

# A Tolerance Interval Approach for Assessment of Agreement With Left Censored Data

Pankaj K. Choudhary<sup>1</sup>

Department of Mathematical Sciences, University of Texas at Dallas

## Abstract

We describe a tolerance interval approach for assessing agreement in method comparison data that may be left censored. We model the data using a mixed model and discuss a Bayesian and a frequentist methodology for inference. A simulation study suggests that the Bayesian approach provides a good alternative to the frequentist one for moderate sample sizes as the latter tends to be liberal. Both may be used for sample sizes 100 or more, with the Bayesian one being slightly conservative. The proposed methods are illustrated with real data involving comparison of two assays for quantifying viral load in HIV patients.

**Key Words:** Concordance correlation; Left censoring; Limit of detection; Limits of agreement; Method comparison; Mixed model; Total deviation index.

**Short title:** Assessing agreement with left censored data.

---

<sup>1</sup>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

We consider the problem of assessing agreement in two competing methods of measuring a continuous variable when the measurements may be subject to left censoring. This type of censoring arises in practice when the methods have lower limits of detection. A measurement is not completely observed if it falls below the detection limit. We assume that the detection limits are known, however, they may be different for the two methods. A method here may be an assay, an instrument, a medical device or a technique or technology. Moreover, the variable being measured typically has clinical importance — its reading on an individual forms the basis of his/her health evaluation. Consider, for example, the viral load data of (Manegold et al., 2000). The goal of their study is to comparatively evaluate two branched-DNA assays — Quantiplex versions 2.0 and 3.0, for quantifying human immunodeficiency virus (HIV) type 1 RNA. Both assays have lower limits of detection: 500 copies/ml for 2.0 and 50 copies/ml for 3.0. The authors say that the concentration of this RNA, also known as the viral load, is strongly correlated with HIV disease progression. We focus on their viral load measurements obtained from 132 plasma samples using the two assays. They range from 50 to 50,000 copies/ml. Among them, the proportion of censored measurements from versions 2.0 and 3.0 are 59.8% and 25%, respectively. Furthermore, 38.6% measurements are completely observed (uncensored) simultaneously on both assays. These data will be analyzed on log (viral load/1000) scale for better adherence with model assumptions. Figure 1 presents the scatter plot of the uncensored measurements. They appear highly correlated and seem to have similar variability. However, the 3.0 measurements are almost always higher than their 2.0 counterparts. We revisit these data in Section 4.

Let the random vector  $(y_1, y_2)$  represent the population of measurement pairs from the two methods. Also, let  $(y_{i1}, y_{i2})$ ,  $i = 1, \dots, n$ , denote a random sample from this population. We model these data using the so-called *Grubbs' model* (Dunn and Roberts, 1999):

$$y_{ij} = \mu_j + b_i + \epsilon_{ij}; \quad i = 1, \dots, n, \quad j = 1, 2; \quad (1)$$

where  $b_i$  is the random effect of the  $i$ -th individual representing his/her true unobservable measurement; and  $\mu_j$  is the fixed mean of the  $j$ -th method. It is assumed that  $b_i \sim$  independent  $\mathcal{N}(0, \sigma_b^2)$ ,  $\epsilon_{ij} \sim$  independent  $\mathcal{N}(0, \sigma_j^2)$ , and the two are mutually independent. The model (1) is a *mixed model* (see Pinheiro and Bates, 2000). It is also known as a *variance components model* (see Searle, Casella and McCulloch, 1992). Further,  $\sigma_b^2$  is the between-individual variance and  $\sigma_j^2$  is the measurement error variance of the  $j$ -th method. These assumptions imply that  $(y_1, y_2)$  has a bivariate normal distribution with mean  $(\mu_1, \mu_2)$ , variance  $(\tau_1^2, \tau_2^2) = (\sigma_b^2 + \sigma_1^2, \sigma_b^2 + \sigma_2^2)$  and correlation  $\rho = \sigma_b^2 / (\tau_1 \tau_2)$ . Let  $d = y_1 - y_2 \sim \mathcal{N}(\mu = \mu_1 - \mu_2, \sigma^2 = \sigma_1^2 + \sigma_2^2)$  denote the population of differences in the measurement pairs. This distribution is free of  $\sigma_b^2$ . Next, let  $(l_1, l_2)$  be the known lower detection limits for  $(y_1, y_2)$  methods. Due to left censoring, we may not observe  $y_{ij}$  completely. Instead we observe  $x_{ij} = \max\{y_{ij}, l_j\}$ ,  $i = 1, \dots, n$ ,  $j = 1, 2$ .

We concentrate on the *tolerance interval* methodology of (Lin, 2000, and Choudhary and Nagaraja, 2007) for the assessment of agreement in  $(y_1, y_2)$ . See (Guttman, 1988) for an introduction to tolerance intervals. In essence, it provides an interval that estimates the range of a specified large proportion of population of measurement differences. The practitioner infers satisfactory agreement if the differences in this interval are not clinically important. There are several other approaches for assessing agreement — prominent among these are

the *limits of agreement* of (Bland and Altman, 1986) and the *concordance correlation* of (Lin, 1989). See the reviews of (Lin et al., 2002) and (Choudhary and Nagaraja, 2004) for their discussion. Taking into account of censoring in agreement analysis is important otherwise the variability may be substantially underestimated. This has been demonstrated in (Barnhart, Song and Lyles, 2005), who have developed a concordance correlation for left censored data. Assuming bivariate normality for the data, they propose inference procedures based on maximum likelihood and generalized estimating equations.

