

**Multiple Imputation by Ordered Monotone Blocks**  
**with Application to the Anthrax Vaccine Research Program**

Fan Li, Michela Baccini, Fabrizia Mealli,

Constantine E Frangakis, Elizabeth R Zell, Donald B Rubin <sup>1</sup>

ABSTRACT. Multiple imputation has become a standard statistical technique for imputing missing values, where imputations are created as random draws from the posterior predictive distribution of the missing data. The CDC Anthrax Vaccine Research Program (AVRP) data created new challenges for multiple imputation due to the large number of variables of different types and the limited sample size. An intuitive method for handling such complex data is to specify, for each variable with missing values, a univariate conditional distribution given all other variables, in the form of a regression model. Such univariate imputation strategies are valid for monotone missing data, but have the theoretical drawback that the fully conditional distributions are generally incompatible when missing data are not monotone. Aiming at reducing incompatibility, we propose the “multiple imputation by ordered monotone blocks” approach to extend the theory for monotone patterns to arbitrary missing patterns. The key idea is to break an arbitrary missing pattern into a collection of smaller but monotone missing patterns. We apply this strategy to impute the missing data in the AVRP data and evaluate its performance by a novel simulation-based approach. A method for creating missing values in the simulated data sets, which mimics the observed missing data patterns, is also proposed.

KEY WORDS: Bayesian, conditional distribution, evaluation, imputation, incompatibility, missing data, monotone blocks.

---

<sup>1</sup>FL is assistant professor, Department of Statistical Science, Duke University, Durham, NC (email: fli@stat.duke.edu); MB is assistant professor and FM is professor, Dipartimento di Statistica, Università di Firenze, Florence, Italy; CEF is professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD; ERZ is mathematical statistician, Division of Bacterial Diseases, CDC, Atlanta, GA; DBR is the John Loeb Professor, Department of Statistics, Harvard University, Cambridge, MA. FL’s research is partially funded by NSF-SES grant 11-31897.

# 1 Introduction

Multiple imputation (MI) (Rubin, 1976, 1987a, 1996; Little and Rubin 2002) has become a standard statistical technique for dealing with missing data, and has been implemented in commercial and open source software packages such as PROC MI in SAS (SAS Institute Inc., 2008), MICE in STATA (StataCorp, 2011), SOLAS (Statistical Solutions Ltd., 2001), IVEware (Raghunathan *et al.*, 2002), and R (van Buuren and Oudshoorn, 1999; Royston, 2004). MI generally involves specifying a joint distribution for all variables in a data set, supplemented by a prior distribution for the parameters of this joint distribution in the Bayesian setting. Multiple imputations of the missing values are then created as random draws from the posterior predictive distribution of the missing values given the observed data.

MI has been successfully applied to many large surveys, including the Consumer Expenditure Survey (Raghunathan and Paulin, 1998), the National Health and Nutrition Examination Survey (Schafer *et al.*, 1998), the National Health Interview Survey (Schenker *et al.*, 2006), and the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium (He *et al.*, 2010), among others. However, MI for missing data in large-scale studies with complex designs remains challenging due to several complications, which are described below using the Anthrax Vaccine Research Program (AVRP) (Marano *et al.*, 2008), conducted by the Centers for Disease Control and Prevention (CDC), as a typical example.

First, the AVRP collects a large number of variables of different types such as continuous, binary, categorical and mixed continuous variables. Data models for MI are often based on the multivariate normal or general location model (e.g., Schafer, 1997), neither of which is appropriate for the AVRP data, as well as for other complex observational and experimental data, which often include non-continuous variables. A commonly adopted strategy for handling such complex data is the “multiple imputation by chained equations (MICE)” (Raghunathan *et al.*, 2001), where one specifies for each variable with missing values a univariate conditional distribution given all other variables, and impute the variables with missing values in a

pre-specified order (e.g., from the most to the least observed variable). Model fitting and imputation steps are performed iteratively, until convergence is achieved under certain criteria. The univariate distributions take the form of regression models, which are straightforward to work with and can accurately reflect different data types. MICE is implemented in several software packages, and has been applied to a number of real studies (e.g., Barnard and Meng, 1999; Taylor *et al.*, 2002; He *et al.*, 2010). However, a potential problem of MICE is the possible incompatibility among the univariate conditional distributions (Arnold and Press, 1989). Theoretically, imputation based on univariate conditionals is valid for monotone missing data, where the univariate distribution for each variable is specified conditional only on the variables that are “more observed”. However, when the missing pattern is not monotone, the collection of fully conditional distributions in MICE may not correspond to any joint distribution for all the variables, i.e., the conditional distributions may be incompatible. In principle, MICE defines a potentially incompatible Gibbs sampler (PIGS). Examples in Li *et al.* (2011) illustrate the dramatic consequences potentially induced by model incompatibility in MICE. Namely, different orders of imputing variables can lead to completely different “convergent distributions”, including non-convergence situations. Despite this, some simulation studies suggest that PIG samplers give satisfactory imputation results in practice (e.g. van Buuren et al, 2006).

Li *et al.* (2011) proposed a general spectrum of imputation strategies - imputation by ordered monotone blocks (IMB), allowing one to choose an imputation strategy to reduce incompatibility while keeping the simplicity of univariate conditional modeling. MICE can be viewed as the simplest case of IMB. IMB extends the theory for monotone patterns of missing data to arbitrary patterns by breaking the problem into a collection of smaller problems where missing data do form a monotone pattern. Focused on theoretical concepts, Li *et al.* (2011) did not discuss many of the key elements in the implementation of IMB.

Second, in the AVRP data, besides the common continuous, semi-continuous, ordinal, categorical, and binary variables, there are several special types of variables for which standard

regression models are not adequate, such as variables with very low variability or defined only on a subset of units (see Section 3.4). Standard generalized linear regression models are no longer directly applicable and specific techniques are required to incorporate such variables in imputation.

Third, while the number of variables is very large, there is a limited number of study participants within each treatment arm. This raises the issue of variable selection when specifying regression models within the multiple imputation procedure.

Fourth, assessment of the imputations is as often important as the imputation itself for valid inference, especially when the imputation stage involves automatized procedure of complex modeling. However, research on this front is limited (Schafer *et al.*, 1998; Tang *et al.*, 2005; He and Zaslavsky, 2011) and evaluation is seldom done in real applications. As such, general guidelines for designing sensible and easy-to-implement evaluation of the MI results need to be developed.

The above challenges are not unique to the AVR P data. Most large-scale surveys and clinical trials with missing data face similar issues. Using the AVR P data as an illustration, our goal is to provide a general template for imputation and evaluation of large complex data and develop corresponding software. In particular, we adopt and refine upon the original IMB proposal. The complete imputation plan for the AVR P will be described. We will propose simulation-based evaluation approaches under both the frequentist and Bayesian paradigms and evaluate performance of the IMB in the AVR P data. The rest of the paper proceeds as follows. Section 2 describes the AVR P, together with its missing data problems. Section 3 presents the complete imputation procedure based on the IMB scheme. Details on the prediction models used for imputation, as well as on methods for variable selection, are provided. Section 4 describes the evaluation plan based on simulations, the method used for creating simulated datasets with missing values, and the evaluation criteria. The evaluation results are discussed in Section 4.2. Section 5 concludes with a discussion.

