Imputation in a Multiformat and Multiwave Survey of Cancer Care

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

Abstract

The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium is a multisite, multimode and multiwave study examining the care delivered to population-based cohorts of newly diagnosed patients with lung and colorectal cancer and assessing predictors and outcomes of that care. As is typical in observational studies, missing data are a serious concern for CanCORS, following complicated patterns that impose severe challenges to investigators and data analysts. We use multiple imputation to deal with block or item nonresponse in the CanCORS surveys. It would be difficult to formulate a joint imputation model that characterizes the underlying relationships among all the variables, especially since the surveys use multiple response formats including nominal, ordinal, and semicontinuous data, with many structured skip patterns. Instead, we applied the sequential conditional regression imputation approach, specifying a collection of models that regress each incomplete outcome on other covariates. We use posterior predictive checking to assess the adequacy of the imputation models.

KEY WORDS:
Cancer; Missing data; Posterior predictive checking; Sequential regression multiple imputation; Survey.

1. Introduction

Large health services and outcome studies, such as that conducted by the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium (Ayanian et al. 2003), provide numerous measurements that provide vast research opportunities for research of health care and policy. However, such studies are subject to the problem of missing data. Enrolled subjects may not have data recorded for all variables of interest; data collection and entry errors can result in missing data values, and subjects can inappropriately skip items from a survey.

The most direct way of dealing with missing data is to exclude incomplete observations (the complete-case ananlysis). This method may give biased results, particularly in models with many variables; even if data are missing for a small percentage of the observations for each variable, with many variables too many observations would become unusable. Other ad-hoc methods, such as mean or median imputation, are easy to implement. But they have well-known disadvantages (Little and Rubin 2002, Chap. 4): variability is underestimated, and relationships among variables are not preserved.

Multiple imputation, a Bayesian model-based approach introduced by Rubin (1987), is a principled method for analysis with missing data. For each missing value, we impute several, say M values, creating M completed datasets. The imputation model, explicit or implicit, is built 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 those M analyses are then combined using standard rules (Rubin 1987) to provide a single inference about the parameters of interest that incorporates uncertainty due to missing data. Multiply imputed databases can be used by various researchers with different analytic objectives (Rubin 1996), as in the CanCORS study, which involves many investigators at different sites.

Although multiple imputation has good statistical properties in principle, it is challenging to implement the method for large datasets with complex study designs, due to the difficulty of specifying a joint imputation model that characterizes the underlying relationships among all the variables with different types and structured skip patterns. A practically appealing strategy in this setting is sequential regression multiple imputation (SRMI) (van Buuren et al. 1999, Raghunathan et al. 2001), which specifies a collection of models that regress incomplete outcomes on other covariates. An additional challenge is to assess the adequacy of imputation models. We propose to use posterior predictive checks (PPC) (Gelman et al. 1996) for this assessment, identifying discrepancies (or lack of) between the original and simulated completed data under the imputation models.

This paper is organized as follows. Section 2 introduces the background of CanCORS and describes the missing data problem. Section 3 presents the SRMI procedures for the CanCORS baseline and follow-up survey data. Section 4 briefly reviews PPC and illustrate its application to multiple imputation. Section 5 presents a simple example. 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 to examine the care delivered to

[*] 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.

population-based cohorts of newly diagnosed patients with lung and colorectal cancer in multiple regions of the country and assess outcomes associated with that care, identifying differences in the delivery of cancer treatment and the reasons for these differences. It consists of seven Primary Data Collection and Research (PDCR) sites (UAB, UCLA/RAND, CRN, NCCC, UI, UNC, and VA) and a Statistical Coordinating Center (SCC). Each PDCR site identifies appropriate samples to obtain combined population-based cohorts of approximately 5000 patients 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.

CanCORS collects data from multiple sources including patient surveys, medical records, and surveys of health care providers. Baseline patient data were collected in a patient interview administered approximately 4 months after diagnosis; follow-up data were collected in a second interview 11-13 months after diagnosis. Medical records were reviewed 15 months after diagnosis. The interviews and medical records 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 health care provider 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 is applied to both the patient and provider surveys, following similar scheme that will be presented in detail in later sections. This paper focuses on the patient survey data because the former has larger sample size, more variables, and more complicated survey design than the latter. The baseline survey obtains information from participants regarding their cancer diagnosis and treatment, quality of life, experience of care, health habits, and other medical conditions, as well as demographic information. Most items for the survey are either copied verbatim or adapted from existing survey instruments that have been used extensively in previous studies. The baseline survey uses five forms, including the full survey for the patient, two brief versions (MD and belief) for patients who cannot complete the full interview, a survey form filled out by a surrogate when the patient is alive but unable to complete the interview, and a survey of surrogates of dead patients. When the respondent drops out before completing the survey, the response is called a partial survey.

