ABSTRACT. Multiple imputation (MI) has become a standard statistical technique for dealing with missing values. The CDC Anthrax Vaccine Research Program (AVRP) dataset created new challenges for MI due to the large number of variables of different types and the limited sample size. A common method for imputing missing data in such complex studies is to specify, for each of \( J \) variables with missing values, a univariate conditional distribution given all other variables, and then to draw imputations by iterating over the \( J \) conditional distributions. Such fully-conditional imputation strategies have the theoretical drawback that the conditional distributions may be incompatible. When the missingness pattern is monotone, a theoretically valid approach is to specify, for each variable with missing values, a conditional distribution given the variables with fewer or the same number of missing values and sequentially draw from these distributions. In this article, we propose the “multiple imputation by ordered monotone blocks” approach, which combines these two basic approaches by decomposing any missingness pattern into a collection of smaller “constructed” monotone missingness patterns, and
iterating. We apply this strategy to impute the missing data in the AVRP interim data. Supplemental materials, including all source code and a synthetic example dataset, are available online.

**KEY WORDS:** Bayesian, conditional distribution, imputation, incompatibility, missing data, monotone blocks, PIGS.
1 Introduction

Multiple imputation (MI) (Rubin, 1978, 1987a-2004, 1996; Schafer, 1997; Little and Rubin, 2002; van Buuren, 2012; Carpenter and Kenward, 2013) 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), the MI macro in MLwiN (Rasbash et al., 2009), the ICE software in STATA (Royston, 2004; Royston and White, 2011), SOLAS (Statistical Solutions Ltd., 2001), IVEware in SAS (Raghunathan et al., 2002), the mice package (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2012) and the pan package in R (Schafer, 2012). See http://www.multiple-imputation.com/ for an overview. Also see Carpenter et al. (2011) for REALCOM-IMPUTE. MI ideally involves specifying a joint distribution for all variables in a dataset, supplemented by a prior distribution for the parameters of this joint distribution. Multiple imputations of the missing values are then randomly drawn 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. MI has also been applied in measurement error problems and for combining data sources (Yucel and Zaslavsky, 2005), and data confidentiality (e.g., Reiter, 2005; Drechsler, 2011). However, MI for missing data in large-scale studies with complex designs remains challenging, as we describe below using the the Centers for Disease Control and Prevention (CDC) Anthrax Vaccine Research Program (AVRP) (Marano et al., 2008) as an example.

First, the AVRP collects a large number of variables of different types, such as continuous, binary, categorical and semi-continuous. Models for MI are often based on the multivariate
normal or general location model (Schafer, 1997), neither of which is applicable to the AVRP data, nor to many other complex datasets. A commonly adopted strategy for handling missing data in such complex studies is “multiple imputation by chained equations (MICE)” (e.g., Raghunathan et al., 2001; van Buuren, 2012), where one specifies, for each variable with missing values, a univariate conditional distribution given all other variables, and then imputes the missing values in a pre-specified order (e.g., from the variable with the most to the least observed data). Model fitting and imputation steps are performed iteratively, until some type of convergence is achieved. The univariate distributions usually take the form of regression models, which are convenient and can accurately reflect different data types. MICE is implemented in several software packages, and has been successfully applied to a number of real studies (e.g., Barnard and Meng, 1999; Ezzati-Rice et al., 1995; Kennickel, 1991; Taylor et al., 2002; Stuart, et al., 2009; He et al., 2010).

A theoretical drawback of MICE is the potential incompatibility among the univariate conditional distributions (Arnold and Press, 1989); that is, the set of the conditional distributions may not correspond to any joint distribution. The incompatibility problem has been widely acknowledged (e.g., van Buuren, 2012), but has not been intensively studied. The limited simulation work that is available suggests that the problem is probably not serious in practice (Drechsler and Rassler, 2008; van Buuren et al., 2006). However, several simple examples in Li et al. (2012) reveal that, when some of the conditional models are misspecified, the imputation error in MICE procedures can (not necessarily will) propagate rapidly over the iterations and lead to meaningless imputations, which suggests that model specifications can be crucial when using MICE strategies. However, in current practice for large datasets with high dimensional variables, the conditional distributions are usually specified in an automatic fashion and model checking is seldom performed.

The model specification issue becomes even more challenging when the sample size is small compared to the number of variables. For example, in the AVRP clinical trial, only 1560
participants were to be randomized into seven treatment groups and over 400 variables were planned to be recorded for each participant. Also, 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. We have attempted to use the available MICE software IVEware, the R package mice (v2.9), as well as the MI marco in MLwiN, to impute the AVRP data and, at least at the time we tried, all failed to return meaningful imputations.

