A Note on Bayesian Inference After Multiple Imputation

Xiang Zhou and Jerome P. Reiter*

Abstract

This article is aimed at practitioners who plan to use Bayesian inference on multiply-imputed datasets in settings where posterior distributions of the parameters of interest are not approximately Gaussian. We seek to steer practitioners away from a naive approach to Bayesian inference, namely estimating the posterior distribution in each completed dataset and averaging functionals of these distributions. We demonstrate that this approach results in unreliable inferences. A better approach is to mix draws from the posterior distributions from each completed dataset, and use the mixed draws to summarize the posterior distribution. Using simulations, we show that for this second approach to work well, the number of imputed datasets should be large. In particular, five to ten imputed datasets—which is the standard recommendation for multiple imputation—is generally not enough to result in reliable Bayesian inferences.

*Xiang Zhou is PhD candidate, Department of Neurobiology, and MA, Department of Statistical Science, Duke University, Durham, NC 27708 (e-mail: x.zhou@duke.edu); and Jerome P. Reiter is Associate Professor, Department of Statistical Science, Duke University, Durham, NC 27708 (e-mail: jerry@stat.duke.edu). This work was supported by the Environmental Protection Agency grant R833293. The authors thank the associate editor and a referee for suggestions on how to explain the bias in the naive approach.

Keywords: Missing, Nonresponse, Sample

1 INTRODUCTION

When some data values are missing, one approach to statistical inference is multiple imputation (Rubin, 1987; Reiter and Raghunathan, 2007). The basic idea is for the analyst to fill in any missing values by repeatedly sampling from the predictive distributions of the missing values. When the posterior distribution of the parameter of interest, or, for likelihood-oriented statisticians, the sampling distribution of the complete-data estimator, is approximately Gaussian, the analyst can obtain inferences by computing point and variance estimates of interest with each dataset and combining these estimates using simple formulas. These formulas serve to propagate the uncertainty introduced by imputation through the analyst's inferences.

When presuming normality of the posterior/sampling distribution is not justifiable, the distribution is not adequately summarized by the mean and variance, so that Rubin's (1987) rules are not appropriate for inference. Nonetheless, some practitioners continue to use Rubin's (1987) rules even when they are theoretically invalid. For example, in a literature review of applications of multiple imputation involving parameters not adequately modeled with normal distributions, Marshall *et al.* (2009) find that, "Rubin's rules without applying any transformations were the standard approach used, when any method was stated." They go on to cite several examples where Rubin's (1987) rules are used to estimate functionals of distributions, such as percentiles of survival distributions.

When normality is not justifiable, Bayesian approaches are viable options for inference. In multiple imputation contexts, the analyst must appropriately utilize the information from the multiple datasets in the inferences; again, simply applying Rubin's (1987) rules to posterior means and variances is generally not correct. An approach suggested by Gelman *et al.* (2004, p. 520) is (i) simulate many draws from the posterior distribution in each imputed dataset, and (ii) mix all the draws. The mixed draws approximate the posterior distribution. Gelman *et al.* (2004) do not evaluate the properties of this approximation, nor do the prominent texts on multiple imputation of Schafer (1997) and Little and Rubin (2002).

In this article, we examine the approximation of Gelman *et al.* (2004, p. 520) using simulation studies. We find that the approach works well with large numbers of multiply-imputed datasets. However, the usual advice for multiple imputation for modest fractions of missing data—that five or ten completed datasets are adequate for inferences—can result in unreliable estimates of posterior distributions. We also point out the pitfalls of incorrectly using Rubin's (1987) rules on functionals of posterior distributions. Specifically, we examine an approach akin to some of those observed by Marshall *et al.* (2009): (i) estimate posterior quantiles in each completed dataset, and (ii) average them across the datasets. We argue and demonstrate that this approach produces unreliable estimates of posterior distributions.

