Published in final edited form as:

Stat Methods Med Res. 2010 December ; 19(6): 653–670. doi:10.1177/0962280208101273.

Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide

Yulei He, Alan M. Zaslavsky, David P. Harrington, Paul Catalano, and Mary Beth Landrum

Y. He, A.M. Zaslavsky, and M.B. Landrum, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave. Boston, MA, 02115; D.P. Harrington and P. Catalano, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 375 Longwood Ave., Boston, MA 02115. The authors are affiliated with the Statistical Coordinating Center of CanCORS.

Abstract

The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium is a multisite, multimode, multiwave study of the quality and patterns of care delivered to population-based cohorts of newly diagnosed patients with lung and colorectal cancer. As is typical in observational studies, missing data are a serious concern for CanCORS, following complicated patterns that impose severe challenges to the consortium investigators. Despite the popularity of multiple imputation of missing data, its acceptance and application still lag in large-scale studies with complicated datasets such as CanCORS. We use sequential regression multiple imputation, implemented in public-available software, to deal with nonresponse in the CanCORS surveys and construct a centralized completed database that can be easily used by investigators from multiple sites. Our work illustrates the feasibility of multiple imputation in a large-scale multiobjective survey, showing its capacity to handle complex missing data. We present the implementation process in detail as an example for practitioners and discuss some of the challenging issues which need further research.

Keywords

Cancer; Imputation diagnostics; Missing data; Sequential regression multiple imputation; Survey

1 Introduction

Large-scale health and social studies, such as that conducted by the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium,¹ can provide numerous measurements that provide opportunities for research on multiple topics. However, such studies are subject to the problem of missing data when enrolled subjects do not have data recorded for all variables of interest: different subjects can have different items collected, data collection and entry errors can result in missing values, and subjects can be ineligible for or refuse to answer some items from a survey.

Ad-hoc missing data methods, such as complete-case (CC) analysis, available-case (AC) analysis, and treating missing data as a separate category in a model, are easy to implement and popular. But they have well-known disadvantages.² These methods tend to yield biased results because relationships among variables are not preserved, and can be inefficient when they reduce sample sizes by removing cases with missing values.

Hot-deck imputation and its variants, which replace missing data with observed values in pre-defined “donor” cells, are often used for large complex surveys.³ They have the advantages of reducing mean squared error for univariate statistics and preserving covariance structure in multivariate datasets. As a single-imputation approach, however, the hot-deck method tends to underestimate the variance of estimates, except with some estimators that correct variance estimates.⁴ In addition, since it is defined implicitly by the procedure that matches subjects with similar observed values to create the donor cells, it is difficult to describe explicitly the statistical models used for missing variables.

Model-based multiple imputation, a Bayesian approach introduced by Rubin,⁵ is a principled method for analysis with missing data. For each missing value, it imputes several, say M , values to create M completed datasets. The imputation model, explicit or implicit, is designed to be appropriate to both the true complete-data distribution and missing-data mechanism. For each of the M completed datasets, standard complete-data methods are used to estimate the parameters of interest and their associated variances. The results of these M analyses are then combined following simple rules² to provide a single inference that incorporates uncertainty due to missing data. For recent reviews of multiple imputation from various perspectives.^{6–10}

Multiple imputation has been very popular for handling incomplete-data problems in “inhouse” applications¹¹ where the imputer is the same party as the analyst, often with a specific analysis goal. The small scale of the data in such studies often affords the analyst more opportunity to tailor the imputation models, assisted by either existing imputation software¹⁰ or statistical programming targeted to specific problems.^{12,13}

On the other hand, although multiple imputation was originally motivated by the need to handle nonresponse in public-use data files or shared databases,¹⁴ there are relatively few such “outside” applications,¹¹ where the imputer is different from the analysts and the imputations are created for general users. Examples include imputation projects for the Fatal Accident Reporting System,¹⁵ census industry and occupation codes,¹⁶ the National Health and Nutrition Examination Survey,¹⁷ and the National Health Interview Survey.¹⁸ However, in these applications, imputation is only applied to a handful of variables which are important to many analyses.

Within the CanCORS consortium, investigators at local centers propose research topics and perform analyses using a centrally-maintained database, acting as the “outside” users for the imputed data. In addition, the consortium intends to release a database for public use. Unlike the aforementioned “outside” applications, the goal of the imputation project in CanCORS is to construct a database with all incomplete variables imputed, without any specific prioritization. The task is further complicated by the complex study design. Such genuine “outside” applications of imputation may rise more frequently as more health and social studies involve collecting large-scale datasets for multiple users.

Our strategy is to apply sequential regression multiple imputation (SRMI, also called “imputation using chained equations” or “imputation by full conditional specification”)^{9,19,20} to build appropriate and feasible imputation models. In this paper, we illustrate the implementation process and discuss some of the practical issues related to data handling and model building, providing a template for practitioners faced with similar projects. Section 2 introduces Can-CORS and describes the missing data problems in the survey. Section 3 briefly reviews SRMI. Section 4 presents the imputation procedure for the CanCORS patient survey data in a step-bystep fashion. Section 5 presents an example of analysis using the imputed data. Finally, Section 6 concludes with a discussion and directions for future research.

2 Study Background

2.1 CanCORS

The CanCORS consortium is funded by the National Cancer Institute and the Veteran's Administration to examine services and outcomes of care delivered to population-based cohorts of newly diagnosed patients (from 2003 to 2005) with lung and colorectal cancer in multiple regions of the country. It consists of seven Primary Data Collection and Research (PDCR) sites and a Statistical Coordinating Center (SCC). Each PDCR site identified appropriate samples to obtain combined cohorts of approximately 5000 patients diagnosed with each cancer. The SCC assists the PDCR sites in the collection of standardized data across the individual research sites and serves as the central repository for the pooled data. In the imputation project, the SCC takes the role of the "imputer", while investigators from PDCR sites and possible users outside the consortium take the role of the "analysts".