In the tolerance interval approach, we take the  $p_0$ -th quantile of absolute differences as the measure of agreement, where  $p_0$  is a specified large probability. In this article, we discuss two such measures. The first is  $q$  — the  $p_0$ -th quantile of  $|d|$ , also called *total deviation index* by (Lin, 2000). Under the model (1), it can be defined as

$$q = \sigma \{ \chi_1^2(p_0, \mu^2/\sigma^2) \}^{1/2}, \quad (2)$$

where  $\chi_1^2(p_0, \Delta)$  represents the  $p_0$ -th quantile of a noncentral chisquare distribution with one degree of freedom and noncentrality parameter  $\Delta$ . A small value of  $q$  indicates a good agreement in  $(y_1, y_2)$ . In (Lin, 2000), it has been stated that inference on  $q$  is cumbersome. So an approximation to it is suggested that works well when  $\mu/\sigma$  is small and allows a simple asymptotic inference. Recently (Choudhary and Nagaraja, 2007) have discussed both exact and asymptotic inferences for  $q$ .

In presence of left censoring with possibly different detection limits, one may also be interested in quantifying the agreement in the conditional population  $(y_1, y_2) | \{y_1, y_2 > l\}$ , where  $l = \max\{l_1, l_2\}$ , rather than the whole bivariate population  $(y_1, y_2)$ . Our second measure  $q_c$  — the  $p_0$ -th quantile of  $|d|$  given  $\{y_1, y_2 > l\}$  is designed for this purpose. It

reduces to  $q$  when there is no lower limit of detection. Unfortunately, it does not have a closed-form expression, but can be computed numerically by solving,

$$Pr(|d| \leq q_c | y_1, y_2 > l) = \frac{1}{S(l, l)} \int_l^\infty \left\{ F_{1|2}(u + q_c) - F_{1|2}(\max\{u - q_c, l\}) \right\} f_2(u) du = p_0, \quad (3)$$

where  $S(u_1, u_2) = Pr(y_1 > u_1, y_2 > u_2)$  is the survivor function of  $(y_1, y_2)$ ,  $F_{1|2}(\cdot)$  is the conditional cumulative distribution function (cdf) of  $y_1 | y_2 \sim \mathcal{N}(\mu_1 + \rho(\tau_1/\tau_2)(y_2 - \mu_2), \tau_1^2(1 - \rho^2))$ , and  $f_2(\cdot)$  is the probability density function (pdf) of  $y_2 \sim \mathcal{N}(\mu_2, \tau_2^2)$ . Alternatively, it can be approximated using Monte Carlo (MC) simulation. The measure  $q$  is free of  $\sigma_b^2$  or the detection limits. On the other hand,  $q_c$  is a function of all five parameters of  $(y_1, y_2)$ , and also depends on the detection limits through  $l$ .

Our goal here is to develop procedures for obtaining upper bounds  $U$  and  $U_c$  that satisfy

$$Pr(q \leq U) = 1 - \alpha, \quad Pr(q_c \leq U_c) = 1 - \alpha, \quad (4)$$

where  $(1 - \alpha)$  is a specified large coverage probability. Then the intervals  $[-U, U]$  and  $[-U_c, U_c]$  can be interpreted as  $p_0$  probability content tolerance intervals for the respective distributions of  $d$  and  $d | \{y_1, y_2 > l\}$ . In other words, we have

$$Pr(H(U) - H(-U) \geq p_0) = 1 - \alpha, \quad Pr(H_c(U_c) - H_c(-U_c) \geq p_0) = 1 - \alpha,$$

with  $H$  as the cdf of  $d$  and  $H_c$  as the cdf of  $d | \{y_1, y_2 > l\}$ . In Section 2, we describe a likelihood based frequentist procedure and a Bayesian procedure for obtaining  $U$  and  $U_c$ . Simulation studies in Section 3 suggest that both can be used for  $n \geq 100$ . However, for smaller values of  $n$ , the Bayesian procedure is preferable as the frequentist one tends to be quite liberal. We illustrate their application to the viral load data in Section 4. We conclude in Section 5 with a short discussion. The Appendix contains some technical details.

## 2 Methodology for computing upper bounds

The observed data here can be represented as the pairs  $(x_{ij}, c_{ij})$ , where  $c_{ij}$  is the censoring indicator:  $c_{ij} = 1$  if  $x_{ij} = y_{ij}$ , i.e., the observation is complete; and  $c_{ij} = 0$  if  $x_{ij} = l_j$ , i.e., the observation is censored;  $i = 1, \dots, n$ ,  $j = 1, 2$ .