2 DESCRIPTION OF THE APPROACHES

In this section, we motivate the approach of Gelman *et al.* (2004, p. 520). We begin with brief reviews of Bayesian inference with incomplete data and of multiple imputation. Let

$Y_{inc} = (Y_{obs}, Y_{mis})$ be the $n \times p$ matrix of data values for n units, where Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing. Let Q be a parameter of interest, for example a regression coefficient. Let Q_α be the value of Q such that

$$\int_{-\infty}^{Q_\alpha} f(Q|Y_{obs})dQ = \int_{-\infty}^{Q_\alpha} \int_{Y_{mis}} f(Q|Y_{obs}, Y_{mis}) f(Y_{mis}|Y_{obs}) dY_{mis} dQ = \alpha, \tag{1}$$

where α is a desired quantile of the posterior distribution of Q. Analysts can approximate this integral with Monte Carlo methods. First, draw Y_{mis} from its posterior predictive distribution. Second, draw a value of Q from its posterior distribution, given the drawn Y_{mis}. Third, repeat these two steps K times, where K is very large. Fourth, sort the K simulated values of Q, and select the (αK)th element of the sorted list. The result is an estimate for Q_α.

In multiple imputation, the analyst creates m completed datasets, $D^{(l)} = (Y_{obs}, Y_{mis}^{(l)})$ where $1 \leq l \leq m$, which are used for analysis. Here, $Y_{mis}^{(l)}$ is a draw from the posterior predictive distribution of $(Y_{mis} \mid Y_{obs})$, or from an approximation of that distribution such as the approach of Raghunathan *et al.* (2001).

Typically, m is much smaller than K would be for Bayesian inference for non-Gaussian distributions. Thus, with small m, drawing one value of Q for each $D^{(l)}$ results in too few draws of Q to get reasonable estimates of Q_α. Instead, we can utilize each completed dataset for more than just one draw of Q. To motivate this, we re-express the integral in (1) as

$$\alpha = \lim_{m \to \infty} \frac{1}{m} \sum_{l=1}^{m} \int_{-\infty}^{Q_\alpha} f(Q|Y_{obs}, Y_{mis}^{(l)}) dQ. \tag{2}$$

This suggests that, for any value of Q in the upper limit of the integral, we can find the associated cumulative probability by (i) sampling J values of Q in each $D^{(l)}$, where J is

large, (ii) finding the percentage of the J draws in each $D^{(l)}$ less than the upper limit value, and (iii) averaging those percentages across all m datasets, where $m \to \infty$. Equivalently, the analyst could mix all of the sampled draws from each dataset, and find the percentage of elements less than the upper limit in the combined draws. This process can be easily adapted to find Q_α: try different upper limits until one reaches the desired α probability.

The approximation of Gelman *et al.* (2004, p. 520), which we denote as $\tilde{Q}_\alpha$, is essentially an approximation of (2) for finite m. Specifically, for each $D^{(l)}$ where $l = 1, \ldots, m$, the analyst simulates J values of Q from $f(Q|D^{(l)})$, where J is large. Let $\hat{f}(Q^{(l)})$ represent the J draws of Q obtained with $D^{(l)}$. The analyst mixes all $\hat{f}(Q^{(l)})$ together to create $\hat{f}(Q^{all})$. The analyst sorts the mJ draws in $\hat{f}(Q^{all})$, and the $\alpha(mJ)$th element of the sorted list is the estimate of Q_α.

We now use simulations to illustrate the properties of $\tilde{Q}_\alpha$. We also use the simulations to emphasize that the naive approach of averaging posterior quantiles can produce poor estimates of Q_α in comparison to $\tilde{Q}_\alpha$. To fix notation for the naive approach, let $Q_\alpha^{(l)}$ be the value of Q in $D^{(l)}$ such that $\int_{-\infty}^{Q_\alpha^{(l)}} f(Q|Y_{obs}, Y_{mis}^{(l)})dQ = \alpha$. Then, $\bar{Q}_\alpha = \sum_{l=1}^{m} Q_\alpha^{(l)}/m$. Clearly, $\bar{Q}_\alpha$ has nothing to do with (2). It is derived from convenience rather than theory.

3 ILLUSTRATIVE SIMULATIONS

The complete data, Y_{inc}, comprise $n = 50$ values generated independently from Bernoulli trials with $\pi = .2$. We introduce missing data by randomly deleting 10%, 30% or 50% of the data completely at random (Rubin, 1976). We use multiple imputation to generate $m = 5$,

$m = 20$, or $m = 100$ completed datasets. We seek Q_α, where $\alpha \in \{.025, .25, .75, .975\}$, for the posterior distribution of π. We generate Y_{inc} and multiple imputations 5000 times to approximate the sampling distributions of $\tilde{Q}_\alpha$ and $\bar{Q}_\alpha$.