Motivated by these challenges, we have two main goals here. First, we propose the “imputation by ordered monotone blocks (IMB)” strategy, to mitigate the incompatibility problem by combining the virtues of the practically feasible MICE strategy and the theoretically-valid sequential imputation strategy for monotone missing data. The key idea is to decompose any missing data pattern into a collection of “constructed” monotone patterns, “constructed” by treating some missing values as observed using their imputed values, then use the sequential strategy to impute the missing values within each monotone block, and iterating. Second, we elaborate the model specification and variable selection procedures for the various types of variables.

Section 2 overviews the missing data problem in the AVRP data. Section 3 presents the IMB strategy as well as on how to obtain the collection of monotone blocks. Section 4 provides details on the IMB algorithm and on the partitioning of missing data into monotone blocks. Details about specifying the conditional models for the various types of variables are provided in Section 5, whereas Section 6 suggests a simple procedure for the selection of predictors in the conditional models. Section 7 concludes.
2 Motivating Application

Inhalation Anthrax is a highly lethal condition in humans and animals. Because of concerns about bioterrorism, US military personnel are now routinely vaccinated against anthrax prior to active service in places where such attacks are considered a threat. In 1970 FDA licensed anthrax vaccine (Anthrax Vaccine Adsorbed - AVA: BioThrax, Emergent BioSolutions, Inc., Rockville, MD) based on a randomized human clinical trial demonstrating protection of mill workers in the 1950s. The CDC AVRP clinical trial, initiated in 2003, 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 AVA delivered by different routes of administration (subcutaneous-SQ versus intramuscular-IM), and different dosing regimens, with sterile saline used as the placebo.

At the time of the interim analysis, whose data we consider here, the AVRP had enrolled 1005 volunteers, healthy adult men and women, 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 six doses followed by two annual boosters); another two groups received saline IM or 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 was omitted from the licensed dosing regimen. There were a total of 25 scheduled visits over a period of 43 months, during which all participants received an injection of vaccine or placebo (eight injections total), had a blood sample drawn (16 total), and had an in-clinic examination for adverse events (22 total). Total antibody (IgG) levels were measured using a standardized and validated Enzyme-Linked ImmunoSorbent Assay (ELISA); the primary immunogenicity endpoints were geometric mean concentration (GMC), geometric mean titer (GMT), proportion of responses with a 4-fold rise in titer (%4 XR), and proportion above a defined threshold value at month 2, 7, 43. More details on the AVRP can be found in Marano
et al. (2008) and Baccini et al. (2010).

The AVRP trial is important because it was designed to provide the basis for reducing negative side effects by changing from SQ to IM and reducing the number of vaccine doses. In fact, the interim AVRP results based on available-case data have already led to changing from SQ to IM in 2008. The final dataset will be used to consider additional reductions in the vaccine priming and booster series. However, the length and the complexity of the study design, which includes more than 2000 variables at the end of the study, pose enormous challenges in statistical analysis, due to the large amount of missing data expected due to missed visits and missing responses. Any comparison that adheres to the intent-to-treat (ITT) principle requires handling of the missing data (Mealli and Rubin, 2002). The simplest complete-data analysis that drops all subjects with any missing data is not generally appropriate. Consider the interim data for example: even though the overall missing data 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. Commonly used ad hoc strategies, including “hotdeck” and “last observation carried forward” imputation, lack theoretical justification and can lead to severe bias. We adopt the theoretically-justified, but complicated, MI approach to handle the missing data.

3 Terminology and Notation

3.1 Monotone Pattern of Missing Data

Suppose there are $N$ units indexed by $i$, and $J$ variables, $V_j$, indexed by $j$, $V = (V_1, \ldots, V_J)$, where the value of $V_j$ for unit $i$ is $v_{ij}$. Also denote the response indicators by $R = \{R_{ij}\}$, where $R_{ij} = 1$ if $v_{ij}$ is observed and $R_{ij} = 0$ if it is missing; let $M_j = \{i : R_{ij} = 0\}$ be the set of units with missing entries for variable $j$, and let $M = \bigcup_{j=1}^{J} M_j$ be the set of all missing entries. For notational simplicity and without loss of generality, order the variables by the number of
observed values: $\sum_i R_{i1} \geq \sum_i R_{i2} \geq \ldots \geq \sum_i R_{iJ}$; and order the units by the number of observed values $\sum_j R_{1j} \geq \sum_j R_{2j} \geq \ldots \geq \sum_j R_{Nj}$. We assume that the missing data are missing a random (MAR) as defined in Rubin (1976); also see Little and Rubin (2002). The missing data pattern is monotone (Rubin, 1974a; Little and Rubin, 2002) if $V_{ij}$ is observed whenever $V_{i,j+1}$ is observed; equivalently, $R_{ij} = 0$ implies that $R_{i,j+1} = \ldots = R_{iJ} = 0$. Thus a monotone pattern creates an irregular staircase separating observed (above or to the left of the staircase) and missing values (below or to the right of the staircase).