CanCORS collected data from multiple sources including patient surveys, medical records, and surveys of physicians who treated CanCORS patients. In the patient surveys, baseline data were collected in a patient interview administered approximately 4 months after diagnosis; follow-up data were collected in a second interview approximately 11–13 months after diagnosis. Medical records were reviewed approximately 15 months after diagnosis. The surveys and medical record reviews collected data about the care received during different stages of illness, including diagnosis, treatment, surveillance for recurrent disease, and palliation, as well as data on various clinical and patient-reported outcomes and patient preferences and behaviors. The physician survey asked physicians about their knowledge and beliefs about care and their practice characteristics. Data from these primary sources are supplemented with cancer registry data and other publicly available datasets such as Medicare claims.

Multiple imputation was applied to both the patient and physician surveys, following similar schemes presented in detail in later sections. This paper focuses on imputation in the patient surveys. These surveys obtain information from participants regarding their cancer diagnosis and treatment, quality of life, experiences of care, care preferences, health habits, other medical conditions, and demographic information. The survey items were organized into 12 to 14 topic-related sections, most of which are identical for lung and colorectal cancer patients.

The baseline survey uses four forms, including the full survey for the patient, a brief version for patients who cannot complete the full interview, a survey for surrogates when the patients are alive but unable to complete the interview, and a survey for surrogates of dead patients. The surrogate surveys contain parts of the full survey and a few additional items that pertain specifically to the surrogate's experiences of the patient's cancer care. The follow-up survey was attempted for all participants who were alive at the time of the baseline survey, but not those who had already died before initial contact.

2.2 Missing data

At the beginning of an imputation project, a good practice is to examine and understand the missingness patterns. Due to the multiformat, multiwave structure of the CanCORS survey data, the patterns of nonresponse are complicated. In general, however, we can identify the following broad categories of missing data, typical of health and social surveys:

1. Unit nonresponse: cases sampled for the survey but not participating in an interview, such as noncontacts and refusers.
2. Block nonresponse: items that do not appear on some survey versions.

3. Item nonresponse:

- a. items that are missing for a subject because structured skip patterns do not call for collecting them. For example, “when was your most recent radiation treatment?” would be skipped by the patients who had not received radiation therapy.
- b. interviewed cases for whom items are missing due to early drop-out from the study.
- c. residual item nonresponse including survey items answered as “don’t know” or “refused”.

For the current release of the survey data, the unit nonresponse rate is about 55%. Different surveys forms vary greatly in the number of items included. When concatenating all surveys together, this leads to a large number of block nonresponses, with an average rate over all variables about 25%. The item nonresponse rates are summarized in Table 1. Within each survey, the item nonresponse rates due to patients’ early drop-out or uncertainty (“don’t know”) or unwillingness (“refused”) about the answers are relatively low, indicating the effectiveness of information collection in the survey interview. On the other hand, a large proportion of item missingness can be attributed to skip patterns. The average missingness rate over all variables is about 66% in the concatenated data.

3 Overview of SRMI

3.1 Theoretical background

Throughout this paper, we assume that missingness is at random (MAR),⁵ meaning that the probability of missingness does not depend on unobserved values conditional on observed data. Under this assumption, it is valid to fit the complete-data model and impute under it, while ignoring the model for the indicator of missingness. The MAR assumption becomes more tenable as the imputation model is made more general, that is, includes more predictors to make it more plausible that missingness depends only on observed characteristics and not on those that are missing.²¹ In some studies, it might be plausible that the probability of missingness is related to unobserved values, so data are not missing at random (NMAR). By definition, the MAR property cannot be falsified from within the dataset at hand but only by imposing modeling assumptions or inferences from other data. Similarly, specification of an NMAR model requires hypothetical assumption from outside the data. Defining an NMAR model is feasible in some applications where attention is focused on particular variables, such as in informative dropout in longitudinal studies. But specifying a general NMAR model for a complex database as CanCORS is practically unachievable because of the assumptions needed concerning the response mechanisms for many incomplete variables, although post-imputation sensitivity analysis might be practical by reweighting the results under MAR.⁷

Many imputation methods are based on assuming a joint distribution for multivariate data.^{2,17} Examples include the multivariate normal or t family for continuous variables, loglinear models for categorical variables, and the general location model for a mixture of continuous and categorical variables. The joint modeling approach is theoretically sound, but may lack the flexibility needed to represent complex data structures arising in surveys. The CanCORS patient survey data consist of a large number of variables having a variety of distributional forms, subject to certain logical or consistency bounds imposed by survey questionnaires, and displaying unsystematic missingness patterns. In such a case, a joint model is difficult to implement because the typical specifications of multivariate distributions are not sufficiently flexible to accommodate these features.

An alternative approach is SRMI, which specifies the multivariate model by separate conditional models for each incomplete variable. van Buuren⁹ presented a comprehensive review, also referring to SRMI as the full conditional approach. Let $Y_j (j = 1, \dots, p)$ denote the variables with missing values, X denote the collection of fully observed variables, and $Y_{j-1}, \dots, Y_{j+1}, Y_p$ denote the $p - 1$ variables in Y 's excluding Y_j . SRMI specifies a conditional model $P(Y_j | Y_{-j}, X; \theta_j)$ for each Y_j , where θ_j denotes the corresponding model parameters. In each iteration of the procedure, it draws θ_j from $P(\theta_j | Y_j^{obs}, Y_{-j}, X)$ using the observed Y_j^{obs} and completed Y_{-j} (from the previous iteration) and X , and then imputes the missing Y_j^{mis} , cycling through all the variables. It is relatively easy to include complex data features in these univariate regression models, following common guidelines of regression modeling applied to the data at hand.