## 2 Motivating Example: the AVRP

Anthrax is a highly lethal acute disease in humans and animals. Because of concerns about the potential for bioterrorism, US military personnel are now routinely vaccinated against anthrax, prior to active service in places where biological attacks are considered a threat. The currently FDA-licensed anthrax vaccine, Anthrax Vaccine Adsorbed (AVA: BioThrax, Emergent BioSolutions, Inc., Rockville, MD), was licensed in 1970 based on a human clinical trial demonstrating protection of mill workers against anthrax. The 1970 licensed regimen for AVA was subcutaneous (SQ) administration of a series of six primary doses (0, 2, 4 weeks and 6, 12, 18 months) followed by annual booster doses. As a consequence of the CDC AVRP interim analysis data (Marano *et al.*, 2008) the licensed use of the vaccine was changed to intramuscular (IM) administration at 0, 1, 6, 12 and 18 months.

The AVRP trial was a 43-month prospective, randomized, double-blind, placebo-controlled trial for the comparison of immunogenicity (i.e., immunity) and reactogenicity (i.e., side effects) elicited by AVRP given by different routes of administration, subcutaneous (SQ) versus intramuscular (IM); and dosing regimens, as many as 8 doses versus as few as 4 doses. In the AVRP trial, sterile saline was used as the placebo. At the time of the interim analysis, the AVRP program had enrolled 1563 participants, healthy adult men and women of 18 to 61 years of age, at five sites in the United States. Participants were randomized into one of seven study groups. One group received AVA as currently licensed (SQ with 6 doses followed by annual boosters); another two groups respectively received saline IM and SQ at the same time points as the currently licensed regimen. The four other groups received AVA IM, one group at the same time points as the currently licensed regimen and the remaining groups in modified dosing regimens; placebo was given when a dose of AVA is omitted from the licensed dosing regimen. There were a total of 25 required visits over a period of 42 months, during which all participants received an injection of vaccine or placebo (8 injections total), had a blood sample drawn (16 total), and have an in-clinic examination for adverse events (22 total). Total anti-

protective antigen IgG antibody (anti-PA IgG) levels were measured using a standardized and validated Enzyme-Linked ImmunoSorbent Assay (ELISA); the primary study endpoints are non-inferiority at month 2, 7, 43 of anti-protective antigen IgG geometric mean concentration (GMC), geometric mean titer (GMT) and proportion of responses with a 4-fold rise in titer (%4 XR). Reactogenicity outcomes were proportions of injection site and systemic AEs. All adverse events, including vaccine reactogenicity, are actively monitored. Several reactogenicity endpoints were assessed. Potential risk factors for adverse events, e.g., sex, pre-injection anti-PA IgG titer, are also recorded. More details on the AVRVP can be found in Marano *et al.* (2008) and Baccini *et al.* (2010).

The AVRVP trial was significant because it is designed to provide the basis in 2008 for a change in the route of AVRVP administration from SQ to IM and of a reduction in the number of vaccine doses. In fact, interim AVRVP results based on available case data have already led to change from SQ to IM regime at FDA. The study final data set will be used in consideration of additional reductions in the vaccine priming and booster series. The study final data set will be used in consideration of additional reductions in the vaccine priming and booster series. However, the length and the complexity of the study design, which creates more than 2000 variables at the end of the study, pose enormous challenge in statistical analysis, due to large amount of missing data generated by dropouts, missed visits and missing responses. Any comparison, such as simple intent-to-treat (ITT), and per-protocol (PP) comparisons, requires proper handling of the missing data (Mealli and Rubin, 2002). The simplest complete data analysis that drops any subjects with missing data is not applicable here. Looking at the interim data, even though the overall missing rate is low (3.4%), only 56 (mainly baseline covariates) among the approximately 400 available variables are fully observed and only 208 out of the 1005 subjects have fully observed variables. Other commonly used *ad hoc* strategies, including “hotdeck” and “last observation carried forward”, lack theoretical justification and are known to potentially lead to severe bias. Thus we adopt the theoretically justified MI approach to

handle the missing data. However, as mentioned above, the complex data structure substantially complicates the implementation of MI in the AVRP trial.

### 3 Multiple Imputation Strategy

The MI strategy developed here helps to satisfy two distinct objectives: (1) reducing algorithmic incompatibility by breaking the arbitrary missing data pattern into monotone blocks; (2) simplifying the modeling and the problem of fewer observations than variables by specifying univariate conditional models within the monotone blocks.

#### 3.1 Imputation by Ordered Monotone Blocks: General Algorithm

The current state-of-the-art procedures for imputing missing data fit fully Bayesian models, assuming some joint probability distribution for the underlying complete data. As a result, to handle general missing data patterns, a principled modeling approach requires, for each model, high level expertise in both statistical computing methodology and software development.

We first introduce some general notation. Suppose there are  $N$  units and  $J$  variables  $Y_j$ ,  $j = 1, \dots, J$ . Denote the value of  $Y_j$  for unit  $i$  by  $y_{ij}$ , and the observed and missing data in  $Y_j$  by  $Y_j^{obs}$  and  $Y_j^{mis}$ , respectively. Let  $Y_{-j} = \{Y_k : k \neq j\}$ . Also denote the response indicators by  $\mathbf{R} = \{R_{ij}\}$ , where  $R_{ij} = 1$  if  $y_{ij}$  is observed and  $R_{ij} = 0$  otherwise; let  $M_j = \{i : R_{ij} = 0\}$ , be the set of missing entries in variable  $j$ , and let  $M = \bigcup_{j=1}^J M_j$ . In this paper we assume the units are exchangeable, and the missing data are missing at random (MAR) as defined in Rubin (1976).

A set of missing data is *monotone* if a permutation  $\{k_j, j = 1, \dots, J\}$  of  $j = 1, \dots, J$  exists such that  $Y_{k_{j+1}}$  is missing whenever  $Y_{k_j}$  is, i.e.,  $M_{k_j} \subset M_{k_{j+1}}$ , for  $j = 1, \dots, J - 1$ . When the missing data pattern is monotone, we may impute missing data variable by variable sequentially as follows:

**Sequential Imputation for Monotone Missing Data.** For  $k = 1, \dots, J$ : (a) Specify a distribution of  $Y_{k_j}$  conditional on the more observed variables  $Y_{k_1}, \dots, Y_{k_{j-1}}$  and a prior distribution  $\pi(\theta)$  for the parameters  $\theta$ , and (b) obtain the posterior distribution of  $\theta$  using only the units with observed  $Y_{k_j}$ , and (c) then impute the missing values  $Y_{k_j}^{mis}$  by random draws from their posterior predictive distribution.

