Research Article

Statistics in Medicine

Received: 2 March 2015, Accepted: 19 October 2015 Published online 12 November 2015 Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6800

Multiple imputation and analysis for high-dimensional incomplete proteomics data

Xiaoyan Yin,[a,b,e*†] Daniel Levy,[a,c] Christine Willinger,[a,c] Aram Adourian[d] and Martin G. Larson[a,b,f]

Multivariable analysis of proteomics data using standard statistical models is hindered by the presence of incomplete data. We faced this issue in a nested case–control study of 135 incident cases of myocardial infarction and 135 pair-matched controls from the Framingham Heart Study Offspring cohort. Plasma protein markers ($K = 861$) were measured on the case–control pairs ($N = 135$), and the majority of proteins had missing expression values for a subset of samples. In the setting of many more variables than observations ($K \gg N$), we explored and documented the feasibility of multiple imputation approaches along with subsequent analysis of the imputed data sets. Initially, we selected proteins with complete expression data ($K = 261$) and randomly masked some values as the basis of simulation to tune the imputation and analysis process. We randomly shuffled proteins into several bins, performed multiple imputation within each bin, and followed up with stepwise selection using conditional logistic regression within each bin. This process was repeated hundreds of times. We determined the optimal method of multiple imputation, number of proteins per bin, and number of random shuffles using several performance statistics. We then applied this method to 544 proteins with incomplete expression data (≤40% missing values), from which we identified a panel of seven proteins that were jointly associated with myocardial infarction. © 2015 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: multiple imputation; stepwise selection; high dimension; imputation quality

1. Introduction

1.1. Study design and objectives

As part of the 'Systems Approach to Biomarker Research (SABRe) in Cardiovascular Disease (CVD) initiative' (http://www.nih.gov/news/health/mar2009/nhlbi-12.htm—accessed December 22, 2014), a nested case–control study was designed with pairs of myocardial infarction cases and controls (N pairs = 135), which were matched using several baseline characteristics. Technical details of the study design and plasma protein detection method have been reported previously [1]. Of the unique protein biomarkers measured ($K = 861$), only 261 were measured in every sample, and no single sample had complete data on all proteins. The experimental aim of the study was to identify proteins that were jointly associated with myocardial infarction. We faced a serious challenge regarding multiple marker analysis: If we ignored biomarkers with missing values, we would waste potentially important information; however, if we removed subjects with any missing values, we would lose all observations. To use the

[a] *The Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA, U.S.A.*
[b] *Department of Biostatistics, School of Public Health, Boston University, Boston, MA, U.S.A.*
[c] *Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Boston, MA, U.S.A.*
[d] *BG Medicine Inc., Waltham, MA, U.S.A.*
[e] *Department of Cardiology, Boston University, Boston, MA, U.S.A.*
[f] *Department of Mathematics and Statistics, Boston University, Boston, MA, U.S.A.*
**Correspondence to: Xiaoyan Yin, Framingham Heart Study, National Heart, Lung, and Blood Institute, 73 Mt Wayte Avenue, Suite 2, Framingham, MA 01702, U.S.A.*
[†] *E-mail: xyin@bu.edu*

available data effectively, we decided to multiply impute missing values. Imputation and subsequent analysis were complicated by the presence of many more variables than observations ($K \gg N$).

We discuss those challenges and our decisions within the following context. To accommodate the pair-matched design, we used conditional logistic regression for statistical analysis. In each model, we maintained a ratio at least 5:1 for pairs/markers. To identify 'important' proteins—that is, proteins associated with myocardial infarction—we used stepwise selection. In all analyses, we strictly adhered to Rubin's rule for multiple imputation [2,3]. We developed an approach in which we randomly shuffled the protein markers into several bins, imputed data, and used stepwise selection for the proteins in each bin. Before applying multiple imputation methods to our incomplete data, we conducted extensive simulations to answer three questions: (1) Which bin size is suitable for these data? (2) Does imputation quality (defined in Section 2.3.2) differ between joint modeling and fully conditional specification (i.e. MCMC vs FCS)? (3) To select a stable panel of proteins associated with myocardial infarction, how many shuffles do we need?

In general, through this work, we developed an approach that applies for high-dimensional data missing completely at random or at random. We demonstrated which parameters in imputation and stepwise selection affect the variability in final model.

1.2. Multiple imputation

Multiple imputation is widely used to cope with missing data [2,4–8]. It fills in missing values to generate multiple complete data sets that preserve the main characteristics of the original data. Plausible values for missing data are drawn randomly from the underlying joint distribution of the variables, and each complete data set is analyzed using standard methods. Across data sets, one combines parameter estimates and their variances, accounting for between- and within-imputation variability, to generate overall estimates and standard errors. This process of estimation, reflecting our uncertainty in missing values, is called 'Rubin's rules'.

There are two general methods to multiply impute multivariate data having an arbitrary pattern of missing values: (1) the joint modeling approach, which uses Markov chain Monte Carlo (MCMC) simulation [5], and (2) the chained equation method—also called fully conditional specification (FCS)—which sequentially imputes one variable at a time [9–11]. The MCMC approach consists of two steps: imputation (I-step) and posterior (P-step). At the I-step, simulated values are generated to replace missing values for each observation independently, conditioning on the current estimate of the mean vector and covariance matrix; at the P-step, the complete data set is used to calculate a posterior distribution (i.e. a mean vector and covariance matrix) for input into the next I-step. These two steps iterate until reaching a stationary distribution, from which values are drawn to replace the missing values in one data set. FCS is also iterative, but missing values are replaced sequentially one variable at a time: At each round of iteration for a given variable, one specifies its conditional distribution with respect to all other variables in the imputation model. Imputed values are drawn from the estimated conditional distribution on the observed value for the variable being considered and on the imputed data for the remaining variables.