### 2.1 The frequentist approach

Let  $f$  and  $F$  be the joint pdf and cdf of  $(y_1, y_2)$  under the model (1). Further, let  $\theta$  denote the column vector of its five parameters  $(\mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \log \sigma_b)$ . The positive parameters are transformed to log scale to get an unconstrained parametrization for maximizing the log likelihood function, say  $L(\theta)$ . It can be expressed as  $L(\theta) = \sum_{i=1}^n \log L_i(\theta)$ , where  $L_i(\theta)$  is the contribution of the observed data on the  $i$ -th individual:

$$L_i(\theta) = \begin{cases} F(l_1, l_2), & \text{if } (c_{i1}, c_{i2}) = (0, 0); \\ (\partial F(l_1, y_2)/\partial y_2)_{y_2=x_{i2}} = \int_{-\infty}^{l_1} f(u, x_{i2}) du, & \text{if } (c_{i1}, c_{i2}) = (0, 1); \\ (\partial F(y_1, l_2)/\partial y_1)_{y_1=x_{i1}} = \int_{-\infty}^{l_2} f(x_{i1}, u) du, & \text{if } (c_{i1}, c_{i2}) = (1, 0); \\ f(x_{i1}, x_{i2}), & \text{if } (c_{i1}, c_{i2}) = (1, 1). \end{cases}$$

Let  $\hat{\theta}$  be the maximum likelihood estimator (MLE) of  $\theta$ . The MLE's  $\hat{q}$  and  $\hat{q}_c$  of the agreement parameters  $q$  and  $q_c$  are found by substituting  $\theta = \hat{\theta}$  in their expressions. Let  $I$  be the *observed information matrix*, and  $G = (\partial \log q / \partial \theta)_{\theta=\hat{\theta}}$  and  $G_c = (\partial \log q_c / \partial \theta)_{\theta=\hat{\theta}}$  be the gradient vectors evaluated at the MLE. The expression for  $G$  is given in the Appendix. Although a closed-form expression for  $G_c$  is not available, it is easy to compute numerically.

From the asymptotic theory of MLE's and the delta method, we have

$$\log \hat{q} \approx \mathcal{N}(\log q, G' I^{-1} G), \quad \log \hat{q}_c \approx \mathcal{N}(\log q_c, G'_c I^{-1} G_c),$$

where  $G'$  is the transpose of  $G$ . Hence the upper bounds  $U$  and  $U_c$  can be computed as

$$U = \exp \{ \log \hat{q} + z(1 - \alpha)(G'I^{-1}G)^{1/2} \}, \quad U_c = \exp \{ \log \hat{q}_c + z(1 - \alpha)(G'_c I^{-1} G_c)^{1/2} \}, \quad (5)$$

with  $z(1 - \alpha)$  as the  $(1 - \alpha)$ -th quantile of a  $\mathcal{N}(0, 1)$  distribution. When  $n$  is large, these bounds satisfy (4) with  $(1 - \alpha)$  as the approximate *level of confidence*. Computing the bounds in (5) first on the log scale and then transforming back leads to better small sample performance than the bounds on the original scale.

## 2.2 The Bayesian approach

In the Bayesian framework, the model (1) is interpreted in a hierarchical fashion as,

$$y_{ij} | (\mu_j, b_i, \sigma_j^2) \sim \text{independent } \mathcal{N}(\mu_j + b_i, \sigma_j^2), \quad b_i | \sigma_b^2 \sim \text{independent } \mathcal{N}(0, \sigma_b^2); \quad (6)$$

$i = 1, \dots, n, j = 1, 2$ . This model has a total of  $(n + 5)$  parameters — including the  $n$  individual random effects. The marginal distributions of  $(y_1, y_2)$  and  $d$ , after integrating out the individual random effect, remain the same as before except that they are now conditional on the parameters  $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_b^2)$ . Consequently, the expressions for the agreement measures  $q$  and  $q_c$  given by (2) and (3) still hold. To complete the Bayesian specification of the model (6), we adopt the following mutually independent prior distributions:

$$[\mu_j] \propto 1, \quad \sigma_j^2 \sim IG(A_j, B_j), \quad j = 1, 2, \quad \sigma_b^2 \sim IG(A_b, B_b), \quad (7)$$

where “[.]” denotes a pdf, and  $IG(A, B)$  represents an inverse gamma distribution, i.e., its reciprocal follows a gamma distribution with mean  $A/B$  and variance  $A/B^2$ . The hyperparameters, i.e., the parameters of prior distributions,  $A$ 's and  $B$ 's of the variance components

are positive and must be specified. Choosing their values near zero lead to largely noninformative prior distributions. These prior distributions are quite standard in the Bayesian mixed model literature (see Ruppert, Wand and Carroll, 2003, ch 16). An alternative to the improper uniform distribution for  $\mu_j$  is a uniform distribution on a large finite interval or a mean zero normal distribution with a large variance, ensuring essentially a noninformative but proper prior distribution. If significant prior information is available, it can also be incorporated by choosing informative values for the hyperparameters.

The resulting joint posterior distribution of parameters, although proper, is not available in a closed-form. So we use a Markov chain Monte Carlo (MCMC) approach for inference. The Appendix describes a Gibbs sampler algorithm for sampling from this distribution. We use it to generate a large number of posterior draws, say  $M$ . Then apply (2) and (3) to every draw to produce  $M$  draws each from the posterior distributions of  $q$  and  $q_c$ . The bounds  $U$  and  $U_c$  are taken as the  $(1 - \alpha)$ -th sample quantiles of these  $M$  draws of  $q$  and  $q_c$ , respectively. They satisfy (4) with  $(1 - \alpha)$  as the approximate *credible probability*. The approximation here is due to the finiteness of  $M$ . The inference is exact with respect to the sample size  $n$ . In addition, from the Bayesian large sample theory (see Gelman et al., 2003, ch 4), the approximate confidence coverage of these bounds is also  $(1 - \alpha)$  when  $n$  is large.