This method is flexible and principled because, in this case, the product of arbitrary conditional distributions produces a well defined joint distribution. However, this method only applies to missing data that conform to a monotone pattern.

*Imputation by ordered monotone blocks* (IMB) (Li *et al.*, 2011) extends the theory for monotone patterns of missing data to arbitrary patterns by breaking an arbitrary pattern into a collection of smaller monotone patterns. Formally, a subset of missing entries composes a monotone missing block  $B_k$  if an ordered list exists of the  $J_k (\leq J)$  variables  $\{Y_{k_j}, j = 1, \dots, J_k\}$  having missing entries belonging to the subset, such that if  $y_{ik_l}$  is missing and belongs to the subset, then for any  $p > l$ ,  $y_{ik_p}$  is missing and belongs to the subset as well.

A *partition* of the missing data into monotone blocks is a collection of  $K$  mutually exclusive monotone blocks  $B_1, \dots, B_K$ , such that,

$$M = \cup_{k=1}^K B_k, \quad \text{and} \quad B_k \cap B_l = \emptyset, \text{ for } k \neq l.$$

An IMB algorithm includes a modeling stage and an imputation stage. The modeling stage consists of the following three elements:

1. *Partitioning all missing entries into monotone blocks.*
2. *Sequential specification of univariate conditional distributions in each monotone block:*  
 Within each block  $B_k$ , conditional models are sequentially specified for the variables with missing data, regarding the most recent imputed values of the missing entries outside the block as “observed”. Suppose the list of variables in  $B_k$  is  $\{Y_{k_j}, j = 1, \dots, J_k\}$

and denote the variables outside  $B_k$  by  $Y_{-B_k}$ . For each  $j$ , one specifies a conditional distribution of  $Y_{k_j}$  conditional on  $Y_{k_1}, \dots, Y_{k_{j-1}}$  and  $Y_{-B_k}$ . When the number of variables is large, such as in the AVR data, the model specification step can be done using variable selection techniques (e.g., stepwise selection).

3. *Order of imputation within and between the monotone blocks:* According to the approach followed with monotone missing data patterns, within each monotone block the variables are imputed by the ascending order of missing proportion within the block. Regarding the order of imputation between the monotone blocks, we propose to impute by the descending order of the total number of missing entries within the blocks, after some simple initial imputations of all the missing data.

In the imputation stage of IMB, after the initial imputation of all the missing data (e.g., by means of or random draws from the observed marginal distribution), we iteratively cycle through the variables and the blocks according to the pre-specified imputing order as follows: For each variable in each monotone block  $B_k$ , fit its specified conditional model given both the observed data and the imputed data outside  $B_k$ ; then impute its missing data in the block by sampling from their posterior predictive distributions. Iterate the imputation steps until the algorithm reaches stationarity.

Because the missing pattern in each  $B_k$  is monotone, the modeling and imputing stages within each block in theory would require a single iteration. However, since there is more than one monotone block, more than one iteration over the blocks is needed to ensure the final imputations to be stable.

An example of a partition into monotone blocks and the IMB algorithm is given in Table 1, where  $N = 10$ ,  $J = 6$ . Panel (a) shows the matrix of response indicators  $R'$  of the original data  $Y'_j$ 's; panel (b) shows  $R$  for the sorted  $Y'_j$ 's (permute both rows and columns of  $R'$ ) where the missing entries are re-arranged to appear in several monotone blocks; panel (c) labels the three monotone blocks  $B_1, B_2, B_3$  in the order of total number of missing entries. After the initial

$Y'_1$	$Y'_2$	$Y'_3$	$Y'_4$	$Y'_5$	$Y'_6$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$
1	1	0	1	1	1	1	0	0	1	1	1	1	$B_2$	$B_2$	1	1	1
1	1	1	0	0	0	1	1	0	1	1	1	1	1	$B_2$	1	1	1
0	1	0	1	1	1	1	1	0	1	1	1	1	1	$B_2$	1	1	1
1	1	1	0	1	0	1	0	1	1	1	1	1	$B_3$	1	1	1	1
1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	$B_1$
0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	$B_1$	$B_1$
1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	$B_1$	$B_1$
1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	$B_1$	$B_1$	$B_1$
1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	$B_1$	$B_1$	$B_1$

(a) unsorted

(b) sorted

(c) sorted with monotone blocks

Table 1: An example of IMB. Matrix of response indicators  $R = \{R_{ij}\}$ .

imputation, an IMB algorithm will first impute the missing data in  $B_1$  in the order of  $Y_4, Y_5|Y_4, Y_6|Y_5, Y_4$ , treating all the data outside  $B_1$  (both observed and most-recently imputed values) as observed; then impute  $B_2$  in the order of  $Y_2, Y_3|Y_2$ , given all the data outside  $B_2$  as observed; last impute  $B_3$  in the same manner. Repeat this procedure until “convergence” in certain sense is reached.

For a given dataset, there are many possible ways of partitioning the missing data into monotone blocks, thus many possible IMB strategies. One of the simplest partition is to take the missing values of each variable as a monotone block, i.e.,  $K = J$  and  $B_k = M_k, k = 1, \dots, J$ . This is exactly the strategy used by MICE. However, it is clear that the general IMB algorithm, same as MICE, defines a *potential incompatible Gibbs* (PIG) sampler. The full conditional distributions of this PIG sampler are the distributions of missing data for each monotone block given the missing data outside the block. There may not exist a joint distribution that has these full conditionals, hence they are called (potentially) *incompatible*. Nevertheless, modeling incompatibility does not necessarily lead to algorithmic incompatibility (i.e., if the PIG algorithm does not uniquely converge). If algorithmic compatibility is approximately achieved, and each conditional model for the monotone block fits the data, then the imputation may be judged as being reasonable (van Buuren *et al.*, 2004). Our strategy is to reduce algorithmic

incompatibility by using a particular combinatorics strategy of missing data partition within the spectrum of IMB algorithms that sequentially maximizes the number of missing entries in each monotone block (discussed in Section 3.2). We refer the readers to Li *et al.* (2011) for more comparison between MICE and IMB. In summary, a carefully selected IMB algorithm can have several advantages over MICE. First, it often avoids error propagation under model mis-specification; second, it reduces the possibility of mis-specification; and third, it reduces the possibility of over-fitting when the number of variables is large. The first advantage is fundamental and directly resulting from the reduced incompatibility in IMB; the last two ones are straightforward, since, by construction, the conditional models in IMB always involve less or equal number of covariates than in MICE. These advantages are especially desirable in imputing complex large-scale datasets, where model specification and imputation is typically done in an automatic fashion, leading to a high chance of mis-specification of conditional models.