3.2 Distributional Assumptions and a Monotone-Distinct Structure on Data

We assume that the distribution of $V$ is row exchangeable, so that with essentially no loss of generality we can model the rows of $V$ as independent and identically distributed (iid) given some global vector parameter $\theta$, lying in a parameter space $\Omega$. Now consider the factorization of the distribution of $V$ into $J$ factors, beginning with the marginal distribution of $V_1$ given its parameter, $\theta_1$, then the conditional distribution of $V_2$ given $V_1$ and its parameter, $\theta_2$, and so forth, and ending with the conditional distribution of $V_J$ given $(V_1, V_2, \ldots, V_{J-1})$ and its parameter $\theta_{J-1(J-1)}$.

Now suppose that the $J$ parameters are distinct (Rubin, 1976) in the sense that they lie in $J$ disjoint parameter spaces, so that, in an obvious notation: $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_{J-1} \times \Omega_J$; furthermore, by distinct we mean that any prior distribution for $\theta$ factorizes into $J$ independent factors corresponding to the factored parameter space. When the missingness pattern for $V$ is monotone and the corresponding parameters are distinct in this sense, then the structure on $V$ is called “monotone-distinct” (Rubin, 1987, Section 5.4).
3.3 Sequential Imputation with a Monotone-Distinct Structure

With such a structure, it is nearly trivial to impute stochastically, or multiply-impute, the missing values, \( V_{\text{mis}} = \{ v_{ij} | R_{ij} = 0 \} \), from \( V_{\text{obs}} = \{ v_{ij} | R_{ij} = 1 \} \) in \( J \) steps. Begin with the most observed variable, \( V_1 \); compute the posterior distribution of \( \theta_1 \) using the units with \( V_1 \) observed, and draw a value from it, say \( \theta_1^* \); then draw a value of the missing part of \( V_1 \), say \( V_{1,\text{mis}}^* \), given \( \theta_1 = \theta_1^* \), and fill it in for \( V_{1,\text{mis}} \). Then move on to \( V_2 \); compute the posterior distribution of \( \theta_{2:1} \) using the units with \( V_2 \) observed, and draw a value from it, say \( \theta_{2:1}^* \), and use this value to draw a value of \( V_{2,\text{mis}} \), say \( V_{2,\text{mis}}^* \), using \( V_{1,\text{obs}} \) and imputed values of \( V_{1,\text{mis}} \). Continue in this way until the \( J^{\text{th}} \) step, which computes the posterior distribution of \( \theta_{J:1...J-1} \), draws a value from it, say \( \theta_{J:1...J-1}^* \), and then uses this to draw a value of \( V_{J,\text{mis}} \), say \( V_{J,\text{mis}}^* \), from its posterior predictive distribution given \( \theta_{J:1...J-1} = \theta_{J:1...J-1}^* \), \( V_{1,\text{obs}}, V_{2,\text{obs}}, \ldots, V_{J-1,\text{obs}} \), and the filled in values of \( V_{1,\text{mis}}, V_{2,\text{mis}}, \ldots, V_{J-1,\text{mis}} \). The collection \( V_{1,\text{mis}}^*, V_{2,\text{mis}}^*, \ldots, V_{J,\text{mis}}^* \) comprises one draw from the posterior predictive distribution of \( V_{\text{mis}} \). Repeating the entire \( J \)-step process independently \( M \) times creates \( M \) multiple imputations.

3.4 An Ordered Partition of the Missing Data

Let \( \{ B_1, \ldots, B_K \} \) be an ordered partition of the set of missing entries, \( M \) (i.e., by the definition of a partition, \( \bigcup_{k=1}^K B_k = M \) and \( B_k \cap B_{k'} = \emptyset \) for \( k, k' = 1, \ldots, K \)), ordered such that \( \text{Card}(B_1) \geq \text{Card}(B_2) \geq \ldots \geq \text{Card}(B_K) \), where \( \text{Card} \) = cardinality, i.e., the number of missing values. Block \( B_K \) is called a monotone block if the variables with entries in \( B_k \) create a monotone pattern of missing data (with the entries not in \( B_k \), for variables represented in \( B_k \), considered observed). If all \( B_k (k = 1, \ldots, K) \) are monotone blocks, then \( \{ B_k, k = 1, \ldots, K \} \) is called an “ordered partition of the missing data into monotone blocks”.

