranking them on the basis of their extent of agreement. Our set-up, however, differs with ANOVA in two crucial respects. Firstly, in ANOVA, one is mainly interested in treatment means. But in method comparison studies, an agreement measure that quantifies the extent of agreement between two methods is the parameter of main concern. Generally it is a function of not just the means of the methods but also their variances and correlation. Secondly, in ANOVA, ranking typically refers to ordering treatments on the basis of their means. In contrast, in method comparison studies, we rank *pairs* of methods on the basis of their extent of agreement.

The ANOVA analogy implies that for multiple comparisons and ranking, we need methodologies for *simultaneous inference* on measures of agreement between the pairs of methods of interest. They, however, are not available in the literature. One can, of course, apply the Bonferroni inequality method to derive a simultaneous inference procedure from the pairwise inference procedures. But such a procedure is well-known to be conservative (Hsu, 1996, chapter 1). To the best of our knowledge, there have been only two lines of work on the topic of comparing multiple methods. The first summarizes the overall level of agreement among all methods in a single index by taking a weighted average of the pairwise agreement measures. The articles of King and Chinchilli (2001) and Barnhart, Haber and Song (2002) fall in this category. Although these articles focus only on the concordance correlation, similar ideas can be used for other scalar measures of agreement. The second line of work assumes a gold standard method in the comparison that serves as a reference, and considers the problem of finding the method that agrees most with it. The articles of St. Laurent (1998), Hutson,

Wilson and Geiser (1998) and Choudhary and Nagaraja (2005a, b) fall in this category. In the first, an intraclass correlation, and in the rest, the mean squared deviation serve as the measure of agreement.

This article attempts to fulfill the aforementioned need by developing frequentist and Bayesian methodologies for multiple comparisons and ranking in method comparison studies. The frequentist paradigm offers an approach based on simultaneous confidence bounds, whereas the Bayesian paradigm offers two approaches — one based on simultaneous credible bounds and the other based on posterior probabilities. The methodologies only require an agreement measure to be scalar. So they can be used with all the measures currently available in the literature except the limits of agreement, which uses two limits to quantify agreement. Incorporating covariates in the analysis is also straightforward.

The rest of this article is organized as follows. In Section 2, we assume a basic model for method comparison data and describe the methodologies for multiple comparisons and ranking of method pairs using an arbitrary scalar measure of agreement. Their properties are examined in Section 3 via simulation. In Section 4, we extend these methodologies to incorporate covariates in the analysis and apply them to a blood pressure data set from the literature. Section 5 concludes with a discussion.

2 Methodology for multiple comparisons and ranking

2.1 A basic model for the data

Suppose there are $J\,(\geq 2)$ methods under comparison and we have m individuals in the study. Let Y_{ijk}, $k = 1, \ldots, n_{ij}$, denote the kth replicate measurement from the jth method on the ith individual. To simplify the presentation of the key ideas, in this section we assume that these data can be modeled as

$$Y_{ijk} = \beta_j + b_{ij} + \epsilon_{ijk}, \quad k = 1, \ldots, n_{ij},\ j = 1, \ldots J,\ i = 1, \ldots, m, \tag{1}$$

where β_j is the fixed effect of the jth method; b_{ij} is the random effect of the ith individual on the jth method, i.e., the individual $\times$ method interaction; and ϵ_{ijk} is the random error. Further, the vector $(b_{i1}, \ldots, b_{iJ})|\Psi \sim$ independent J-variate normal distribution with mean zero and an arbitrary covariance matrix Ψ, $\epsilon_{ijk}|\sigma_j^2 \sim$ independent $\mathcal{N}(0, \sigma_j^2)$ and are mutually independent of the interaction terms. (All vectors in this article are column vectors unless specified otherwise.) When the measurements are not replicated, the interaction term b_{ij} in (1) is replaced by a random individual effect b_i, where $b_i|\psi^2 \sim$ independent $\mathcal{N}(0, \psi^2)$, otherwise the model is not identifiable. Extensions of this model is discussed in Section 4.

2.2 Measuring agreement

Let the random vector $(Y_1, \ldots, Y_J)$ represent the population of measurements from J methods. In essence, this vector denotes the measurements from J methods on a randomly selected individual from the population, measured once by each method. Also let $\underline{\gamma}$ be the vector of parameters in the model (1). It follows from the assumptions of this model that the joint distribution of $(Y_1, \ldots, Y_J)|\underline{\gamma}$ is J-variate normal with

$$E(Y_j) = \beta_j,\ var(Y_j) = \psi_{jj} + \sigma_j^2,\ cov(Y_j, Y_l) = \psi_{jl},\ j \neq l = 1, \ldots, J, \tag{2}$$

where ψ_{rs} is the (r, s)th element of the matrix Ψ. This distribution is well-defined irrespective of the values of the number of replicates n_{ij}.

Let θ be an arbitrary scalar measure of agreement between two methods whose either large or small values indicate good agreement. When referring specifically to the agreement between methods (j, l), θ will be denoted as θ_{jl}. By definition it is a function of the parameters (2) of the bivariate normal distribution of (Y_j, Y_l).

Consider, for example, the popular agreement measure CCC (Lin, 1989). It is defined as

$$\theta_{jl} \equiv \rho_{jl} = \frac{2cov(Y_j, Y_l)}{(E(Y_j - Y_l))^2 + var(Y_j - Y_l) + 2cov(Y_j, Y_l)} = \frac{2\psi_{jl}}{\mu_{jl}^2 + \tau_{jl}^2 + 2\psi_{jl}}, \tag{3}$$

where $\mu_{jl} = E(Y_j - Y_l) = \beta_j - \beta_l$ and $\tau_{jl}^2 = var(Y_j - Y_l) = \psi_{jj} + \psi_{ll} - 2\psi_{jl} + \sigma_j^2 + \sigma_l^2$. It is basically a measure of the mean squared difference between Y_j and Y_l, $E(Y_j - Y_l)^2$, which is scaled to lie between $(0, 1)$ (Lin, 1989). A large value for CCC indicates good agreement between the two methods. The methods have perfect agreement — i.e., $\mu_{jl} = 0 = \tau_{jl}^2$ or equivalently $Pr(Y_j = Y_l) = 1$, in the limiting case when $\rho_{jl} = 1$.

