Published in final edited form as:

Stat Med. 2012 January 13; 31(1): 1–18. doi:10.1002/sim.4413.

Diagnosing imputation models by applying target analyses to posterior replicates of completed data[‡]

Yulei He^{*,†} and Alan M. Zaslavsky

Department of Health Care Policy, Harvard Medical School, Boston, MA, 02115, USA

Abstract

Multiple imputation fills in missing data with posterior predictive draws from imputation models. To assess the adequacy of imputation models, we can compare completed data with their replicates simulated under the imputation model. We apply analyses of substantive interest to both datasets and use posterior predictive checks of the differences of these estimates to quantify the evidence of model inadequacy. We can further integrate out the imputed missing data and their replicates over the completed-data analyses to reduce variance in the comparison. In many cases, the checking procedure can be easily implemented using standard imputation software by treating re-imputations under the model as posterior predictive replicates. Thus, it can be applied for non-Bayesian imputation methods. We also sketch several strategies for applying the method in the context of practical imputation analyses. We illustrate the method using two real data applications and study its property using a simulation.

Keywords

completed data; conservatism; imputation software; missing data; model uncongeniality; normal approximation; posterior predictive p -value; re-imputation

1. Introduction

Missing data commonly pose severe challenges to data analysts. Multiple imputation [1] is a popular approach to analysis with missing data. This Bayesian model-based approach consists of two major steps. In the simulation step, missing values are drawn from their posterior predictive distributions to create several, say M , completed datasets. In the analysis step, proposed analyses are applied to each completed dataset (the observed data and imputed unobserved data) using complete-data procedures, and the M sets of results are then combined into a single inference, using simple rules to account for imputation uncertainty. Various statistical models have been developed for imputation. See, for example, reviews [2, 3]. General practitioners can easily implement these methods using readily available software [4, 5].

[‡]Supporting information may be found in the online version of this article.

© Copyright © 2011 John Wiley & Sons, Ltd.

*Correspondence to: Yulei He, Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, USA.

[†]he@hcp.med.harvard.edu

Despite the popularity of multiple imputation, checking of imputation models appears to be far from universally practiced approach. In many methodological studies, Monte Carlo simulations assess the performance of imputation models on incomplete data produced under an artificial data-generating mechanism. They may not provide adequate guidance to analysts confronted with real data generated under an unknown process.

A few studies have checked imputation models in empirical applications. Gelman *et al.* [6] used cross-validation to assess the fit of a hierarchical Bayesian model for incomplete data arising from multiple surveys. Gelman *et al.* [7] proposed to use posterior predictive replicates of completed data for model checking with missing or latent data, focusing on completed-data plots and exploratory data analysis (see also [8, 9]).

In an analysis using imputed data, the analysis procedure applied to the completed data may be ‘uncongenial’ to the imputation model, that is, inconsistent in a way that affects the validity of the multiple imputation analysis [10]. In one of our examples (Section 4), the intended complete-data analysis is a linear regression with two predictors and their interactions, whereas the imputation model assumes a regression with only main effects. In another example, a Cox regression is fitted using imputations from linear and generalized linear models. Uncongeniality can lead to inconsistent point or interval estimates in theory, without necessarily leading to practically unsatisfactory results. Motivated by such facts, we argue that a goal of checking imputation model adequacy is to assess whether the analysis results using the completed data conform to those might have been expected in the absence of missing data.

In this paper, we propose a practical approach to imputation model diagnostics using posterior predictive checking (PPC) [11, 12]. We apply ‘target’ analyses of interest to both the completed data and their posterior replicates simulated under the model and evaluate the difference of the estimates as large differences may suggest model inadequacy. This distinguishes our approach from most diagnostics focusing on general goodness-of-fit statistics. Section 2 introduces the background. Section 3 describes the methods and the practical implementation strategy. Section 4 consists of two applications. Section 5 presents simulation studies to evaluate the performance of the approach. Finally, Section 6 suggests future research directions.

2. Background

2.1. Bayesian model-based imputation

Suppose for concreteness that the complete data are $Y = (y_{ij})$, where y_{ij} represents the observation of variable j for case i ; the methods apply equally to more complex data structures. Following the notation of [13], let $R = (r_{ij})$ denote the missing data indicators, where $r_{ij} = 1$ if y_{ij} is observed and 0 otherwise. Then, the complete data are $Y = Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$, where ‘com’ signifies complete data and ‘obs’ and ‘mis’ represent the collections of indices for which $r_{ij} = 1$ and $r_{ij} = 0$, respectively. In what follows, we let Y_{obs} represents both the random variables and their observed values, whereas Y_{mis} represents the unobserved random variables.

Let θ and $f(\theta)$ denote the parameters of the complete-data imputation model, $f(Y_{\text{com}}|\theta)$, and their prior distribution, ϕ and $f(\phi|\theta)$ the parameters of the missingness model, $f(R|Y_{\text{com}}, \phi)$, and their prior distribution. Then, the joint posterior distribution is $f(\theta, \phi, Y_{\text{mis}}|Y_{\text{obs}}, R) \propto f(\theta) f(\phi|\theta) f(Y_{\text{com}}|\theta) f(R|Y_{\text{com}}, \phi)$. Under the ignorable missingness assumption [1], $f(\theta, \phi|Y_{\text{obs}}, R) \propto f(\theta|Y_{\text{obs}}) f(\phi|Y_{\text{obs}}, R)$. Because the posterior inference for θ does not depend on the missingness model, Bayesian model-based imputation draws from the posterior predictive distribution $f(Y_{\text{mis}}|Y_{\text{obs}})$ under the imputation model $f(Y_{\text{com}}|\theta) f(\theta)$. On the other hand, it is sometimes difficult to implement a fully Bayesian imputation model because of its complexity and computational burden. Imputation methods that are not fully Bayesian might approximate draws from $f(Y_{\text{mis}}|Y_{\text{obs}})$ and be simpler to implement in practice [14, chapter 10].

2.2. Posterior predictive checking

With complete data, PPC [11] compares the observed Y with draws of replicates Y^{rep} from their posterior predictive distribution using a discrepancy function Q , a scalar function of the data, and possibly unknown model parameters θ . One commonly used diagnostic summary in PPC is the posterior predictive p -value [12, 15], defined as the probability that the replicated data would be more extreme than the observed data, as measured by Q ,

$$p_B = P(Q(Y^{\text{rep}}, \theta) \geq Q(Y, \theta) | Y) = \int \int I_{Q(Y^{\text{rep}}, \theta) \geq Q(Y, \theta)} f(Y^{\text{rep}}, \theta | Y) d\theta dY^{\text{rep}} \quad (1)$$

where I is the indicator function. The use of p_B accounts for uncertainty about θ in evaluating the similarity between Y and Y^{rep} . An extreme p_B -value (close to 0 or 1) suggests that the discrepancy between Y and Y^{rep} may not be easily explained by chance and hence casts suspicion on the model fit [16, chapter 6].

Posterior predictive checking has been widely used for model checking in applied Bayesian analysis (see, for example, [16, chapter 6]). To test the general fit of the model for complete data, the choices of discrepancy functions may include the following: (a) common descriptive statistics for the data, such as means, variances, quantiles, and correlations; (b) summaries of model fit, such as a χ^2 discrepancy; (c) graphs of residuals measuring discrepancies between the model and data; and (d) features of the data not directly addressed by the probability model. We can calculate p_B by simulation. For each of L_1 draws $\{\theta^l\}$ from $f(\theta|Y)$ ($l = 1, \dots, L_1$), we draw one $Y^{\text{rep}, l}$ from $f(Y^{\text{rep}}|\theta = \theta^l)$, hence obtaining L_1 draws from $f(Y^{\text{rep}}, \theta|Y)$. The PPC compares the realized test quantities $Q(Y, \theta^l)$ with the predictive distribution of test quantities, $Q(Y^{\text{rep}, l}, \theta^l)$; the estimated p_B -value is the proportion of these L_1 simulations for which $Q(Y^{\text{rep}, l}, \theta^l) \geq Q(Y, \theta^l)$.

From a formal hypothesis-testing perspective, the use of PPC in Bayesian model checking has been criticized for its conservatism because the p_B -value is not uniformly distributed under the null model [17–20]. The observed data are used twice insofar as they are checked against replicated data generated from parameters drawn conditional on the same observed data. This might make it difficult to criticize the model being checked. Alternative model-checking strategies with uniformly distributed ‘ p -values’ under the null model have been proposed [18, 19, 21–25]. But these methods are often complicated, requiring either identifying statistics of special type, complex algebraic derivation, or intensive computation.

In this paper, we do not use imputation diagnostics to ‘accept’ or ‘reject’ the model from a hypothesis-testing perspective. Our goal is rather to discover whether aspects of the data that do not fit the model well cause discrepancies in statistical inferences. We use PPC in the context of practical imputation analyses because of the following: (a) multiple imputation can be regarded as a form of replication; (b) the posterior predictive p -value quantifies the evidence that the discrepancy is not due to chance alone; and (c) the calculation is easy to implement and suitable for the practitioners who may lack the statistical resources to apply alternative Bayesian model-checking strategies. Our motivation also follows the supporting arguments in [26–30] for use of PPC with complete data in applied Bayesian analyses. We recognize that the posterior predictive p -values referred to the uniform distribution will tend to be conservative and explore this issue further in Sections 3.5 and 5.3.