Notice that when considering monotone block, \( B_k \), if all missing entries in \( M \) that are not included in \( B_k \) are considered observed, the missing entries in \( B_k \) can be imputed using sequential imputation as in Section 3.3 assuming the parameters in the corresponding univariate
regressions are distinct. This idea forms the basis for our algorithm.

4 The Imputation by Monotone Blocks (IMB) Algorithm

4.1 Basic Iterations

Starting Values and Block \( B_1 \)

Start with block \( B_1 \), and impute the missing values in \( B_2, \ldots, B_K \) by drawing at random from the marginal distributions of the corresponding variables. Then, \( B_1 \), combined with the observed values in \( V \) and with the imputed values outside \( B_1 \), creates a monotone pattern of missing data. Specify a sequence of univariate conditional distributions with distinct parameters, and use the sequential imputation method of Section 3.3 to create imputed values for all missing values in \( B_1 \).

Block \( B_2 \)

Now consider \( B_2 \), accepting the starting imputed values for \( B_3, \ldots, B_K \) and the just imputed values for \( B_1 \), but discarding the previously imputed values for \( B_2 \). Doing so creates a monotone missing data pattern with values in \( B_2 \) missing, but every other entry in \( V \) considered observed. Specify a sequence of conditional univariate distributions to impute \( B_2 \) assuming a monotone-distinct structure.

Block \( B_3 \), etc.

Move on to \( B_3 \), treating previously imputed values in \( B_1, B_2, B_4, \ldots, B_K \) as observed, but treating the entries in \( B_3 \) as missing, and proceed as before. Continue to impute through \( B_K \) in the same manner. Then return to \( B_1 \) and iterate. Continue to iterate until some approximate convergence criterion is satisfied, as discussed in Section 4.3.
4.2 Small Example

Because the missing data pattern in each $B_k$ is monotone, if there were only one monotone block, this algorithm would require a single iteration to converge to a draw from the posterior predictive distribution of $V_{mis}$. However, because there are multiple monotone blocks, more than one iteration over the blocks is needed for final imputations that are stable.

An example of a partition into monotone blocks and the IMB algorithm is given in Figure 1, where $N = 10$, $J = 6$. Panel (a) shows the matrix of response indicators of the sorted variables and units; panel (b) labels the three monotone blocks $B_1, B_2, B_3$ in the order of their total number of missing entries. After the initial imputations, an IMB algorithm will first impute the missing data in $B_1$ in the order of $(V_2|V_1, V_3, V_4), (V_5|V_2, V_1, V_3, V_4), (V_6|V_5, V_2, V_1, V_3, V_4)$, treating all the data outside $B_1$ (both observed and most-recently imputed values) as observed; then IMB will impute $B_2$ in the order of $(V_3|V_1, V_2, V_5, V_6), (V_4|V_3, V_1, V_2, V_5, V_6)$, treating all the data outside $B_2$ as observed; and finally, IMB will impute $B_3$ in the order of $(V_3|V_1, V_2, V_4, V_5, V_6)$.

4.3 Convergence Issues

As with many iterative MI strategies, there appears to be no obvious method for assessing whether an IMB algorithm has converged. A common diagnostic tool is to plot one or more parameters against the iteration number and assess convergence by how different the variance between different sequences is relative to the variance within each individual sequence, similar to the Gelman-Rubin statistic (Gelman and Rubin, 1992) used in Markov chain Monte Carlo (MCMC) diagnostics. In the AVRP data, five parallel MCMC chains were run with independent multiple initializations of the missing values, and we judged the convergence of the chains based on G-R for relevant statistics of completed data.

For a particular dataset, there may be many ways to partition the missing data into monotone blocks, and consequently many possible IMB strategies. One of the simplest partitions is to take the missing values of each variable as a monotone block, i.e., $K = J$ and $B_k = M_k, k =$
Figure 1: An example of IMB. Matrix of response indicators $R = \{R_{ij}\}$ in (a) and monotone blocks in (b).

1, \ldots, J. This is exactly the strategy used by MICE. However, as mentioned earlier, general IMB algorithms define potentially incompatible Gibbs samplers (PIGS), because the univariate conditional models across the monotone blocks may not correspond to any joint distribution.