1.3. Number of variables per imputation model (binning)

Before performing multiple imputation, we had to determine how many variables to include in one imputation model. Collins et al. [12] suggested that including as many variables as possible would increase estimation efficiency and reduce bias. Graham 2009 [6] suggested using a maximum of 100 variables in multiple imputation, even with a sample size as large as 1000, and fewer variables if the sample size is smaller.

Like us, Emerson et al. [13] had too many variables for a single imputation model ($N = 677$ samples vs $K = 42$ variables). They divided variables into bins of nine or fewer (arbitrary choice) and imputed data within each bin. We also needed to shuffle variables into bins. Our initial experience was that including too many variables in a bin often produced a singular covariance matrix for MCMC, and the expectation maximization (EM) algorithm in bootstrap resampling for priors often failed to converge. The reason for the latter result is that the number of parameters to be estimated by the EM algorithm increases rapidly because K means and $K(K-1)/2$ covariance elements must be estimated. By simulations, we were able to tune both bin size and number of shuffles.

1.4. Stepwise selection with imputed data

Stepwise selection of proteins with imputed data is likewise challenging. If we run stepwise selection on each imputed data set, the variables selected will likely differ among m imputed data sets, so the parameter estimates cannot be combined following Rubin's inference rules. To follow Rubin's rules, we must run stepwise selection simultaneously on m imputed data sets and decide whether to include/remove a variable using parameter estimates and standard errors combined across m imputed data sets. This method is computationally intensive, but it is the natural way to obtain unbiased estimates with correct precision. Following Wood et al. [3], we call this method the 'RR' (abbreviated from Rubin's rules) approach.

With more candidate variables than subjects, another issue is the number of variables in one analysis model. In logistic regression, maintaining five to ten events per variable (EPV) is recommended [14]. We had 135 events, which allowed for 27 biomarkers to be evaluated in one model. Performing stepwise selection within the same bins as those used in imputation, a relatively small bin size conformed with EPV guidelines.

1.5. Number of shuffles and criteria for importance

Results of imputation and stepwise selection depend on which variables are contained within each bin. During imputation, imputed values are generated using variables in the same bin; during variable selection, p values for individual variables are affected by which of the other variables appear in the analysis model. By repeatedly shuffling variables into bins, we impute and test each variable in the presence of most other variables. Here, we compare selection results for different numbers of random shuffles of biomarkers into bins.

2. Methods

2.1. Plasma protein data collection and analysis

Our sample was comprised of 135 pairs of myocardial infarction cases and controls identified from the Framingham Heart Study Offspring cohort [15]. As a nested case–control design, the cases and controls were matched for age, sex, smoking status, and statin use at baseline exam cycles 5, 6, 7, or 8. Details about the study design have been described previously [1]. In total, 861 protein markers were measured, including 261 with complete expression data, 283 incomplete markers with $\leq$40% missing values, and 317 markers with $>$40% missing values. We dropped proteins with $>$40% missing values for two reasons: (1) We may be not able to collect these markers in practice and (2) high missingness leads to poor imputation quality and unstable inference (we revisit this issue in the Section 4).

We assumed that the incomplete data were missing completely at random. Plasma samples were measured in sets of eight, including three matched pairs and two pooled reference aliquots. For any protein with missing data in the reference samples, all six experimental samples also had missing values. This accounted for the vast majority ($>$90%) of missing data in cases and controls. Furthermore, means and standard deviations of protein expression values were fairly constant across a wide spread of missing percentages (Supplementary Figure 1). We rank normalized [16] marker data to avoid problems caused by non-normality. All imputation models and analysis models included known CVD risk factors: body mass index (BMI), current diabetes status, plasma high density lipoprotein (HDL) cholesterol, hypertension treatment, systolic blood pressure, and total plasma cholesterol. We used matching factors as auxiliary variables in imputation models.

2.2. General methods and notation

The overall strategy was as follows: (1) randomly shuffle biomarkers into bins; (2) within each bin, perform imputation m times; and (3) perform stepwise selection on m data sets from each bin. In phase 2, we took markers with high inclusion frequency and pooled them in a single bin for imputation and stepwise selection to establish a final multiple-protein model.

As a natural analysis method for a matched case–control design, we used conditional logistic regression with case status as the response variable, matched pair identifier as the stratification factor, and protein biomarkers as the main predictors, adjusting for clinical risk factors.

Let i index individuals and h index case–control pairs (pair ID). The model then takes the following form:

$$\log\left(\frac{\pi_{hi}}{1-\pi_{hi}}\right) = \alpha_h + \boldsymbol{x}_{hi}^{'}\boldsymbol{\beta} + \boldsymbol{y}_{hi}^{'},$$

where $\boldsymbol{\pi_{hi}}$ is the probability of being a case for the ith individual in matched pair h with covariates $\boldsymbol{y_{hi}}$ and proteins $\boldsymbol{x_{hi}}$ and $\boldsymbol{\alpha_h}$ is a pair-specific intercept which is to be conditioned out. Clinical covariates were forced in.