3. Methods

3.1. Posterior predictive checking on completed data

Following the scheme in [7], we assess the posterior distribution of the completed-data discrepancy, the difference between function of completed data and their replicates under the model, that is, $Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{com}}^{\text{rep}}$ and Y_{mis} are drawn from $f(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}}) = \int f(Y_{\text{com}}^{\text{rep}} | \theta) f(Y_{\text{mis}}, \theta | Y_{\text{obs}}) f(\theta | Y_{\text{obs}}) d\theta$. We summarize this distribution by a credible interval estimated either by the empirical distribution across L_1 simulations or by a computationally efficient approximation (Appendix A). The mean, median, or other summary of such an interval (i.e., $\bar{Q}(Y_{\text{com}}^{\text{rep}}) - \bar{Q}(Y_{\text{obs}}, Y_{\text{mis}})$) quantifies the average discrepancy, which can be viewed as the bias introduced by the model for some Q of analytic interest (Section 3.2).

In addition, we can calculate the posterior predictive p -value for the imputation model as

$$p_{B,\text{com}} = P(Q(Y_{\text{com}}^{\text{rep}}) \geq Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}) = \int \int I_{Q(Y_{\text{com}}^{\text{rep}}) \geq Q(Y_{\text{obs}}, Y_{\text{mis}})} f(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}}) dY_{\text{com}}^{\text{rep}} dY_{\text{mis}}. \quad (2)$$

We can estimate $p_{B,\text{com}}$ by simulation. For $l = 1, \dots, L_1$, we draw θ^l from $f(\theta | Y_{\text{obs}})$ and impute Y_{mis}^l from $f(Y_{\text{mis}} | Y_{\text{obs}}, \theta = \theta^l)$, as in a typical multiple imputation for Y_{mis} , and simulate $Y_{\text{com}}^{\text{rep},l}$ from $f(Y_{\text{com}}^{\text{rep}} | \theta = \theta^l)$. We can estimate $p_{B,\text{com}}$ as the proportion of the L_1 draws for which $Q(Y_{\text{com}}^{\text{rep},l}) \geq Q(Y_{\text{obs}}, Y_{\text{mis}}^l)$. To avoid the computational burden of simulation in practice, we can approximate $p_{B,\text{com}}$ on the basis of a modest number of replicates (Appendix A).

3.2. Checking targeted to analytic inferences in imputation

In a typical imputation project, analysts are ultimately concerned with substantive analysis results using imputations. Therefore, choosing discrepancy functions Q that are estimands of these analyses links the model diagnostics with the analytic objectives. For example, if a proposed analysis is a regression, analysis-specific discrepancy functions Q might include the regression coefficients, t -statistics (standardized regression coefficients), or significance levels, all of which are functions of the data. The magnitude or range of the completed-data discrepancy, $Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}})$, quantifies the bias of the imputation analysis if the

entire dataset were generated from the assumed model. The associated $p_{B,\text{com}}$ -value helps us to decide if the discrepancy is caused by chance alone. The discrepancies can be considered in the context of the scientific objective of the analysis. For instance, a very extreme $p_{B,\text{com}}$ -value may be of little concern if the average discrepancy is ‘practically’ insignificant for the analysis of interest. This focus distinguishes our strategy from previous approaches using generic goodness-of-fit summaries.

Model diagnostics targeted to analytic inferences directly address the ‘usability’ of the imputation model. In many cases, sensible analysis results can be obtained under uncongenial imputation models [10]. Note that we suppress θ in $Q()$ of completed-data discrepancy to stress that completed-data analyses may have little to do with the imputation model because of possible uncongeniality. The purpose of applying PPC, therefore, is not to reject or invalidate ‘wrong’ or ‘uncongenial’ imputation models but rather to detect relevant model deficiencies and areas for improvement. We elaborate this point further in our examples (Section 4). Note that our methods are intended to diagnose the adequacy of the imputation model to support particular post-imputation analyses, not to assess the adequacy of those analyses, which rest on the analyst’s scientific knowledge, further informed by conventional diagnostics.

Many applications of multiple imputation are in ‘one-analyst-one-goal’ studies, where the imputer is also the analyst, often with a specific analysis goal. In these ‘in-house’ applications [31], the analysis procedures or models are known to the imputer/analyst and hence can be used in evaluating the discrepancy. In ‘outside’ applications, where imputations to public databases will be used for ‘many-analysts-many-goals’, the imputer may not know all potential analyses. However, the imputer could apply checks on the basis of his or her substantive knowledge and information provided by analysts about the potential analyses. On the other hand, if the imputer routinely provides both imputed data and replicates, analysts can check the adequacy of imputation models for their own analyses.

3.3. Practical implementation by re-imputation

More generally, PPC can be viewed as diagnosing models by re-imputing completed data using any desired method for multiple imputation. In Equation (2), $Y_{\text{com}}^{\text{rep}}$ and Y_{mis} can be obtained as imputations in the concatenated dataset $(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}}, Y_{\text{obs}})$. Suppose $Y = (Y_1, Y_2)$, where Y_1 includes incomplete variables whose imputation method (conditional on Y_2) is to be assessed while Y_2 is complete. As shown in Figure 1, we can first create a duplicated dataset in which Y_2 is retained, but the incomplete Y_1 is made completely missing. Then, we

can concatenate the original and duplicate sets together as $\begin{pmatrix} (Y_{1,\text{obs}}, Y_{1,\text{mis}}), & Y_2 \\ Y_{1,\text{com}}^{\text{rep}}, & Y_2 \end{pmatrix}$, where $Y_{1,\text{mis}}$ and $Y_{1,\text{com}}^{\text{rep}}$ are missing. Imputations for $Y_{1,\text{mis}}$ and replicates $Y_{1,\text{com}}^{\text{rep}}$ can be obtained by applying the imputation method for the original dataset to the concatenated set, and the comparison follows naturally. Note that duplication of Y_2 does not affect estimation of the conditional distribution $f(Y_1|Y_2)$ but has the operational advantage of allowing the imputation program to automatically produce $Y_{1,\text{com}}^{\text{rep}}$. This re-imputation strategy facilitates checking by practitioners who implement imputation through standard software, only requiring re-

application of the imputation code after simple data manipulation but requiring no specific knowledge of the form of the model or the algorithms used in imputation, which sometimes can be rather sophisticated.

The same re-imputation strategy can be used to check procedures that are not fully Bayesian, such as hot deck imputation [14, chapter 10] and sequential regression imputation (example 2, Section 4), because the checking method does not require $Y_{\text{com}}^{\text{rep}}$ and Y_{mis} to be posterior predictive draws defined under a fully Bayesian model. For the models that are not fully Bayesian, although the formal theory of PPC may not apply, the produced diagnostics still measure the predictive difference between the analytic inferences of the completed data and their replicates (re-imputations) under the model.

3.4. Checking imputation models for subsets of variables

Imputation analysis often involve multiple incomplete variables. Checking might be focused on several key missing variables, and therefore, it might be unnecessary to replicate the variables that are completely or almost completely observed or that are incomplete yet of little interest for a particular analysis. On the other hand, if the goal is to diagnose imputations for all the variables, it may be challenging to successfully replicate all variables simultaneously, perhaps much more difficult than the actually required imputation if only a few variables are missing for any case. In the latter case, segmenting the diagnostics by replication and checking of only a few variables at a time conditional on other variables is more manageable and might facilitate localization of model deficiencies. These deficiencies can then be addressed by such methods as transformation of the offending variable or modifying the conditional specification in a ‘chained equation’ approach. For the same reasons, Abayomi *et al.* [8] applied plot diagnostics to imputations of each variable fixing other variables in the data.

Let $Y = (Y_1, Y_2)$, where Y_1 includes the incomplete variables of interest in an assessment and $Y_2 = (Y_{2,\text{obs}}, Y_{2,\text{mis}})$ includes the remaining variables possibly incomplete. Missing data in Y_2 can first be ‘filled in’ by multiple imputation $Y_{2,\text{mis}}^m (m=1, \dots, M)$ drawn from $f(Y_{2,\text{mis}} | Y_{1,\text{obs}}, Y_{2,\text{obs}})$. Then, PPC can be performed to compare $Q(Y_{1,\text{com}}^{\text{rep},m}, Y_{2,\text{obs}}, Y_{2,\text{mis}}^m)$ with $Q(Y_{1,\text{obs}}, Y_{1,\text{mis}}^m, Y_{2,\text{obs}}, Y_{2,\text{mis}}^m)$ for each of the M datasets, where $Y_{1,\text{com}}^{\text{rep},m}$ and $Y_{1,\text{mis}}^m$ are drawn from $f(Y_{1,\text{com}}^{\text{rep}}, Y_{1,\text{mis}} | Y_{1,\text{obs}}, Y_{2,\text{obs}}, Y_{2,\text{mis}}^m)$. Thus, Y_1 is replicated and checked conditional on completed Y_2 assuming the imputations of Y_2 are adequate. We can then average the diagnostics (e.g., the average discrepancy or $p_{B,\text{com}}$ -value) across the M datasets.