The TDI (Lin, 2000) is another agreement measure, which is defined as

$$\theta_{jl} \equiv q_{jl} = p_0\text{th quantile of } |Y_j - Y_l| \text{ for a specified } p_0 = \tau_{jl}\left\{\chi_1^2(p_0, \mu_{jl}^2/\tau_{jl}^2)\right\}^{1/2}, \qquad (4)$$

where $\chi_1^2(p_0, \Delta)$ represents the p_0th quantile of a chi-squared distribution with one degree of freedom and noncentrality parameter Δ. Generally p_0 is large (≥ 0.80) in applications. So the TDI measures how large the absolute difference between Y_j and Y_l can be in a large proportion of population. This measure is positive and a small value for it indicates good agreement between the methods. Perfect agreement results when $q_{jl} = 0$. Expressions for other measures of agreement can be derived in a similar manner.

2.3 Multiple comparisons of method pairs

The goal of multiple comparisons inference is to get simultaneous bounds for all θs of interest with probability, say, $(1 - \alpha)$. In the frequentist approach, these bounds are simultaneous *confidence* bounds with $(1-\alpha)$ coverage probability, whereas in the Bayesian approach, they are simultaneous *credible* bounds with $(1 - \alpha)$ posterior probability. The bounds are used to compare the extent of agreement among the method pairs, and also to infer which pairs, if any, have sufficient agreement for their interchangeable use. When there is a reference method (say, method 1), we would like to compare the extent of agreement of the other methods with the reference. This leads to a total of $(J - 1)$ comparisons involving $\theta_{1j}, j = 2, \ldots, J$. On the other hand, when there is no reference, we are interested in all-pairwise comparisons. In this case, we have a total of $\binom{J}{2}$ comparisons involving $\theta_{jl}, j < l = 1, \ldots, J$.

If the agreement measure θ is such that its small value implies good agreement (e.g., TDI), we compute simultaneous *upper* bounds, say, $\{U_{1j}, j = 2, \ldots, J\}$ for multiple comparisons with a reference, and $\{U_{jl}, j < l = 1, \ldots, J\}$ for all-pairwise comparisons. Conversely, if θ is such that its large value implies good agreement (e.g., CCC), we compute simultaneous *lower* bounds, say, $\{L_{1j}, j = 2, \ldots, J\}$ for multiple comparisons with a reference, and $\{L_{jl}, j < l = 1, \ldots, J\}$ for all-pairwise comparisons.

We now consider how to compute these bounds. For convenience, we label the method pairs of interest as $1, \ldots, K$; the associated θs as $\theta_1, \ldots, \theta_K$; their upper bounds as $U_1, \ldots, U_K$; and their lower bounds as $L_1, \ldots, L_K$. In particular, $K = J - 1$ for comparisons with a reference and $K = \binom{J}{2}$ for all-pairwise comparisons. Also, let the vector $\underline{\theta}$ denote $(\theta_1, \ldots, \theta_K)$.

2.3.1 The frequentist approach

Let $\hat{\underline{\theta}}$ be the maximum likelihood estimator (MLE) of $\underline{\theta}$ obtained by fitting model (1) to the observed data. When the number of individuals m is large, the distribution of $\hat{\underline{\theta}}$ is approximately normal with mean $\underline{\theta}$. This suggests the following confidence bounds for $\theta_1, \ldots, \theta_K$:

$$L_k = \hat{\theta}_k - c_{1-\alpha,K}\, v_{kk}^{1/2}, \quad U_k = \hat{\theta}_k - d_{\alpha,K}\, v_{kk}^{1/2}, \quad k = 1, \ldots, K, \tag{5}$$

where v_{kk} is the asymptotic variance of $\hat{\theta}_k$, and $c_{1-\alpha,K}$ and $d_{\alpha,K}$ are critical points that ensure $(1-\alpha)$ coverage probability in the limit as m tends to infinity for each simultaneous bound. Appendix A provides formulas for v_{kk} and discuss two methods — a "standard" approach and a bootstrap approach — for computing the critical points.

2.3.2 The Bayesian approach

We now present a Bayesian method for computing simultaneous credible bounds for $\underline{\theta}$. Although this procedure can be used with any appropriate choice of prior distributions for the parameters in model (1), nevertheless we assume the following priors as they have become

quite common in practice (see, e.g., Spiegelhalter et al., 2003, Carlin, 1996):

$$\beta_j \sim \mathcal{N}(0, V_j^2),\ 1/\sigma_j^2 \sim \text{Gamma}(A_j, B_j),\ j = 1, \ldots, J,\ \Psi^{-1} \sim \text{Wishart}(\nu_0, R). \qquad (6)$$

These distributions are mutually independent and their parameters need to be specified. See Carlin (1996) for an example of how to elicit them. The joint posterior distribution of parameters in (1) with priors (6) is not available in a closed-form. Therefore, a Markov chain Monte Carlo (MCMC) approach is needed to simulate draws from the joint posterior. For this purpose, one can employ a Gibbs sampler algorithm (Gelman et al., 2003, chapter 11) or use the `WinBUGS` package of Spiegelhalter et al. (2003). Furthermore, instead of the inverse gamma distributions as priors for error variances in (6), one can also use the recent alternatives of Brown and Draper (2006) and Gelman (2006).