### 3.2 Partitioning missing data into monotone blocks

Intuitively, the more missing entries the major monotone block includes, the closer an IMB is to imputation of the data with a fully monotone missing data pattern, so that the possibility of incompatibility is lower. The idea of defining blocks which include as many missing entries as possible is a natural extension of the *imputation by major monotone pattern* strategy proposed by Rubin (2003), which exploits a single major monotone block. Here we propose the following procedure to sequentially obtain the (approximately) largest first block. We first identify the variable with the most missing entries, say  $Y_{(1)}$ ; then select the variable, say  $Y_{(2)}$ , that has the most missing entries overlapping with  $Y_{(1)}$ ; then select the third variable that has the most overlapping missing entries with both  $Y_{(1)}$  and  $Y_{(2)}$ . Continue the process until there is no variable with overlapping missing entries with all previously selected variables. These missing entries composes the first monotone block  $B_1$ . The second monotone block  $B_2$  can be obtained by starting from the variable with the most missing entries excluding the missing

entries in  $B_1$  and then applying the same procedure. Repeat the same procedure until all the missing values have been allocated to a monotone block. There can be cases where this procedure does not give the partition with the largest possible first monotone block. However, our experience suggests that as long as the first few blocks contain most of the missing data (as in the AVR data), the results are very similar across different partitions.

We applied the above partitioning procedure separately for each treatment arm in the AVR data. Information about the missing data and the monotone blocks of the interim data is shown in Table 2. Even though the total number of monotone blocks can be large, the first monotone block usually dominates, covering a large proportion of missing data. On average, the first 3 monotone blocks include more than 85% of the missing values in each arm. In fact, most of the blocks after the fifth contain no more than 10 missing values in two variables.

Treatment Arm	Number of Subjects	Number of Missing values	Number of Blocks	Percent in 1st monotone block	Percent in First 3 monotone blocks
0	165	927	15	45	75
1	170	1372	13	74	84
2	168	1558	13	65	85
3	166	1383	15	79	90
4	167	1325	15	74	89
5	85	252	7	74	91
6	84	334	9	87	93

Table 2: Summary of missing data and monotone blocks by treatment arm in AVR data used for the interim analysis.

Once the monotone blocks are obtained, we impute the missing data within each arm using the sequential imputation procedure in Section 3.1. In the AVR data, we run 5 parallel Markov chain Monte Carlo (MCMC) chains and judge the convergence of the chains based on the criterion of potential scale reduction (Gelman and Rubin, 1992) for statistics on relevant immunogenicity and reactogenicity variables. Independent multiply imputed datasets are created by repeating the process with independent multiple initializations of the missing values.

### 3.3 Specification of conditional and predictive distributions

Modeling univariate conditional distributions instead of large joint distributions allows one to easily specify and fit models for different types of outcomes. We classify the outcome variables in the AVR data into the following five types: (1) binary; (2) categorical with either three (ordered or unordered) levels, or four unordered levels; (3) ordered categorical with at least four but at most eleven observed levels having a natural ordering; (4) continuous - defined here as an ordered outcome with more than eleven observed levels and with no extreme level having an observed frequency of at least 20%; (5) mixed continuous-binary - an ordered outcome with more than eleven observed levels and with one of the two extreme levels having an observed frequency of at least 20%.

For unit  $i$  we denote an outcome variable by  $Y_i$  (the subscript  $j$  indicating which variable is dropped here since we focus on one variable) and the set of predictors by the vector  $\mathbf{X}_i$ . In our conditional modeling approach, the outcome in one model can be used as a predictor in the model for another outcome, so that a variable can be denoted by  $Y_i$  in one situation but be included in the  $\mathbf{X}_i$  vector in another one. Here, we first describe the conditional models we propose using, and then discuss variable selection.

*Binary.* For binary outcomes, we propose a logistic regression model with a noninformative prior for the regression coefficients:

$$\text{logit}\{\Pr(Y_i = 1|\mathbf{X}_i, \boldsymbol{\beta})\} = \mathbf{X}_i' \boldsymbol{\beta}, \quad (1)$$

with noninformative prior on the coefficients  $\pi(\boldsymbol{\beta}) \propto 1$ . A draw from the posterior distribution of  $\boldsymbol{\beta}$  is approximated by the Sampling Importance Resampling (SIR) method (Rubin, 1987). This is done by (i) simulating a pool of “candidates” as a large number (e.g., 1000) of draws from a normal distribution centered at the maximum likelihood estimate (MLE) of  $\boldsymbol{\beta}$ , with covariance matrix set to the inverse of the observed Fisher information; (ii) calculating, for each

draw, the importance ratio of the actual posterior density to the approximate normal density, and (iii) sampling one of those draws with probability proportional to the importance ratios. In (i) the MLE of  $\beta$  is always included in the pool, in order to avoid the final draw to be extreme merely from not having any candidates with a high importance ratio in the pool. Based on this final draw of  $\beta$ , the missing values of  $Y$  are imputed independently across subjects according to the logistic model.

*Categorical with three levels or four or more unordered levels.* Categorical variables with  $K$  levels are modeled with  $K - 1$  sequential binary regressions and noninformative prior for the regression coefficients:

$$\text{logit} \{ \Pr(Y_i = k \mid Y_i \geq k, \mathbf{X}_i^{(k)}, \beta^k) \} = \{ \mathbf{X}_i^{(k)} \}' \beta^k, \quad (2)$$

where  $\mathbf{X}^{(k)}$  and  $\beta^k$  for  $k = 1, \dots, K - 1$  are the selected predictors and corresponding coefficients for the  $k$ th level regression, with noninformative prior  $\pi(\beta^1, \dots, \beta^{K-1}) \propto 1$ . Drawing from the posterior distribution of the parameters of each logistic regression is performed using the approach described for binary outcomes. A missing value for  $Y_i$  is then imputed by simulating sequentially the indicators for the events  $\{Y_i = 1\}, \dots, \{Y_i = K - 1\}$  until one indicator is drawn as 1. If all the indicators are drawn as 0, then  $Y_i$  is set to  $K$ .

*Categorical with four or more ordered levels.* Ordered categorical variables are treated as the continuous variables (see below). The imputed values are rounded to the nearest level observed in the data. This modeling here is preferred to a proportional odds or probit approach for reasons of computational stability.

*Continuous.* Continuous outcomes are modeled with normal linear regressions and noninformative priors for the parameters:

$$f(Y_i \mid \mathbf{X}_i, \beta, \sigma^2) = N(\mathbf{X}_i' \beta, \sigma^2), \quad \pi(\beta, \sigma^2) \propto 1/\sigma^2.$$

where  $N(a, b)$  is the normal density with mean  $a$  and variance  $b$ . The posterior distribution of  $\sigma^2$  is such that  $df(s^2/\sigma^2)$  has a  $\chi^2$  distribution with  $df$  degrees of freedom, where  $df$  are the residual degrees of freedom. The posterior distribution  $f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{X}, Y)$  is normal with mean equal to the least squares estimate of  $\boldsymbol{\beta}$  and covariance matrix equal to  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ , where  $\mathbf{X}$  is the design matrix.