3.5. Variance reduction by integrating out missing data in comparison

Let $Y_{\text{com}}^{\text{rep}} = (Y_{\text{obs}}^{\text{rep}}, Y_{\text{mis}}^{\text{rep}})$, where ‘obs’ and ‘mis’ index corresponding to positions in the original and replicated data. In evaluating the completed-data discrepancy, Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$ are generated from the same model but are imputed independently (conditional on θ), and including them adds variance to the comparison, reducing its apparent power. It might appear to be more preferable to compare the observed data with their replicates (observed-data discrepancy, e.g., $Q(Y_{\text{obs}}^{\text{rep}}) - Q(Y_{\text{obs}})$) to identify the model inadequacy. But unless the data are missing completely at random, $Q(Y_{\text{obs}})$ can be biased from the completed-data

inference $Q(Y_{\text{com}})$ under the more general missingness at random mechanism. In addition, directly calculating $Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}})$ might be difficult with general missingness patterns for multiple incomplete variables when the observed-data pattern is not conducive to calculation of Q requiring a complete rectangular data matrix.

We propose to reduce the variance of the distribution of the completed-data discrepancy by integrating out the missing data and their replicates through the decomposition

$f(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}}) = f(Y_{\text{mis}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}) f(Y_{\text{obs}}^{\text{rep}} | Y_{\text{obs}})$. Let $D_{\text{com}} = Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}})$, a draw of completed-data discrepancy, and $D_{\text{ecom}} = E(Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}})$, the expectation conditional on $Y_{\text{obs}}^{\text{rep}}$. By the rule of expectation, $E(D_{\text{com}} | Y_{\text{obs}}) = E(D_{\text{ecom}} | Y_{\text{obs}})$ and $\text{var}(D_{\text{com}} | Y_{\text{obs}}) \geq \text{var}(D_{\text{ecom}} | Y_{\text{obs}})$. So, the expectations are the same, but $f(D_{\text{com}} | Y_{\text{obs}})$ might be tighter as a consequence of variance reduction.

We can calculate the posterior predictive p -value corresponding to this expected completed-data discrepancy, denoted by $p_{B,\text{ecom}}$, as

$$p_{B,\text{ecom}} = P(E[Q(Y_{\text{com}}^{\text{rep}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}] \geq E[Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}] | Y_{\text{obs}}) \\ = \int I_{E[Q(Y_{\text{com}}^{\text{rep}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}] \geq E[Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}]} f(Y_{\text{obs}}^{\text{rep}} | Y_{\text{obs}}) dY_{\text{obs}}^{\text{rep}}, \quad (3)$$

so $p_{B,\text{com}} = P(D_{\text{com}} \geq 0 | Y_{\text{obs}})$ and $p_{B,\text{ecom}} = P(D_{\text{ecom}} \geq 0 | Y_{\text{obs}})$. Under normal approximations (Appendix A),

$p_{B,\text{com}} \approx \Phi(E(D_{\text{com}} | Y_{\text{obs}}) / \sqrt{\text{var}(D_{\text{com}} | Y_{\text{obs}})})$, $p_{B,\text{ecom}} \approx \Phi(E(D_{\text{ecom}} | Y_{\text{obs}}) / \sqrt{\text{var}(D_{\text{ecom}} | Y_{\text{obs}})})$, and hence $p_{B,\text{ecom}}$ is closer to the two tails than $p_{B,\text{com}}$ and tend to be less conservative than $p_{B,\text{com}}$ when both are referred to a uniform distribution.

In practice, we can integrate out Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$ using a multiple imputation simulation strategy. We treat Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$ as ‘missing data’ and Y_{obs} and $Y_{\text{obs}}^{\text{rep}}$ as ‘observed data’ and impute the former in $(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}}, Y_{\text{obs}})$ after drawing $Y_{\text{com}}^{\text{rep}}$ and Y_{mis} from $f(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}})$. The following is a simulation-based procedure for calculating $p_{B,\text{ecom}}$.

1. For $l_1 = 1, \dots, L_1$, we impute $Y_{\text{mis}}^{l_1}$ and simulate $Y_{\text{com}}^{\text{rep}, l_1}$ from $f(Y_{\text{com}}^{\text{rep}}, Y_{\text{mis}} | Y_{\text{obs}})$ as in the procedure for estimating $p_{B,\text{com}}$ (Figure 1).

2. For the l_1 th replicate, we need to calculate

$E(Q(Y_{\text{com}}^{\text{rep}, l_1}) | Y_{\text{obs}}^{\text{rep}, l_1}, Y_{\text{obs}}) - E(Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1}) | Y_{\text{obs}}^{\text{rep}, l_1}, Y_{\text{obs}})$ or the discrepancy averaging out Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$. For the concatenated data $(Y_{\text{com}}^{\text{rep}, l_1}, Y_{\text{mis}}^{l_1}, Y_{\text{obs}})$ we set $Y_{\text{mis}}^{l_1}$ and $Y_{\text{mis}}^{\text{rep}, l_1}$ as missing values and re-impute them using L_2 -independent draws $Y_{\text{mis}}^{\text{rep}, l_1 l_2}$, $l_2 = 1, \dots, L_2$ under the imputation model (Figure 2). For each of the L_2 imputations, we calculate the difference:

$D_{l_1 l_2} = Q(Y_{\text{com}}^{\text{rep}, l_1}, Y_{\text{mis}}^{\text{rep}, l_1 l_2}) - Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1 l_2})$ and then obtain the average

$$\bar{D}_{l_1} = \sum_{l_2=1}^{L_2} D_{l_1 l_2} / L_2 \text{ for the } l_1 \text{th replicate. Note that as } L_2 \rightarrow \infty, \\ \bar{D}_{l_1} \rightarrow E(Q(Y_{\text{com}}^{\text{rep}, l_1}) | Y_{\text{obs}}^{\text{rep}, l_1}, Y_{\text{obs}}) - E(Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1}) | Y_{\text{obs}}^{\text{rep}, l_1}, Y_{\text{obs}}).$$

3. Calculate the $p_{B, \text{ecom}}$ as the proportion of the L_1 draws for which $D_{l_1}^- = 0$.

As can be seen, this integration is performed as a double (nested) simulation; the inner simulation integrates out $Y_{\text{mis}}^{\text{rep}}$ and Y_{mis} by averaging over L_2 draws of each, and the outer simulation estimates a $p_{B, \text{ecom}}$ -value over L_1 draws of $Y_{\text{obs}}^{\text{rep}}$. The advantage of variance reduction comes at the price of substantial additional computational effort, as it requires a nested simulation. We also provide an approximation formula for $p_{B, \text{ecom}}$ with a modest number of replicates in Appendix A.

3.6. Model discrepancy and imputation inference: a simple example

The completed-data discrepancy characterizes the deviation between completed-data statistics, and their replicates, originally observed and missing, are generated under the candidate model. When the fraction of missing data is small, this is likely to be larger than the deviation between the completed-data statistics, assuming the missing data are imputed under the correct model. As Rubin and Schenker [32] stated, ‘A 80% OK model applied to a dataset with 30% missing information leads to only 6% deficiency ($6\% = (1-80\%) \times 30\%$)’, but this may yield 14% completed-data discrepancy ($14\% = (1-80\%) \times (1-30\%)$).

For a simple illustration of the relationship between the model discrepancy and the bias of an imputation inference, consider a univariate sample of observed Y_{obs} of size n_{obs} , with data missing completely at random and $r = n_{\text{mis}}/n_{\text{obs}}$. The observed-data sample mean is $\bar{Y}_{\text{obs}}$. We check the imputation model $Y_{\text{mis}} \stackrel{\text{i.i.d.}}{\sim} N(\mu, 1)$, where μ is a predetermined constant, so a difference between μ and $\bar{Y}_{\text{obs}}$ indicates a location shift of the model.

Let $Q(Y) = \bar{Y}$, then $Q(Y_{\text{obs}}) = \bar{Y}_{\text{obs}}$, $Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}} \sim N((\bar{Y}_{\text{obs}} + r\mu)/(1+r), r/((1+r)^2 n_{\text{obs}}))$, $Q(Y_{\text{com}}^{\text{rep}}) | Y_{\text{obs}} \sim N(\mu, 1/((1+r)n_{\text{obs}}))$. We then have the following discrepancies and p -values conditional on Y_{obs} :

- A. $Q(Y_{\text{obs}}^{\text{rep}}) - Q(Y_{\text{obs}}) \sim N((\mu - \bar{Y}_{\text{obs}}), 1/n_{\text{obs}})$, $p_B = \Phi((\mu - \bar{Y}_{\text{obs}}) \sqrt{n_{\text{obs}}})$.
- B. $Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}}) \sim N((\mu - \bar{Y}_{\text{obs}}) / (1+r), (2r+1) / ((1+r)^2 n_{\text{obs}}))$, $p_{B, \text{com}} = \Phi((\mu - \bar{Y}_{\text{obs}}) \sqrt{n_{\text{obs}}} / \sqrt{2r+1})$.
- C. $E(Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}) \sim N((\mu - \bar{Y}_{\text{obs}}) / (1+r), 1/((1+r)^2 n_{\text{obs}}))$, $p_{B, \text{ecom}} = P_B = \Phi((\mu - \bar{Y}_{\text{obs}}) \sqrt{n_{\text{obs}}})$.