Let l be the number of non-overlapping bins into which B proteins are shuffled, let n be the number of shuffles, and let m be the number of imputed data sets. We shuffled proteins into l bins completely at random, and the B proteins in each bin (together with auxiliary variables) went into one imputation model, generating m imputed data sets. Stepwise selection of proteins was performed using the m imputed data sets for that bin. Every step was based on p values inferred from m imputed data sets. For example, at step 1, each protein was evaluated using every data set: Each protein's m parameter estimates were combined using RR to generate a single p value that determined whether the protein entered the model.

From l bins, we obtained l panels of proteins to form the bin-specific panels. Pooling these l panels of proteins, we obtained one shuffle-specific panel, so that from n shuffles we obtained n shuffle-specific panels. We assumed that more important biomarkers have higher inclusion frequency in shuffle-specific panels.

2.3. Simulations

Through simulations, we explored the roles of key factors and tuned the imputation process. We began with the observed complete data for 261 proteins and proceeded in four steps: (1) introduce missing values via simulation; (2) perform multiple imputation with varying bin sizes, using FCS and MCMC with or without priors to evaluate their effects on imputation quality; (3) determine the number of shuffles needed to obtain a stable panel of biomarkers; and (4) compare the results with those obtained from complete data. The simulation process is summarized in Figure 1.

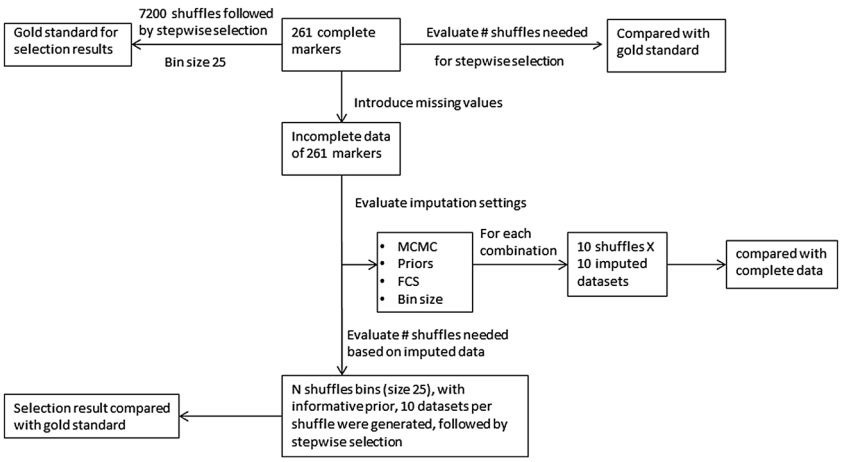

Figure 1. Flow chart of simulation process. Using 261 biomarkers with complete data, we introduced missing values and studied parameters for successful imputation. The imputation parameters included the number of variables per bin, FCS versus MCMC, use of prior information in MCMC, and the number of shuffles. From evaluating these parameters, we determined that 25 was the 'best' bin size and that MCMC with prior information worked better than MCMC without prior information or FCS. Using these settings, we imputed data and studied the parameters for stepwise selection.

2.3.1. Introducing missingness in complete biomarker data. The simulated missingness should reflect the distribution of missing values across all biomarkers with missing values clustered in case–control pairs. Therefore, we masked some data completely at random, by pairs, to mimic the distribution in the set of 544 biomarkers that we would ultimately analyze. The histogram of the numbers of non-missing values in pairs for observed and simulated data is displayed in Supplementary Figure 2.

2.3.2. Imputation quality measures. As pointed out in the imputation literature, proper imputation should generate accurate and plausible values and preserve the variability and correlation matrix of the original data [17–19]. We used three measures for gauging imputation quality: distance ratio (DR; a Mahalanobis distance statistic), mean absolute difference in correlation coefficients (MADC), and frequency of extreme values.

We defined a modified Mahalanobis distance, DR, which measures the cumulative distance between the imputed data and complete data. Smaller mean DR indicates better imputation quality. Details are in Appendix A.

The MADC indicates how well the original correlations among variables are preserved in imputed data sets. We calculated correlation coefficient matrices for the complete data ($\boldsymbol{R}$) and for each imputed data set ($\boldsymbol{R}_q^{'}$ for data set q). The MADC is the average of the absolute element-wise differences between the two correlation matrices, where p is the number of variables:

$$MADC_q = mean\left|\boldsymbol{R} - \boldsymbol{R}_{\boldsymbol{q}}^{'}\right| = \frac{2}{p(p-1)} \sum_{for\ a<b} \sum \left| r(a,b) - r_q^{'}(a,b) \right|$$

Smaller MADC indicates a better-preserved correlation matrix.

In imputation, implausible (extreme) values may occur and too many extreme values indicate poorly imputed data. We used the frequency of extreme values as a complementary measure for imputation quality (refer to Supplementary Materials).

2.3.3. Imputation method. We used procedure MI of SAS software version 9.3 (Copyright, SAS Institute Inc., Cary, NC, USA) to generate imputed data sets. We used random seeds in single-chain MCMC and regression imputation in FCS with 20 iterations for burn-in.

We evaluated the effects of the number of variables per bin (bin size) and the number of independent allocations of variables to bins (shuffles) on imputation quality. We also investigated whether using informative priors (i.e. mean vector and covariance matrices from masked data) for multivariate distributions in MCMC affects imputation quality, and we compared MCMC with FCS.