Based on the draw of  $\boldsymbol{\beta}, \sigma^2$ , the missing values of  $Y$  are imputed independently across units according to the normal regression model. Any imputed value outside the range of the observed values is set to the nearest observed value.

*Mixed continuous-binary.* Our modeling strategy for variables of mixed type assume that the extreme value with at least 20% of observations is 0, and the remaining values are positive. We specify a logistic regression for  $Y_i^*$ , taking on value 1 if  $Y_i > 0$  and 0 otherwise, and a log-normal regression for the log of the positive values of  $Y$ . That is:

$$\text{logit}\{\Pr(Y_i^* = 1 \mid \mathbf{X}_i, \boldsymbol{\beta}^{(1)})\} = \{\mathbf{X}_i^{(1)}\}'\boldsymbol{\beta}^{(1)}, \quad (3)$$

$$f(\log(Y_i) \mid Y_i > 0, \mathbf{X}_i, \boldsymbol{\beta}^{(2)}, \sigma^2) = N(\{\mathbf{X}_i^{(2)}\}'\boldsymbol{\beta}^{(2)}, \sigma^2) \quad (4)$$

with uninformative prior  $\pi(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma^2) \propto 1/\sigma^2$ . A draw from the posterior distribution of the parameters is obtained for the two models separately, according to the procedures described for binary and continuous outcomes. A missing value of  $Y$  is then imputed by first imputing the indicator of the variable being 0 or positive and, if positive, imputing a value using the log-normal regression.

### 3.4 Modeling variables of special types

In addition to the above data types, there are several special type of outcomes.

1. *Variables with some portion being constant.* If a variable is constant in its observed values for all treatment arms, it will not be considered in the imputations. Otherwise, if a

variable is constant in the observed  $n_{obs}$  values out of  $n$  intended values in one arm, specific imputation strategies are used depending on the variable type: (1) if the variable is binary, the observed value, say 1, will be imputed with probability  $1 - 1/(2n_{obs})$ , which is the least extreme probability for which a 1 would be expected for all observed values; (2) if the variable is categorical with three levels or four unordered levels, the observed value, say 1, will be imputed with probability  $1 - 1/(2n_{obs})$ , and the remaining categories will be imputed with probabilities obtained by splitting  $1/(2n_{obs})$  equally across those categories; (3) if the variable is continuous or ordered categorical, the observed value will be imputed for all missing values; and (4) if the variable is mixed and the observed value is a 0, the missing values will be imputed as in (1) and the variable will be considered binary; if the observed value is different from 0, the missing values will be imputed as in (3).

2. *Variables with low variability.* If a variable is not constant but has very low variability, it can potentially make estimated model parameters reaching the boundary of the parameter space, if the model is not linear in the parameters. These may create computational problems. To address this difficulty, the current outcome-covariate pattern  $(Y, \mathbf{X})$  can be augmented by adding two terms of pseudodata  $(\mathbf{1}, \mathbf{X}_{aug})$  and  $(\mathbf{0}, \mathbf{X}_{aug})$ , respectively, where  $\mathbf{X}_{aug}$  is a matrix equal to  $\mathbf{X}$  and  $\mathbf{1}$  and  $\mathbf{0}$  are vectors of 1's and 0's of length equal to the rows of  $\mathbf{X}$ . The terms are assigned weights  $p_1$  and  $p_0$  so that  $p_1/p_0$  equals the observed marginal odds of  $Y = 1$  and  $p_1 + p_0 = 1/100$ . Upon checking, this adjustment stabilized estimation without essentially affecting predictions.

Variables with low variability are also the indicators of rare event, e.g. an immunogenicity value above a pre-specified threshold. The usual analysis is conducted by deriving those indicators based on the imputed continuous values. With a small number of imputations, this type of analysis may cause an underestimation of rare proportions. There is more than one solution to such problems if they arise. One is to increase the number of imputations for the continuous variables. Using different numbers of imputations per variable is called nested multiple

imputation (Shen, 2000; Rubin 2003), and requires using nested combining rules, but does not require modifying the complete-data analysis (see also Reiter and Raghunathan, 2007). Another approach consists, instead, to keep the number of imputations fixed for all variables, but change the complete-data analysis, in non-standard ways. Analysis should be conducted first on the continuous variable, and estimates of rare proportions derived from its estimated distribution, rather than creating the binary indicators of the rare events. Some discussion on this issue will be provided in Section 4.2

3. *Variables defined only on a subset of units.* Sometimes, there may be variables that are defined only for subsets of units. In the AVRPP some variables are defined only for women and include a menopause indicator and an indicator of use of oral contraception. When used as predictors, these variables are assigned the value 0 for men and the fully observed female indicator is always included among the predictors; this corresponds to include an interaction term of this variable with the female indicator. If these variables have missing values, then the model for their conditional distribution is fitted using only women with observed values of these variables.

4. *Avoiding imputation of inconsistent values for immunogenicity variables.* Immunogenicity measurements cannot dramatically increase their values (i.e., they can increase their values only within a certain range of natural variability) after a missed vaccine injection or after an injection not containing AVA (saline injection). Therefore, if these variables have missing values, imputed values should be consistent with the administered injections. In particular, we have to assure: a) consistency of imputed immunogenicity values with observed immunogenicity values; b) consistency among imputed immunogenicity values. In order to assure consistency of an imputed value of an immunogenicity variable at visit  $t$ ,  $W_t$ , say ELISA antibody concentration every time we sample from the posterior predictive distribution of  $W$ , we check consistency of the drawn value with the observed and imputed values of  $W_j$ ,  $j < t$ , on the same subject. If the value is larger than the last observed or imputed value plus  $\sigma$  when no AVA

injection was given, we sample again from the posterior predictive distribution until a consistent value is drawn. We set the value  $\sigma$  equal to the standard deviation of  $W_t$  in the placebo arm.

### 3.5 Specification of predictors

Since there are around 400 variables and only 1005 subjects divided into seven treatment groups (80-170 subjects in each treatment) in the AVRP interim data, we must constrain the number of predictors that enter the conditional model for each outcome. The predictor selection takes place before the imputation procedure. It is based on a preliminary imputation of all the missing values from their empirical marginal distributions. We allow the predictors for each outcome variable to differ across different arms and monotone patterns.

Demographic variables (age and sex) are fully observed and are always included in the model. For each outcome with missing values, the potential predictors are all the variables that are *more* observed (i.e., with less missing values) than the outcome in a particular monotone block. We use a stepwise choice procedure to choose the predictors for univariate conditional model as follows: (1) fit regression models of the outcome given each single potential predictor, age and sex, and (2) sort the predictors according to the corresponding Akaike's information criterion (AIC) (Akaike, 1974) and select the 20 predictors with the smallest AIC values. Finally, we check the *fittability* of the conditional model which simultaneously includes all the selected predictors on the complete cases. *Fittability* is defined as invertibility of the corresponding design matrix. It is checked sequentially on the subsets of predictors sorted by AIC in a backward fashion. If one predictor is not *fittable*, that means it does not contain enough information on the outcome. Therefore this predictor is dropped from the subset and the same checking goes on to the next selected predictor until the last one. The checking procedure is done separately within each treatment arm and each monotone block. Theoretically, the stepwise choice procedure may not be as desirable as the non-greedy variable selection methods

such as the Bayesian stochastic search variable selection (George and McCulloch, 1993), but it does provide computationally manageable solution with generally satisfying results in our application.