Note that, in this case, the p -value of discrepancy (A), which can be calculated directly from Y_{obs} and $Y_{\text{obs}}^{\text{rep}}$, is equivalent to that of discrepancy (C) obtained by integrating out Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$ because of the linearity of Q . As $r \rightarrow \infty$ while μ is fixed, the comparison using the observed-data discrepancy (A) or expected completed-data discrepancy (C) is unaffected, but $p_{B, \text{com}} \rightarrow 0.5$. This illustrates that both the similarity between Y_{mis} and $Y_{\text{mis}}^{\text{rep}}$ and the associated variance can offset the discrepancy from the observed data with a high proportion

of missing values if $p_{B,\text{com}}$ is used, whereas $p_{B,e,\text{com}}$ is less affected by the amount of missing data in its power of detecting model inadequacy.

We now quantify the deviation between the imputation inference and the correct inference using $E(Q(Y_{\text{obs}}, Y_{\text{mis}}^{\text{imp}}) - Q(Y_{\text{obs}})) = (r/(1+r))(\mu - \bar{Y}_{\text{obs}})$, where $Y_{\text{mis}}^{\text{imp}}$ is an imputed value under the assumed model. When $r \rightarrow 0$, both $p_{B,\text{com}}$ and $p_{B,e,\text{com}} \rightarrow \Phi((\mu - \bar{Y}_{\text{obs}}) / \sqrt{n_{\text{obs}}})$, which could be extreme if μ differs greatly from $\bar{Y}_{\text{obs}}$. But $(r/(1+r))(\mu - \bar{Y}_{\text{obs}}) \rightarrow 0$, showing that the inadequacy of the model has little effect on the imputation inference with little missing data. This illustrates that a large completed-data discrepancy does not necessarily suggest a large deviation between the imputation and the correct inferences with a small amount of missing data.

4. Examples

4.1. A biometric dataset with trivariate normal imputation

The first example uses a dataset (fitmiss) from the online documentation of SAS PROC MI and PROC MIANALYZE (http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_mi_sect002.htm) (SAS Institute Inc., Cary, NC, USA). PROC MI implements multiple imputation under the multivariate normal model, and PROC MIANALYZE combines multiple completed-data analysis results. These data on 31 men in a physical fitness course at North Carolina State University include three variables, oxygen (oxygen intake, milliliter per kilogram body weight per minute), runtime (time to run 1.5 miles in minutes), and runpulse (heart rate while running), each of which has a few missing values. A complete-case (CC) analysis involving all three variables would discard 11 of 31 observations.

A trivariate normal imputation model is assumed in the SAS documentation, which also illustrates a multiple imputation analysis of the linear regression of oxygen on runtime and runpulse as well as the regression of oxygen on runtime, runpulse, and their interaction. The former analysis is congenial with the normal imputation model, but the latter is more general and thus uncongenial.

We check whether the imputation model is adequate for the regression analysis with the interaction, which is not included in the imputation model. We perform PPC using both completed data ($L_1 = 5000$) and expected completed-data discrepancies ($L_1 = 5000$ and $L_2 = 50$) by simulation. Table I shows that the magnitude of the average discrepancies for the coefficients is nontrivial. The posterior predictive p -values are somewhat near 0 and 1, and $p_{B,e,\text{com}}$ -values are more extreme than $p_{B,\text{com}}$ -values. These results cast some suspicion on the model. Therefore, we applied an expanded imputation model with the interaction and included the corresponding coefficients in Table I for comparison. The difference of coefficients under two imputation models (i.e., $Q(Y_{\text{obs}}, Y_{\text{mis}})$) are around one-fourth of their standard errors (not shown) and do not affect the judgments of statistical significance. Thus, the original model appears to have some misfit for the data but does not appear to have a large impact on the imputation inference. Given the fact that imputation under the multivariate normal model is easier to implement using standard software than the model

with interactions, the results suggest that the former one is adequate for practitioners. In addition, we note that both p_{com} -values and $p_{B,\text{ecom}}$ -values are close to 0.5 under imputation model with interaction. This is expected because the imputation model is identical to the analysis model, and therefore, no adequacy will be detected under PPC (see more discussion in Section 5).

4.2. Multiple imputation for missing covariates in a Cox model

The second example uses a dataset from Eastern Cooperative Oncology Group study E9486, a randomized trial of three treatments for previously untreated multiple myeloma [33–35]. The main analysis of interest for this dataset assesses the effect on survival of a deletion on the short arm of chromosome 13 (denoted by $13q^-$) in a subgroup of 325 subjects on whom this abnormality was measured by fluorescent in situ hybridization [35].

The dataset (Table 8 of the Supporting Information) includes 13 covariates that have been associated with prognosis in myeloma: quantities that are measured in the blood or serum (albumin, β_2 -microglobulin, creatinine, the presence of the cytoplasmic-immunoglobulin heavy chain [IgA or IgG], and light chain [kappa], the percent of plasma cells in the bone marrow, and hemoglobin); characteristics of the circulating myeloma cells (plasma cell labeling index, IL-6 receptor status, the level of C-reactive protein, and $13q^-[a13q14 = 1$ if present; 0 otherwise]); and the traditional Durie–Salmon stage (1 for stage I or II and 0 for III). The variables have relatively low rates of missing data, ranging from 0% to 3.7%. A Cox regression analysis for the survival time using only the complete cases would result in a loss of 30 subjects (almost 10% of the sample). In this analysis, all continuous variables are dichotomized using previously published threshold values: $\text{album.c} = 1$ if albumin > 3 ; $\text{b2m.c} = 1$ if β_2 -microglobulin > 2.7 ; $\text{creat.c} = 1$ if creatinine > 2 ; $\text{pcli.c} = 1$ if %plasma cell labeling index > 1 ; $\text{il6.c} = 1$ if IL-6 receptor status is > 270 ; and $\text{crp.c} = 1$ if C-reactive protein > 2 .

We apply four imputation models to the incomplete variables in the original form, in order of increasing completeness:

- I.** Imputation of covariates under a multivariate normal model for binary, continuous, and semicontinuous variables. This model illustrates the common error of failing to include outcomes as predictors when imputing missing covariates.
- II.** Like model I but expanded to include survival time and the censoring indicator as predictors.
- III.** Like model II except that some of the continuous or semicontinuous variables are transformed to bring their distributions closer to normality.
- IV.** Sequential regression imputation [36, 37] for each variable conditional on the other variables. Some of the variables are transformed as in model III. Unlike models I–III, imputations of the binary variables are now by logistic regression.

Imputations produced from all four models are modified to make them consistent with the formats of the observed values, including rounding the continuous imputations for binary variables (in models I–III) and truncating continuous imputations outside the ranges of the

observed data. Table II shows the results from the CC analysis and the four imputation models. Each of the imputation models increases precision of the estimates relative to CC. The significance level of the predictor *a13q14* appears to be stronger from the multiple imputation analyses. However, there is overall little difference among the results from different imputation models because of a rather small proportion of missing data.

We perform diagnostics for models I–III using PPC with the library ‘norm’ in R (<http://www.r-project.org>). Although model IV is not fully Bayesian, we check it using the re-imputation strategy (Section 3.3) with the library ‘mice’, another R package. Table III shows the Cox model regression coefficients on the complete data and their replicates as well as their associated $p_{B,com}$ -values from all models, calculated over $L_1 = 5000$ simulations. Under model I, the regression coefficients of the incomplete covariates in the replicates (i.e.,

$\bar{Q}(Y_{com}^{rep})$) are biased towards 0 because of ignoring the survival outcome in the imputation, which also affects coefficients of the fully observed covariates in the replicates. For example, $\hat{\beta}_{a13q14}$ tends to be significantly higher in the replicates than the completed data with imputations (.379 vs. .302). The large average discrepancies are unlikely due to chance alone as many of the associated $p_{B,com}$ -values are rather extreme. By including the survival time and censoring indicator in the imputation, model II clearly improves on model I, as $\bar{Q}(Y_{com}^{rep})$ and $Q(Y_{obs}, Y_{mis})$ are getting closer, and the $p_{B,com}$ -values are moving to the center of the distribution. Applying transformations in model III appears to further improve the fit, especially for the variables that have been transformed (e.g., creat.c and pcli.c). Diagnostic results from model IV do not differ greatly from those of model III. For models III and IV, the discrepancies of the t -statistics (results not shown) do not indicate a change of significance for the predictors using replicates. Diagnostic results (not shown) from expected completed-data discrepancy estimated using simulation with $L_1 = 5000$ and $L_2 = 50$ are very similar, as can be expected in this example of very little missing data.