Let $\underline{\gamma}_1, \ldots, \underline{\gamma}_M$ denote a large number of draws from the joint posterior of the model parameter vector $\underline{\gamma}$. The corresponding draws from the posterior of θ_k, say $\{\theta_k^1, \ldots, \theta_k^M\}$, are obtained by using the fact that θ_k is a function of $\underline{\gamma}$, $k = 1, \ldots, K$. Next, let $\{\theta_k^{(1)}, \ldots, \theta_k^{(M)}\}$ denote the ordered values of $\{\theta_k^1, \ldots, \theta_k^M\}$ and $\{r_k^1, \ldots, r_k^M\}$ be their ranks. When upper bounds are desired, define $u_{1-\alpha}$ as the $(1-\alpha)$th sample quantile of $\{\max_{k=1}^K r_k^1, \ldots, \max_{k=1}^K r_k^M\}$. Since M is large, the posterior probability of the simultaneous event $\{\theta_k \leq \theta_k^{(u_{1-\alpha})}, k = 1, \ldots, K\}$ is approximately $(1 - \alpha)$ (Besag et al., 1995). Hence we can take $\{U_k = \theta_k^{(u_{1-\alpha})}, k = 1, \ldots, K\}$ as the simultaneous upper credible bound. Analogously, when lower bounds are desired, define l_α as the αth sample quantile of $\{\min_{k=1}^K r_k^1, \ldots, \min_{k=1}^K r_k^M\}$, and take $\{L_k = \theta_k^{(l_\alpha)}, k = 1, \ldots, K\}$ as the simultaneous lower credible bound.

2.4 Ranking of method pairs

We now describe two approaches for ranking the method pairs $\{1, \ldots, K\}$, say, in decreasing order of agreement. The first is based on the simultaneous bounds of the previous section. Both frequentist and Bayesian bounds can be used. Specifically, when upper bounds are of

interest, the ranks of $\{U_1, \ldots, U_K\}$ in ascending order, and when lower bounds are of interest, the ranks of $\{L_1, \ldots, L_K\}$ in descending order can be taken as the inferred ranks. We refer to this induced ordering as the *bound-based ordering*. There are, however, two flaws in this procedure. Firstly, it is rather ad hoc and provides no control over the probability of correct inference, i.e., inference that the induced ordering of θs matches with their true unknown ordering. This probability may be low or high depending upon whether the bivariate distributions of the measurement pairs are similar or not and whether the sample is small or large. In particular, this probability is $1/K!$ in the extreme case when the measurements from different methods have exchangeable distributions and hence all the θs are equal. Secondly, this procedure does not directly convey how uncertain the induced ordering is. Although we expect the uncertainty to be high if some of the bounds have similar magnitudes, it is helpful to have an explicit knowledge of this uncertainty.

The uncertainty issue can be resolved by using the posterior probabilities of all possible $K!$ orderings. The posterior probability of an ordering is simply the proportion of posterior draws of $\underline{\theta}$ that satisfy the given ordering. The ordering with the highest posterior probability can be inferred as the induced ordering of the K method pairs. We refer to it as the *probability-based ordering*. The uncertainty in the inferred ordering is small if its probability is near one. On the other hand, if two or more orderings have similar posterior probabilities, some method pairs probably have similar extent of agreement.

3 Monte Carlo Simulation study

In this section we describe the results of a simulation study to evaluate the performance of the inference procedures discussed in the previous section. Since these procedures can be used with any scalar measure of agreement, we also compare the results for two popular measures, CCC and TDI.

To keep the computations manageable, we restrict attention to a special case of model (1), wherein the random individual $\times$ method interaction b_{ij} is written as $b_{ij} = b_i + b^*_{ij}$, $j = 1, \ldots, J,\ i = 1, \ldots, m$. Here $b_i|\psi^2 \sim$ independent $\mathcal{N}(0, \psi^2)$, $b^*_{ij}|\phi^2 \sim$ independent $\mathcal{N}(0, \phi^2)$, and the two are mutually independent. This structure of b_{ij} implies that the diagonal elements of Ψ equal $\psi^2 + \phi^2$ and its off-diagonal elements equal ψ^2. When all $n_{ij} = 1$, b^*_{ij} is dropped from the model as it gets confounded with the error term. Further, in the Bayesian case, Gamma(A_ψ, B_ψ) and Gamma(A_ϕ, B_ϕ) distributions are adopted as priors for $1/\psi^2$ and $1/\phi^2$, respectively.

This simulation study assumes the following: number of methods $J = 3$, number of comparisons $K = 3$ (i.e., all-pairwise comparisons), $1-\alpha = 0.95$, p_0 for TDI $= 0.80$, $m \in \{30, 60\}$, $n \in \{1, 2, 3\}$, and three settings of model parameters, which are summarized in Table 1. Setting 1 has exactly the same extent of agreement between the method pairs. In settings 2 and 3, the method pairs have differing extent of agreement, with the differences being more substantial in setting 3 than setting 2. Further, in the Bayesian case, "noninformative" values are assigned to the parameters of the prior distributions. In particular, the variances of β_js are taken to be ∞, and the parameters As and Bs of the variance components are given a common value of 0.001. The details of the simulation procedure are provided in Appendix B. Here we only summarize the results.

Table 2 presents the estimated simultaneous coverage probabilities of the bounds for TDI and CCC computed using two frequentist approaches (standard and bootstrap) and a Bayesian approach. The coverage probabilities observed with the bootstrap approach are generally close to the nominal 0.95 level for both TDI and CCC, even with $m = 30$. These probabilities do not seem to depend on the parameter setting. The coverage probabilities of the Bayesian bounds range between 0.95-0.98. Thus, these bounds appear conservative and the TDI bounds seem more conservative than the CCC bounds. Both become more accurate as the methods become more dissimilar or as m increases. The coverage probabilities of

the standard frequentist bounds mostly lie between 0.91-0.94 in case of TDI and between 0.95-0.97 in case of CCC. Thus, the bounds appear anticonservative in the TDI case and conservative in the CCC case. These probabilities do not seem to be impacted by the parameter settings. There does not seem to be any marked effect of n on the coverage probabilities with any method. Overall, the bootstrap bounds appear the best as confidence bounds, albeit they also take the most time to compute. The Bayesian bounds are the next best except only in the case of CCC with $n = 1$, when the standard bounds appear a bit less conservative than them. In the light of this result, we only consider the bootstrap and the Bayesian bounds in the remainder of this section.