To create each completed dataset, we first sample a value of π from the appropriate Beta distribution based on Y_{obs}. We use a uniform prior distribution for π. We then draw Y_{mis} from a Bernoulli distribution using the sampled π. After the imputation steps, in each $D^{(l)}$ we draw $J = 10000$ values of π from Beta$(\sum(Y_{obs} + Y_{mis}^{(l)}) + 1, n - \sum(Y_{obs} + Y_{mis}^{(l)}) + 1)$, which is the posterior computed with $D^{(l)}$. To get $\tilde{Q}_\alpha$, we mix and sort the mJ draws of π, and select the $\alpha(mJ)$th element. To get $\bar{Q}_\alpha$, we compute the α-quantile in each $D^{(l)}$ and average them across the m datasets.

Figure 1 shows the distributions of $\tilde{Q}_\alpha - Q_\alpha$ and $\bar{Q}_\alpha - Q_\alpha$ for $\alpha \in \{.025, .25\}$ across the 5000 replications with $m = 100$. Here, Q_α is computed from $f(\pi|Y_{obs})$. For each scenario, $\tilde{Q}_\alpha$ is nearly centered on Q_α. There do not appear to be any trends with the percentage of missing data, apart from the expected increase in variability as the percentage of missing data increases. However, in additional simulations with $m = 5$ and $m = 20$, typically $\tilde{Q}_\alpha > Q_\alpha$ for small α, and $\tilde{Q}_\alpha < Q_\alpha$ for large α. This is evident in Figure 2, which shows that when $m = 5$ and to a lesser extent when $m = 20$, the posterior intervals based on $\tilde{Q}_\alpha$ tend to be tighter than warranted for modest m. This problem disappears when $m = 100$.

The inaccuracy when $m = 5$ merits closer inspection, because often practitioners only create five multiple imputations for analysis. Across all missing data scenarios, the median lengths of the 50% and 95% posterior intervals are smaller when $m = 5$ than when $m = 100$.

Put another way, analysts appear to obtain sharper inferences by using five imputations than using one hundred imputations. This does not imply that analysts should use small m for Bayesian inference after multiple imputation; on the contrary, it implies that approximations $\tilde{Q}_\alpha$ based on small m are not reliable. Hence, analysts planning on Bayesian inference after multiple imputation should generate a large number of completed datasets.

What about $\bar{Q}_\alpha$? As evident in Figure 1, $\bar{Q}_\alpha$ can differ substantially from Q_α, and its performance worsens as the percentage of missing values increases. More often than not, $\bar{Q}_\alpha > Q_\alpha$ for small α, and $\bar{Q}_\alpha < Q_\alpha$ for large α. Hence, as also evident in Figure 2, analysts who construct posterior intervals based on $\bar{Q}_\alpha$ tend to have tighter ranges than warranted.

What is wrong with the naive approach of averaging posterior quantiles? Each $Q_\alpha^{(l)}$ is a summary of the posterior distribution of Q estimated as if $D^{(l)}$ was in fact genuine data with n records. However, the observed data comprise fewer than n records, so that the actual posterior distribution of Q is more dispersed than the complete-data posterior distribution. Thus, each $Q_\alpha^{(l)}$ is biased towards the median, and so is their average.

4 APPLICATION TO BIOASSAY DATA

To illustrate Bayesian inference after multiple imputation on genuine data, we modify data from a bioassay experiment that appears in Gelman *et al.* (2004, p. 88–93), who took them from Racine *et al.* (1986). The data comprise two measurements on $n = 20$ animals. Let x_i be the natural logarithm of the dose of a chemical compound administered to animal i, where $x_i \in \{-.86, -.30, -.05, .73\}$. There were five animals at each dose level. Let $y_i = 1$ if