Although none of the imputation models is strictly congenial with the Cox model, the checking results clearly identify inadequacies of models I and II and suggest that both models III and IV are more adequate than the former. Given the fact that the discrepancies of models III and IV are not evident and the imputations can be implemented through standard software, they can be chosen as preferred ones to use for practical analyses. Further effort for improving imputations is possible, for example, by drawing missing covariates from a Cox model consistent with the survival analysis, but requires sophisticated statistical and computational expertise.

5. Simulation study

We use a simple simulation study to illustrate the performance of the proposed diagnostic methods, varying several factors including the complete-data sample size, missingness proportion, and different choices of the imputation and analysis models. We also assess the frequentist operating characteristics of posterior predictive p -values for imputation models from a formal hypothesis testing perspective.

5.1. Design

We consider two variables, an incomplete Y_1 and a fully observed Y_2 . The data-generating model is $Y_2 \sim \text{Unif}(-3, 3)$, $Y_1|Y_2 \sim N(1+Y_2+0.5Y_2^2, 1)$. The following describes the factorial simulation design:

1. Sample size: $n_{\text{com}} = 1000$ with 300 simulations; $n_{\text{com}} = 50$ with 600 simulations.
2. Missingness mechanism: Y_1 is missing at random with $\text{logit}(P(Y_1 \text{ is missing}|Y_2)) = \beta_0 + 0.5Y_2$, where β_0 is chosen so that approximately 20%, 40%, 60%, or 80% of cases are missing.
3. Imputation models: Y_1 is specified as a linear or quadratic function of Y_2 .
4. Completed-data analyses (choices of Q):
 - 4.A Mean and variance of Y_1 .
 - 4.B Percentiles of Y_1 (5%, 25%, 50%, 75%, and 95%).
 - 4.C The coefficient estimates and t -statistics from each of the following least-squared regression analyses, some of which (e.g., 4.C.C and 4.C.D) differ from models that generate the data and are uncongenial with the imputation models.
 - 4.C.A A quadratic regression of Y_1 on Y_2 .
 - 4.C.B A linear regression of Y_1 on Y_2 .
 - 4.C.C A quadratic regression of Y_2 on Y_1 .
 - 4.C.D A linear regression of Y_2 on Y_1 .

For both imputation models, we apply the Bayesian regression imputation procedure [1, chapter 5] to impute missing data and simulate replicates of Y_1 conditional on Y_2 , assuming noninformative prior distributions for the coefficients and the logarithm of the error variance. We estimate all completed-data discrepancies by simulation with $L_1 = 5000$ and calculate $p_{B,\text{ecom}}$ for analyses 4.B, 4.C.A, and 4.C.C by simulation with $L_1 = 500$ and $L_2 = 20$. R sample code of this simulation can be found in the supporting information.

5.2. Posterior predictive checking results for signaling model inadequacy

When the imputation model is quadratic and therefore matches the data-generating model, the average discrepancies are small and the associated posterior predictive p -values are very close to 0.5 across different scenarios, as expected (results not shown). Table IV shows the results using the completed-data discrepancy when the imputation model is a linear regression, with $n_{\text{com}} = 1000$ and 20% and 80% missingness rates. We present the average of statistics from completed data and their replicates, namely, $Q(Y_{\text{obs}}, Y_{\text{mis}})$ and $\bar{Q}(Y_{\text{com}}^{\text{rep}})$ as well as the average of the associated $p_{B,\text{com}}$ -values across simulations in each scenario. For the marginal mean and variance of Y_1 , the average statistics for the completed data and their replicates are indistinguishable and the associated $p_{B,\text{com}}$ -values are close to 0.5, showing little evidence of model inadequacy. This is because both the marginal mean and variance

are sufficient statistics of the model and tend to be replicated well by the linear regression imputation.

With $p_{\text{mis}} \approx 20\%$, average discrepancies for some of the percentiles (i.e., 5%, 50%, 75%, and 95%) are substantial, showing that the model fails to replicate well the marginal distribution of Y_1 . This is because the data are asymmetrically distributed, but the linear regression imputes from a symmetric distribution. The associated $p_{B,\text{com}}$ -values are also very extreme, suggesting the discrepancy is unlikely to be due to chance.

For both linear regression analyses, the regression coefficients are essentially the same in the completed data and replicates, showing no sign of model inadequacy. This is because the linear regression imputation preserves the correlation between the two variables. The results for t -statistics are similar (not shown).

From a model selection point of view, these diagnostics would suggest that a nonlinear term between Y_1 and Y_2 might be necessary to improve the model adequacy.

For the quadratic regression of Y_1 on Y_2 , the coefficients for the intercept and quadratic term are very different between the completed data and their replicates, clearly identifying the curvature that is omitted from the imputation model. The correspondingly $p_{B,\text{com}}$ -values are rather extreme. The quadratic regression analysis of Y_2 on Y_1 also suggests evidence of model inadequacy for the linear and quadratic terms, although this analysis is incoherent from the true data-generating model. Diagnostic results for the t -statistics (not shown) are consistent with those from the regression coefficients.

With more (80% instead of 20%) missing data, the diagnostics are less likely to detect imputation model inadequacy, with smaller average discrepancies of all statistics and less extreme $p_{B,\text{com}}$ -values. Nonetheless, some diagnostics still signal model inadequacy, such as those from the quadratic regression of Y_1 on Y_2 .

Smaller total sample sizes reduce the power for detecting model inadequacy, as expected. Comparing the results between $n_{\text{com}} = 50$ (not shown) and $n_{\text{com}} = 1000$ with $p_{\text{mis}} \approx 20\%$, the magnitudes of the average discrepancies are similar, but the $p_{B,\text{com}}$ -values of the former are less extreme because of the greater variation of the statistics.

The PPC results using the expected completed-data discrepancy show similar trends to those using the completed-data discrepancy with respect to the impact of total sample size and choices of analysis. The $p_{B,\text{ecom}}$ -values (Table IV) are either similar or more extreme than the corresponding $p_{B,\text{com}}$ -values, the latter occurring especially with the higher missingness rate (80%) but not apparent with small or moderate amount of missing data (e.g., 20%).

These results also illustrate the importance of the choice of Q . Letting Q be the quadratic regression coefficient of Y_1 on Y_2 or vice versa detects the inadequacy of linear regression imputation model because it identifies a key feature, the curvature of the relationship, omitted from the imputation model. On the other hand, an analysis more congenial with the same imputation model, such as the linear regression of Y_1 on Y_2 or vice versa, does not detect the discrepancy. Therefore, practitioners should try to identify statistics Q

corresponding to features of analytic interest that are not reflected in imputation models. This is related to the notion that sufficient statistics are generally not the optimal discrepancy functions in PPC [16].

5.3. Frequentist operating characteristics of posterior predictive p -values for imputation models

Under the quadratic (correct) regression imputation model, the distributions of the posterior predictive p -values are not uniform but rather concentrated around one-half, illustrating their conservatism under the correct model. Table V shows the selected percentiles of the p -values for some statistics. The extent of the conservatism, however, varies for different discrepancy functions. For example, p -values for the mean, linear regression coefficients, and quadratic regression coefficients of Y_1 on Y_2 are more concentrated around one-half. Those for the percentiles and quadratic regression coefficients of Y_2 on Y_1 are less so. For these same discrepancy functions, the distribution of $p_{B,ecom}$ -values is closer to uniform than that for $p_{B,com}$ -values, suggesting that the former is less conservative. For other statistics such as mean and linear regression coefficients, the distributions of the two p -values appear to be similar under the correct model. See also Figures 3 and 4 (Supporting Information) for the histograms of the p -values and their quantile plots versus the uniform distribution.

If we consider the model checking from a formal hypothesis-testing perspective, that is, controlling the type I error at a fixed level, we can calculate the power of the calibrated posterior predictive p -values in ‘accepting’ or ‘rejecting’ the model and make a comparison between $p_{B,com}$ and $p_{B,ecom}$. We first estimate the 5th and 95th percentiles of the p -values under the correct model from the simulation and then count the number of the p -values falling out the estimated interval under the incorrect (linear regression) model across 500 simulations. The relative frequency is an estimate of the power when controlling the type I error at .10.

Table VI shows the results with $n_{com} = 1000$. For the quadratic regression coefficients, the powers of both p -values are very high and similar with small or moderate missingness rate. The advantage of $p_{B,ecom}$ over $p_{B,com}$ is only apparent with relatively high missingness rates (e.g., 60% or 80%). For the statistics of mean, variance, and linear regression analysis, which have been shown to be ineffective in detecting the model inadequacy, the powers of $p_{B,ecom}$ and $p_{B,com}$ are usually less than .2, and the former can be either greater or smaller than the latter. For the percentile statistics (i.e., 5%, 50%, 75%, and 95%), with 80% missing, $p_{B,ecom}$ shows an obvious advantage of power over $p_{B,com}$, but they are similar with lower missingness rates. The pattern of power analysis for $n_{com} = 50$ is overall similar for the mean, variance, and regression estimates (results not shown). The power of $p_{B,ecom}$ can be either greater or smaller than that of $p_{B,com}$ for the quantile estimates, which, however, may be more likely to subject to simulation error with small sample size.