Table 3 reports the probability of correct inference for the three ranking procedures — two based on bounds and one based on posterior probabilities. The results are presented only for settings 2 and 3 since there is no "correct" ordering in case of setting 1. The probabilities for CCCs tend to be a bit higher than those for TDIs. The Bayesian procedures have somewhat higher probabilities than the bootstrap procedure for $n = 2, 3$ and the converse is true for $n = 1$. Between the two Bayesian procedures, the probability-based procedure has slightly higher probabilities than the bound-based procedure. The probabilities are practically one in case of setting 3. In case of setting 2 with $m = 30$, the probabilities for TDIs lie between 0.72-0.83 for the bootstrap procedure. Further, as expected, the probabilities increase until they reach one as m or n increase or methods become more dissimilar.

We next consider the probability that the two Bayesian procedures, one based on credible bounds and the other based on posterior probabilities, lead to the same ordering. The estimates of this probability, not presented, are practically identical for TDI and CCC. They are fairly high — between 0.82-0.86 for setting 1, between 0.92-0.96 for setting 2, and practically one in case of setting 3. Setting 1 has the smallest probabilities, possibly because there is no correct ordering to agree upon in this case. The probabilities increase until they reach one as m increases or as the methods become more dissimilar. As for n, they increase

from $n = 1$ to $n = 2$ and then remain practically constant.

Table 4 presents the estimated probabilities that TDI and CCC lead to the same ordering. In general, these results suggest that, except when the methods have similar pairwise agreement, it does not matter much whether TDI or CCC is used for ordering method pairs as they are likely to produce the same inference. The probabilities are almost one for the posterior probability-based procedure, no matter what m, n or the parameter settings are. Among the bound-based procedures, the bootstrap tends to have higher probabilities than the Bayesian procedure. Here also setting 1 with $m = 30$ has the smallest probabilities — ranging between 0.90-0.98 for the bootstrap procedure and 0.84-0.88 for the Bayesian procedure. The other two settings have probabilities of 0.94 or more. Further, the probabilities for $n = 2$ and $n = 3$ seem similar. But they tend to be less than those for $n = 1$ in case of setting 1 and at least as large as those in case of settings 2,3. As expected, the probabilities increase with m or as the methods become more dissimilar, until they reach one.

4 Extension and application

The model (1) assumed in Section 2 is quite basic. Its extensions are needed to incorporate covariate effects and complex data structures. A general mixed-model framework for data from $J = 2$ methods and how to express θ — the measure of agreement between them — as a function of the model parameter vector $\underline{\gamma}$ have been discussed in Choudhary (2008). It is straightforward to extend this modeling approach for data from $J > 2$ methods and to express $\theta_1, \ldots, \theta_K$ — the values of θ for all pairs of methods of interest — as functions of $\underline{\gamma}$. Once these expressions are available, the frequentist and Bayesian methodologies developed in Section 2 can be employed by replacing the model (1) with the new model. This is illustrated below using an example. We will focus only on TDI and CCC. The other

agreement measures can be handled analogously.

Example (Blood pressure study): In this study (Torun et al., 1998, Barnhart and Williamson, 2001, Barnhart et al., 2002), systolic blood pressure (SBP) and diastolic blood pressure (DBP) are measured using four methods — three observers using a mercury sphygmomanometer (MS), say, methods 1, 2, 3, and one observer using an inexpensive digital sphygmomanometer (DS), say, method 4. The DS method is easier to use than the MS method. All four methods are used once on each of 228 individuals. Thus, we have 8 measurements (4 SBP and 4 DBP) from every individual. They range between 82-236 mmHg for SBP, and between 50-148 mmHg for DBP. These data have been analyzed in the above articles using pairwise and overall CCCs. Here we compare the extent of agreement between the three MS observers, and the agreement between the MS observers and the DS observer.

Let Y_{ijt} be the BP measurement of the tth type (type 1 = SBP, type 2 = DBP) from the jth method on the ith individual. After a preliminary analysis, we model these data as

$$Y_{ijt} = \beta_{jt} + b_i + b^*_{ij} + b_{it} + \epsilon_{ijt},\ i = 1, \ldots, 228,\ j = 1, \ldots, 4,\ t = 1, 2, \tag{7}$$

where β_{jt} is the fixed mean for the jth method and tth type; b_i is the random effect of ith individual; b^*_{ij} is the random individual $\times$ method interaction; b_{it} is the random individual $\times$ type interaction; and ϵ_{ijt} is the random error. Here $b_i|\psi^2 \sim$ independent $\mathcal{N}(0, \psi^2)$, $b^*_{ij}|\phi^2 \sim$ independent $\mathcal{N}(0, \phi^2)$, $b_{it}|\xi^2 \sim$ independent $\mathcal{N}(0, \xi^2)$, $\epsilon_{ijt}|\sigma^2_{jt} \sim$ independent $\mathcal{N}(0, \sigma^2_{jt})$, and the random variables are mutually independent. We now proceed along the lines of Section 2.2 to deduce that the CCC and the TDI for evaluating agreement between methods (j, l) for measuring tth type of BP can be expressed as

$$\rho_{jlt} = \frac{2(\psi^2 + \xi^2)}{\mu^2_{jlt} + \tau^2_{jlt} + 2(\psi^2 + \xi^2)}, \quad q_{jlt} = \tau_{jlt}\left\{\chi^2_1(p_0, \mu^2_{jlt}/\tau^2_{jlt})\right\}^{1/2}, \tag{8}$$

respectively, where $\mu_{jlt} = \beta_{jt} - \beta_{lt}$ and $\tau^2_{jlt} = 2\phi^2 + \sigma^2_{jt} + \sigma^2_{lt}$, $j < l = 1, \ldots, 4$, $t = 1, 2$.