When  $M$  is large, directly using (3) for computing  $q_c$  requires a great deal of computation time since it involves solving a somewhat nonstandard integral equation. Substantial savings in time can be achieved by using the following MC approach:

1. For a given setting of  $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_b^2)$ , simulate a random sample of large size, say  $N$ , from the bivariate normal distribution of  $(y_1, y_2)$ . Take their differences to get the



corresponding sample from the normal distribution of  $d = y_1 - y_2$ .

2. Approximate  $q_c$  as the  $p_0$ -th sample quantile of only those absolute differences from Step 1 that correspond to  $\{y_1, y_2 > l\}$ . One can also approximate  $q$  as the  $p_0$ -th sample quantile of all the absolute differences from Step 1.

This approach can also be used for approximating  $\hat{q}_c$  in (5) in the previous section. Computing  $q_c$  via simulation is not recommended for numerical computation of the gradient vector  $G_c$  in (5) as the results may not be accurate. Since we need to solve (3) only a few times for computing the estimate and its standard error (SE) in (5), we prefer this exact approach in the frequentist case.

### 3 Simulation study

The frequentist bounds given by (5) are valid only when  $n$  is large. A large  $n$  is also needed for the frequentist interpretation of the Bayesian bounds to hold. So in this section we use MC simulation to estimate and compare the true confidence coverage of the proposed bounds for moderate values of  $n$ . We also study the impact of ignoring censoring — i.e., treating the censored observations as complete observations.

We perform this investigation at the following settings:  $p_0 = 0.80$ ,  $\alpha = 0.05$ ,  $(\mu_1, \sigma_1) = (0, 1)$ ,  $\mu_2 \in \{0, 1\}$ ,  $\sigma_2 \in \{0.5, 1\}$ ,  $\sigma_b \in \{2, 4\}$ , common censoring rate  $r \in \{25\%, 50\%\}$ , and  $n \in \{30, 60, 100\}$ . The starting points for likelihood maximization and Gibbs sampler are chosen as the true parameter values plus a random  $\mathcal{N}(0, 0.01)$  noise. The quantiles  $q$  and  $q_c$  in the frequentist case are estimated using (2) and (3). However, in the Bayesian case, we use the MC approach described in the previous section with  $N = 10,000$ . Furthermore,

all the six hyperparameters in (7) are given a common noninformative value of  $10^{-3}$ . The Gibbs sampler is run for 50,000 iterations, the first half of the chain is discarded as burn-in, and only every fifth iteration is saved (thinning) to keep further computations manageable. Thus, a total of 5,000 MCMC draws are saved.

At a given combination of settings, we first compute the detection limit  $l_i$  as the  $r$ -th quantile of  $y_i$ ,  $i = 1, 2$ , and obtain the true values of  $(q, q_c)$ . Then we simulate a random sample of size  $n$  from the bivariate normal distribution of  $(y_1, y_2)$  and censor them by setting as  $l_i$  those values of  $y_i$  that fall below  $l_i$ ,  $i = 1, 2$ . Next, we apply the methodology of previous section to this sample and compute the bounds  $U$  and  $U_c$ . Finally, we verify whether they are correct by checking  $\{q \leq U\}$  and  $\{q_c \leq U_c\}$ . This process is repeated 1,000 times separately for the Bayesian and the frequentist approaches. The proportion of times a bound is correct gives an estimate of its true confidence coverage probability. The computations are programmed in the statistical software R (R development core team, 2006).

Tables 1 and 2 report the respective coverage probability estimates for  $U$  and  $U_c$ . The results for the two bounds appear quite similar. The Bayesian bounds tend to be conservative whereas the frequentist ones tend to be liberal. The larger censoring rate seems to produce slightly more conservative results in the Bayesian case, and slightly more liberal results in the frequentist case. The results also seem somewhat consistent across different parameter combinations. For the Bayesian bounds with 25% censoring rate, the empirical coverage is about 2% above the nominal level of 95% for  $n \leq 60$ , and it decreases by about 0.5% for  $n = 100$ . In contrast, the coverage of frequentist bounds with 25% censoring rate is about 3% below the nominal level for  $n = 30$ , about 1% below for  $n = 60$ , and quite close to 95% for  $n = 100$  except for a few exceptions. Overall, for  $n < 100$ , the frequentist bounds are not

recommended due to their liberal nature and the Bayesian bounds should be used in this case. Both may be used for  $n \geq 100$ , keeping in mind that the Bayesian ones may be a little conservative for  $n$  near 100.

We perform a similar investigation to study the impact of ignoring censoring on the coverage probabilities. In this case, all observations in the censored  $(y_1, y_2)$  sample are treated as complete observations when fitting the model, and only 100 MC replications are used. Table 3 presents the estimates for  $n = 60$ . They range from 0-77% for a nominal level of 95% — clearly demonstrating that the variability is severely underestimated for both the approaches. The results gets worse as the censoring rate increases. Further simulations with  $n = 30$  and  $n = 100$  (results not presented) suggest that the estimates for  $n = 30$  are only somewhat better than  $n = 60$  and are substantially worse for  $n = 100$ .