Despite the increasing popularity of SRMI,^{6,9,18–20} its statistical properties are not fully understood, although some discussion can be found.^{9,22} One issue is the possible incompatibility among the conditional models, that is, the possibility that there exists no joint distribution with the conditionals of the assumed forms. The second is the behavior of the Gibbs-like chain of drawn parameters and missing values of each conditional model in the sequential imputation. With incompatible models, these draws may not converge to a single stationary distribution. Correspondingly, there are no rules for assessing the convergence of the Gibbs-like chain. Despite the incompleteness of theoretical justification for SRMI, van Buuren et al.²³ presented a simulation study in which the apparent incompatibility did not greatly affect the final imputation inferences. In addition, empirical findings from Ambler, Omar, and Royston⁶ and van Buuren⁹ suggested that a few iterations (< 20) of the Gibbs-like chain might work well for problems with modest fraction of missing data (< 10 – 15% missingness), but the convergence behavior is largely unknown in more demanding problems.

3.2 Software: IVEware

Three statistical packages for implementing SRMI are the SAS (SAS Institute Inc, Cary, NC, USA)-callable “impute” module of IVEware,²⁴ the library “MICE” in R,²⁵ and the routine “ICE” in STATA (STATA CORP, TX, USA). We used the first in this work because the SCC stores, processes and releases survey data in SAS format.

The following regression imputation models are implemented in IVEware for a missing outcome Y_j . Predictors can include the main effects and interactions among Y_{-j} and X ; similar models are implemented in the two other software packages.

1. A normal linear regression model, if Y_j is continuous.
2. A logistic regression model, if Y_j is binary.
3. A polytomous or generalized logit regression model, if Y_j is categorical with more than two categories.
4. A Poisson loglinear model, if Y_j is a count.
5. A two-part model, if Y_j is semi-continuous, where a logistic regression is used to model the zero/non-zero status for Y_j , and a normal linear regression is used to model the value of Y_j conditional upon its being non-zero.

4 Implementation of Imputation

Clear and complete documentation of the imputation procedures, including data processing and model building, is crucial in preparation of large databases. It helps analysts to understand the mechanisms behind the imputation and conduct with their analysis, allows

the imputer to modify or update the imputation program when necessary, and maintains continuity over the lifetime of a complex project.

4.1 Initial data processing

The released database had to be deidentified before being accessed and used by the analysts, as in many other studies. Prior to imputation, all potential identifiers, such as the contact information of patients, surrogates, and care providers, were removed from the datasets. The subject's birth date was replaced by age groups. Dates of post-diagnosis medical procedures were transformed to the relative time (in days) to the patient's diagnosis date and the latter was removed from the datasets as well. Furthermore, prior to model-based imputation, we deterministically imputed values known from their logical relationship with other responses.

4.2 Strategies for block missing and skipped items

Although the survey variables are organized into topic-related sections, imputing sections (or other divisions) of the survey separately may ignore important associations among variables, as the best predictors for a variable can be in both the same and different sections. Imputing the survey as a whole also makes the MAR assumption more plausible. In addition, although the study used several different survey forms, there are a large number of overlapping items and hence it would be inefficient to impute them separately. We concatenated all of the surveys to create a combined rectangular dataset and imputed all included items simultaneously.

Based on input from clinical investigators, concatenation and imputation were carried out separately for patients with lung and colorectal cancer to avoid complex interactions, since the two groups have different disease etiology and care patterns. We did not further stratify the imputation because interactions with other factors did not appear to be equally important.

The concatenated datasets contain 5149 subjects with 793 variables for lung cancer and 4921 subjects with 798 variables for colorectal cancer. The concatenation procedure causes block missingness in the combined dataset for the variables that are not used in all survey forms. Some of these missing data may not be used in any meaningful analysis, such as followup survey items for patients who died before attempted contact. Others may be of potential use, depending on the context of the analysis. For example, questions about patients' incomes were omitted from the brief survey, but the missing values are meaningful for the brief survey participants and could be imputed and used if the sample for an analysis involving the income variables includes that group. We first imputed all block missing values, and after imputation, we restored missingness for the block missing data that are not eligible for analyses or not used by investigators. The validity of such massive imputations can be checked by comparison to nonresponse weighting²⁶ (Section 5.3).

Values for which item nonresponse is due to skip patterns in surveys would be excluded from most analyses, since skip patterns are designed to avoid collecting information that is not meaningful, such as the severity of radiation therapy side effects for a patient who did not have radiation therapy. In contrast, values for which item nonresponse due to drop-out or is coded as "don't know" or "refused" in general could be included in analyses and hence should be imputed.