Next, we obtain maximum likelihood and Bayesian fits of this model. The `nlme` package (Pinheiro et al., 2006) in `R` (R Development Core Team, 2007) is used for the former. For

the latter, we use `WinBUGS` by calling it in `R` through the `R2WinBUGS` package of Sturtz, Ligges and Gelman (2005), and assume the following mutually independent vague priors — $\mathcal{N}(0, 10^4)$ for βs and Gamma$(10^{-3}, 10^{-3})$ for reciprocals of each of the variance components. Table 5 presents the MLEs and posterior means of 12 TDIs (with $p_0 = 0.80$) and 12 CCCs, as defined in (8), and their respective simultaneous bounds with $1 - \alpha = 0.95$. The two sets of point estimates are very similar. Moreover, the bounds from the three methods — standard, bootstrap and Bayesian — are remarkably identical. Further, since all the lower bounds for CCC are near one, the four methods have sufficiently high agreement for their interchangeable use. The same inference is reached on the basis of TDI upper bounds as it appears from White et al. (1993) that differences in BP measurements of about 10 mmHg is considered clinically acceptable.

The bounds also reveal that the methods tend to agree more in case of DBP ($t = 2$) than SBP ($t = 1$). The posterior probability of this event, i.e., $\{\rho_{jl2} > \rho_{jl1}, \text{ for all } j < l = 1, \ldots, 4\}$ or $\{q_{jl2} < q_{jl1}, \text{ for all } j < l = 1, \ldots, 4\}$, approximately equals 0.96 in both cases. Moreover, the agreement between any two of the three MS observers (i.e., methods 1, 2, 3) tends to be higher than the agreement between an MS observer and the DS observer. This event, i.e., $\{\min(\rho_{12t}, \rho_{13t}, \rho_{23t}) > \max(\rho_{14t}, \rho_{24t}, \rho_{34t})\}$ or $\{\max(q_{12t}, q_{13t}, q_{23t}) < \min(q_{14t}, q_{24t}, q_{34t})\}$, has posterior probability 1 for SBP and 0.74 for DBP.

We now consider ordering the method pairs. Among the MS observers, both TDI and CCC bounds are practically identical for all three pairs of observers, suggesting that it is difficult to accurately rank the method pairs. In fact, in case of SBP, the two most probable orderings are $\{(1,3) \succ (1,2) \succ (2,3)\}$ and $\{(1,3) \succ (2,3) \succ (1,2)\}$, with respective posterior probabilities 0.54 and 0.33. (Here "$\succ$" means "is better than.") For DBP, they are $\{(2,3) \succ (1,3) \succ (1,2)\}$ and $\{(1,3) \succ (2,3) \succ (1,2)\}$, with respective probabilities 0.41 and 0.36. Since none of these probabilities is near one, the uncertainty in these orderings is rather large. A closer examination reveals that, among the three pairs, the observers $(1,3)$

for SBP agree the most while the observers $(1, 2)$ for DBP agree the least. These events have fairly high posterior probabilities — 0.87 and 0.77, respectively. Thus, it is clear that the uncertainty in the orderings is due to similar extent of agreement in the remaining MS observer pairs, namely $(1, 2)$ and $(2, 3)$ for SBP, and $(1, 3)$ and $(2, 3)$ for DBP.

When the DS observer is compared with the MS observers, the orderings induced by both TDI and CCC bounds are $\{(\text{DS}, \text{MS3}) \succ (\text{DS}, \text{MS1}) \succ (\text{DS}, \text{MS2})\}$ for SBP and $\{(\text{DS}, \text{MS3}) \succ (\text{DS}, \text{MS2}) \succ (\text{DS}, \text{MS1})\}$ for DBP. The probability-based orderings concur with these bound-based orderings. Moreover, since their posterior probabilities — 0.85 in case of SBP and 0.70 in case of DBP — are high, these inferences are fairly precise.

5 Discussion

In this article, we discussed frequentist and Bayesian methodologies for multiple comparisons and ranking of method pairs on the basis of their extent of agreement. For ranking, we discovered that it does not matter whether a TDI or a CCC is used to measure agreement since they normally produce the same ordering. We also discovered that, from a frequentist viewpoint, the bootstrap procedure is somewhat better than a Bayesian procedure with noninformative priors for producing simultaneous bounds with coverage probabilities close to the nominal level in samples of moderate sizes, and both are generally better than the standard frequentist procedure.

The Bayesian paradigm, on the other hand, has two advantages over a bootstrap procedure — it additionally offers an inference procedure based on posterior probabilities, which is especially useful for ordering method pairs, and there is also the flexibility of being able to take into account of any available prior information. The Bayesian procedure also takes less time to implement than the bootstrap procedure. Nevertheless, to use the Bayesian approach one has to consider issues such as prior specification and sensitivity, and implementing an

MCMC algorithm and diagnosing its convergence. But there are standard ways to deal with them for the models discussed in this article (see Gelman et al., 2003, chapter 11).

We have assumed normality for the data as the measurements in method comparison studies tend to be highly dependent and it is difficult to model them flexibly outside the framework of mixed models. However, the methodologies described here can be used with other parametric models as well by appropriately modifying the model specification. Some authors (e.g., Barnhart and Williamson, 2001, and Barnhart, Song and Haber, 2005) have used a semiparametric GEE approach to model method comparison data. But since in this case one models the moments of the distribution, it is useful only for performing inference on agreement measures that are defined in terms of moments (e.g., a CCC), and not on a measure such as a TDI, which is a quantile.