The brief survey contains a subset of the items from the full survey. The surrogate surveys also contain subsets of the items of the full survey, and a few additional items that pertain specifically to the surrogate's experiences of the patient's cancer care. The brief MD and belief surveys are very similar, but the former includes the items from the full survey needed to identify physicians and hospitals providing care to the patient, and the latter has fewer questions about treating physicians, but includes additional items on patient beliefs and preferences regarding treatment options.

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 being contacted. The follow-up surveys include a survivor survey to patients who were alive at the time of scheduled follow-up, and a decedent surrogate survey for survivors of patients who had died since the baseline interview. The purpose of the former is to collect details of treatment received after the baseline survey, cancer recurrence or progression, and changes in quality of life, functional status, symptoms, experiences of care, and changes in financial sources. The purpose of the decedent surrogate follow-up survey is to collect data on the quality of end-of-life care, especially symptom management and hospice care.

2.2 Missing data

The term "missing data" refers to the difference between a dataset with all desired items completed for all desired subjects, and the data that are actually obtained. Thus it is only meaningful in relation to some definition of what constitutes "complete data". The ideal CanCORS survey dataset would have the baseline and follow-up surveys including different subtypes as described in Section 2.1. On the other hand, for any particular analysis, only a subset of the variables are needed and the definitions of complete and missing data would be adjusted accordingly.

Due to the multiformat and multiwave structure of the survey data, the patterns of nonresponse are complicated. In general, however, we can identify the following broad categories of missing data:

(1) Unit nonresponse: cases sampled for the survey but not participating in interview, such as noncontacts and refusers.

(2) Block nonresponse: interviewed cases for whom blocks of items are missing due to early drop out from the survey (partial surveys) or use one of the short survey forms (e.g. the brief baseline survey).

(3) Item nonresponse: (a) items that are missing for a case because structured skip patterns do not call for collecting it; (b) residual item nonresponse including survey items that were refused or answered as "don't know".

The estimated unit nonresponse rate for the current release of the CanCORS survey data is around 53%. Table 1 lists the crude estimates of block and item nonresponse rates in which both block missingness and skip patterns are coded as "not applicable" in the database.

Table 1: Block and item nonresponse rates for items in the CanCORS patient surveys

Survey	Not Applicable		Don't know/Refused	
	Range	Mean	Range	Mean
Full Baseline	0-99%	49%	0-36%	1.23%
Brief Baseline	0-99%	57%	0-17%	0.65%
Surrogate Live	0-99%	49%	0-27%	1.58%
Surrogate Death	0-99%	61%	0-27%	1.27%
Follow-up	0-99%	61%	0-32%	0.53%

Note: The range and mean are across the variables.

2.3 Multiple Imputation

Multiple imputation, originally proposed by Rubin (1987) for handling nonresponse problems in public-use survey datasets, is becoming a popular approach to incomplete-data problems across different research fields. In general, the use of multiple imputation approach fall within two classes, the "outside" and the "in-house" applications (Barnard and Meng 1999). In the former case, the purpose is to produce imputations for public-use data files to fit for the "many-analysts-many-goals"; the imputer is typically different from the analysts. In the latter case, the typical and smaller "one-analyst-one-goal" studies, the imputer is the same party as the analyst with a specifical analysis goal. Example literature for "outside" applications include imputation projects for census industry and occupation codes (Schenker et al. 1993), key survey variables in NHANES III (Schafer 1997; Chap. 6), income data in NHIS (Schenker et al. 2006). Much more literature can be found for "in-house" applications, such as Tu et al. (1993), Raghunathan et al. (1996), Gelman et al. (1998).

Within the CanCORS consortium, various substantive research topics have been proposed by different involved investigators and these involve using different parts of databases. In addition, an important goal of the consortium is to construct a feasible study cohort not only for insider users, but possibly also for outside/public users. Clearly, the use of multiple imputation in this work belongs to the "outside" applications. The CanCORS SCC takes the role of the "imputer" while investigators from PDCR sites are the "analysts".