### 4.4 Modeling Incompatibility and Algorithmic Incompatibility

Here we make a distinction between modeling incompatibility (under MAR) and algorithmic incompatibility. Modeling incompatibility refers to the situation where there is no joint distribution corresponding to the conditional models, whereas algorithmic incompatibility means the incompatible conditional distributions that drive the resulting incompatible Gibbs sampler will not uniquely converge. In general, model incompatibility is a necessary but not sufficient condition for algorithmic incompatibility. To see this difference, consider the following simple...
example. Let \( J = 2 \) and \( N = 3 \), where \( v_{11} \) and \( v_{22} \) are missing but all other entries of the dataset are observed. Partition the missing data into two monotone blocks, and specify the conditional models \((V_1|V_2) \sim N(v_{21}^2, 1)\) and \((V_2|V_1) \sim N(v_{12}^2, 1)\), where \( N(a, b) \) is the normal density with mean \( a \) and variance \( b \). Clearly, the conditional models are incompatible, but due to the particular missing data pattern, the resulting IMB algorithm is algorithmically compatible: At each iteration, first impute \( v_{11} \) by drawing from \( v_{11} \sim N(v_{21}^2, 1) \), and then impute \( v_{22} \) by drawing from \( v_{22} \sim N(v_{21}^2, 1) \). This strategy converges immediately. If algorithmic compatibility is approximately achieved, and each conditional model for the monotone block fits the data well, then the imputation may be judged as being reasonable (van Buuren et al., 2006).

IMB is designed to reduce algorithmic incompatibility by sequentially maximizing the number of missing entries in each monotone block. This particular strategy has at least two potential advantages over MICE. First, MICE requires cycling through the univariate conditional models for all the variables with missing values in each iteration. If one conditional model, say \((V_1|V_2, \ldots, V_J)\), is misspecified, then the errors in imputing missing values in \( V_1 \) may propagate into the imputation of the next variable, say \( V_2 \), and then \( V_3 \), and so on. This is particularly problematic in the presence of extrapolation. In some situations, such as Example 3 in Li et al. (2012), the error due to one misspecified model propagates rapidly and leads to meaningless imputations after only a few iterations. In contrast, using IMB, errors when imputing one variable within one monotone block do not necessarily affect the next monotone block. Therefore, error propagation over iterations may be avoided with IMB. Second, by construction, IMB involves specifying conditional models with at most the number of predictors as MICE and thus potentially reduces the possibility of model mis-specification. These conclusions are intuitively reasonable but seem difficult to formalize into theorems.
4.5 A Simple Method for Partitioning Missing Data into Approximately-Ordered Monotone Blocks

Intuitively, the more missing entries in each monotone block, the closer IMB is to the sequential imputation for monotone missing data, so that the degree of incompatibility is less. The idea of defining blocks to include as many missing entries as possible is a natural extension of the imputation by the major monotone pattern strategy proposed by Rubin (2003), which exploits a single major monotone block. Here we propose a sequential procedure to obtain (approximately) the largest blocks, illustrated by the example in Figure 1.

1. Identify the variable that has the largest number of missing entries. For example, $V_6$ in Figure 1. Let all the missing entries in this variable belong to the first monotone block, $B_1$.

2. Select the second variable that has the largest number of units with missing entries in both it and the variable selected in Step 1. For example, this is $V_5$ in Figure 1. Let the missing entries that this variable has in common with the variable selected in Step 1 also belong to $B_1$.

3. Select the third variable that has the largest number of units with missing entries in both it and in the two variables selected in Steps 1 and 2. For example, this is $V_2$ in Figure 1. Let the missing entries that this variable has in common with the variables selected in Steps 1 and 2 also belong to $B_1$. Continue this process until no variable can be found with units that have missing entries both in it and in all the variables selected in previous steps. This completes the identification of the missing entries belonging to the first monotone block, $B_1$.

4. Treat all the missing entries in $B_1$ as observed, and obtain the second monotone block, $B_2$, among the remaining missing entries by applying Steps 1-3.
5. Repeat Step 4 until all the missing entries have been allocated to a monotone block.

There can be cases where this procedure does not give the correct partition with the ordering of monotone blocks defined in Section 3.3. However, our experience with the AVRP data suggests that as long as the first few blocks contain most of the missing data, the results are similar when using different partitions into monotone blocks.

The result of applying this partitioning procedure independently to each treatment arm in the AVRP interim data is shown in Table 1. Even though the total number of monotone blocks can be large, the first three monotone blocks contain a large proportion of the missing data. In the AVRP, the first three 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 1% of the missing values.