A limitation of the proposed methodologies is that they are not reliable for very small sample sizes, unless a Bayesian approach is used with informative priors. They do, however, seem to work well when 30 or more individuals are in the sample. Finally, our focus here has been only on the analysis of method comparison studies involving multiple methods. We do not consider how to design such studies, in particular, how to compute sample sizes for planning such studies. This issue is currently under investigation.

Appendix A: Computation of the simultaneous confidence bounds

Let $L(\underline{\gamma})$ be the likelihood function of the parameter vector $\underline{\gamma}$ assuming model (1) for the observed data. Further, let $\hat{\underline{\gamma}}$ be the MLE of $\underline{\gamma}$. When m is large, it is well-known that the approximate distribution of $\hat{\underline{\gamma}}$ is $\mathcal{N}\left(\underline{\gamma}, I^{-1}\right)$, where $I = -(\partial^2 \log L(\underline{\gamma})/\partial \underline{\gamma}^2)|_{\underline{\gamma}=\hat{\underline{\gamma}}}$ is the observed Fisher information matrix (Lehmann, 1999, chapter 7). Assuming that $\underline{\theta}$ is a differentiable function of $\underline{\gamma}$, it follows from the delta method (Lehmann, 1999, chapter 5) that the approximate distribution of $\hat{\underline{\theta}}$ is $\mathcal{N}\left(\underline{\theta}, GI^{-1}G'\right)$, where $G = (\partial \underline{\theta}/\partial \underline{\gamma})|_{\underline{\gamma}=\hat{\underline{\gamma}}}$ is the Jacobian matrix and G' is the transpose of G.

Thus, the asymptotic variance v_{kk} that appears in the confidence bounds (5) is the kth diagonal element of $GI^{-1}G'$. Moreover, it can be easily seen that the critical points $c_{1-\alpha,K}$ and $d_{\alpha,K}$ in (5) are the $(1-\alpha)$th quantile of $Z_{\max} = \max_{k=1}^{K} Z_k$ and the αth quantile of $Z_{\min} = \min_{k=1}^{K} Z_k$, respectively, where $Z_k = (\hat{\theta}_k - \theta_k)/v_{kk}^{1/2}$, and the distribution of $(Z_1, \ldots, Z_K)$ is multivariate normal with mean zero and covariance matrix obtained by pre- and post-multiplying $GI^{-1}G'$ with a diagonal matrix $\text{diag}\{v_{11}^{-1/2}, \ldots, v_{KK}^{-1/2}\}$.

One way to compute the critical points is to directly use their definitions and apply the method of Hothorn, Bretz and Westfall (2008), which uses an algorithm of Genz (1992) for efficient computation of multivariate normal probabilities. We call it the "standard" approach. It is expected to work well when the normal approximation for $\underline{\hat{\theta}}$ is good, but it may require m to be quite large. A better alternative for moderate m is to employ the studentized bootstrap method (Davison and Hinkley, 1997). It consists of the following steps:

(i) Simulate resampled data Y_{ijk}^*, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, J$, $i = 1, \ldots, m$, from model (1) taking $\underline{\gamma} = \underline{\hat{\gamma}}$.

(ii) Fit model (1) to the resampled data using maximum likelihood. Let $\underline{\hat{\gamma}}^*$ be the resulting MLE of $\underline{\gamma}$ and $\underline{\hat{\theta}}^* = (\hat{\theta}_1^*, \ldots, \hat{\theta}_K^*)$ be the MLE of $\underline{\theta}$. Also, let v_{kk}^* be the asymptotic variance of $\hat{\theta}_k^*$. Compute $Z_k^* = (\hat{\theta}_k^* - \hat{\theta}_k)/v_{kk}^{*1/2}$, $k = 1, \ldots, K$.

(iii) If lower bounds are desired, compute $Z_{\max}^* = \max_{k=1}^{K} Z_k^*$, and if upper bounds are desired, compute $Z_{\min}^* = \min_{k=1}^{K} Z_k^*$.

(iv) Repeat the steps (i)-(iii) a large number of times, say B, to get B draws of $Z_{\max}^*$ or $Z_{\min}^*$. Take the $(1-\alpha)$th sample quantile of the draws of $Z_{\max}^*$ as the critical point $c_{1-\alpha,K}$, and the αth sample quantile of the draws of $Z_{\min}^*$ as the critical point $d_{\alpha,K}$.

Since the confidence bounds in (5) are large-sample approximations, their accuracy can generally be improved by first applying a normalizing transformation to the agreement mea-

sure, computing the bounds on the transformed scale, and then applying the inverse transformation to get the bounds on the original scale. In particular, the log transformation in case of TDI (Lin, 2000) and the Fisher's-z transformation in case of CCC (Lin, 1989) are known to produce more accurate bounds than their original scale counterparts.

Appendix B: Details of the Monte Carlo simulation procedure

The probability estimates reported in Section 3 are computed using the following steps at each combination of settings: (a) simulate data from the assumed model; (b) fit it to the simulated data using maximum likelihood or the Bayesian approach; (c) compute simultaneous upper bounds (U_{12}, U_{13}, U_{23}) for the three TDIs (q_{12}, q_{13}, q_{23}) and simultaneous lower bounds (L_{12}, L_{13}, L_{23}) for the three CCCs $(\rho_{12}, \rho_{13}, \rho_{23})$; (d) deduce the bound-based ordering in case of the frequentist approach, and both the bound-based and the probability-based orderings in case of the Bayesian approach; (e) check whether the event of interest happens; (f) repeat steps (a)-(e) 1,000 times in case of the bootstrap approach and 2,000 times in case of others, and calculate the proportion of times the event of interest happens. This proportion is the desired estimated probability. For posterior simulation in (b), an adaptation of the Gibbs sampler algorithm in Yin et al. (2008) is employed. The Gibbs sampler is run for 2,000 iterations and the first 500 iterations are discarded as burn-in. Moreover, 500 resamples are used for the computation of bootstrap critical points in (c). Finally, for frequentist bounds in (c), we apply the log transformation of TDI and the Fisher's-z transformation of CCC.