Imputation in the presence of skip patterns presents particular challenges due to the fact that although the skipped item is MAR (missingness always depends on the observed skip response), skipping is deterministic and therefore the imputation model for the skipped items is inestimable without further assumptions. Such assumptions should be based on a scientifically justifiable understanding of the relationship between the screener item and the

skipped item. For example, considering the following five questions: (i) have you had surgery for your cancer? (ii) how many surgeries have you had for your cancer? (iii) how involved were you in the decision about surgery? (iv) have you had any problems with your self-care? (v) how satisfied are you with your communication with your doctor? Questions (ii) and (iii) are skipped for patients who have had no surgery yet, i.e., answering NO on (i). For these patients, we deterministically impute 0 for question (ii) before the other imputation calculations. In effect we regarded items (i) and (ii) as together eliciting a response for the number of surgeries received. If we treated these skips as missing data and imputed positive values, ignoring the relationship with (i), it would bias imputed responses to question (iv) for the skipped patients if patients receiving more surgery have more problems with self-care. On the other hand, relationships of patient characteristics to satisfaction with doctor communication might be mediated in part through involvement in decisions about surgery. If we remove this mediated effect by deterministically imputing a single value to item (iii) for patients having no surgery, we would not accurately reproduce the marginal relationship underlying prediction of missing values of item (v) from patient characteristics. While we could posit that this relationship is different for those who did not have surgery, it would be impractical with limited data and analytic resources to investigate all such interactions. Instead, we treat responses to (iii) that are missing due to skips as missing data and impute them. After imputation, these imputed values are restored to skips in the analytic datasets. While a few imputations were deterministic as for item (ii), most of the skipped items were handled similarly to item (iii).

4.3 Specifying imputation models

The flexibility of SRMI allows imputing variables of various distributional types. We classified each variable as “categorical”, “continuous”, “mixed”, or “transferred”, following IVEware’s syntax rules; examples of these are shown in Table 2. The variables involved in the imputation modeling (i.e., those which are imputed and/or act as predictors) belong to the first three categories (Section 3.2), while the ones excluded from the imputation process are “transferred”. “Categorical” variables are nominal variables whose response levels do not have an obvious ordering. “Continuous” variables include both truly continuous variables (in the original or transformed scale) and ordinal ones. We treated the latter as “continuous” because IVEware has no option for directly modeling ordinal variables. Treating them as “categorical” would lose the ordering of responses, and it can be difficult to fit the general logit model for ordinal variables with more than a few response categories. The distribution of a “mixed” (semi-continuous) variable consists of a point mass at zero and a continuous positive part. For “continuous” variables, we forced the draws of imputations to fall within the ranges shown from the observed data, using the “bounds” option in IVEware. For ordinal variables, we rounded the fractional imputed numbers to the nearest integer after imputation to make them consistent with the original data formats.

Due to the large number of variables included in the datasets, it is virtually impossible to include all of them in each prediction equation of SRMI. IVEware has options for automatic model selection. First, users can set thresholds for the minimum marginal R^2 increment in the stepwise selection, meaning that a variable will only be selected if the increase of R (ΔR^2) is greater than the threshold. With a smaller ΔR^2 criterion, more predictors will be selected. But the models will be more complex, producing less stable fits and slowing the imputation computations.

The IVEware documentation does not provide guidance for setting the ΔR^2 criterion. We tested different values (i.e., $\Delta R^2=0.1$, 0.01, and 0.001) and compared the corresponding model selection and fitting results. In our data, setting $\Delta R^2 = 0.1$ tended to select a very small number (1 or 2) of predictors for most of the missing outcomes, and therefore may underidentify important predictors in the imputation. On the other hand, setting $\Delta R^2 = 0.001$

selected many more predictors, but many of the regression coefficient estimates were rather extreme (e.g., violating the rule of thumb that logistic regression coefficients should be within the interval $(-5, 5)^{27}$ with large standard errors and unstable across iterations. Some of the large logistic coefficients indicated likely deterministic relationships between variables, such as speaking Mandarin at home and not of Latino and Hispanic origin, and thus helped us to identify items that should be imputed deterministically. Apart from these cases, large logistic coefficients could be due to overfitting models to small samples. Model selection with $\Delta R^2 = 0.01$ appeared to yield more reasonable results than either of the above alternatives and hence was used in our implementation. In addition, we limited the maximum number of predictors for the variables with small number of observed cases to prevent overfitting models for these variables, using the “maxpred” option, which we set equal to 1 for variables with less than 50 observed values.

We examined the selected predictors for missing outcomes and the associated regression coefficients at different iterations (i.e., 3, 5, 10, 15, and 20) of the Gibbs-like chain and found most of them remain similar after running the program for several (3 to 5) iterations. We also applied some simple analyses using multiply imputed datasets obtained from those iterations, and the results were close to each other. In addition, we applied SRMI to only those variables that were involved in these simple analyses and collected imputations after a long run of the Gibbs-like chain (e.g., several thousand iterations). The corresponding analysis results were similar to those obtained from the imputations of the whole dataset after a few iterations. These provided some evidence for the plausibility of the imputed values of the latter. The multiple imputations released to the consortium investigators were obtained from running the program with different initial seeds for 5 iterations.

4.4 Simple diagnostics

There are few diagnostic tools for imputation modeling, especially for large-scale datasets. However, we performed some simple ad-hoc diagnostics to obtain an overall summary of the performance of multiple imputation. An example of model assessment targeted to specific analyses appears in Section 5.2. We estimated the marginal means and pairwise correlation coefficients of all continuous and binary variables after applying SRMI, and compared them with those obtained using AC. In general, the differences of point estimates can go in either direction. On the other hand, standard errors for the means from multiple imputation are not expected to be significantly larger than those from AC, while multiple-imputation standard errors for the correlation coefficients are most likely to be smaller than those from AC because using AC removes cases with either variable missing. Table 3 summarizes the distribution of the standard error differences across all means and correlations for the two cancer types. For means, SRMI reduced the standard error for about 50% of the variables compared to AC. But for correlations, SRMI reduced the standard error for about 75% of pairs of the variables compared to AC. The summary results overall met our expectation.