<table>
<thead>
<tr>
<th>treatment arm</th>
<th>number of subjects</th>
<th>number of missing values</th>
<th>number of blocks</th>
<th>percent in first monotone block</th>
<th>percent in first three monotone blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>165</td>
<td>927</td>
<td>15</td>
<td>45</td>
<td>75</td>
</tr>
<tr>
<td>1</td>
<td>170</td>
<td>1372</td>
<td>13</td>
<td>74</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>168</td>
<td>1558</td>
<td>13</td>
<td>65</td>
<td>85</td>
</tr>
<tr>
<td>3</td>
<td>166</td>
<td>1383</td>
<td>15</td>
<td>79</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>167</td>
<td>1325</td>
<td>15</td>
<td>74</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>85</td>
<td>252</td>
<td>7</td>
<td>74</td>
<td>91</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
<td>334</td>
<td>9</td>
<td>87</td>
<td>93</td>
</tr>
</tbody>
</table>

Table 1: Summary of missing data and monotone blocks by treatment arm in AVRP data used for the interim analysis.


5 Specification of Conditional and Predictive Distributions

Modeling univariate conditional distributions instead of multivariate joint distributions allows the easy specification and fitting of models for different types of outcomes. We classify the outcome variables in the AVRP data into nine types described in this section.

For unit $i$ we denote an outcome variable by $Y_i$ and the set of predictors by $X_i$ (both $Y$ and $X$ are subsets of $V$). In the 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 model but be included in the $X_i$ vector in another one. We describe the conditional models, and then discuss variable selection in Section 6.

5.1 Binary Variables

For binary outcomes, we use a logistic regression model:

$$\logit \{ \Pr(Y_i = 1 | X_i, \beta) \} = X_i' \beta,$$

with prior distribution for the regression coefficients, $\pi(\beta) \propto 1$. The posterior distribution for $\beta$ is not in closed-form and is approximated with Sampling Importance Resampling (SIR) (Rubin, 1987b), done by (1) simulating a large number of “candidates” from a normal distribution centered at the maximum likelihood estimate (MLE) of $\beta$, with covariance matrix set to the inverse of the observed Fisher information; (2) calculating, for each draw, the importance ratio of the actual posterior density to the approximate normal density; and (3) sampling one of those draws with probability proportional to the importance ratios. In Step (1), it is possible to have an extreme draw of $\beta$ when all the candidates in the pool have low importance ratios. To avoid this, the MLE of $\beta$, which leads to the largest importance ratio, is always included in the pool of draws. Based on the final draw of $\beta$, the missing values of $Y$ are imputed independently across subjects according to the logistic model. It has subsequently come to our attention that there may exist an improved method for drawing from posterior distributions for
logistic regressions (e.g., Polson et al., 2013).

5.2 Ordinal Variables with Three Levels, or Unordered Categorical Variables

Ordinal (or ordered categorical) variables with three levels are modeled with two sequential logistic regressions:

\[
\text{logit} \left\{ \Pr(Y_i = m \mid Y_i \geq m, \mathbf{X}^{(m)}_i, \beta^{(m)}) \right\} = \mathbf{X}^{(m)}_i \beta^{(m)},
\]

where \( \mathbf{X}^{(m)} \) and \( \beta^{(m)} \) for \( m = 1, 2 \) are the selected predictors and corresponding coefficients for the \( m \)th level regression, with prior distribution \( \pi(\beta^{(1)}, \beta^{(2)}) \propto 1 \). Drawing from the posterior distribution of the parameters of each logistic regression is performed using the SIR approach described in Section 5.1. A missing value for \( Y_i \) is then imputed by simulating sequentially the indicators for the events \( \{Y_i = 1\} \) and \( \{Y_i = 2\} \) until one indicator is drawn to be 1. If both indicators are drawn to be 0, then \( Y_i \) is set to 3.

A similar procedure applies to unordered categorical variables with \( K \) levels (\( K \) is an integer larger than 2), where \( K - 1 \) sequential logistic models similar to those in (1) are used. Here the category indices \( k \) are arbitrarily assigned and the order, when setting up the sequential models, does not affect the imputation.

5.3 Ordinal Variables with at Least Four but at Most Eleven Levels

Ordinal variables with between four and eleven levels are treated as continuous variables, except that imputed values are rounded to the nearest level observed in the data. This approach is preferred over the proportional odds model mainly for computational reasons.
5.4 Continuous Variables

Continuous outcomes are modelled using normal linear regressions, after an appropriate transformation of outcomes to be approximately conditionally normal:

\[ Y_i | X_i, \beta, \sigma^2 \sim N(X'_i \beta, \sigma^2), \quad \text{with } \pi(\beta, \sigma^2) \propto 1/\sigma^2. \]

The posterior distribution of \( \sigma^2 \) is such that \( (df \cdot s^2/\sigma^2) \) has a \( \chi^2 \) distribution with \( df \) degrees of freedom. The posterior distribution of \( \beta \) given \( \sigma^2 \) is normal with mean equal to the least squares estimate of \( \beta \) and covariance matrix equal to \( (\sum X_iX'_i)^{-1}\sigma^2 \). Based on the draws of \( \beta \) and \( \sigma^2 \), the missing values of \( Y \) are imputed independently across units according to the normal regression model. Any imputed value outside the range of all the observed values is set equal to the nearest observed value.