2.3.4. Imputation tuning: bin size, FCS, and MCMC with or without informative priors. We tried bin sizes of 40 and 50, but MCMC often failed with or without informative priors. Therefore, we tested smaller bin sizes: $B = 10$, 15, 20, 25, or 30. The 261 masked biomarkers were randomly shuffled into roughly equally sized bins. The biomarkers in each bin entered one imputation model together with auxiliary variables. To study imputation quality, we computed mean DR and MADC for ten imputed data sets from each imputation model per bin size per shuffle. In addition, we randomly shuffled the biomarkers ten times given each bin size, so that for each bin size, there were ten values of mean DR and ten values of MADC.

Figures 2 and 3 display imputation quality in terms of mean DR and MADC, respectively, for combinations of bin sizes, informative priors, and FCS versus MCMC methods. We found that imputation quality improved with increasing bin size. In the absence of an informative prior, with a bin size of 25 or 30, more than half of the imputations did not converge. However, using a prior led to smaller mean DR and MADC, even with small bin sizes. Compared with a bin size of 30, a bin size of 25 provided similar imputation quality with more successful imputations in the presences of an informative prior. All other aspects being equal, FCS had poorer quality than MCMC, that is, greater mean DR and greater MADC. Using the frequency of extreme values, we drew the same conclusion (Supplementary Table 1). We decided to impute with a bin size of 25 using MCMC with an informative prior.

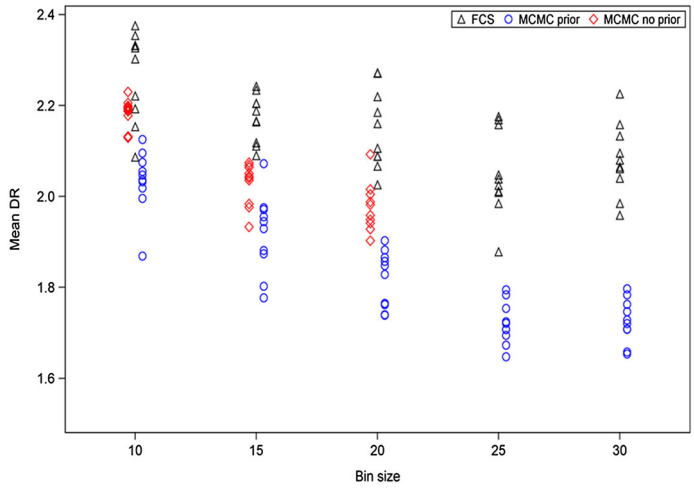

Figure 2. Imputation quality metric: mean distance ratio (*DR*) by bin size and imputation method. Results are based on simulated data. For each combination of bin size and prior information, ten shuffles were performed and ten datasets were imputed within each bin. Bin-specific datasets were merged within each shuffle. Mean DR was calculated across the ten datasets within each shuffle. (Smaller DR indicates less distance between the imputed data and original data.) Mean DR decreased with increasing bin size. MCMC with prior information generated data with smaller DR than either FCS or MCMC without priors. Failure to finish imputation was common with bin sizes of 25 and 30 when no prior was used in MCMC.

2.3.5. Assessing stepwise selection following imputation. We first used the complete data to assess how many shuffles were needed to produce a stable panel in stepwise selection. The frequency that a marker was selected ($p < 0.05$ to enter and stay) in bin-specific models was defined as the inclusion frequency. We created three 'gold standards' by running 7200 shuffles of 261 markers with a bin size of 25, one for each inclusion frequency threshold: 40, 50, and 60%. We used Cohen's kappa statistic to quantify agreement between the gold standard and the panel selected under various numbers of shuffles (20 to 220 by 40). Details are in Appendix B. We found that 140 shuffles were sufficient to obtain stable selection results with 261 complete biomarkers (Figure 4).

To evaluate the process with our masked data, we imputed 20 data sets using a bin size of 25 and the MCMC approach with an informative prior. We performed stepwise selection based on the RR method within each bin on the 20 stacked data sets. For this, we revised a SAS macro from Chen (https://www.biostat.wisc.edu/sites/default/files/tr_217.pdf, accessed July 30, 2012) to perform stepwise selection in conditional logistic models. At all inclusion frequency thresholds, agreement increased until ~140 shuffles,

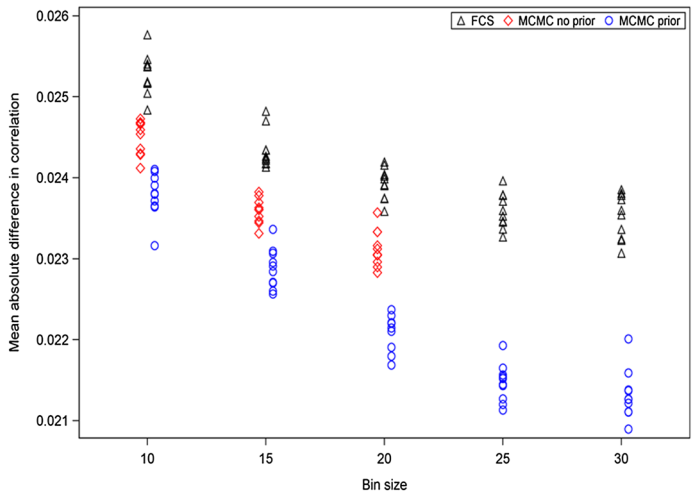

Figure 3. Imputation quality metric: mean absolute difference in correlations (*MADC*) by bin size and imputation method. MADC decreased with increasing bin size. MCMC with prior information generated data with smaller differences in correlation matrices than either FCS or MCMC without priors. The mean and standard deviation of the absolute correlation coefficients in complete data are 0.13 and 0.10.