The comparison also helped us to identify a handful of variables for which using SRMI led to substantial increases in variance estimates of either the mean or correlation. Further examination showed that they generally fall into two classes.

1. Categorical variables with a very unbalanced distribution of classes (e.g., a 1/0 binary variable of which over 95% of the cases are 0). For these variables, the data are often too sparse to support conventional maximum likelihood estimation of the logistic regression model used in IVEware. As a result, the regression estimates are often unstable, introducing large between-imputation variability, and hence greatly increase the total variance. A solution might be to use a fully Bayesian logistic

imputation approach which uses an informative prior distribution for the regression parameters and draws from the exact posterior distribution.^{27,28}

2. Variables that need to be massively imputed because they are not used in all survey forms but need to be used in analyses involving data from multiple surveys. These imputations tend to be more sensitive to model misspecifications.

There appears to have no easy one-step solution for improving the imputations for all these variables following the generic SRMI scheme. Instead, our strategy is to document these “troubling” variables, and to handle them in an analysis-specific manner. That is, if such variables are involved in a specific analysis and the use of the generic SRMI causes any severe problem, as noticed and reported by the analysts, we would use a more tailored imputation model for them within the context of that analysis. This requires communication and collaboration between the imputer and analysts. We believe such strategy is especially important for imputing large-scale data since it is often necessary for practitioners to find a balance between model validity and feasibility to meet the time line of the whole project.

4.5 Processing and releasing multiply imputed datasets

The raw imputations were further processed before release, including creation of some composite variables (e.g., SF-12 scales) and merging of imputations with data from other sources. The released datasets include five completed sets and the original incomplete one, as well as the imputation documentation for easy check and comparison.

New imputations are updated as raw data collection and processing are updated periodically, as is typical in multi-site cohort studies. In addition, the analysts send the imputer their questions and inputs encountered in performing specific imputation analyses, and these are considered in updates of the imputation program.

5 Imputation Analysis Example

5.1 Hospice care analysis

We present a multiple imputation analysis by one of the PDCR investigators as an example of application by an “outside” user. The objective of this study was to examine patterns of cancer hospice care, which includes a broad array of palliative and support services for individuals with terminal illness. It identified patient characteristics and preferences that are associated with patient reports in the baseline survey that they had discussed hospice with a care provider and had used hospice. The study subsample ($n = 2474$) consists of all advanced lung cancer patients (stage IIIB or IV). The outcome variables are patients’ hospice discussion and use, and covariates include patients’ clinical and sociodemographic characteristics.

A simplified illustrative analysis for discussion of hospice is presented here, while a more detailed analysis with the full description and conclusions will be presented in a clinical journal. Table 4 describes the variables from the analytic subsample; some of them have a substantial amount of missing data. To better understand the missingness pattern, we performed separate logistic regression analyses using cases with no missing data, with an indicator variable of nonresponse for each incomplete variable as the outcome, and the remaining variables in the sample as the predictors. The results (not shown) indicated that the missingness of income and insurance are significantly (at the 5% level) related to other variables. Therefore, the complete cases of the subsample may not be treated as a simple random sample of the original data, that is, the missing completely at random (MCAR)⁵ assumption is likely violated.

We performed a logistic regression analysis using the multiply imputed data for “hospice discussion” and predictors including all other variables in the subsample. We also applied the CC approach and the method that treats missing data as a distinct category. Table 5 shows the results from each method. The regression estimates from CC and SRMI are somewhat different, and the latter produced smaller standard errors than the former for all regressors. At the 5% level, predictors associated with Hispanic ethnicity and divorced/separated marital status were non-significant under CC but significant under SRMI, while the predictor associated with a history of myocardial infarction (significant under CC) became non-significant under SRMI. In this case, CC discards close to 30% of the subjects. When the assumption of MCAR is violated, as in our example, CC removed cases in a non-random fashion and could distort the joint distribution among the variables. As a result, it could produce both biased point estimates and inflated standard errors, and thus misidentify significant predictors. The results from the missing data indicator method are overall similar to those from SRMI, although the former also discarded around 4% of the subjects with the missing outcome variable (i.e., hospice discussion). But in general, the missingness indicator method can produce severe bias even when data are MCAR.

5.2 Imputation model assessment and sensitivity analysis

Since it is difficult to assess the adequacy of all imputation models for all variables in the survey, a plausible strategy is to focus on the imputation model for the variables involved in specific analyses. In the hospice subsample and assuming MAR, we performed posterior predictive checking^{29–31} to examine the deviation of selected analysis results Q computed from the completed data with imputations compared to the values of Q calculated from simulated copies of the completed data under the model. The corresponding posterior predictive (Bayesian) p -value for the completed data is defined as

$$p_{B,com} = P(Q(Y_{com}^{rep}) \geq Q(Y_{obs}, Y_{mis}) | Y_{obs}),$$

where Y_{com}^{rep} (replicated completed data) and Y_{mis} (imputations for missing data) are drawn from their posterior predictive distribution given the observed data Y_{obs} , that is, $P(Y_{com}^{rep}, Y_{mis} | Y_{obs})$. If the imputation model is adequate, the results from applying the same analysis to the Y_{com}^{rep} and (Y_{obs}, Y_{mis}) should be similar, and the resulting $p_{B,com}$ -value should be close to 0.5. On the other hand, a $p_{B,com}$ -value that is too big or too small (e.g., $p_{B,com} > 0.95$ or < 0.05) suggests a considerable difference between the two and hence indicates a potential significant impact on analysis results from the misfit of the imputation model.