Overall, the results confirmed the fact that posterior predictive p -values tend to be conservative for checking imputation models in a hypothesis testing framework. They also appear to suggest that $p_{B,ecom}$ may have some power advantages over $p_{B,com}$ for discrepancy functions that effectively characterize the model inadequacy. The advantage is more salient with higher missingness rates.

6. Discussion

To check the adequacy of imputation models, we propose using simulation to create both imputations and replicates of completed data, applying analyses of interest to both the completed data and their replicates, and finally comparing analysis results from completed data with those obtained from replicates under the model. The methods can be implemented if the imputer and analyst are the same party. If the imputer and analyst are distinct, the imputer can distribute multiple replicates to be used by the analyst to check the imputation procedure against his or her own analyses. These methods could also be applied in related settings; for example, when confidentiality is protected by providing multiple synthetic datasets [38–40], the data producer can make similar comparisons to assess the adequacy of the data synthesis model.

Although the discrepancy measures, Q , are often generic summaries of fit in a typical PPC exercise, we advocate choosing analytic estimators of interest as Q in the context of imputation diagnostics. The model inadequacy, quantified by $Q(Y_{\text{com}}^{\text{rep}}) - Q(Y_{\text{obs}}, Y_{\text{mis}})$ or other derived statistics, is therefore directly related to the quantity of scientific interest. In addition, we have shown that the diagnostic measure is more effective by using Q 's from the target analysis that might not be directly considered in the imputation model. This strategy also fits the practical scenarios, where many implemented imputation models are standard yet uncongenial with post-imputation analyses.

Checking the adequacy of models is an important step of model refining and selection process. As we have illustrated with both simulated and real examples, inadequacies detected by extreme discrepancies in Q 's of interest can help us identify important features of data (e.g., nonlinear relationship) that are omitted from imputation models. We can then use conventional strategies such as transformation or include more variables to incorporate these features. The model diagnostics can also be helpful in the context of sensitivity analysis in which we apply several possible models as the better model would yield better diagnostics. In addition, imputations can be calibrated outside the model on the basis of the discrepancy measures [8, 41].

Our procedures check imputation models under the ignorability assumption by comparing completed data with simulated replicates. Nonignorable missing data models involve hypothetical assumptions about the missing values that are not based on the data and therefore cannot be supported or invalidated by comparisons of replicates with completed data. Nonetheless, models for nonignorably missing data are often specified by modifying a baseline ignorable model, for example, by hypothesizing a scale shift relative to ignorable imputations or a 'not missing at random' selection model. The proposed method can then be used to check the appropriateness of the baseline model as a starting point for developing a nonignorable model.

The availability of imputation software facilitates the use of multiple imputation, but it might encourage a false impression among practitioners that imputation is a purely mechanical process. Checking of imputation models is important because it can identify model defects and facilitate model improvement. The proposed checking procedure can be

implemented with simple enhancements of the original imputation code, in most cases, using standard multiple imputation software. Our methods are also naturally complementary to plotting techniques [7, 8] that can be used in exploratory detection of discrepancies between models and data. Therefore, we recommend that producers of imputation software add features to existing packages so that they can automatically produce replicates (as re-imputations) and use them to check imputation model adequacy for specific analyses. Typically, thousands of replicates are drawn to calculate $p_{B,com}$ or $p_{B,ecom}$ values. Our experiences show that this is generally not an issue for small or moderate sample size. With large sample sizes, we can reduce the number of simulations or use the approximation formula on the basis of a modest number of replicates. We believe that these methods should work reasonably well.

Consistent with previous literature for checking Bayesian models with complete data, our simulation study demonstrates that posterior predictive p -values tend to be conservative for checking imputation models. Among the methods proposed to remedy this conservatism, calibration of p -values (against their null distributions) [23] is a promising strategy. This approach, however, requires proper prior distributions specified for a fully Bayesian model and a double simulation scheme. Further efforts might be directed at adapting the calibration strategy to practical imputation analysis or determining when the generic PPC is close enough to calibration to be useful.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to express our appreciation to David Harrington, the associate editor, and two anonymous reviewers for their helpful critiques and suggestions. The National Cancer Institute grant U01 CA093344 supported this research.

APPENDIX A

Approximating posterior predictive p -values from a modest number of replicates

Simulation-based PPC procedures typically calculate p -values on the basis of a large number (often thousands) of replicates to obtain accurate estimates of tail probabilities. We propose to approximate the p -values on the basis of a modest number of replicates.

Suppose the posterior predictive distribution of the scalar Q is approximately normal, possibly after applying a transformation. For example, the normal assumption might hold well for regression model coefficients with moderate or large sample size, but transformations might be necessary for variance estimates or predicted probabilities. With complete data as in Section 2.2, we estimate the distribution of $Q(Y^{rep})|Y$ from the empirical distribution of L_1 -drawn values of $Q(Y^{rep,l})$. If $f(Q(Y^{rep,l})|Y)$ is normal and the normal model parameters have noninformative priors, then $Q(Y^{rep})|\{Q(Y^{rep,l})\} \sim t_{L_1-1}(\bar{Q}, \bar{Q}(1 + 1/L_1))$

$(Q(Y^{\text{rep},l}) - Q)^2 / (L_1 - 1)$), where $\bar{Q} = \sum_{l=1}^{L_1} Q(Y^{\text{rep},l}) / L_1$. The p_B -value can thus be approximated based on the cumulative distribution function of the t -distribution. Suppose $p_B = \alpha$ with infinitely many replicates, then with L_1 replicates and using the t -distribution, an approximate estimate of p_B is $F_{L_1-1}(\Phi^{-1}(\alpha)(1 + 1/L_1)^{-1/2})$. The following table shows the approximations under different numbers of replicates L_1 for some α . It appears that $L_1 = 50$ gives an overall good approximation and even $L_1 = 20$ might be adequate.

Table A.1

Approximations of p_B with different numbers of replicates.

α	L_1					
	5	10	20	50	100	1000
.01	.0436	.0254	.0172	.0127	.0113	.0101
.025	.0668	.0456	.0351	.0290	.0270	.0252
.05	.0968	.0739	.0621	.0548	.0524	.0502

To approximate $p_{B,\text{com}}$, let $D_l = Q(Y_{\text{com}}^{\text{rep},l}) - Q(Y_{\text{obs}}, Y_{\text{mis}}^l)$, $l = 1, \dots, L_1$; we can construct a t -distribution with the mean D as the average of D_l 's, the variance as $1 + 1/L_1$ times the variance of D_l 's, and $L_1 - 1$ degrees of freedom. Then, $p_{B,\text{com}}$ can be approximated by the probability of being non-negative under this t -distribution, that is, $Y_{\text{mis}}^{\text{rep},l_1 l_2}$.

The approximation to $p_{B,\text{ecom}}$ is based on a single-factor random-effects model [42, chapter 24]. Suppose L_1 imputations for Y_{mis} and L_1 replicates for $Y_{\text{com}}^{\text{rep}}$ are drawn, and for each $(Y_{\text{com}}^{\text{rep},l_1}, Y_{\text{obs}}, Y_{\text{mis}}^{l_1})$, we draw L_2 imputations of $Y_{\text{mis}}^{\text{rep},l_1 l_2}$ and $Y_{\text{mis}}^{l_1 l_2}$. We use σ_B^2 to denote the between-replicate variance for the population

$\{E(Q(Y_{\text{com}}^{\text{rep},l_1}) | Y_{\text{obs}}^{\text{rep},l_1}, Y_{\text{obs}}) - E(Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1}) | Y_{\text{obs}}^{\text{rep},l_1}, Y_{\text{obs}})\} (l_1=1, \dots, \infty)$ and σ_W^2 to denote the mean within-replicate variance for the populations

$\{Q(Y_{\text{obs}}^{\text{rep},l_1}, Y_{\text{mis}}^{\text{rep},l_1 l_2}) - Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1 l_2})\} (l_2=1, \dots, \infty)$. Again, let

$$D_{l_1 l_2} = Q(Y_{\text{obs}}^{\text{rep},l_1}, Y_{\text{mis}}^{\text{rep},l_1 l_2}) - Q(Y_{\text{obs}}, Y_{\text{mis}}^{l_1 l_2}), \bar{D}_{l_1} = \sum_{l_2=1}^{L_2} D_{l_1 l_2} / L_2, \text{ and}$$

$$\bar{D}_{..} = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} D_{l_1 l_2} / (L_1 L_2).$$

If σ_B^2 and σ_W^2 are known and the normal assumption holds, then the posterior distribution of $E(Q(Y_{\text{com}}^{\text{rep}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}}) - E(Q(Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}}^{\text{rep}}, Y_{\text{obs}})$, given the replicates and imputations, is $N(\bar{D}_{..}, (1 + 1/L_1)\sigma_B^2 + \sigma_W^2 / (L_1 L_2))$. When σ_B^2 and σ_W^2 are estimated from the replicates and imputations using unbiased moment estimates as

$$s_B^2 = \sum_{l_1=1}^{L_1} (\bar{D}_{l_1} - \bar{D}_{..})^2 / (L_1 - 1) - \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} (D_{l_1 l_2} - \bar{D}_{l_1})^2 / (L_1 L_2 (L_2 - 1)) \text{ and}$$

$$s_W^2 = \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} (D_{l_1 l_2} - \bar{D}_{l_1})^2 / (L_1 (L_2 - 1)), \text{ then the target distribution can be}$$

approximated by $t_{L_1-1}(\bar{D}_{..}, (1 + 1/L_1)s_B^2 + s_W^2 / (L_1 L_2))$. The $p_{B,\text{ecom}}$ can be approximated by the probability of being non-negative under this t -distribution.