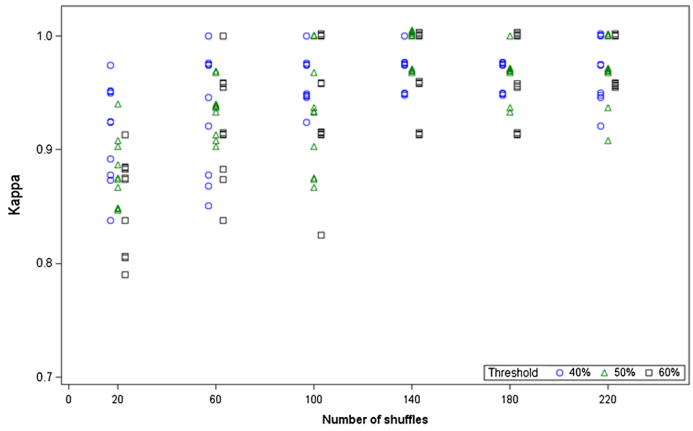

Figure 4. Kappa statistics for model selection results based on complete data of 261 biomarkers by number of shuffles into bins of size 25. Kappa quantifies the agreement between selection results based on specific numbers of shuffles versus a gold standard obtained from 7200 shuffles. We ran stepwise selection within bins holding ~25 biomarkers. Given n shuffles, we called a biomarker 'important' if it was selected at least 40, 50, or 60% among n times. From these results, we concluded that 140 shuffles were sufficient for a stable panel.

above which kappa values were ≥0.8. As expected, kappa values for imputed data were smaller than for complete data (Figure 5).

Table 1 shows the numbers of biomarkers selected per bin and per shuffle using complete data and imputed data. The median number of selected biomarkers per bin was 2, whereas the median number

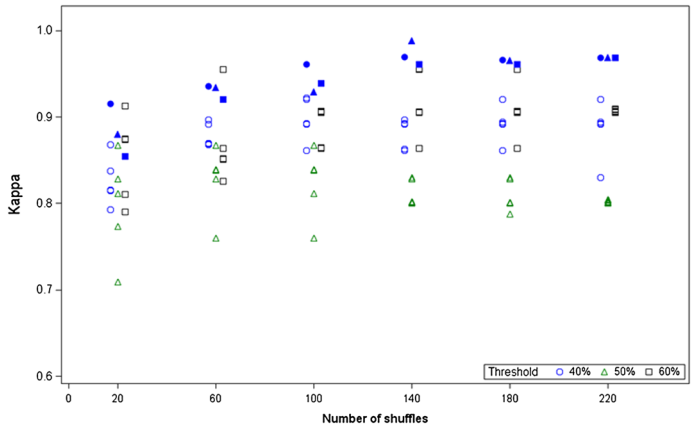

Figure 5. Kappa statistics based on imputed data for the 261 biomarkers after introducing missing values. Kappa values compare the presence/absence of biomarkers in the gold standard model (complete data, filled symbols) versus the imputed data (20 imputations, open symbols). For each combination of number of shuffles (n) and selection threshold (t), kappa from the complete data is the average from ten replications. Kappa values were calculated by comparing selection results with a gold standard (as in Figure 4). Kappa values are generally smaller than those based on the complete data. It appears that kappa does not improve with more than 140 shuffles.

Table I. Number of biomarkers selected in conditional logistic models per bin and per shuffle.

Statistic	Observed complete data		Imputed data[a]	
	Per bin	Per shuffle	Per bin	Per shuffle
Min	0	13	0	11
Q1	1	18	1	16
Median	2	19	2	18
Q3	3	21	2	20
Max	10	28	9	26

[a]The imputed data are based on 135 case–control pairs, 261 biomarkers (≤40% missingness), a bin size of 25, and imputation using MCMC with priors.

per shuffle was 18 to 19. Typically, one or two fewer biomarkers were selected with imputed data than with complete data.

2.3.6. Conclusions from simulation study. From these simulation studies, we drew three primary conclusions. First, bigger bin sizes improved imputation quality, but the imputation model failed to converge with very large bins. A bin size of 25 was the optimal choice given our sample size. Second, the MCMC approach with informative priors produced higher quality imputation results than other approaches. Third, with 261 biomarkers (135 case–control pairs, ≤40% missing data per biomarker), we needed at least 140 shuffles to obtain a stable panel of selected biomarkers.

3. Application to actual data

Characteristics of the study sample have been published [1]. Briefly, 135 pairs of myocardial infarction cases and controls were matched for age, sex, smoking, and statin use: 65 ± 9 years old, 34% women, 24% current smokers, and 19% statin users. More cases than controls were diabetic (28 vs 7%) or on hypertension treatment (47 vs 37%). Other characteristics (cholesterol, BMI, and blood pressure) were similar between cases and controls.

We used 544 biomarkers (261 biomarkers with complete data and 283 with ≤40% missing values). Biomarkers were shuffled randomly into 22 bins of target size ~25. Biomarkers in each bin—together with case status, matching factors, and clinical covariates—were used to impute missing values by the MCMC approach (observed mean and covariance matrices were used for priors). For each bin, 20 imputed data sets were generated and stepwise conditional logistic regression was implemented using the RR approach.