Particularly in this analysis, we chose Q to be the logistic regression coefficients and the associated standard errors, t -statistics, and p -values because these are of main interest to the investigator. Deviations in these reflect lack of fit of the imputation model that affects the analytic inferences. The posterior predictive p -values were calculated by simulation. All incomplete variables were replicated under the model, conditioning on the remaining variables in the analytic subsample. Table 6 listed the $p_{B,com}$ -values for the aforementioned quantities based on 1000 simulations. Except for a few quantities associated with the standard errors, most of the $p_{B,com}$ -values fall within the acceptable range (.05, .95), suggesting that the imputation model is adequate for the logistic regression analysis.

A good practice in imputation is to carry out sensitivity analysis using alternative modeling strategies, perhaps implemented in different software packages. To this end, we applied a general location model for the hospice subset. Specifically, we treated race, marital status, and insurance as nominal variables and assumed they follow a loglinear model with

conditional independence. We treated other variables (binary or ordinal) as continuous with multivariate normal distributions conditional on the categorical variables, and rounded the fractional imputations prior to completed-data analysis. This approximation to the joint distribution¹⁷ was implemented using the library “MIX” in R. Estimates for the logistic model (not shown) were rather similar to those obtained using SRMI, thus increasing confidence on our results.

5.3 Weighting analysis

We demonstrate an example of using nonresponse weighting for checking the accuracy of multiple imputations. As indicated in Section 4.2, this is specifically targeted to the block missingness that were massively imputed.

Section 8 of the patient baseline survey covered a wide variety of items about the patient’s quality of life, including activity/energy level, physiological/psychological status, and the level of pain, discomfort, anxiety, and other cancer specific symptoms. Many of them are to be used to construct symptom scales to predict other health outcomes. However, most of these items were used only in the full survey but not in the surrogates and brief surveys.

We examined the validity of the imputations for these items by comparing complete-data summaries with estimates obtained by nonresponse weighting. We illustrated with an analysis of the lung cancer sample. We fitted a logistic regression model for inclusion in the full survey with demographics, cancer stage, and most of the common variables used in all surveys as predictors. Nonresponse weights for block missingness of Section 8 were calculated as the inverse of the predicted probability from the logistic model. We ignored the small proportion of item missingness from the full survey in the weighting analysis.

Table 7 lists the descriptive statistics for some of the scales calculated using CC (using only the full survey patients), the weighting method, and SRMI based on 5 imputed datasets. These scales are in the range 0–100 and lower scores indicate worse symptoms. Surrogates and brief surveys were more often obtained from patients who were sicker or deceased (including disproportionately more stage IV patients), so mean scores by CC are biased upwards. The bias of CC is largest within the stage IV subgroup, which contains the largest proportion of imputed cases. Compared to CC, the adjustments from SRMI and weighting are in the same direction, but SRMI estimates showed larger adjustments and smaller standard errors than the weighting method, possibly due to the limited predictive power of the nonresponse model predictors we selected and the large variation of the weights from the heterogeneous sample. We would expect the advantage of SRMI over weighting to be even more pronounced in an analysis that is primarily based on variables with no block missingness, since it makes better use of all the available information on those variables.

6 Discussion and Future Directions

This paper demonstrates the use of multiple imputation to construct a centralized database for multiple users with different analytic objectives. As health or social policy decisions are often based on analyses of large incomplete databases, our work may serve as a template for similar applications. By showing the details of implementation, we aim to encourage practitioners to apply multiple imputation for complex large-scale datasets.

SRMI is used to tackle the challenges arising from the complex survey data in this work, and we believe it is a suitable solution for other missing data problems with similar scale and complexity. However, we have still experienced a variety of challenging issues in implementation. These problems include how to build models which can effectively exclude unused block nonresponses or skipped items, how to do model selection in the presence of a

large number of predictors, how to decide when to stop the chain and collect imputations, and how to perform model validation and diagnostics. We adopted some ad-hoc solutions, but we caution against the mechanical use of SRMI despite its increasing popularity. Echoing other literature in this area,^{7–9} we call for more methodological research on the properties of SRMI which would help answer the above questions. Given current limitations on the practical and theoretical development of SRMI, we also encourage practitioners to do sensitivity analysis using alternative imputation strategies based on joint models, the properties of which are better understood, as illustrated in the hospice care example.

We used IVEware to implement SRMI, but found its current version to have significant limitations. Ambler, Omar, and Royston⁶ illustrated the use of other SRMI software packages, but we surmise that they would also have similar limitations when applied to complex largescale datasets. Our experience calls for the enhancement of the current SRMI software to allow more flexible imputation modeling and data handling, such as adding models for ordinal data and allowing the option of correcting impossible imputations after each conditional model is fitted. In addition, to avoid overfitting of the models for small samples, the software should allow specification of prior distributions for regression parameters.

An emerging trend in health and social studies is to use information from different sources in data analysis. In CanCORS, in addition to the survey data, data from medical records and Medicare claims are also collected for the enrolled subjects. Therefore, variables in different sources can be used simultaneously in substantive analyses. In the ideal centralized database, imputation for the missing data should incorporate information from all sources. This will present a more challenging task since the combined data would include more variables and more complex relationships among them. Data reduction techniques, such as latent variable modeling,³² coupled with SRMI, might be a promising methodological tool to extend the “variable-by-variable” fashion to a “block(of variables)-by-block” one. Relevant research is underway for the CanCORS data.