Our survey of the multiple imputation literature also show that in most of the applications, the number of targeted variables for imputation is relatively small. This allows finer tailoring of imputation models. In most of the "in-house" applications, the concerned variables are usually the main outcome or predictors in a specific analysis. For examples of "outside" applications, the targeted variables are often some key variables that are important to many analyses but suffer sizeable proportions of missingness. In our work, however, the general goal is to construct a database that ideally has all incomplete data being "filled-in" without any specific prioritization of variables. Therefore, the challenge of the modeling task lies in the large scale of datasets as well as the complexity of survey structure. Our strategy is to construct a sensible imputation model that incorporates as much as possible the available data and our knowledge about the missing-data mechanism, but at the same time keeps the model building and fitting feasible. As describe in the following sections, we use SRMI to tackle this problem.

3. SRMI for CanCORS Patient Survey Data

3.1 Background

A brief description of SRMI follows; see Raghunathan et al. (2001) for details. Let X denote the fully-observed variables; let Y_1, Y_2, …, Y_p denote p variables with missing values, ordered by the amount of missingness, from least to most. The imputation process for Y_1, Y_2, …, Y_p proceeds as follows. In the first round, Y_1 is regressed on X, and the missing values of Y_1 are imputed; then Y_2 is regressed on X and Y_1 (including the imputed values of Y_1), and the missing values of Y_2 are imputed, then Y_3 is regressed on X, Y_1, and Y_2, and the missing values of Y_3 are imputed; and so on, until Y_p is regressed on X, Y_1, Y_2, …, Y_{p-1}, and the missing values of Y_p are imputed. Starting from the second round, the imputation process carried out in round 1 is repeated, except that now, in each regression, all variables except for the variables to be imputed are included as predictors. Thus, Y_1 is regressed on X, Y_2, Y_3, …, Y_p, and the missing values of Y_1 are re-imputed; then Y_2 is regressed on X, Y_1, Y_3, …, Y_p, and the missing values of Y_2 are re-imputed; and so on. The parameters drawn from each of the regression models over the whole process constitute iterations from a Gibbs-like chain for SRMI. To obtain multiple imputations, we can either collect final draws of missing values from several independently produced chains or collect multiple draws spaced out within a single chain to avoid serial correlations.

Raghunathan et al. (2001) developed an imputation software, IVEware, for implementing the SRMI. This software is a free SAS-callable routine, which provides some features for model selection with options including the number of predictors used in each prediction equation and the criteria for stepwise regression selection. A detailed manual can be found at http://www.isr.umich.edu/src/smp/ive.

For the regressions in the SRMI procedure, the following models are implemented in IVEware:

(1) A normal linear regression model, if the Y-variable is continuous.

(2) A logistic regression model, if the Y-variable is binary.

(3) A polytomous or generalized logit regression model, if the Y-variable is categorical with more than two categories.

(4) A Poisson loglinear model, if the Y-variable is a count.

(5) A two-part model, if the Y-variable is mixed (i.e., semi-continuous), where logistic regression is used to model the zero/non-zero status for Y, and normal linear regression is used to model the value of Y conditional upon its being non-zero.

A similar SRMI software package is the "MICE" library in R (van Buuren et al. 1999). We use IVEware for multiple imputation in this work because the CanCORS SCC stores and processes survey data in SAS format.

The advantage of adopting the SRMI strategy in this work is to meet the requirement of *practical objectivity and generality* (Meng 1995), meaning that the imputation model should not be in serious conflict with common analytic models used for analyzing the data files. For example, the imputation scheme for one variable is conditional upon all other variables in the datasets, as the associations among variables are of common interest to analysts. In addition, the above listed regression models are the typical ones used by analysts.

Because SRMI requires only the specification of individual regression models for each of the Y-variables, it does not necessarily imply a joint model for all of the Y-variables conditional on X. More discussion about such incompatibility between the set of conditional regression models and the joint model can be found in van Buuren et al. (2007). For simplicity, however, we treat the imputations produced from these conditional regression models as if from a single imputation "model" (or procedure).

3.2 Implementation of SRMI

3.2.1 Constructing a working dataset

It would be inefficient to impute different surveys separately. Our strategy is to concatenate all of the surveys to create a combined rectangular dataset and then to impute all surveyed variables simultaneously. This uses information about variable associations from surveys in which not all variables were asked and thus increases the effective sample size. The imputation is carried out separately for the samples of patients with lung and colorectal cancer since these groups of patients have different characteristics and the survey variables used for each type of cancer are slightly different.