Whereas our initial tuning used 261 biomarkers, we now analyzed 544 protein markers, which should presumably require a larger number of shuffles. We used a single surrogate complete data set and found that we needed 260 shuffles to yield stable selection results. Details are in the Supplementary Materials. Therefore, we repeated the process above for 260 shuffles. Protein markers chosen with at least 40% frequency are listed in Supplementary Table 2. The number of biomarkers chosen at frequency thresholds of 40, 50, and 60% were 33, 26, and 24, respectively.

Using one bin with 26 biomarkers chosen with frequency ≥50%, we performed a final round of multiple imputation plus stepwise selection. To ensure stable regression parameter estimates, we generated 50 imputed data sets. Imputation was followed by stepwise selection with criterion $p < 0.05$ to enter and stay in the model. From this process, seven proteins were jointly associated with myocardial infarction status: glycoprotein 5 (gene: *GP5*), cluster of differentiation 5 molecule (CD5) antigen-like (*CD5L*), alpha-amylase 1 (*AMY1A*), myoglobin (*MB*), collagen α-1 (XVIII) chain (*COL18A1*), protein kinase C inhibitor protein 1 (*YWHAZ*), and multimerin-2 (*MMRN2*) (Table 2). With the exception of *MB*, an established diagnostic marker of myocardial infarction [20,21], these data present a novel set of proteins as potential biomarkers of myocardial infarction. *GP5* is part of a group of surface glycoproteins that

Table II. Final panel of biomarkers jointly predicting myocardial infarction status in a multiple-marker conditional logistic model.

Marker (gene name)	Missing values	Inclusion[a] frequency	Single marker *p* value	Final model using RR approach[b]		
				Odds ratio	95% CI	*P* value
GP5	11%	100%	0.0010	0.40	(0.18, 0.88)	0.023
CD5L	0%	99%	0.0012	0.50	(0.29, 0.86)	0.013
AMY1A	40%	98%	0.019	0.40	(0.20, 0.81)	0.012
MB	9%	97%	0.0053	0.38	(0.17, 0.83)	0.017
COL18A1	0%	90%	0.036	2.43	(1.25, 4.70)	0.009
YWHAZ	8%	85%	0.0056	0.34	(0.16, 0.75)	0.008
MMRN2	0%	79%	0.038	2.10	(1.19, 3.70)	0.011

[a]Inclusion frequency was calculated from 260 shuffles of 544 markers (stage 1) and rounded to the nearest integer percentage.
[b]The final model (stage 2) was based on 135 case–control pairs and 50 imputed datasets with the 26 most frequently included markers. The model was adjusted for BMI, diabetes status, HDL cholesterol, hypertension treatment, systolic blood pressure, and total cholesterol.

mediate the adhesion of platelets to injured vascular walls, and knockout models of *GP5* in mice have implicated this protein in thrombin-induced platelet activation [22]. *CD5L* is involved in the innate immune response. Roles for CD5 in the development of atherosclerotic lesions and in the inflammatory response underlying metabolic syndrome have been described [23,24]. *AMY1A* is a salivary enzyme that contributes to carbohydrate metabolism and has been associated with metabolic syndrome, obesity, and diabetes [25,26]. *COL18A1*, a structural extracellular matrix protein, can be cleaved to produce endostatin that is an endogenous anti-angiogenic protein that may have a protective effect on the progression of atherosclerosis [27–29]. *YWHAZ* is a cell signaling protein that has been shown to contribute to cardiac hypertrophy and fibrosis in diabetic mice and may function in the development of cardiovascular complications of diabetes in humans [30,31]. *MMRN2* is another extracellular matrix glycoprotein that has been shown to impair angiogenesis and may function as a key negative regulator of vascularization during normal vascular development [32,33].

4. Discussion

We developed an approach for identifying important predictors of binary outcome status when missing values exist in hundreds of candidate predictor variables and when sample size is small relative to the number of variables. In this context, we sought to identify a parsimonious multiple marker model. Because data appeared to be missing completely at random, we could not recover extra information regarding single predictor associations with the outcome. The individual protein p values based on imputed data and based on partially observed data are similar in scale overall (Supplementary Figure 5a) or across missing percentage groups (Supplementary Figure 5b). Rather, the purpose of imputation was to make multimarker analyses tractable. Given our assumption that data were missing completely at random, this process would not work for data with non-random missingness.

We performed simulation based on a complete data subset and randomly masked values for some predictors. We shuffled variables into bins and performed imputation immediately followed by stepwise selection, strictly adhering to RR. We examined choices of bin size, number of random shuffles of biomarkers to bins, FCS versus MCMC imputation methods (with or without prior information on multivariate distributions), and different inclusion frequency thresholds for selecting important predictors. We applied our simulation findings to real data with 544 biomarkers (mean missingness = 9% and maximum missingness = 40% per marker) on 135 myocardial infarction case–control pairs.

We found that MCMC with informative priors provided the best imputation quality. Typically, FCS and MCMC generate similar analysis results for arbitrarily missing data [34,35]; however, this might not hold true when the number of variables is large relative to sample size. As mentioned by Carpenter and Kenward [10], with a large number of variables, providing a stabilized covariance matrix is necessary, and this was only feasible for our data with MCMC. We encountered convergence failures as we tuned imputation and model selection parameters. Likely, others will too, depending on sample size, amount of missing data, imputation method, and modeling approach. We used failure to converge as a red flag to signal dysfunctional settings. Therefore, we chose parameters so that the process was free of convergence issues. Specifically, we used prior information for the missing data and bin size 25 (during imputation and model fitting). Other researchers may explore using/ignoring prior information and may try several bin sizes.