## 4 Application

In case of the viral load data, we have  $(y_1, y_2) = \log\{(2.0 \text{ count}, 3.0 \text{ count})/1000\}$ , the detection limits  $(l_1, l_2) = \log\{(500, 50)/1000\}$  and  $n = 132$ . We fit the Grubbs' model (1) to these data using both the frequentist ML and the Bayesian approaches. The computations were programmed in **R**. Additionally, for the Bayesian fitting, we used the **WinBUGS** package of (Spiegelhalter et al., 2003) by calling it from **R** through the **R2WinBUGS** package of (Sturtz, Ligges and Gelman, 2005). Further, all the six hyperparameters in (7) are given a common value of  $10^{-3}$ . Moreover since **WinBUGS** does not allow improper prior distributions, we modified the distributions for  $(\mu_1, \mu_2)$  in (7) as independent uniforms on the interval  $(-10^3, 10^3)$ . These choices lead to essentially noninformative prior distributions. We run

three parallel Gibbs sampler chains for 30,000 iterations each. The overdispersed starting points for  $(\mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \log \sigma_b)$  are chosen as their MLE and  $\text{MLE} \pm$  twice their SE's. The starting points for  $b_1, \dots, b_n$  are taken as independent draws from a  $\mathcal{N}(0, \sigma_b^2)$  with the values of  $\sigma_b^2$  chosen earlier. The first half of each chain is discarded as burn-in. We use the Gelman-Rubin scale reduction factor (see Gelman et al., 2003, ch 11) for diagnosing convergence. It compares the between-chain variation with the within-chain variation and should be near one for convergence. The values of this factor computed from the second halves of the chains for all the 137 parameters in the model are less than 1.16 — indicating that adequate convergence has been reached. We also use a thinning rate of 10 for each chain so that only a total of 4,500 MCMC draws are saved for posterior summarization. Here we employ WinBUGS mainly to illustrate its use for fitting the censored data model. The posterior simulation approach described in the Appendix, which we also used for the simulation study in previous section, produces similar results.

Table 4 presents the MLE, its SE, the posterior mean and the posterior standard deviation (SD) of selected parameters. The two sets of parameter estimates appear similar with the exception of  $\log \sigma_2$ , whose posterior mean is substantially smaller than its MLE. The MLE's and posterior means of  $(\mu_1, \mu_2, \tau_1, \tau_2, \rho, \mu, \sigma)$  are  $(-1.31, -0.48, 3.34, 3.43, 0.96, -0.83, 0.92)$  and  $(-1.39, -0.50, 3.44, 3.47, 0.96, -0.89, 0.95)$ , respectively. They are also quite similar. In either case,  $y_1$  and  $y_2$  appear highly correlated, and seem to have similar variabilities. However,  $y_2$ 's mean is about 0.85 higher than  $y_1$ 's mean. The probability of  $\{y_1, y_2 > l = \log(500/1000)\}$  is estimated as 0.42 in both cases. The upper bounds  $(U, U_c)$ , with  $(1 - \alpha) = 0.95$  and  $p_0 = 0.80$ , are  $(1.71, 1.70)$  in the frequentist case and are  $(1.99, 1.84)$  in the Bayesian case. The two sets of bounds differ to some extent, but the direction of their

difference is consistent with our earlier finding that, for  $n$  near 100, the Bayesian bound tends to be slightly conservative, whereas the frequentist bound tends to be quite accurate. Note also that  $U_c$  is slightly lower than  $U$  in the former case and the two are roughly the same in the latter case.

Consider now the tolerance interval interpretation of the Bayesian bounds. It says that 80% of  $d = y_1 - y_2$  population is estimated to lie within  $[-U, U] = [-1.99, 1.99]$ . Furthermore, given  $\{y_1, y_2 > l\}$ , the same percentage of  $d$  population is estimated to lie within  $[-U_c, U_c] = [-1.84, 1.84]$ . These conclusions hold individually with 95% credible probability (or approximate confidence level). This extent of differences in log (viral load) counts from Quantiplex versions 2.0 and 3.0 is too large for their interchangeable use, particularly since it appears from (Manegold et al., 2000) that differences of more than 1.15 may be important from a clinical point of view. To assess the prior sensitivity of these results, we refit the model (6) with the common hyperparameter in (7) as 0.1 and 1, and compute  $(U, U_c)$ . The resulting bounds, (1.99, 1.84) and (2.03, 1.88), show that the conclusion is not sensitive to the choice of hyperparameters, provided they are noninformative or moderately informative. The frequentist bounds also lead to a similar conclusion regarding insufficient agreement.

Since the methods are highly correlated and have similar variabilities, it is clear that the mean difference of about 0.85 in the two is the main cause of disagreement. To investigate this further, we subtract 0.85 from all the uncensored values of  $y_2$  and redo the above analysis. The resulting frequentist and Bayesian bounds  $(U, U_c)$  are (1.12, 1.06) and (1.29, 1.20), respectively. The extent of agreement between the two Quantiplex versions now is better than before. In addition, if 1.15 is used as the threshold for agreement, the frequentist bounds do allow the inference of sufficient agreement between the methods after the transformation.

This perhaps is a more accurate inference than the Bayesian inference of still insufficient agreement since the latter may be a bit conservative for  $n = 132$ .

## 5 Discussion

In this article, we consider method comparison data that may be left censored, and describe a frequentist and a Bayesian inference procedure for assessing agreement using tolerance intervals. The results are derived under the assumption of a normal theory mixed model for the data, possibly after a suitable transformation. However, the resulting bivariate normality may not always hold in practice. In fact, in case of the viral load data, the posterior predictive checks recommended by (Gelman et al., 2003, ch 6) give some evidence that the assumed model does not adequately capture the observed correlation — although it does fit well to other aspects of the data. Sometimes there may also be nonlinear dependence in methods that cannot be modelled in the usual normal theory framework. Alternative models are currently being investigated to deal with these scenarios.