The concatenation procedure causes structure missingness in the combined dataset for the variables which are not used in all surveys, a form of block missingness. These missing values may be of potential use if the analysis is targeted to a population that used multiple survey forms. For example, questions about patients' income were omitted from the brief survey but they are meaningful for the survey participants and could be imputed if the population for an analysis including income includes the brief survey patients. In this case, the massive imputation of income in the brief survey is largely dependent on the data and model derived from the full survey. On the other hand, values for which item nonresponse is due to skip patterns would generally not be included in analyses, unlike those coded with "don't know" or "refused", since skip patterns are designed to avoid collecting information that is not meaningful, such as the severity of chemotherapy side effects for a patient who did not have chemotherapy. However, IVEware is not able to exclude these complicated skip patterns from the imputation. Our strategy is first to impute both block and item nonresponses using IVEware. After imputation, missingness can be restored for the skip patterns, as well as block missing data, depending on the context of analyses. In addition, data editing procedures can be easily applied to the imputed data so that the logical patterns among survey variables can be retained.

Following IVEware's syntax, we classify all variables into four classes: categorical, continuous, mixed, and "transferred" (carried forward without modification). The variables involved in the imputation modeling (those which are imputed and/or act as predictors) fall into the first three categories, while the ones excluded from the imputation process, e.g. patient ID, are transferred. The categorical variables are nominal variables whose response levels do not have an obvious ordering. The continuous variables include both truly continuous variables and ordinal variables. The reason for treating the latter as continuous is that IVEware has no option for directly modeling ordinal variables, and if the data were treated as categorical (nominal), then the multinomial logistic regression would lose the ordering relationship and would be difficult to for more than a few response categories. The distribution of a mixed variable consists of a point mass at zero and a continuous positive part. For continuous variables, we force the imputations to fall within the ranges shown from the observed data, using the "bounds" option in IVEware. We round the fractional imputed numbers to the nearest integer after imputation to make their formats consistent with the original survey data.

3.2.2 Building imputation models

In the current implementation, we consider marginal effects of survey variables but not interactions. We include the indicators for survey types and PDCR sites as predictors to model different patients' characteristics across those factors.

The combined dataset contains around 5000 observations and around 800 variables for either cancer type and hence it is virtually impossible to include all survey variables in each prediction equation. IVEware has programming options for automatic model selection. First, we specify the maximum number of imputation predictors for each variable based on its number of observed values, as specified in Table 2. Our empirical experience show that if no limit is imposed on the number of predictors , IVEware may overfit models for variables with small numbers of observed cases. In addition, we set the minimum marginal R-square increment in the stepwise selection as 0.001, meaning that a variable will only be selected if the increase of R-square is greater than 0.001. Such criteria can be adjusted, depending on the purpose of analysis. With a smaller R-square criterion, more predictors will be selected, slowing the imputation computations.

Table 2: Limits on number of predictors (excluding the intercept)

No. of Observed Values	Maximum No. of Predictors
< 200	1
$(200, 300)$	2
$(300, 400)$	3
$(400, 500)$	4
> 500	5

There is no established rule for assessing the convergence of the Gibbs-like chain from SRMI; see van Buuren (2007) for more related discussion. With the aforementioned model selection procedures, most of the regression coefficient estimates and selection of predictions seem stable after running the program for several (3 to 5) iterations. This is consistent with empirical findings from other applications in Raghunathan et al. (2001) and van Buuren et al. (2006), where they showed that running the Gibbs-like chain for a few iterations appear to achieve stable imputations for SRMI.

4. Assessing the Adequacy of Imputation Models

Despite the popularity of multiple imputation in practical research, the literature on model assessment appears to be limited. Some studies use Monte Carlo simulation to evaluate the performance of imputation inferences. These typically create missing values for the original data (complete or incomplete), and evaluate the performance of imputation methods on inferences of some population quantities or pre-specified model parameters, using the before-deletion results as the yardstick. The evaluation procedures are applied to simulated rather than real datasets. On the other hand, PPC can be directly used for the dataset to which the model is applied. In this section, we outline the strategy of using PPC for imputation model assessment.

4.1 General concepts

Rubin (1984) proposed use of PPC, as a Bayesianly justifiable approach that monitors model fit while conditioning on observed data in simulating predictive values. Meng (1994) and Gelman et al. (1996) formally defined the posterior predictive P-value and illustrated the procedure through practical applications. Gelman et al. (2004, Chap. 6) gave more examples.