5.5 Semi-continuous Variables

For semi-continuous variables, we specify a logistic regression for the indicator \( 1(Y_i > 0) \), and a normal regression for the positive values of \( Y_i \), also after an appropriate transformation of positive values to be approximately conditionally normal:

\[
\text{logit}\{\Pr(1(Y_i > 0) = 1 | X_i^{(1)}, \beta^{(1)})\} = \{X_i^{(1)}\}'\beta^{(1)}, \\
Y_i | Y_i > 0, X_i^{(2)}, \beta^{(2)}, \sigma^2 \sim N(\{X_i^{(2)}\}'\beta^{(2)}, \sigma^2),
\]

with \( \pi(\beta^{(1)}, \beta^{(2)}, \sigma^2) \propto 1/\sigma^2 \), and \( X^{(1)}, X^{(2)} \) being the covariates in the logistic and the normal regression, respectively. A draw from the posterior distribution of the parameters is obtained for the two models independently, 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, by imputing a value using the normal regression.
5.6 Variables with Substantial Portions Constant

If all observed values of a variable are the same for all treatment arms, it will be excluded from the imputations. Otherwise, if a variable is constant, with value \( c \), among the observed \( n_{\text{obs}} \) values out of \( n \) possible values in one treatment arm, specific imputation strategies are used depending on the variable type. (1) If the variable is binary, the missing entries are imputed by the value \( c \), say 1, with probability \( 1 - 1/(2n_{\text{obs}}) \), which is the least extreme probability for which a 1 would be expected for all observed values \(^2\). (2) If the variable is ordinal or unordered categorical with \( M \) levels, then the missing entries will be imputed by the value \( c \) with probability \( 1 - 1/(2n_{\text{obs}}) \), and by one of the remaining categories each with probability equal to \( 1/(2(K-1)n_{\text{obs}}) \). (3) If the variable is continuous, all the missing entries are imputed by \( c \). (4) If the variable is mixed and the observed value \( c \) is 0, the variable is treated as binary and the missing values are imputed as in (1), and if \( c \) is different from 0, the missing values are imputed as in (3).

5.7 Variables with Low Variability

If a variable is not constant but has very low variability, it can create model parameters estimated to be on the boundary of their parameter spaces, which can create computational problems. To address this issue, the outcome-covariate matrix \((Y, X)\) is augmented by adding two weighted pseudo-data matrices: \((1, X)\) and \((0, X)\), where \(1\) and \(0\) are vectors of 1’s and 0’s of dimensions equal to the rows of \( X \). These two matrices are respectively assigned weights \( p_1 \) and \( p_0 \) so that \( p_1/p_0 \) equals the observed marginal odds of \( Y = 1 \) and \( p_1 + p_0 = 1/100 \). That is, the pseudo data preserve the marginal odds of \( Y \) and are weighted at most one percent of the original data. In our application, we found that this adjustment stabilizes estimation essentially.

---

\(^2\)Let \( x_{\text{obs}} \) be the number of 1’s observed in the \( n_{\text{obs}} \) observations, then \( E(x_{\text{obs}}) = n_{\text{obs}} \cdot p \), where \( p \) is the probability of \( x = 1 \). When \( x_{\text{obs}} = n_{\text{obs}} \), the least extreme value of \( p \), such that \( n_{\text{obs}} \cdot p \) is rounded to the closest integer \( n_{\text{obs}} \), satisfies \( E(x_{\text{obs}}) = n_{\text{obs}} \cdot p = n_{\text{obs}} - 1/2 (\approx n_{\text{obs}}) \). Solving this gives \( p = 1 - 1/(2n_{\text{obs}}) \).
without negatively affecting predictions.

5.8 Variables Defined Only on a Subset of Units

In the AVRP data, some variables are defined only for women: for example, the menopause indicator or the indicator for use of oral contraception. If these variables have missing values, then the model for their conditional distribution is estimated using only women with observed values of these variables. When used as predictors, these variables are assigned the value 0 for men and interaction terms of these variables with the fully observed female indicator are always included in the model.

5.9 Variables with Temporal Constraints

Some immunogenicity variables have temporal constraints because they cannot logically dramatically increase after a missed vaccine injection or after an injection not containing AVA (a saline injection). That is, the values can increase only within a certain range of natural variability. Therefore, if these variables have missing values due to missed visits, the imputed values should be consistent with the administered injections. In order to accomplish this, we check consistency of the drawn value at visit $t$ for unit $i$, $Y_{i,t}$, with the observed and imputed values of $Y_{i,j}$, $j < t$, on the same unit. If the drawn value is larger than the last observed or imputed value plus a pre-fixed value, $\delta$, when no AVA injection was given at visit $t$, we draw again from the posterior predictive distribution until a consistent value is drawn; we set the value $\delta$ equal to the standard deviation of that variable in the placebo arm.