The computations are programmed in `R` (R Development Core Team, 2007). We use the `nlme` package (Pinheiro et al., 2006) to get MLEs, the `numDeriv` package (Gilbert, 2006) to get the derivatives needed for their asymptotic variances, and the `multcomp` package (Hothorn et al., 2008) to get standard critical points in (5). Analysis of one simulated data set with $(m, n) = (60, 3)$, on a Linux machine with 2.33 GHz processor and 2 GB RAM, takes about 17 seconds for the standard approach, 4 minutes for the Bayesian approach and

21 minutes for the bootstrap approach.

Acknowledgment

The authors thank Professors Huiman Barnhart and Michael Haber for providing the blood pressure data. They also thank the reviewers and the Associate Editor for their comments that greatly improved this article.

References

[1] Barnhart, H. X. and Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibiltiy. *Biometrics* **57**, 931–940.

[2] Barnhart, H. X., Haber, M. J. and Lin, L. I. (2007). An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* **17**, 529–569.

[3] Barnhart, H. X., Haber, M. J. and Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* **58**, 1020–1027.

[4] Barnhart, H. X., Kosinski, A. S. and Haber, M. J. (2007). Assessing individual agreement. *Journal of Biopharmaceutical Statistics* **17**, 697–719.

[5] Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3–41.

[6] Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310.

[7] Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1**, 473 – 514.

[8] Carlin, B. P. (1996). Hierarchical Longitudinal Modelling. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 303-319. Chapman and Hall/CRC, Boca Raton.

[9] Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* **138**, 1102–1115.

[10] Choudhary, P. K. and Nagaraja, H. N. (2005a). Selecting the instrument closest to a gold standard. *Journal of Statistical Planning and Inference* **129**, 229–237.

[11] Choudhary, P. K. and Nagaraja, H. N. (2005b). A two-stage procedure for selection and assessment of agreement of the best instrument with a gold standard. *Sequential Analysis* **24**, 237–257.

[12] Choudhary, P. K. and Nagaraja, H. N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* **137**, 279–290.

[13] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, New York.

[14] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515 – 534.

[15] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edn. Chapman and Hall/CRC, Boca Raton.

[16] Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–149.

[17] Gilbert, P. (2006). *numDeriv: Accurate Numerical Derivatives*. R package version 2006.4-1.

[18] Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

[19] Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, Boca Raton, FL.

[20] Hutson, A. D., Wilson, D. C. and Geiser, E. A. (1998). Measuring relative agreement: Echocardiographer versus computer. *Journal of Agricultural, Biological, and Environmental Statistics* **3**, 163–174.

[21] King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**, 2131 – 2147.

[22] Lehmann, E. L. (1999). *Elements of Large Sample Theory*. Springer-Verlag, New York.

[23] Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268. Corrections: 2000, 56:324-325.

[24] Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* **19**, 255–270.

[25] Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* **97**, 257–270.

[26] Moura, C., Fontes-Sousa, A. P., Teixeira-Pinto, A., Areias, J. C. and Leite-Moreira, A. F. (2009). Agreement between echocardiographic techniques in assessment of the left ventricular myocardial performance index in rabbits. *American Journal of Veterinary Research* **70**, 464–471.

[27] Pinheiro, J., Bates, D., DebRoy, S. and Sarkar, D. (2006). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-76.

[28] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

[29] Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. http://www.mrc-bsu.cam.ac.uk/bugs.

[30] St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics* **54**, 537–545.

[31] Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* **12**, 1–16.

[32] Torun, B., Grajeda, R., Mendez, H., Flores, R., Martorell, R. and Schroeder, D. (1998). Evaluation of inexpensive digital sphygmomanometers for field studies of blood pressure. *Federation of American Societies of Experimental Biology Journal* **12**, S5072.

[33] Welinder, A. E., Wagner, G. S., Horacek, B. M., Martin, T. N., Maynard, C. and Pahlm, O. (2009). EASI-Derived vs standard 12-lead electrocardiogram for Selvester QRS score estimations of chronic myocardial infarct size, using cardiac magnetic resonance imaging as gold standard. *Journal of electrocardiology* **42**, 145–51.

[34] White, W.B., Berson, A.S., Robbins, C., Jamieson, M.J., Prisant, L.M., Roccella, E. and Sheps, S.G. (1993). National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers. *Hypertension* **21**, 504–509.

[35] Yin, K., Choudhary, P. K., Varghese, D. and Goodman, S. R. (2008). A Bayesian approach for sample size determination in method comparison studies. *Statistics in Medicine* **27**, 2273–2289.

Setting	$(\beta_1, \beta_2, \beta_3)$	$(\sigma_1, \sigma_2, \sigma_3)$	(ψ, ϕ)	(q_{12}, q_{13}, q_{23})	$(\rho_{12}, \rho_{13}, \rho_{23})$
1	$(0, 0, 0)$	$(1, 1, 1)$	$(4, 0.5)$	$(2.03, 2.03, 2.03)$	$(0.93, 0.93, 0.93)$
2	$(0, 0.5, 1)$	$(1, 0.75, 1.25)$	$(4, 0.5)$	$(1.95, 2.60, 2.18)$	$(0.93, 0.89, 0.92)$
3	$(0, 1, 2)$	$(1, 0.5, 2)$	$(4, 0.5)$	$(2.15, 4.02, 3.08)$	$(0.92, 0.77, 0.85)$

Table 1: Parameter settings used in the simulation study and the resulting true values of pairwise TDIs (qs) and CCCs (ρs).