The Bayesian modelling approach described here may also be used for agreement analysis using concordance correlation. Once we have the draws from the posterior distribution of model parameters, we can get the posterior draws of concordance correlation and use them for inference. Since the frequentist procedures of (Barnhart et al., 2005) seem to require  $n \geq 100$  to work well, this Bayesian approach may provide a good alternative for smaller values of  $n$ . Further investigation is needed to confirm this.

When an observation falls below the detection limit, we use the limit itself as the associated censored observation. However, some authors use half the detection limit in its place

(see Barnhart et al., 2005). The methodology described here can be easily adapted to handle this situation. Sometimes the methods also have an upper detection limit in addition to a lower limit — leading to measurements that are interval censored. This basic idea of this article can still be applied by appropriately modifying the likelihood function.

Finally, the numerical computations required for implementing the proposed methodology can be easily programmed, e.g., in R. One can also use the WinBUGS package of (Spiegelhalter et al., 2003) for the Bayesian fitting. Software code used for analyzing the viral load data is available from the author.

## Appendix

**Expression for  $G = (\partial \log q / \partial \theta)_{\theta = \hat{\theta}}$  in (5):** Let  $\phi$  and  $\Phi$  denote the pdf and the cdf of a  $\mathcal{N}(0, 1)$  distribution. The quantile  $q$  given by (2) can also be defined as the solution of  $\Phi((q - \mu)/\sigma) - \Phi((-q - \mu)/\sigma) = p_0$ . Upon using implicit differentiation, and letting  $z_l = (-q - \mu)/\sigma$ ,  $z_u = (q - \mu)/\sigma$  and  $s = \phi(z_l) + \phi(z_u)$ , we have the following:

$$\begin{aligned} \frac{\partial \log q}{\partial \mu_1} &= \frac{1}{sq} (\phi(z_u) - \phi(z_l)); & \frac{\partial \log q}{\partial \mu_2} &= \frac{1}{sq} (\phi(z_l) - \phi(z_u)) \\ \frac{\partial \log q}{\partial \log \sigma_j} &= \frac{\sigma_j^2}{sq\sigma} (z_u \phi(z_u) - z_l \phi(z_l)), & j = 1, 2; & \frac{\partial \log q}{\partial \log \sigma_b} = 0. \end{aligned}$$

Now since  $\theta$  is the column vector  $(\mu_1, \mu_2, \log \sigma_1, \log \sigma_2, \log \sigma_b)$ ,  $G$  is simply obtained by evaluating the above expressions at  $\theta = \hat{\theta}$ .

**A Gibbs sampler for posterior simulation:** For  $i = 1, \dots, n$ ,  $j = 1, 2$ , let  $[x_{ij} | \text{others}] = \{\phi(z_{ij})/\sigma_j\}^{c_{ij}} \{\Phi(z_{ij})\}^{1-c_{ij}}$ , where  $z_{ij} = (x_{ij} - \mu_j - b_i)/\sigma_j$ , denote the likelihood of the observed pair  $(x_{ij}, c_{ij})$  given all the model parameters in (6). Under the prior distributions

(7), we have the following full conditional distributions — i.e., the conditional posterior distributions of a group of parameters given all the remaining parameters:

$$\begin{aligned}
[\mu_1, \mu_2 | \text{others}] &\propto \prod_j \prod_i [x_{ij} | \text{others}]; \\
[b_1, \dots, b_n | \text{others}] &\propto \prod_j \prod_i [x_{ij} | \text{others}] \prod_i \phi(b_i / \sigma_b); \\
[\sigma_1^{-2}, \sigma_2^{-2} | \text{others}] &\propto \prod_j \prod_i [x_{ij} | \text{others}] \prod_j \sigma_j^{-2(A_j-1)} \exp(-B_j / \sigma_j^2); \\
\sigma_b^2 | \text{others} &\sim IG(A_b + n/2, B_b + \sum_i b_i^2 / 2).
\end{aligned}$$

In each iteration of the Gibbs sampler, we draw  $(\mu_1, \mu_2)$ ,  $(b_1, \dots, b_n)$ ,  $(\sigma_1^{-2}, \sigma_2^{-2})$  and  $\sigma_b^2$ , in the given order, from their respective full conditionals. Repeating this iteration a large number of times until convergence generates a Markov chain whose stationary distribution is the desired posterior distribution. The first three conditional densities above do not have standard forms. So to sample from them we use a Metropolis algorithm, with independent normals as the proposal distributions. The means of these proposals are taken as the parameter values in the previous iteration and the variances are pre-specified. A common variance is used for the parameters in each group. They are chosen by trial and error so that the acceptance rate of the proposals is approximately 25% (see Gelman et al., 2003, ch 11). The parameters  $(\sigma_1^{-2}, \sigma_2^{-2})$  are simulated on log scale to normalize their distributions.

## Acknowledgements

The author thanks Dr. Huiman Barnhart for the invitation to participate in this special issue. He is also thankful to Dr. C. Manegold for providing the viral load data, and to Dr.



Swati Biswas and the reviewers for several helpful comments.