## 4 Evaluation

Much work has been devoted to propose and apply imputation methods, but remarkably little has been done in corresponding evaluations (Schafer *et al.*, 1996; Tang *et al.*, 2005; He and Zaslavsky, 2011). Comparing the imputed values to the observed values would be a most intuitive evaluation, but is neither generally possible nor valid. Here we propose a general template to evaluate MI procedures by simulation.

MI has been proved to be randomization valid if imputations are drawn from an (approximately) correct Bayesian model; but not all available MI procedures are appropriate in a specific study with real data. Their propriety depends on the posited response mechanism, on the (implicit or explicit) imputation models specified for the data, and on the complete-data analysis the ultimate user performs. We propose to simulate missing data in a fully observed subset of the dataset, to mimic the observed missing data patterns, and impute these created missing entries. We then propose comparing inferences based on the imputed dataset to inferences based on the original, complete dataset. A similar approach was taken in NHANES III imputation evaluation by Gelman and Rubin (1996) and Schafer *et al.* (1998), as well as in Raghunathan and Rubin (1998), Tang *et al.* (2005), Bernaards *et al.* (2007). A fully Bayesian evaluation approach based on posterior predictive checks was developed in He and Zaslavsky (2011).

### 4.1 Evaluation Proposal

Our evaluation proposal consists of the following steps:

1. Start with an actual dataset with missing data, and have a plan of a class of relevant analyses to be performed on relevant variables in the data.
2. Create a *population* from the complete cases on the relevant variables. Call this population *truth*.
3. Generate missing data in the *truth* that mimic the missing data patterns of those relevant variables in the actual data.
4. Impute the missing data by a chosen MI method, e.g., IMB.
5. Obtain diagnostics (e.g., bias, coverage) on selected relevant estimands by comparing the inference based on the imputed datasets to that based on the *truth*.
6. If problems detected, go back to modify the model/analysis as revealed by the diagnostics.

We now specify in more detail our proposal for performing these steps, and use the evaluation of the IMB method applied to the AVRPP trial interim data as an illustrative example.

#### **4.1.1 Choosing relevant variables and analysis**

In complex datasets, as the AVRPP data, there are many variables and a large number of analysis that can be conducted on those variables. It is therefore useful to limit the evaluation to the important analyses, the survey or the experiment was planned for.

For example, in the AVRPP trial, the antibody level comparison of primary interest is the full dose SQ versus the reduced dose IM, and the primary immunogenicity analysis submitted to the FDA was based on Intention-To-Treat (ITT) comparisons. Therefore the most relevant variables to the MI are the two immunogenicity variables, ELISA concentration and titers.

### 4.1.2 Constructing a Population

In order to construct a credible artificial population, the *truth*, on which to impose missing patterns, we propose to extract the subset of units with complete cases on the most relevant variables in the AVRVP trial interim data, identified in the previous step. Eventual missing values on other variables are left missing. This procedure should result with a population of cases that is more realistic than any probability model could have invented, because it is made of actual cases in the data. We then simulate a population that closely mimics the AVRVP trial interim data. In particular we created a population with three treatment groups: one receiving eight AVA doses SQ, one receiving eight AVA doses IM, and one combining units belonging to the three arms receiving either four or five or seven AVA doses intramuscularly. This population was divided into two subpopulations:

- SET A: Records complete on key immunogenicity variables (the *truth*)
- SET B: Records incomplete on at least one of the key immunogenic variables.

### 4.1.3 Creating Patterns of Nonresponse

Once a population has been created,  $K_1$  replicates of it must be generated by imposing missing patterns that mimic those observed in the real data. The amount of missingness imposed may vary, depending on the real case study. For example, in the AVRVP trial data, the imposed amount of missingness varies because the observed amount at the time of the interim analysis is assumed to underestimate the amount of missingness that will be observed in the full trial. In order to design a missing data process which is able to create real, or at least realistic, missing data patterns, we propose the following procedure.

We first identify the missing patterns of the  $t$  key variables in the dataset. A missing data pattern is a unique vector of  $t$  missing indicators for the  $t$  variables; there are at most  $2^t$  missing data patterns. We then simulate  $K_1$  copies of the *truth* with varying missing data patterns (and

proportion of missingness) using the following steps:

1. Count the numbers of units that belong to each pattern. Rank the missing patterns by their sizes in decreasing order (pattern 1 is the most prominent missing pattern).
2. For pattern  $k = 0, \dots, K$  (0 means fully observed), estimate logistic regression using a pseudocount prior distribution (see Appendix in Rubin (2004)) to model the probability of being in pattern  $k$  versus being in patterns  $k + 1, \dots, K$  given all fully observed covariates  $\mathbf{X}$ . Denote the estimated intercept and coefficients  $(\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k)$ , then the estimated probability of a unit  $i$  being in pattern  $p_{i,k}$  is  $\text{logit}^{-1}((\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k)(1, \mathbf{X}_i)^T)$ .
3. Choose  $\alpha_0$  that gives the overall proportion of missing data approximately equal to the observed one, or a different one depending on the aim of the analysis.
4. For each unit  $i$  in the *truth*, calculate its probability of being fully observed:

$$\text{logit}^{-1}((\alpha_0, \hat{\boldsymbol{\beta}}_0)(1, \mathbf{X}_i)^T)$$

Then randomly assign an indicator of being fully observed to the units based on these probabilities. Next, for each unit that is assigned to be missing (versus fully observed), calculate its probability of belonging to missing pattern 1:  $\text{logit}^{-1}((\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1)(1, \mathbf{X}_i)^T)$  and randomly assign it to pattern 1 (versus patterns 2 to  $K$ ). Continue this procedure through pattern  $K - 1$ . At last, each unit in the “truth” belongs to a missing pattern (including the pattern of fully observed).

5. Repeat Step 4  $K_1$  times.

In order to perform a random-response randomization evaluation (see Rubin, 1987a), from each of the  $K_1$  replicates, one should draw  $K_2$  subsamples of the *truth* to reflect sampling variability. In total,  $K_1 K_2$  samples are drawn from the “population”.

For the evaluation of IMB in the AVRP trial we generated  $K_1 = 200$  replicates of the *truth* by randomly imposing observed missing data patterns, but no samples were drawn from these replicates. The reason for this is that the estimands of interest are the before-deletion causal statistics in the finite population, so that evaluation is carried out only from a random-response perspective, ignoring sampling variability.