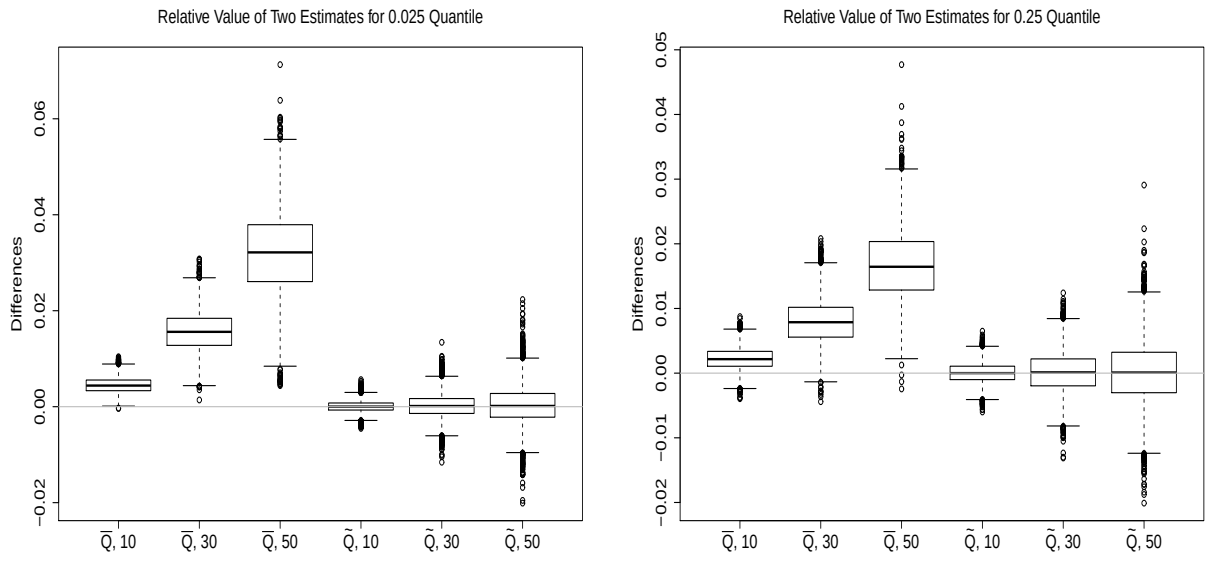

Figure 1: Box plots of $\bar{Q}_\alpha - Q_\alpha$ and $\tilde{Q}_\alpha - Q_\alpha$ in different settings with $n = 50$, $m = 100$, and 10%, 30%, or 50% missing data. The first three plots in each panel are for $\bar{Q}_\alpha - Q_\alpha$, and the second three plots in each panel are for $\tilde{Q}_\alpha - Q_\alpha$. The labels on the horizontal axis show the percentage of missing data. Generally, $\bar{Q}_\alpha$ is substantially different than Q_α, whereas $\tilde{Q}_\alpha$ estimates Q_α reasonably well.

animal i dies shortly after receiving the dose, and let $y_i = 0$ otherwise. There are no missing data in the study. Therefore, we deleted a randomly selected 20% of the y_i values.

The goal of the analysis is to learn about the toxicity of the compound, which we do with a logistic regression of Y on X. Because of the small sample size, it is doubtful that the sampling distributions of the estimated regression coefficients are well-approximated by normal distributions. Following Gelman *et al.* (2004), we therefore use a Bayesian logistic regression model to learn about the toxicity of the compound, so that $y_i|\pi_i \sim Bernoulli(\pi_i)$ where $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. We use the non-informative prior distribution $f(\beta_0, \beta_1) \propto 1$.

The primary targets of scientific interest are the posterior distributions of β_0 and β_1.

To multiply-impute the four missing values, we first draw a value of (β_0, β_1) from its approximate posterior distribution using grid sampling. We substitute the drawn values into the equation for π_i for each of the four animals with missing data. We then draw values of Y_{mis} from each animal's Bernoulli distribution to create the completed dataset, $D^{(l)}$, where $l = 1, \ldots, m$. We examine three scenarios: $m = 5$, $m = 20$, and $m = 100$.

For each $D^{(l)}$, we determine quantiles of $f(\beta_0, \beta_1 | D^{(l)})$ by using grid sampling again. We sample $J = 10,000$ values from the joint distribution for each completed dataset. By mixing the mJ draws of (β_0, β_1), we can compute values of $\tilde{Q}_\alpha$.