## References

- Barnhart, H. X., Song, J., Lyles, R. H. (2005). Assay validation for left-censored data. *Statistics in Medicine* 24:3347–3360.
- Bland, J. M., Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i:307–310.
- Choudhary, P. K., Nagaraja, H. N. (2004). Measuring agreement in method comparison studies - A review. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, pp. 215–244, Balakrishnan, N., Kannan, N. and Nagaraja, H. N. (editors), Boston: Birkhauser.
- Choudhary, P. K., Nagaraja, H. N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 137:279–290.
- Dunn, G., Roberts, C. (1999). Modelling method comparison data. *Statistical Methods in Medical Research* 8:161–179.
- Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall/CRC.
- Guttman, I. (1988). Statistical tolerance regions. In *Encyclopedia of Statistical Sciences*, Vol. 9, pp. 272-287, Kotz, S., Johnson, N. L. and Read, C. B. (Editors), New York: John Wiley.

- Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268. Corrections: 2000, 56:324–325.
- Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19:255–270.
- Lin, L. I., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* 97:257–270.
- Manegold, C., Krempe, C., Jablonowski, H., Kajala, L., Dietrich, M., Adams, O. (2000). Comparative evaluation of two branched-DNA human immunodeficiency virus type 1 RNA quantification assays with lower detection limits of 50 and 500 copies per milliliter. *Journal of Clinical Microbiology* 38:914–917.
- Pinheiro, J. C., Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Ruppert, D., Wand, M. P., Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Searle, S. S., Casella, G., McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley.

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

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

Figure 1: Scatter plot of the uncensored viral load data. Also included is the 45-degree line through origin. Viral load is measured as copies/ml.

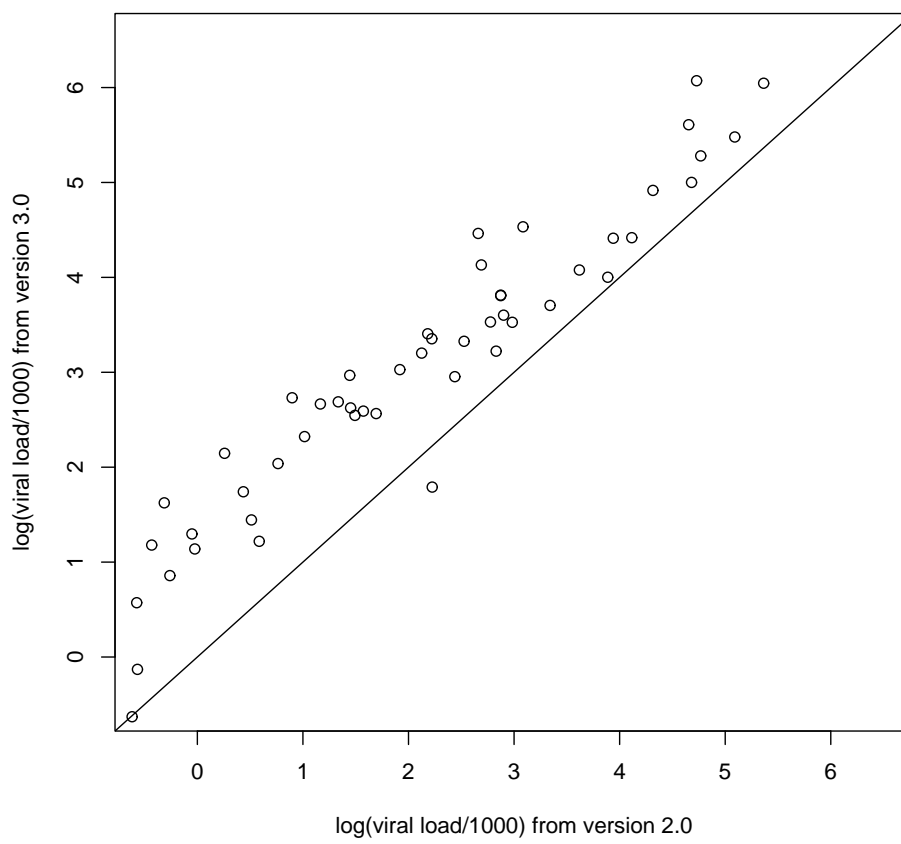


Table 1: Estimated confidence coverage (in %) of a 95% upper bound  $U$  for the measure of agreement  $q$ . Here  $(p_0, \mu_1, \sigma_1) = (0.80, 0, 1)$ . The estimates are based on 1,000 replications and have a standard error of 0.7.

		25% censoring						50% censoring					
		Bayesian			Frequentist			Bayesian			Frequentist		
		$n$											
$(\mu_2, \sigma_2, \sigma_b)$	$q$	30	60	100	30	60	100	30	60	100	30	60	100
(0, 0.5, 2)	1.43	97.8	97.3	96.9	92.2	94.3	92.5	99.1	97.1	98.0	93.4	94.5	95.4
(0, 0.5, 4)	1.43	97.3	97.6	96.9	92.0	93.9	93.3	98.9	97.2	96.6	91.9	93.5	95.4
(0, 1, 2)	1.81	98.1	97.8	97.8	92.7	94.2	94.7	99.1	97.9	98.3	91.9	93.8	96.0
(0, 1, 4)	1.81	97.5	97.3	97.7	92.8	93.7	95.7	98.0	97.2	98.1	91.6	94.4	95.6
(1, 0.5, 2)	1.96	95.2	95.8	95.4	91.4	93.2	94.3	95.5	93.5	92.4	90.7	92.0	93.2
(1, 0.5, 4)	1.96	95.4	95.6	95.7	91.0	93.3	94.2	95.4	93.7	93.1	89.9	92.6	91.7
(1, 1, 2)	2.25	97.2	97.6	96.8	93.1	94.8	94.8	97.8	96.0	96.0	92.5	93.8	95.3
(1, 1, 4)	2.25	96.6	96.5	96.1	92.5	94.5	95.5	97.4	96.3	96.5	90.4	94.4	94.6

Table 2: Estimated confidence coverage (in %) of a 95% upper bound  $U_c$  for the conditional measure of agreement  $q_c$ . Here  $(p_0, \mu_1, \sigma_1) = (0.80, 0, 1)$ . The estimates are based on 1,000 replications and have a standard error of 0.7.