#### 4.1.4 Evaluating Imputation Results

Once samples with missing values have been obtained, each one should be imputed with the MI method whose performance we want to evaluate. The analysts performing this imputation step should be blinded to the *truth*, and so they should only be given the  $K_1 K_2$  samples with missing data. For each of them,  $m$  complete samples should be created by multiply imputing the missing values. For the AVRP, we used the IMB strategy to create  $m = 5$  imputed datasets for each of the 200 copies. Finally,  $(1-\alpha\%)$  confidence intervals for a set of statistics involving key variables are computed from the imputed datasets and then compared to the corresponding intervals based on the original, complete dataset.

Specifically, on each of the multiply imputed samples, compute  $(1-\alpha)\%$  confidence intervals (CIs), for each estimand of interest  $Q$ , according to the following:

$$\bar{Q}_m \pm t_v(\alpha/2)T_m^{1/2}$$

where  $\bar{Q}_m$  is the average of the  $m$  complete-data estimates,  $T_m = \bar{U}_m + (1 + m^{-1})B_m$  is the total variance,  $\bar{U}_m$  is the average of the  $m$  complete data variances  $B_m$  is the between variance,  $v$  is the degrees of freedom equal to  $(m - 1)(1 + r_m^{-1})^2$  or to the Barnard and Rubin (1999) adjusted formula. If evaluation is performed only from a random-response perspective, as is the case for our AVRP trial evaluation, the previous formulas should be used with  $\bar{U}_m = 0$ .

Coverage can be computed in at least two different ways. A frequentist coverage can be calculated by counting the number of times that confidence intervals cover the *truth*, i.e., the

value of  $Q$  in the *truth*. A Bayesian coverage can be calculated by approximating the posterior distribution of the parameter  $Q$  from the true complete data with a normal distribution; the area under this distribution for the  $(1-\alpha)\%$  confidence intervals is then computed and averaged over the samples. When estimands of interests are causal estimands, as is the case for the AVRP, we suggest to use the Bayesian coverage, which reflects uncertainty about causal quantities even in the before-deletion finite population.

Evaluations like this are rarely done but often give rather disappointing results for standard (non missing data) procedures; see, for example, Rao *et al.* (2003) in a survey context, where 95% nominal intervals can dip below 60% coverage.

## 4.2 Evaluation of Imputations in the AVRP Trial Data

In order to evaluate the performance of the IMB imputation procedure on the AVRP data, we selected the ELISA concentration and the ELISA titer measured at 8 weeks and 7 months as key immunogenicity variables. In the analysis phase, dummy variables were also created, which are equal to 1 if the original ELISA measure (observed or imputed) was greater than a fixed threshold and 0 elsewhere. For each of the three treatment groups (see section 4.1.2), 200 datasets with varying missing data patterns were created from the *truth*, according to the procedure described in section 4.1.3. Different amounts of missingness were assumed: 10%, 20%, 30%, 40%, and 50%. In particular 40 datasets were generated for each percentage of missing data. After detecting the missing data patterns characterizing the key variables in the real data, Bayesian logistic regression models were specified in order to obtain the posterior probability of being in a specific pattern versus being in less prominent patterns, given the fully observed variables: age group (< 30, 30 – 39, 40 – 49, 50+), sex (male, female premenopause, female postmenopause), enrollment site (Baylor, Emory, Mayo, Wrair, UAB), race (white, black, other), education (high school or less, some college, more than 3 years of college, graduate School), health status as compared to 5 years ago (excellent, very good, good/fair/poor). Then,

the simulated datasets were obtained by randomly assigning a specific missing data pattern to each unit of the *truth* according to the posterior probabilities from the Bayesian model. The intercept of the logistic regression model was tuned in order to obtain the desired overall proportion of missing data. For each of the 200 simulated replicates, 5 imputed datasets were generated using the IMB approach. We estimated the geometric mean of the ELISA concentration and the geometric mean of the ELISA titer at 8 weeks and 7 months from these datasets and the corresponding 95% and 80% confidence intervals according to the procedure described in section 4.1.4. In the evaluation, we also considered the proportion of individuals with ELISA concentration greater or equal to 20 micrograms/ml anti-PA IgG and the proportion of individuals with ELISA titer greater or equal to 1 : 200. Confidence intervals' frequentist coverage and Bayesian coverage of the true values of the parameters of interest were calculated.

	arm	titer		concentration	
		FC	BC	FC	BC
8 weeks (95%)	1	0.840	0.948	0.877	0.948
	2	0.927	0.952	0.902	0.946
	3	0.878	0.930	0.854	0.941
7 months (95%)	1	0.829	0.921	0.902	0.927
	2	0.902	0.944	0.890	0.948
	3	0.854	0.902	0.867	0.934
8 weeks (80%)	1	0.691	0.793	0.728	0.794
	2	0.707	0.799	0.780	0.791
	3	0.610	0.758	0.634	0.794
7 months (80%)	1	0.463	0.695	0.610	0.757
	2	0.756	0.776	0.622	0.782
	3	0.378	0.696	0.634	0.767

Table 3: Frequentist coverage (FC) and Bayesian coverage (BC) of the 95% and 80% confidence intervals for the geometric mean of ELISA titer and concentration.

The evaluation results are reported in Tables 3 and 4 . In all the analyses, the frequentist coverage was lower than the Bayesian coverage, which better reflects the underlying inferential state of knowledge contained in the “truth”. The coverage of the confidence intervals at

	arm	titer		concentration	
		FC	BC	FC	BC
8 weeks (95%)	1	0.948	0.953	0.948	0.952
	2	0.952	0.944	0.946	0.933
	3	0.930	0.954	0.941	0.951
7 months (95%)	1	0.921	0.865	0.927	0.957
	2	0.944	0.877	0.948	0.911
	3	0.902	0.937	0.934	0.961
8 weeks (80%)	1	0.793	0.803	0.794	0.800
	2	0.799	0.789	0.791	0.804
	3	0.758	0.808	0.794	0.803
7 months (80%)	1	0.695	0.665	0.757	0.810
	2	0.776	0.702	0.782	0.746
	3	0.696	0.769	0.767	0.828

Table 4: Frequentist coverage (FC) and Bayesian coverage (BC) of the 95% and 80% confidence intervals for the proportions of ELISA titer and concentration above the threshold.

8 weeks was close to the nominal one, while a certain degree of undercoverage was observed at 7 months, in particular for the ELISA titer. This undercoverage induces undercoverage in the threshold estimates based on these continuous variable, which is sometimes even more pronounced. In interpreting these evaluation results, one should consider whether the standard approach for calculating proportions based on imputed continuous variables is the correct approach. This stimulates the exploration of better complete data approach for estimating rare proportions.