We also use the simulation study in Section 5 to study the performance of the finite-replicate approximation of $p_{B,com}$. Tables 9 and 10 (Supporting Information) show selected results using normal approximation with $L_1 = 50$, 20 and 5. Overall, the $p_{B,com}$ -values approximated from $L_1 = 50$ and $L_1 = 20$ replicates are very close to those calculated from simulation with $L_1 = 5000$ across all scenarios and fairly close even with $L_1 = 5$.

References

1. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
2. Harel O, Zhou XH. Multiple imputation: review of theory, implementation, and software. *Statistics in Medicine*. 2007; 26:3057–3077. [PubMed: 17256804]
3. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007; 16:219–242. [PubMed: 17621469]
4. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression model. *The American Statistician*. 2007; 61:79–90. [PubMed: 17401454]
5. Yu LM, Burton A, Riverto-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*. 2007; 16:243–258. [PubMed: 17621470]
6. Gelman AE, King G, Liu C. Not asked and not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association*. 1998; 93:846–857.
7. Gelman AE, Mechelen IV, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*. 2005; 61:74–85. [PubMed: 15737080]
8. Abayomi K, Gelman AE, Levy M. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2008; 57:273–291.
9. Rizopoulos D, Verbeke G, Molenberghs G. Multiple imputation based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*. 2010; 66:20–29. [PubMed: 19459832]
10. Meng XL. Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science*. 1994; 10:538–573.
11. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*. 1984; 12:1151–1172.
12. Gelman AE, Meng XL, Stern HS. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistical Sinica*. 1996; 6:733–807.
13. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
14. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. New York: Wiley; 2002.
15. Meng XL. Posterior predictive p -values. *Annals of Statistics*. 1994; 22:1142–1160.
16. Gelman, AE.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. London: Chapman and Hall; 2004.
17. Dey DK, Gelfand AE, Swartz TB, Vlachos PK. A simulation-intensive approach for checking hierarchical models. *Test*. 1998; 7:325–346.
18. Bayarri MJ, Berger JO. P values for composite null models (with discussion). *Journal of the American Statistical Association*. 2000; 95:1127–1142.
19. Robins JM, van der Vaart A, Ventura V. The asymptotic distribution of p -values in composite null models (with discussion). *Journal of the American Statistical Association*. 2000; 95:1143–1172.
20. Sinharay S, Stern HS. Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*. 2003; 111:209–221.
21. Bayarri MJ, Berger JO. Quantifying surprise in the data and model verification (with discussion). *Bayesian Analysis*. 1999; 6:53–82.
22. Robins, JM. Comment on “Quantifying surprise in the data and model verification (with discussion)”. In: Bayarri, MJ.; Berger, JO., editors. *Bayesian Analysis*. Vol. 6. 1999. p. 67–74.

23. Hjort NL, Dahl FA, Steinbakk GH. Post-processing posterior predictive p values. *Journal of the American Statistical Association*. 2006; 101:1157–1174.
24. Johnson VE. Bayesian model assessment using pivotal quantities. *Bayesian Analysis*. 2006; 2:719–734.
25. Bayarri MJ, Castellanos ME. Bayesian checking of the second levels of hierarchical model (with discussion). *Statistical Science*. 2007; 22:323–343.
26. Meng, XL.; Gelman, AE. Comment on “Quantifying surprise in the data and model verification (with discussion)”. In: Bayarri, MJ.; Berger, JO., editors. *Bayesian Analysis*. Vol. 6. 1999. p. 75–76.
27. Van Dyk, DA. Comment on “Quantifying surprise in the data and model verification (with discussion)”. In: Bayarri, MJ.; Berger, JO., editors. *Bayesian Analysis*. Vol. 6. 1999. p. 77–78.
28. Stern, HS. Comment on “The asymptotic distribution of p -values in composite null models (with discussion)”. In: Robins, et al., editors. *Journal of the American Statistical Association*. Vol. 95. 2000. p. 1157–1159.
29. Gelman AE. A Bayesian formulation of exploratory analysis and goodness-of-fit testing. *International Statistical Review*. 2003; 71:369–382.
30. Gelman, AE. Comment on “Bayesian checking of the second levels of hierarchical model”. In: Bayarri, MJ.; Castellanos, ME., editors. *Statistical Science*. Vol. 22. 2007. p. 349–352.
31. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*. 1999; 8:17–36. [PubMed: 10347858]
32. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine*. 1991; 10:585–598. [PubMed: 2057657]
33. Liu P, Leong T, Quam L, Billadeau D, Kay N, Greipp P, Kyle R, Van Ness B. Activating mutation of N- and K-ras in multiple myeloma show different clinical associations: Analysis of the Eastern Cooperative Oncology Group Phase III trial. *Blood*. 1996; 88:2699–2706. [PubMed: 8839865]
34. Trendle M, Leong T, Kyle R, Kathzman J, Oken M, Kay N, Van Ness B, Greipp P. Prognostic significance of the s-phase fraction of light-chain-restricted cytoplasmic immunoglobulin (clg) positive plasma cells in patients with newly diagnosed multiple myeloma enrolled on Eastern Cooperative Oncology Group treatment trial E9486. *American Journal of Hematology*. 1999; 61:232–237. [PubMed: 10440908]
35. Fonseca R, Harrington D, Oken M, Gordon W, Bailey R, Van Weir S, Henderson K, Blood E, Rajkumar S, Kay N, Van Ness B, Greipp PR. Biological and prognostic significance of interphase fluorescence *in situ* hybridization detection of chromosome 13 abnormalities (13) in multiple myeloma: An Eastern Cooperative Oncology Group Study. *Cancer Research*. 2002; 62:715–720. [PubMed: 11830525]
36. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 1999; 18:681–694. [PubMed: 10204197]
37. Raghunathan TE, Lepkowski JM, VanHoewyk J. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001; 27:85–95.
38. Rubin DB. Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata. *Journal of Official Statistics*. 1993; 9:461–468.
39. Reiter JP. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*. 2002; 18:531–543.
40. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*. 2003; 19:1–16.
41. Yucel RM, He Y, Zaslavsky AM. Using calibration to improve rounding in imputation. *The American Statistician*. 2008; 62:125–129.
42. Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. *Applied Linear Statistical Models*. 4th Ed.. Illinois: Irwin; 1996.

Figure 1.
Schematic plot of implementation of posterior predictive checking for imputation models,
where $Y_{1,obs}$ are italicized and missing values are indicated by 'M'.

Figure 2.

Schematic plot of double simulation for calculating $p_{B,ecom}$, where $Y_{1,obs}$ and $Y_{1,rep}$ are italicized and missing values are indicated by 'M'.

Table I

Diagnostic results for the regression analysis with interaction under the imputation models for the fitmiss dataset.

Analysis	$Q(\bar{Y}_{\text{obs}}, Y_{\text{mis}})$	$\bar{Q}(Y_{\text{com}}^{\text{rep}})$	$p_{B,\text{com}}$	$p_{B,\text{ecom}}$
Imputation with no interaction				
Intercept	158.68	93.43	.179	.104
Runtime	-9.65	-3.19	.830	.906
Runpulse	-.46	-.08	.819	.892
Runtime $\times$ runpulse	.038	0	.174	.094
Imputation with interaction				
Intercept	177.48	175.35	.501	.500
Runtime	-11.43	-11.23	.495	.496
Runpulse	-.56	-.55	.494	.495
Runtime $\times$ runpulse	.048	.047	.505	.504

Table II

Cox regression analysis results for the myeloma dataset.

Predictors	CC		Model I		Model II		Model III		Model IV	
	EST (SE)	p-value	EST (SE)	p-value	EST (SE)	p-value	EST (SE)	p-value	EST (SE)	p-value
al3q14	.247 (.134)	.066	.298 (.126)	.019	.290 (.126)	.022	.297 (.126)	.018	.298 (.126)	.018
album.c	-.166 (.153)	.280	-.112 (.147)	.448	-.112 (.147)	.448	-.113 (.147)	.444	-.109 (.147)	.457
b2m.c	.293 (.147)	.046	.276 (.143)	.053	.296 (.141)	.036	.290 (.141)	.040	.291 (.141)	.039
creat.c	.474 (.184)	.010	.515 (.173)	.003	.492 (.173)	.004	.498 (.175)	.004	.505 (.174)	.004
iga	.181 (.220)	.410	.206 (.210)	.326	.191 (.211)	.365	.197 (.211)	.349	.202 (.213)	.343
igg	.077 (.201)	.700	.132 (.191)	.488	.126 (.192)	.514	.138 (.191)	.472	.127 (.192)	.508
kappa	-.192 (.149)	.200	-.185 (.141)	.188	-.193 (.141)	.171	-.190 (.142)	.182	-.191 (.143)	.182
pc.pct.c	.333 (.132)	.012	.356 (.126)	.005	.355 (.125)	.005	.351 (.126)	.005	.355 (.125)	.005
hemo.c	.025 (.146)	.860	.020 (.139)	.887	.020 (.139)	.887	.011 (.139)	.938	.018 (.139)	.895
pcli.c	.476 (.136)	.000	.453 (.129)	.000	.459 (.130)	.000	.458 (.130)	.000	.461 (.129)	.000
il6.c	.262 (.148)	.076	.238 (.142)	.092	.241 (.141)	.088	.224 (.141)	.113	.233 (.140)	.096
crp.c	.275 (.184)	.140	.273 (.171)	.109	.298 (.173)	.085	.286 (.178)	.107	.270 (.176)	.125
stage	.068 (.137)	.620	.050 (.132)	.703	.055 (.132)	.676	.053 (.132)	.688	.057 (.132)	.663

CC, complete-case analysis; models I–IV, multiple imputation analyses.