Table 1 displays several quantiles for β_0 and β_1. When $m = 100$, the values of $\tilde{Q}_\alpha$ are close to the corresponding values of Q_α. As expected, the differences between $\tilde{Q}_\alpha$ and Q_α are largest when $m = 5$. For both β_0 and β_1, the posterior intervals are too narrow when $m = 5$. The Table also displays estimates based on $\bar{Q}_\alpha$. Once again, they are less reliable than those based on $\tilde{Q}_\alpha$.

To see if the results in Table 1 for $m = 5$ are unusual, we repeated the posterior simulation 100 times. In 57% of these replications, $\tilde{Q}_{.975} - \tilde{Q}_{.025}$ for β_1 with $m = 5$ was shorter than $Q_{.975} - Q_{.025}$ for β_1 from the observed data; roughly the same trend held for the interquartile range for β_1 and for the intervals involving β_0. The lengths of the one hundred $\tilde{Q}_{.975} - \tilde{Q}_{.025}$ for β_1 with $m = 5$ ranged from 15.8 (1.5 to 17.3) to 21.4 (3.0 to 24.4), as compared to a length of 19.5 for $Q_{.975} - Q_{.025}$. Thus, there are substantial chances of estimating inappropriately short posterior intervals with $\tilde{Q}_{.975} - \tilde{Q}_{.025}$, although the risks appear to be random rather

than systematic. Given the potential for overstating accuracy, we would be reluctant to recommend or use $m = 5$ for this analysis.

We also repeated the analysis using an imputation model that differs from the analysis model. Specifically, for imputations we assume that $f(y_{ij}) \sim Bernoulli(\pi_j)$ in each of the $j = 1, \ldots, 4$ dosage strata. The results show the themes seen in Table 1.

5 CLOSING REMARKS

As both multiple imputation and Bayesian inference grow in popularity, we anticipate that practitioners will commonly use Bayesian inference after multiple imputation. We hope that this article reduces the number of practitioners who naively and incorrectly average posterior quantiles and other functionals, and encourages practitioners to use the approach of Gelman *et al.* (2004, p. 520) with large m.

References

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London: Chapman & Hall.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Marshall, A., Altman, D. G., Holder, R. L., and Royston, P. (2009). Combining estimates

of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology* **9:57**.

Racine, A., Grieve, A., Fluhler, H., and Smith, A. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics* **35**, 93–150.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