We limited our investigation to 544 biomarkers with ≤40% missing data. When we tried to relax this threshold to 50% (580 markers), the imputation model often failed to converge. Even when imputation was successful, the quality deteriorated (MADC increased ~20%, Supplementary Figure 4). We attempted to maximize bin size while maintaining at least five ‘cases’ per variable. Larger bin sizes caused convergence problems during the imputation step. We kept the same bins in imputation and analysis to adhere to RR.

We performed many tests in each round of stepwise selection, and we used $p < 0.05$ as the criterion for entering and staying in a (bin- and shuffle-specific) model. As a result, the probability approaches 1.0 of at least one type I error. Simulations under known conditions could be conducted to examine false-positive and true-positive rates of protein selection; however, those are beyond the scope of this investigation. We note that with more than 500 candidate predictors, the ‘best’ multiple marker model is likely not unique: Slight changes in imputation parameters or analysis may produce different models. Nevertheless, our process of studying imputation and analysis under various conditions (here, the

imputation method, the number of markers per bin, the number of shuffles, and imputation quality measures) can be straightforwardly applied to a range of statistical models and outcomes.

Appendix A: Mahalanobis distance ratio

Mahalanobis distance measures the dissimilarity between two vectors and is often used to detect outlying multivariate data points [36]. In the simulations, we knew the true value of each missing data point; therefore, we used Mahalanobis distance to measure distance from a vector to another with known mean and covariance matrix.

For subject j, the Mahalanobis distance is $D_j = (\mathbf{X}_j - Y_j)\boldsymbol{S}^{-1}(\mathbf{X}_j - Y_j)^T$ where X_j is the vector of biomarkers from the imputed data, Y_j is the vector of observed complete data for the same subject, and $\boldsymbol{S}$ is the covariance matrix obtained from the observed complete data. The closer the imputed values are to the original values, the better the imputation quality and the smaller the value of D. However, we cannot directly use the D statistic, because the missing pattern varies across individuals: Some subjects have missing values in marker A, some in marker B, and so on. Thus, X_j is a mixture of observed values and imputed values. If marker k is observed for subject j, the corresponding element $(X_{jk} - Y_{jk})$ is 0; otherwise, it is the difference between the imputed and the original values. D tends to increase with the number of missing biomarkers for each subject. To overcome this issue, we created the distance statistic $D_{uj} = U_j\boldsymbol{S}^{-1}U_j^T$, where U has the same dimensions as X and Y, and its element takes a value of 0 if the corresponding $x = y$ and equals 1 otherwise. The ratio D/D_u, which we call distance ratio (DR), is a weighted average of the sum of the squared difference accounting both for covariance between the corresponding elements of Y and for the pattern of missing values in X and Y for each subject.

Appendix B: Kappa statistic

We performed n random shuffles with $n = 20$, 60, 100, 140, 180, and 220, respectively. The inclusion frequency of each marker being selected was calculated as (no. of times chosen) divided by n. We also tested $n = 7200$, an arbitrarily large number, to obtain the 'gold standard' for selection frequency of each marker. We explored how the results varied depending on the importance threshold (t) for inclusion frequency, which is defined as the proportion of times that a biomarker appears in a shuffle-specific panel, with $t = 40$, 50, and 60%.

For n shuffles, we obtained a list of 'important' biomarkers at each threshold t. We used Cohen's kappa statistic to quantify agreement between chosen biomarkers from various n and the gold standard. Larger values of kappa denote more stable selection results. Kappa = 1 means that the selected variables agree perfectly with the gold standard. For each n and t, the process was repeated ten times.

The following is an example of a kappa calculation for $n = 140$:

Chosen by gold standard	$t = 40\%$, kappa = 0.92			$t = 50\%$, kappa = 0.97			$t = 60\%$, kappa = 0.95		
	No	Yes	Total	No	Yes	Total	No	Yes	Total
No	239	2	242	244	1	245	249	1	250
Yes	1	19	20	0	16	16	0	11	11
Total	240	21	261	244	17	261	249	12	261

Acknowledgements

This work is supported by NIH contract (N01-HC-25195) and Cooperative Research and Development Agreement (CRADA) between BG Medicine, Inc., the National Heart, Lung, and Blood Institute, and Boston University.

References

1. Yin X, Subramanian S, Hwang SJ, O'Donnell CJ, Fox CS, Courchesne P, Muntendam P, Gordon N, Adourian A, Juhasz P, Larson MG, Levy D. Protein biomarkers of new-onset cardiovascular disease: prospective study from the systems approach to biomarker research in cardiovascular disease initiative. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2014; **34**:939–945.
2. Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley: New York, 1987.

3. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Statistics in Medicine* 2008; **27**:3227–3246.
4. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 1995; **142**:1255–1264.
5. Schafer JL. *Analysis of incomplete multivariate data*. Chapman & Hall: London, New York, 1997.
6. Graham JW. Missing data analysis: making it work in the real world. *Annual Review of Psychology* 2009; **60**:549–576.
7. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 2006; **59**:1087–1091.
8. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *American Statistician* 2007; **61**:79–90.
9. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007; **16**:219–242.
10. Carpenter JR, Kenward MG. *Multiple imputation and its application* (1st edn). John Wiley & Sons: Chichester, West Sussex, 2013.
11. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**:681–694.
12. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
13. Emerson JW, Dolled-Filhart M, Harris L, Rimm DL, Tuck DP. Quantitative assessment of tissue biomarkers and construction of a model to predict outcome in breast cancer using multiple imputation. *Cancer Informatics* 2009; **7**:29–40.
14. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
15. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. *American Journal of Epidemiology* 1979; **110**:281–290.
16. Blom G. *Statistical estimates and transformed beta-variables*. Wiley: New York, 1958.
17. Brand JPL, van Buuren S, Groothuis-Oudshoorn K, Gelsema ES. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica* 2003; **57**:36–45.
18. Denk M, Weber M. Avoid filling swiss cheese with whipped cream: imputation techniques and evaluation procedures for cross-country time series. *IMF Working Paper* 2011; **11**:1–27.
19. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *Journal of Royal Statistical Society* 2008; **57**:273–291.
20. Braunwald E, Antman EM, Beasley JW, Califf RM, Cheitlin MD, Hochman JS, Jones RH, Kereiakes D, Kupersmith J, Levin TN, Pepine CJ, Schaeffer JW, Smith EE 3rd, Steward DE, Theroux P, Gibbons RJ, Alpert JS, Faxon DP, Fuster V, Gregoratos G, Hiratzka LF, Jacobs AK, Smith SC Jr. ACC/AHA 2002 guideline update for the management of patients with unstable angina and non-ST-segment elevation myocardial infarction--summary article: a report of the American College of Cardiology/American Heart Association task force on practice guidelines (Committee on the Management of Patients With Unstable Angina). *Journal of the American College of Cardiology* 2002; **40**:1366–1374.
21. Gibler WB, Gibler CD, Weinshenker E, Abbottsmith C, Hedges JR, Barsan WG, Sperling M, Chen IW, Embry S, Kereiakes D. Myoglobin as an early indicator of acute myocardial infarction. *Annals of Emergency Medicine* 1987; **16**:851–856.
22. Ramakrishnan V, Reeves PS, DeGuzman F, Deshpande U, Ministri-Madrid K, DuBridge RB, Phillips DR. Increased thrombin responsiveness in platelets from mice lacking glycoprotein V. *Proceedings of the National Academy of Sciences* 1999; **96**:13336–13341.
23. Arai S, Shelton JM, Chen M, Bradley MN, Castrillo A, Bookout AL, Mak PA, Edwards PA, Mangelsdorf DJ, Tontonoz P, Miyazaki T. A role for the apoptosis inhibitory factor AIM/Spalpha/Api6 in atherosclerosis development. *Cell Metabolism* 2005; **1**:201–213.
24. Miyazaki T, Kurokawa J, Arai S. AIMing at metabolic syndrome. Towards the development of novel therapies for metabolic diseases via apoptosis inhibitor of macrophage (AIM). *Circulation Journal* 2011; **75**:2522–2531.
25. Nakajima K, Nemoto T, Muneyuki T, Kakei M, Fuchigami H, Munakata H. Low serum amylase in association with metabolic syndrome and diabetes: a community-based study. *Cardiovascular Diabetology* 2011; **10**:1475–2840.
26. Nakajima K, Muneyuki T, Munakata H, Kakei M. Revisiting the cardiometabolic relevance of serum amylase. *BMC Research Notes* 2011; **4**:419.
27. Isobe K, Kuba K, Maejima Y, Suzuki J, Kubota S, Isobe M. Inhibition of endostatin/collagen XVIII deteriorates left ventricular remodeling and heart failure in rat myocardial infarction model. *Circulation Journal* 2010; **74**:109–119.
28. Iribarren C, Herrinton LJ, Darbinian JA, Tamarkin L, Malinowski D, Vogelman JH, Orentreich N, Baer D. Does the association between serum endostatin, an endogenous anti-angiogenic protein, and acute myocardial infarction differ by race? *Vascular Medicine* 2006; **11**:13–20.
29. Moulton KS, Heller E, Konerding MA, Flynn E, Palinski W, Folkman J. Angiogenesis inhibitors endostatin or TNP-470 reduce intimal neovascularization and plaque growth in apolipoprotein E-deficient mice. *Circulation* 1999; **99**:1726–1732.
30. Gurusamy N, Watanabe K, Ma M, Zhang S, Muslin AJ, Kodama M, Aizawa Y. Inactivation of 14-3-3 protein exacerbates cardiac hypertrophy and fibrosis through enhanced expression of protein kinase C beta 2 in experimental diabetes. *Biological and Pharmaceutical Bulletin* 2005; **28**:957–962.
31. Watanabe K, Thandavarayan RA, Gurusamy N, Zhang S, Muslin AJ, Suzuki K, Tachikawa H, Kodama M, Aizawa Y. Role of 14-3-3 protein and oxidative stress in diabetic cardiomyopathy. *Acta Physiologica Hungarica* 2009; **96**:277–287.
32. Lorenzon E, Colladel R, Andreuzzi E, Marastoni S, Todaro F, Schiappacassi M, Ligresti G, Colombatti A, Mongiat M. MULTIMERIN2 impairs tumor angiogenesis and growth by interfering with VEGF-A/VEGFR2 pathway. *Oncogene* 2012; **31**:3136–3147.

33. Burgisser PE, Tempel D, Cheng C, Pasterkamp G, Duckers H. Multimerin2 influences vascularization through modulation of the extracellular matrix. *Circulation* 2014; **130**:A20152–A20152.
34. Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JA. Joint modelling rationale for chained equations. *BMC Medical Research Methodology* 2014; **14**:28.
35. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 2010; **171**:624–632.
36. Rousseeuw PJ, Zomeren BCV. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 1990; **85**:633–639.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