Table III

Diagnostics results from imputation models for the myeloma dataset.

Predictors	Model I			Model II			Model III			Model IV		
	$\bar{Q}(\bar{Y}_{\text{obs}}, Y_{\text{mis}})$	$\bar{Q}(Y^{\text{rep}}, Y_{\text{com}})$	$p_{B,\text{com}}$	$\bar{Q}(\bar{Y}_{\text{obs}}, Y_{\text{mis}})$	$\bar{Q}(Y^{\text{rep}}, Y_{\text{com}})$	$p_{B,\text{com}}$	$\bar{Q}(\bar{Y}_{\text{obs}}, Y_{\text{mis}})$	$\bar{Q}(Y^{\text{rep}}, Y_{\text{com}})$	$p_{B,\text{com}}$	$\bar{Q}(\bar{Y}_{\text{obs}}, Y_{\text{mis}})$	$\bar{Q}(Y^{\text{rep}}, Y_{\text{com}})$	$p_{B,\text{com}}$
al3ql4	.302	.379	.028	.300	.337	.196	.298	.359	.126	.297	.361	.169
album.c	-.111	-.008	.258	-.111	-.159	.614	-.113	-.120	.534	-.113	-.107	.487
b2m.c	.279	.054	.947	.281	.159	.809	.285	.307	.436	.286	.312	.426
creat.c	.509	.707	.001	.505	.634	.038	.501	.583	.135	.501	.588	.155
iga	.197	.007	.822	.196	.196	.493	.192	.192	.499	.193	.285	.371
igg	.129	.011	.727	.128	.067	.632	.124	.063	.629	.123	.146	.470
kappa	-.176	-.008	.130	-.178	-.173	.474	-.178	-.162	.463	-.181	-.223	.587
pc.pct.c	.354	.450	.005	.354	.421	.065	.354	.416	.096	.354	.413	.150
hemo.c	.015	-.074	.977	.015	-.039	.848	.013	-.009	.650	.013	-.008	.629
pcli.c	.457	.008	.999	.457	.174	.985	.457	.308	.864	.457	.280	.856
fl6.c	.228	.014	.939	.231	.168	.680	.232	.122	.761	.233	.108	.864
crp.c	.283	.017	.969	.289	.212	.713	.287	.360	.375	.286	.327	.432
stage	.052	-.008	.940	.053	.060	.468	.056	.095	.226	.056	.094	.300

Note: $\bar{Q}$ and $p_{B,\text{com}}$ -values are estimated from $L_1 = 5000$ simulations.

Simulation results for completed-data discrepancy with $n_{\text{com}} = 1000$, 20% and 80% missingness when the imputations are based on a linear regression of Y_1 on Y_2 .

Table IV

Analysis	$p_{\text{mis}} \approx 20\%$			$p_{\text{mis}} \approx 80\%$		
	$Q(\tilde{Y}_{\text{obs}}, Y_{\text{mis}})$	$\tilde{Q}(Y_{\text{rep}}^{\text{com}})$	$P_{B,\text{com}}$	$Q(\tilde{Y}_{\text{obs}}, Y_{\text{mis}})$	$\tilde{Q}(Y_{\text{rep}}^{\text{com}})$	$P_{B,\text{com}}$
Mean of Y_1	2.41	2.41	.500	2.20	2.20	.500
Variance of Y_1	4.88	4.88	.495	3.35	3.35	.492
5% of Y_1	-.52	-1.21	1	-.62	-.84	.874
25% of Y_1	.81	.86	.332	.92	.95	.401
50% of Y_1	1.97	2.41	0	2.10	2.20	.152
75% of Y_1	3.71	3.95	.029	3.39	3.45	.294
95% of Y_1	6.71	6.02	1	5.32	5.20	.757
Linear regression Y_1 on Y_2						
Intercept coefficient	2.40	2.40	.500	2.19	2.19	.500
Slope coefficient	.85	.85	.501	.55	.55	.500
Linear regression Y_2 on Y_1						
Intercept coefficient	-1.26	-1.25	.488	-1.06	-1.06	.488
Linear coefficient	.52	.52	.513	.48	.48	.509
Quadratic regression Y_1 on Y_2						
Intercept coefficient	1.33	2.40	0	1.90	2.19	.003
Linear coefficient	.85	.85	.492	.55	.55	.497
Quadratic coefficient	.36	0	1	.10	0	1
Quadratic regression Y_2 on Y_1						
Intercept coefficient	-1.18	-1.26	.878	.911	-1.03	.666
Linear coefficient	.42	.52	.011	.005	.42	.130
Quadratic coefficient	.02	0	.986	.995	.01	.891

Note: Diagnostics are estimated by simulation with $L_1 = 5000$ $P_{B,\text{com}}$ are only calculated for percentiles of Y_1 and quadratic regressions of Y_1 (Y_2) on Y_2 (Y_1).

Table V

Distribution of $p_{B,com}$ and $p_{B,ecom}$ -values under the correct imputation model with $n_{com} = 1000$ and 20% missingness.

Statistics	p -values	5%	25%	50%	75%	95%
Mean of Y_1	$p_{B,com}$	.467	.488	.502	.517	.535
	$p_{B,ecom}$	.465	.487	.504	.519	.537
75th percentile of Y_1	$p_{B,com}$	.175	.349	.505	.661	.864
	$p_{B,ecom}$	.098	.298	.505	.708	.934
Linear regression	$p_{B,com}$	.465	.486	.500	.515	.538
Y_1 on Y_2 : linear coefficient	$p_{B,ecom}$	.464	.485	.500	.514	.537
Linear regression	$p_{B,com}$	.474	.493	.509	.524	.542
Y_2 on Y_1 : linear coefficient	$p_{B,ecom}$	.472	.494	.509	.523	.542
Quadratic regression	$p_{B,com}$	.470	.491	.504	.517	.534
Y_1 on Y_2 : quadratic coefficient	$p_{B,ecom}$	.467	.492	.506	.518	.535
Quadratic regression	$p_{B,com}$	.194	.374	.503	.630	.792
Y_2 on Y_1 : quadratic coefficient	$p_{B,ecom}$	.143	.351	.502	.659	.848

Table VI

Power of $p_{B,com}$ and $p_{B,ecom}$ -values under linear regression imputation with $n_{com} = 1000$

Analysis	$p_{mis} \approx 20\%$		$p_{mis} \approx 40\%$		$p_{mis} \approx 60\%$		$p_{mis} \approx 80\%$	
	$p_{B,com}$	$p_{B,ecom}$	$p_{B,com}$	$p_{B,ecom}$	$p_{B,com}$	$p_{B,ecom}$	$p_{B,com}$	$p_{B,ecom}$
Mean of Y_1	.143	.113	.123	.087	.113	.090	.107	.117
Variance of Y_1	.100	.100	.140	.120	.113	.087	.130	.177
5% of Y_1	1	1	1	1	1	1	.763	.963
25% of Y_1	.240	.227	.443	.347	.477	.480	.043	.213
50% of Y_1	1	1	1	1	1	1	.853	.990
75% of Y_1	.987	.987	.983	.990	.943	.947	.010	.640
95% of Y_1	1	1	1	1	1	1	.600	.960
Linear regression Y_1 on Y_2								
Intercept coefficient	.160	.123	.143	.083	.110	.107	.113	.100
Linear coefficient	.123	.100	.133	.073	.120	.097	.113	.107
Linear regression Y_2 on Y_1								
Intercept coefficient	.133	.157	.097	.107	.123	.133	.123	.157
Linear coefficient	.127	.153	.117	.073	.087	.120	.117	.113
Quadratic regression Y_1 on Y_2								
Intercept coefficient	1	1	1	1	1	1	1	1
Linear coefficient	.850	.867	.813	.843	.723	.857	.590	.760
Quadratic coefficient	1	1	1	1	1	1	1	1
Quadratic regression Y_2 on Y_1								
Intercept coefficient	.983	.987	.993	.993	.983	.990	.567	.963
Linear coefficient	.993	.993	.997	1	.997	.997	.900	.997
Quadratic coefficient	.993	.993	.997	1	.997	.997	.903	.990