		25% censoring						50% censoring						
		Bayesian			Frequentist			Bayesian			Frequentist			
		$n$			$n$			$n$			$n$			
$(\mu_2, \sigma_2, \sigma_b)$	$q_c$	30	60	100	30	60	100	$q_c$	30	60	100	30	60	100
(0, 0.5, 2)	1.33	97.8	97.2	96.3	92.0	94.1	92.9	1.29	98.4	96.4	95.8	91.2	92.3	93.0
(0, 0.5, 4)	1.39	97.1	97.5	96.8	92.0	94.0	93.5	1.36	98.1	96.9	95.9	90.6	92.7	94.9
(0, 1, 2)	1.67	97.5	97.6	97.3	92.0	93.9	94.6	1.57	98.2	97.4	97.0	90.6	91.8	94.9
(0, 1, 4)	1.74	97.1	97.4	96.8	92.3	93.1	95.0	1.69	97.8	97.1	96.9	90.2	92.8	94.3
(1, 0.5, 2)	1.65	98.1	97.0	96.7	93.9	94.0	94.0	1.50	98.4	97.9	96.7	93.7	93.7	94.1
(1, 0.5, 4)	1.83	96.9	97.4	97.7	92.7	94.8	95.4	1.75	97.6	96.8	98.2	92.8	94.0	95.0
(1, 1, 2)	1.99	96.9	95.2	93.6	91.4	92.8	94.1	1.84	96.6	95.2	94.5	89.8	92.0	93.0
(1, 1, 4)	2.14	96.8	96.2	96.7	92.4	93.1	95.0	2.06	97.2	95.9	97.2	89.8	93.2	93.9

Table 3: Estimated confidence coverage (in %) of 95% upper bounds when the censored observations are treated as complete observations. Here  $(n, p_0, \mu_1, \sigma_1, ) = (60, 0.80, 0, 1)$ . The estimates are based on 100 replications.

$(\mu_2, \sigma_2, \sigma_b)$	25% censoring				50% censoring			
	Bayesian		Frequentist		Bayesian		Frequentist	
	$U$	$U_c$	$U$	$U_c$	$U$	$U_c$	$U$	$U_c$
(0, 0.5, 2)	53.0	55.0	35.0	38.0	2.0	5.0	3.0	6.0
(0, 0.5, 4)	62.0	62.0	41.0	43.0	7.0	7.0	4.0	4.0
(0, 1, 2)	48.0	51.0	37.0	41.0	5.0	6.0	0.0	2.0
(0, 1, 4)	58.0	57.0	42.0	45.0	10.0	10.0	1.0	5.0
(1, 0.5, 2)	68.0	76.0	63.0	71.0	6.0	44.0	4.0	41.0
(1, 0.5, 4)	71.0	73.0	65.0	73.0	21.0	40.0	12.0	39.0
(1, 1, 2)	71.0	73.0	66.0	71.0	13.0	43.0	10.0	37.0
(1, 1, 4)	75.0	77.0	70.0	71.0	22.0	35.0	15.0	33.0

Table 4: Estimates of selected parameters for the viral load data. The mean and SD refer to the posterior distribution.

	Frequentist		Bayesian	
	MLE	SE	Mean	SD
$\mu_1$	-1.31	0.07	-1.39	0.32
$\mu_2$	-0.48	0.07	-0.50	0.30
$\log \sigma_1$	-1.20	0.08	-1.09	1.03
$\log \sigma_2$	-0.14	0.06	-0.67	0.92
$\log \sigma_b$	1.20	0.06	1.22	0.08
$\log q$	0.48	0.04	0.53	0.09
$\log q_c$	0.47	0.04	0.48	0.08



## List of figures and tables

Figure 1. Scatter plot of the uncensored viral load data. Also included is the 45-degree line through origin. Viral load is measured as copies/ml.

Table 1. Estimated confidence coverage (in %) of a 95% upper bound  $U$  for the measure of agreement  $q$ . Here  $(p_0, \mu_1, \sigma_1) = (0.80, 0, 1)$ . The estimates are based on 1,000 replications and have a standard error of 0.7.

Table 2. Estimated confidence coverage (in %) of a 95% upper bound  $U_c$  for the conditional measure of agreement  $q_c$ . Here  $(p_0, \mu_1, \sigma_1) = (0.80, 0, 1)$ . The estimates are based on 1,000 replications and have a standard error of 0.7.

Table 3. Estimated confidence coverage (in %) of 95% upper bounds when the censored observations are treated as complete observations. Here  $(n, p_0, \mu_1, \sigma_1, ) = (60, 0.80, 0, 1)$ . The estimates are based on 100 replications.

Table 4. Estimates of selected parameters for the viral load data. The mean and SD refer to the posterior distribution.