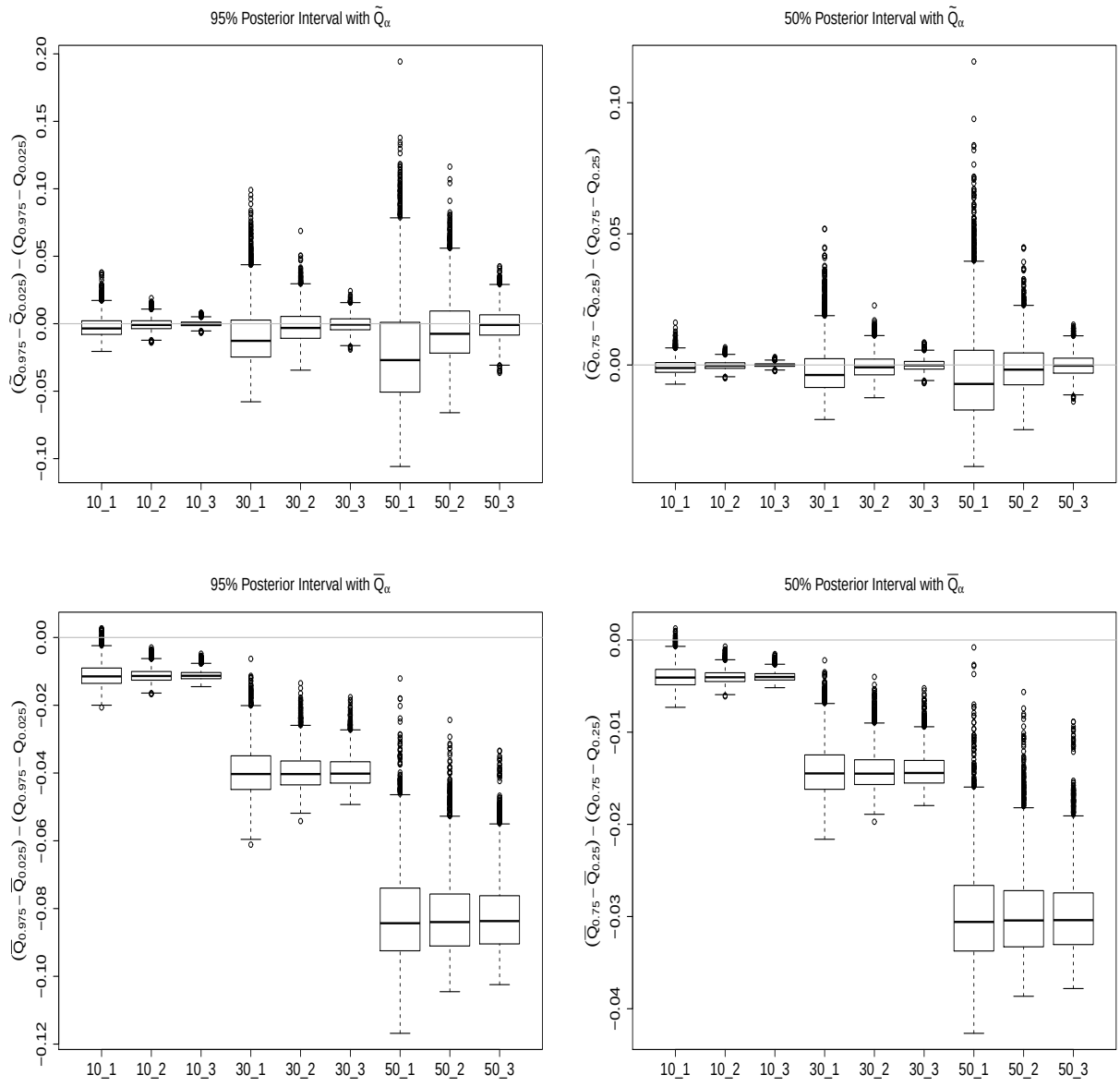

Figure 2: Box plots of differences in lengths of the approximate and true 50% and 95% posterior intervals with $n = 50$. The labels on the horizontal axis show the percentage of missing data followed by the value of m, where 1 represents $m = 5$, 2 represents $m = 20$, and 3 represents $m = 100$. The intervals based on $\tilde{Q}_\alpha$ (top panel) are relatively poor approximations for modest m but good for large m. The intervals based on $\bar{Q}_\alpha$ (bottom panel) are always unreliable.

		$\tilde{Q}_\alpha$ for MI with			$\bar{Q}_\alpha$ for MI with		
	Q_α	$m = 5$	$m = 20$	$m = 100$	$m = 5$	$m = 20$	$m = 100$
Posterior quantiles for β_0							
α =.025	-1.72	-1.87	-1.72	-1.77	-1.53	-1.21	-1.33
α =.25	-0.39	-0.60	-0.35	-0.38	-0.37	-0.08	-0.20
α =.75	1.12	0.70	1.20	1.16	0.88	1.22	1.06
α =.975	2.84	2.14	3.00	2.97	2.26	2.71	2.50
95% interval length	4.56	4.01	4.72	4.74	3.79	3.92	3.83
Posterior quantiles for β_1							
α =.025	2.37	2.62	2.73	2.24	2.55	2.75	2.65
α =.25	5.47	5.58	5.84	5.29	5.46	5.98	5.69
α =.75	11.73	11.56	12.27	11.69	11.21	12.31	11.68
α =.975	21.82	21.08	22.52	21.88	20.33	22.01	20.98
95% interval length	19.45	18.46	19.79	19.64	17.78	19.26	18.33

Table 1: Quantile estimates for the bioassay data from one set of multiple imputations using $m = 5$, $m = 20$, and $m = 100$. The imputation model is the same as the analysis model. Here, Q_α is estimated with the observed data. Inferences based on $\tilde{Q}_\alpha$ are reliable with $m = 100$ but less so with $m = 5$ or $m = 20$. Inferences based on $\tilde{Q}_\alpha$ are generally unreliable.