References

1. Ayanian JZ, Chrischilles EA, Fletcher RH, et al. Understanding cancer treatment and outcomes: the Cancer Care Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology* 2003;22:2292–2296.
2. Little, RJA.; Rubin, DB. *Statistical Analysis of Missing Data*. 2nd ed.. New York: Wiley; 2002.
3. Marker, DA.; Judkins, DR.; Winglee, M. Large-scale imputation for complex surveys. In: M, GR.; Dillman, DA.; Eltinge, JL.; Little, RJA., editors. *Survey Nonresponse*. New York: Wiley; 2002. p. 329–341.
4. Rao JNK. On variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association* 1996;91:499–520.
5. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
6. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing data predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research* 2007;16:277–298. [PubMed: 17621472]
7. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007;16:259–275. [PubMed: 17621471]
8. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007;16:199–218. [PubMed: 17621468]
9. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007;16:219–242. [PubMed: 17621469]

10. Yu LM, Burton A, Riverto-Arias O. Evaluation of software for multiple imputation of semicontinuous data. *Statistical Methods in Medical Research* 2007;16:243–258. [PubMed: 17621470]
11. Barnard J, Meng XL. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* 1999;8:17–36. [PubMed: 10347858]
12. Gelman AE, King G, Liu C. Not asked and not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association* 1998;93:846–857.
13. Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of American Statistical Association* 2005;100:1123–1132.
14. Rubin DB. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association* 1978:20–34.
15. Heitjan DF, Little RJA. Multiple imputation for the Fatal Accident Reporting System. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1991;40:13–29.
16. Schenker N, Treiman DJ, Weidman L. Analyses of public use decennial census data with multiply imputed industry and occupation codes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1993;42:545–556.
17. Schafer, JL. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall; 1997.
18. Schenker N, Raghunathan TE, Chiu PL, Makuc DM, Zhang G, Cohen AJ. Multiple imputation for missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* 2006;101:924–933.
19. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999;18:681–694. [PubMed: 10204197]
20. Raghunathan TE, Lepkowski JM, VanHoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001;27:85–95.
21. Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* 1996;91:473–489.
22. Gelman AE. Parameterization and Bayesian modeling. *Journal of the American Statistical Association* 2004;99:537–545.
23. van Buuren S, Brand JPL, Groothuis-Oudshoorn K, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006;76:1049–1064.
24. IVEware: imputation and variance estimation software. [updated 2007 Aug 2; cited 2008 Nov 25]. Available from: <http://www.isr.umich.edu/src/smp/ive/>
25. R: a language and environment for statistical computing. [cited 2008 Nov 25]. Available from: <http://www.r-project.org/>
26. Cochran, WG. *Sampling Techniques*. New York: Wiley; 1977.
27. Gelman AE, Jakulin A, Grazia MP, Su Y. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*. 2008 To appear.
28. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 1991;86:68–78.
29. Gelman, AE.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. 2nd ed.. London: Chapman and Hall; 2004.
30. Gelman AE, Mechelen IV, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* 2005;61:74–85. [PubMed: 15737080]
31. He Y, Zaslavsky AM. Posterior predictive checking of imputation models. Department of Health Care Policy, Harvard Medical School. 2008 Unpublished technical document.
32. Song J, Belin TR. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine* 2004;23:2827–2843. [PubMed: 15344189]

Table 1

Item nonresponse rates for individual patient surveys

Survey	No. of variables	Sample Size	Skip pattern (%)		Drop-out/Don't know/Refused (%)	
			Mean	Range	Mean	Range
Full baseline	535	5763	50	0-99	1.23	0-37
Brief baseline	156	1397	48	0-99	2.49	0-22
Surrogate live	413	1083	50	0-99	1.61	0-28
Surrogate death	473	1827	62	0-99	1.46	0-29
Follow-up	384	6087	65	0-99	0.55	0-33

Notes: 1. The means and ranges are across the variables within each survey, ignoring the small differences of the survey items between the lung and colorectal cancer patients.
 2. We calculate the item nonresponse rate statistics separately for skip pattern (3a) and dropout/don't know/refused (3b or 3c).

Table 2

Examples of variable types in IVEware

Variable type in IVEware	Survey question	Response
Categorical	Are you receiving or have you had radiation therapy for your cancer?	1=Yes, 2=No.
Continuous	How tall are you without shoes?	Feet/inches or centimeters.
	How would you rate the quality of health care you received for your cancer?	1=Excellent, 2=Very good, 3=Good, 4=Fair, 5=Poor.
Mixed	During the 4 weeks right before your cancer was first discovered, how many days did your physical or emotional problems keep you in bed for all or most of the day?	Number of days, and 0 if none.
Transferred	Patient ID	_____

Table 3

Relative differences (%) in standard error estimates

Percentiles	Mean	Corr
5th	-13.70	-46.10
25th	-.62	-10.50
50th	0	-1.80
75th	.59	0
95th	38.91	4.31

$${}_{\theta}\Delta SE = \frac{se(\hat{\theta}_{SRMI}) - se(\hat{\theta}_{AC})}{se(\hat{\theta}_{AC})} \times 100\%$$

Note: 1. For each mean or correlation, we calculated ${}_{\theta}\Delta SE = \frac{se(\hat{\theta}_{SRMI}) - se(\hat{\theta}_{AC})}{se(\hat{\theta}_{AC})} \times 100\%$, where θ is either a marginal mean or the Fisher transformation of a correlation coefficient.