	Standard			Bootstrap			Bayesian		
	n			n			n		
Setting	1	2	3	1	2	3	1	2	3
Total deviation index ($m = 30$)									
1	92.0	93.7	93.3	95.5	95.1	94.6	98.0	98.1	97.8
2	90.6	93.2	92.4	95.8	93.4	95.3	97.1	96.6	96.7
3	91.6	92.2	92.4	95.1	93.9	94.7	96.5	95.7	95.7
Total deviation index ($m = 60$)									
1	91.5	94.6	94.3	94.8	96.2	96.0	96.4	96.7	96.8
2	92.5	92.4	92.8	95.2	94.5	95.0	97.2	96.0	95.7
3	92.0	94.2	93.0	95.7	95.1	96.6	96.2	95.1	95.4
Concordance correlation coefficient ($m = 30$)									
1	95.4	96.2	97.2	94.5	95.8	95.9	97.0	96.7	96.8
2	94.9	96.8	97.2	95.4	94.9	94.7	96.5	96.7	96.2
3	95.8	96.0	96.9	95.4	94.5	94.0	95.5	96.5	95.6
Concordance correlation coefficient ($m = 60$)									
1	94.6	96.6	96.0	94.4	95.6	94.4	96.9	95.7	95.6
2	95.0	96.2	96.4	95.6	95.2	94.6	96.3	95.4	95.0
3	95.8	96.4	97.0	95.7	94.8	95.3	96.2	95.1	95.0

Table 2: Estimated coverage probabilities (%) of simultaneous bounds with $1 - \alpha = 0.95$. Here "m" and "n" respectively stand for number of individuals and number of replicate measurements per individual. The parameter settings are defined in Table 1.

		Bootstrap			Bayesian					
		bound-based			bound-based			prob.-based		
		n			n			n		
m	Setting	1	2	3	1	2	3	1	2	3
Total deviation index										
30	2	72	81	83	69	83	86	69	84	89
30	3	99	100	100	99	100	100	99	100	100
60	2	83	90	94	84	91	94	84	92	96
60	3	100	100	100	100	100	100	100	100	100
Concordance correlation coefficient										
30	2	73	84	88	69	84	89	69	85	89
30	3	99	100	100	99	100	100	99	100	100
60	2	84	92	95	84	93	95	84	93	96
60	3	100	100	100	100	100	100	100	100	100

Table 3: Estimated probabilities (%) of inferring the correct ordering, which is $q_{12} < q_{23} < q_{13}$ for TDIs and $\rho_{12} > \rho_{23} > \rho_{13}$ for CCCs. Here "m" and "n" respectively stand for number of individuals and number of replicate measurements per individual.

	Bootstrap			Bayesian					
	bound-based			bound-based			prob.-based		
	n			n			n		
Setting	1	2	3	1	2	3	1	2	3
$m = 30$									
1	98	90	91	88	84	84	99	99	99
2	97	97	95	94	96	96	99	99	100
3	100	100	100	99	100	100	100	100	100
$m = 60$									
1	98	93	96	90	87	85	100	100	100
2	99	98	99	97	98	99	99	99	100
3	100	100	100	100	100	100	100	100	100

Table 4: Estimated probabilities (%) that the orderings induced using TDI and CCC are the same. Here "m" and "n" respectively stand for number of individuals and number of replicate measurements per individual.

	TDI	CCC
	$(q_{12}, q_{13}, q_{23}, q_{14}, q_{24}, q_{34})$	$(\rho_{12}, \rho_{13}, \rho_{23}, \rho_{14}, \rho_{24}, \rho_{34})$
Systolic blood pressure		
Point estimates		
MLE	(6.29, 5.88, 6.39, 8.97, 9.24, 8.32)	(0.980, 0.982, 0.979, 0.960, 0.957, 0.965)
Bayesian	(6.36, 5.93, 6.44, 9.05, 9.32, 8.39)	(0.979, 0.982, 0.979, 0.960, 0.957, 0.965)
Simultaneous bounds		
Standard	(7.07, 6.61, 7.18, 10.05, 10.27, 9.32)	(0.973, 0.976, 0.972, 0.947, 0.945, 0.954)
Bootstrap	(7.11, 6.65, 7.22, 10.10, 10.33, 9.37)	(0.973, 0.977, 0.973, 0.948, 0.945, 0.955)
Bayesian	(7.07, 6.62, 7.24, 10.13, 10.48, 9.37)	(0.973, 0.976, 0.973, 0.948, 0.944, 0.955)
Diastolic blood pressure		
Point estimates		
MLE	(5.43, 5.17, 5.11, 6.43, 6.29, 5.66)	(0.985, 0.986, 0.986, 0.979, 0.980, 0.984)
Bayesian	(5.48, 5.21, 5.15, 6.50, 6.35, 5.72)	(0.985, 0.986, 0.986, 0.979, 0.980, 0.983)
Simultaneous bounds		
Standard	(6.05, 5.85, 5.72, 7.18, 7.03, 6.30)	(0.980, 0.982, 0.982, 0.972, 0.973, 0.978)
Bootstrap	(6.09, 5.89, 5.75, 7.22, 7.07, 6.34)	(0.980, 0.982, 0.982, 0.973, 0.974, 0.979)
Bayesian	(6.15, 5.93, 5.81, 7.25, 7.10, 6.32)	(0.980, 0.982, 0.982, 0.973, 0.974, 0.978)

Table 5: MLEs and posterior means of 12 TDIs and 12 CCCs (6 for systolic and 6 for diastolic BP) and their respective simultaneous bounds with $1 - \alpha = 0.95$ in case of blood pressure study. The bounds are upper bounds in case of TDI and are lower bounds in case of CCC.