## 5 Summary and Remarks

We have provided a general description on how to implement sequential multiple imputation methods to large scale complex data as the AVRIP data and proposed a Monte Carlo simulation based method to evaluate the imputation results. Motivated by the missing data problem arising from the AVRIP, we have developed a general approach, IMB, to handle non-monotone missing

data from data sets that have large number of variables of various types. The IMB breaks any arbitrary missing pattern into blocks of separate patterns each of which is monotone and can be handled with sequential modeling and imputation. By design, IMB simplifies the modeling process: for each variable having missing values in a monotone block, the set of possible predictors is reduced to the variables that are “more observed” than that variable. In some applications this reduction may be sufficient to handle the problem of fewer observations than predictors. For the AVR data, the number of variables was so large, compared to the number of observations, that an additional variable selection procedure was required. A computationally feasible variable selection algorithm was proposed in Section 3.5.

IBM should also reduce the incompatibility typical of sequential imputation strategies. To the best of our knowledge, no measure of incompatibility exists, so that we cannot quantify the reduction of the incompatibility compared to other sequential imputation strategies such as MICE. We can however judge the quality of our imputations. First, convergence (checked as described in Section 3.2) was achieved both when using the real data set, as well as when implementing the evaluation strategy using simulated data. Again, convergence is not a proof of achieved compatibility, but non-convergence problems are more likely to be observed with incompatible distributions. Second, the proposed evaluation procedure applied to the AVR data has shown acceptable coverages of our multiple imputations. However, it also highlighted some problems for binary variables, derived as indicators of imputed continuous variables being above a specified threshold. Only with evaluations of this kind can we be aware of this and similar problems. This has stimulated the exploration of better complete data approaches for estimating rare proportions that can be applied in the final analysis of the data.

## References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716-723.

2. Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association* **84**, 152-156.
3. Baccini, M., Cook, S., Frangakis, C.E., Li, F., Mealli, F., Rubin, D.B., and Zell E.Z. (2010). Multiple imputation in the Anthrax Vaccine Research Program. *Chance* **23(2)**, 16-23.
4. Barnard J., and Meng X.L. (1999) Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, **8**, 17-36.
5. Barnard, D., and D.B. Rubin. (1999) Small-sample degrees of freedom with multiple imputation. *Biometrika*, **86**, 948-955.
6. Bernaards C.A., Belin, T.R., and Schafer J.L. (2007) Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, **26(6)**, 1368-82.
7. Gelman, A. E., Carlin, J. B., Stern, H. S. and Rubin, D.B. (2004). Bayesian data analysis. Boca Raton: Chapman and Hall.
8. Gelman, A.E. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457-472.
9. Gelman A. E. and Rubin, D. B. (1996). Markov Chain Monte Carlo Methods in Biostatistics. *Statistical Methods in Medical Research* **5(4)**, 339-355.
10. George, E. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.
11. He, Y., Zaslavsky, A.M., Harrington, D.P., Catalano, P. and Landrum, M.B. (2010). Multiple imputation for a large-scale complex survey: a practical guide. *Statistical Method in Medical Research* **19(6)**, 653-670.
12. He, Y. and Zaslavsky, A.M. (2011). Diagnosing imputation models by applying target analyses to posterior replicates of complete data. *Statistics in Medicine*. Forthcoming.

13. Li, F., Yu, Y., and Rubin, D.B. (2012). Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines. *Duke University Department of Statistical Science Discussion Paper, 11-24*.
14. Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New Jersey: Wiley.
15. Marano, N., Plikaytis, B.D., Martin, S.M., Rose, C., Semenova, V.A., Martin, S.K., Freeman, A.E., Li, H., Mulligan, M.J., Parker, S.D., Babcock, J., Keitel, W., El Sahly, H., Poland, G.A., Jacobson, R.M., Keyserling, H.L., Soroka, S.D., Fox, S.P., Stamper, J.L., McNeil, M.M., Perkins, B.A., Messonnier, N., Quinn, C.P., for the Anthrax Vaccine Research Program Working Group. (2008) Effects of a Reduced Dose Schedule and Intramuscular Administration of Anthrax Vaccine Adsorbed on Immunogenicity and Safety at 7 Months A Randomized Trial, *Journal of the American Medical Association*, **300(13)**, 1532-1543.
16. Mealli F., and Rubin, D.B. (2002) Assumptions when Analyzing Randomized Experiments with Noncompliance and Missing Outcomes, *Health Services and Outcomes Research Methodology*, **3(4)**, 225–232.
17. Raghunathan, T. E. and Paulin, G. S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics of the American Statistical Association*, 1-10.
18. Raghunathan, T. E. and Rubin, D. B. (1998). Roles for Bayesian Techniques in Survey Sampling. *Proceedings of the Silver Jubilee Meeting of the Statistical Society of Canada*, 51-55.
19. Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85-95.
20. Raghunathan, T.E., Solenberger, P.W. and van Hoewyk, J. (2002). IVEware: Imputation and Variance Estimation Software User Guide. Survey Research Center, Institute for Social Research University of Michigan.

21. Rao, J. N. K., Jocelyn, W. and Hidioglou N. A. (2003). Confidence coverage properties for regression estimators in uniphase and two phase sampling. *Journal of Official Statistics* **19**, 17–30.
22. Reiter, J. P. and Raghunathan, T. E. (2007), The multiple adaptations of multiple imputation, *Journal of the American Statistical Association* **102**, 1462 - 1471.
23. Royston, P. (2004) Multiple imputation of missing values. *The Stata Journal* **3**, 227-241.
24. Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
25. Rubin, D.B. (1987a). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
26. Rubin, D.B. (1987b) The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association* **82**, 543–546.
27. Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–517.
28. Rubin, D.B. (2003). Nested Multiple Imputation of NMES via Partially Incompatible MCMC. *Statistica Neerlandica* **57(1)**, 3-18.
29. Rubin, D.B. (2004). Multiple imputation for nonresponse in surveys (4th Ed) . New York: Wiley.
30. SAS Institute Inc. (2008). *SAS 9.2 Reference*. Cary, NC: SAS Institute Inc.
31. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
32. Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., and Rubin, D.B. (1998). The NHANES III Multiple Imputation Project. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 28-37.
33. Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924-933.

34. Shen, Z. (2000). Nested Multiple Imputation. Unpublished Ph.D. dissertation. Harvard University, Department of Statistics.
35. StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.
36. Statistical Solutions Ltd. (2001). SOLAS for Missing Data Analysis. Statistical Solutions, Cork, Ireland.
37. Tang, L., Song, J., Belin, T.R., and Unuetzer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* **24**, 2111–2128.
38. Taylor, J. M. G., Cooper, K. L., Wei, J. T., Sarma, R. V., Raghunathan, T. E. and Heeringa, S. G. (2002) Use of Multiple Imputation to Correct for Nonresponse Bias in a Survey of Urologic Symptoms among African-American Men. *American Journal of Epidemiology* **156**, 774-782.
39. van Buuren, S. and Oudshoorn C.G.M. (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/PG 99.054.
40. van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C.G.M. and Rubin D. B. (2006). Fully Conditional Specifications in Multivariate Imputation. *Journal of Computational and Graphical Statistics* **76(12)**, 1049-1064.