2. The multiple imputation inferences are obtained from 10 imputations.

Table 4

Variables for hospice care analysis

Variable	Label and classification	% missing
Hospice discussion	1=hospice discussed, 0=no	3.56
Race/ethnicity	1=white, 2=black, 3=hispanic, 4=asian, 5=other	0.16
Income	1= <20k, 2= 20–40k, 3=40–60k, 4=>60k	19.48
Marital status	1= married/live with partner, 2=widowed, 3=divorced/separated, 4=never married	2.79
Insurance	1=medicare, 2=medicaid, 3=private, 4=other	8.61
Cancer stage	1=stage IV, 0=stage IIIB	0
Chemotherapy	1=chemo, 0=no	1.78
Deceased	1=deceased within 1 yr of dx, 0=no	0
Myocardial infarction	1=heart attack, 0=no	12.37
Depression	1=depression, 0=no	12.37

Table 5

Hospice care analysis results

Predictor	Ref.	CC			Missingness indicator			SRMI		
		EST	SE	p-value	EST	SE	p-value	EST	SE	p-value
Intercept		-1.97	.51	.00	-2.08	.43	.00	-2.17	.42	.00
Race	White									
Black		.03	.21	.90	.03	.17	.86	-.02	.17	.89
Hispanic		-.51	.31	.10	-.68	.24	.00	-.66	.24	.01
Asian		.24	.29	.42	.20	.25	.41	.21	.25	.40
Other		.46	.27	.08	.46	.23	.04	.42	.22	.06
Marital status	Married/Partner									
Widowed		-.24	.18	.19	-.10	.15	.50	-.07	.15	.66
Divorced/Separated		.24	.17	.17	.29	.14	.05	.29	.15	.05
Never Married		.66	.32	.04	.57	.25	.03	.58	.27	.03
Missing					.09	.42	.83			
Received chemo	No									
Yes		-1.25	.12	.00	-1.01	.10	.00	-1.01	.10	.00
Missing					.07	.58	.91			
Income	< 20K									
20K - 40K		.12	.15	.45	.13	.14	.35	.12	.14	.41
40K - 60K		-.07	.19	.71	-.04	.18	.83	-.02	.18	.91
60K+		-.38	.21	.07	-.32	.19	.09	-.30	.21	.16
Missing					-.15	.20	.45			
Insurance	Medicare									
Medicaid		.09	.31	.77	.26	.27	.33	.28	.25	.26
Private		.04	.20	.86	.01	.17	.97	.01	.18	.98
Other		.81	.34	.02	.91	.28	.00	.85	.28	.00
Missing					.19	.23	.41			
Myocardial infarction	No									
Yes		-.41	.17	.02	-.24	.15	.11	-.23	.16	.13
Missing					-.62	.56	.27			

Predictor	Ref.	CC			Missingness indicator			SRMI		
		EST	SE	p-value	EST	SE	p-value	EST	SE	p-value
Depression	No									
Yes		.47	.14	.00	.36	.12	.00	.35	.13	.01
Missing					-.31	.58	.59			
Deceased with 1 yr of dx	No									
Yes		1.85	.16	.00	1.59	.13	.00	1.69	.13	.00
Cancer Stage	Stage IIIB									
Stage IV		.72	.13	.00	.76	.11	.00	.77	.11	.00

Note: CC=Complete-case analysis; SRMI=Sequential regression multiple imputation.

Table 6Posterior predictive *p*-values for the logistic regression analysis

Predictor	Ref.	$p_{B,com}(\beta)$	$p_{B,com}(SE)$	$p_{B,com}(t)$	$p_{B,com}(p\text{-value})$
Intercept		.52	.06	.69	.31
Race	White				
Black		.20	.89	.15	.65
Hispanic		.48	.34	.53	.47
Asian		.53	.47	.52	.17
Other		.45	.22	.41	.40
Marital status	Married/Partner				
Widowed		.68	.93	.68	.10
Divorced/Separated		.50	.02	.46	.45
Never Married		.51	.99	.75	.75
Received chemo	No				
Yes		.55	.46	.56	.44
Income	< 20K				
20K - 40K		.42	.15	.41	.36
40K - 60K		.50	.22	.49	.21
60K+		.63	.99	.56	.40
Insurance	Medicare				
Medicaid		.55	.52	.55	.26
Private		.47	.52	.46	.47
Other		.17	.85	.26	.26
Myocardial infarction	No				
Yes		.41	.35	.41	.50
Depression	No				
Yes		.52	.37	.52	.52
Deceased with 1 yr of dx	No				
Yes		.50	.44	.47	.47
Cancer Stage	Stage IIIB				
Stage IV		.49	.71	.52	.52

Table 7

Estimates of Quality-of-life Scales

Sample	Scale	CC		Weighting		SRMI	
		Mean	SE	Mean	SE	Mean	SE
All	EORTC_LC	79	.31	77	.58	75	.26
	EORTC_QLQ	77	.39	74	.60	71	.28
	DYSPNEA	63	.59	59	.96	59	.50
Stage I/II	EORTC_LC	82	.48	81	.69	78	.42
	EORTC_QLQ	80	.60	79	.83	75	.51
	DYSPNEA	64	.95	62	1.34	61	.73
Stage IV	EORTC_LC	78	.62	76	1.07	73	.40
	EORTC_QLQ	75	.81	73	1.18	68	.43
	DYSPNEA	63	1.23	59	2.00	58	.82

Notes: 1. CC=Complete-case analysis; SRMI=Sequential regression multiple imputation.
2. EORTC (European Organization for Research and Treatment of Cancer)_LC (Lung cancer specific): coughing, sore mouth, trouble swallowing and pain.
3. EORTC_QLQ (quality of life questionnaire): insomnia, nauseated, vomited, etc.
4. DYSPNEA: short of breath when resting, walking, or climbing stairs.