6 Selection of Predictors and Their Temporal Order

Because there are approximately 400 variables and only 1005 subjects divided into seven treatment groups (80-170 subjects in each) in the AVRP interim data, for computational simplicity,
we constrain the number of predictors that enter the conditional model for each outcome. The predictor selection takes place before the imputation procedure using data completed by a preliminary imputation drawing from each variable’s observed marginal distribution. We allow the predictors for each outcome variable to differ across different arms and monotone blocks.

Demographic variables (age and sex) are fully observed and are always included in the models. For each outcome with missing values, the potential predictors are all the variables that are *more* observed (i.e., with fewer missing values) than the outcome in a particular monotone block. We use a stepwise procedure to choose the predictors for univariate conditional models as follows. Fit regression models of the outcome given each single possible predictor, and age and sex, sort the predictors according to the corresponding Akaike’s information criterion (AIC) (Akaike, 1974), and then 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, where *fittability* is defined as invertibility of the corresponding design matrix and is checked sequentially on the subsets of predictors sorted by AIC. If one predictor is not *fittable*, that suggests that it does not contain enough extra information about the outcome, and so this predictor is dropped from the subset and the same checking continues to the next selected predictor, until the last one. The checking is conducted independently within each treatment arm and each monotone block.

The AVRP data include longitudinal measurements of several variables. Partitioning the missing data into monotone blocks breaks the temporal order of these variables, but this is usually of no concern for the following reasons. First, imputation is a stochastic prediction problem: If one variable is highly correlated with another variable, then including the former into the conditional model of the latter improves the prediction, and thus the quality of the imputation, regardless of their chronological order. Second, because we perform variable selection when specifying conditional models, if variables are highly correlated, then lagged or future values of the same variable will usually be selected as predictors, and therefore corre-
lations between such variables are implicitly used in the imputation. Also, the variation over short time periods of physiological variables and bio-markers may not be well modeled using imputation methods for time series data, which account for trends and seasonality (e.g., Hopke, Liu, and Rubin, 2001).

7 Conclusion

Imputing missing data in large complex studies is often challenging due to the large (sometimes larger than the sample size) number and different types of variables. The widely-used MICE approach is flexible, but has the theoretical drawback that the conditional distributions being specified may be incompatible. The sequential imputation method for monotone missing data is theoretically valid, but is limited to monotone missing data patterns. Motivated by the missing data arising from the AVRP trial data, we have proposed a new imputation approach, IMB, which attempts to combine the virtues of these two approaches. The key idea is to decompose any missing data pattern into a collection of constructed monotone patterns and impute the missing data within each monotone block sequentially. By design, IMB simplifies the modeling process: For each variable having missing values in a monotone block, the set of predictors is reduced to the variables that are more observed than that variable, either actually observed or imputated at a previous step. In some applications this reduction may be sufficient to handle the problem of fewer observations than predictors. For the AVRP 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.

IBM does not solve the incompatibility of MICE-based imputation approaches. Our goal is to provide an approach that may mitigate the problem. To the best of our knowledge, as yet no accepted measure of incompatibility exists, so that we cannot accurately quantify the reduction of the incompatibility compared to other sequential imputation strategies such as MICE. Moreover, due to the complex structure and the special types of variables in the AVRP data,
we were not able to impute the AVRP data using any of the existing MICE packages and thus cannot provide a direct comparison between IMB and MICE in this application. Nevertheless, as discussed, the key to the success of sequential imputation strategies lies in reasonable specifications of the conditional models, and there are several intuitive advantages of IMB over MICE in terms of reducing the possibilities for model mis-specification and thereby reducing error propagation in the presence of mis-specification. Convergence suggests algorithmic compatibility.

**Supplemental Materials**

**R/Fortran code and simulated data:** The supplemental materials provide the R and Fortran routines for implementing the IMB algorithm. We provide a synthetic dataset that closely mimics the dimension and structure of the AVRP data as the public use file is not yet available. Please read the file README.pdf in the zip file for more details. (IMBsuppl.zip, zip archive).

**References**


40. Schafer, J.L. (2012). Pan: Multiple imputation for multivariate panel or clustered data. URL: [http://cran.r-project.org](http://cran.r-project.org), R Package Version 0.5.


44. StataCorp. (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.