To evaluate the fit of a Bayesian model, the observed (complete) data Y can be compared to the posterior predictive distribution Y^{rep} using a test quantity Q, which can be a function of the unknown model parameters θ as well as data because it is evaluated over draws from the posterior distribution of both θ and Y. The posterior predictive (Bayesian) P-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by Q,

$$P_B = P(Q(Y^{rep}, \theta) \geq Q(Y, \theta)|Y) \tag{1}$$

where the probability is taken over the posterior distribution of θ and the posterior predictive distribution of Y^{rep}, that is, the joint distribution $P(Y^{rep}, \theta|Y)$:

$$P_B = \int\int I_{Q(Y^{rep},\theta)\geq Q(Y,\theta)} P(Y^{rep}|\theta)P(\theta|Y)dY^{rep}d\theta,$$

where I is the indicator function. This formula uses the property that $P(Y^{rep}, \theta|Y) = P(Y^{rep}|\theta, Y)P(\theta|Y) = P(Y^{rep}|\theta)P(\theta|Y)$.

The posterior predictive distribution of Y^{rep} can be computed by simulation. Given L draws from the posterior distribution of θ, we draw one Y^{rep} from the predictive distribution given each simulated θ; we now have L draws from the joint posterior distribution $P(Y^{rep}, \theta|Y)$. The PPC compares the realized test quantities $Q(Y, \theta^l)$ and the predictive test quantities, $Q(Y^{rep,l}, \theta^l)$. The estimated P-value is just the proportion of these L simulations for which the test quantities equals to or exceeds its realized value; that is, for which $Q(Y^{rep,l}, \theta^l) \geq Q(Y, \theta^l)$, $l = 1, \ldots, L$. A P-value that is too big or too small, e.g. $P_B > 0.95$ or < 0.05, indicates evidence of lack of fit.

4.2 PPC for multiple imputation

4.2.1 General strategy

If data are not fully observed, then the observed data Y_{obs} are characterized by (Y_{com}, R), where Y_{com} is the complete data and R is the missingness pattern matrix. Let θ denote the parameters involved in the complete-data model, $P(Y_{com}|\theta)$, and ϕ denote the parameters involved in the missingness model, $P(R|Y_{com}, \phi)$. Under the assumption of ignorable missingness (Rubin 1987), one can simulate replicates of the completed data Y_{com} without having to model the missing-data mechanism, since the inference for θ does not depend on the missingness model.

With missing data, the most general form of the test quantity is $Q(Y_{com}, R, \theta, \phi)$, the corresponding posterior predictive replication being $Q(Y_{com}^{rep}, R^{rep}, \theta, \phi)$. Gelman et al. (2005) suggested performing PPC for missing data using test quantities of the completed data, $Q(Y_{com})$, because (1) inferences under the complete-data model are often of the substantive interest; (2) Y_{com}^{rep} can be easily simulated under the ignorability assumption, while simulating Y_{obs}^{rep}, a deterministic function of Y_{com}^{rep} and R^{rep}, generally requires knowledge of the missingness mechanism, which is usually unknown or not of main interest.

4.2.2 PPC targeted to imputation analysis

Gelman et al. (2004, Chap. 6) gave several examples of PPC for complete data in which choices of Q include (1) common descriptive statistics for the data, such as means, variances, quantiles, and correlations; (2) summaries of model fit, such as χ^2 discrepancy; (3) graphs of residuals measuring discrepancies between the model and the data; and (4) features of the data not directly addressed by the probability model. For incomplete data, Gelman et al. (2005) and Abayomi et al. (2007) used complete-data graphs to detect lack of fit of complete-data models.

These types of testing functions are focused on the general fit of the model. To assess the adequacy of imputation models, however, we propose also including quantities of interest from the practitioners' analyses. The primary interest of analysts centers on completed-data inferences for such quantities rather than the general fit of the underlying imputation model. Hence the deviation between the inferences from the original and simulated completed data under the model, that is, $[Q(Y_{obs}^{rep}, Y_{mis}^{rep}) - Q(Y_{obs}, Y_{mis})|Y_{obs}]$, informs analysts about the effect of lack of fit of the model on the analytic inferences. For example, if a proposed analysis is to regress outcome Y on covariates X, then the analysis-specific Q's might include the regression coefficients, associated standard errors, t-statistics, and corresponding P-values. The implementation is a straightforward application of the analysis code to the simulated completed datasets.

4.2.3 Simulating predictive values using multiple imputation devices

The posterior predictive P-value for the completed data is

$$P_{B,com} = P(Q(Y_{obs}^{rep}, Y_{mis}^{rep}) \geq Q(Y_{obs}, Y_{mis})|Y_{obs}), \qquad (2)$$

where

$$P(Y_{obs}^{rep}, Y_{mis}^{rep}, Y_{mis}|Y_{obs}) \quad = \int P(Y_{obs}^{rep}, Y_{mis}^{rep}|\theta)P(Y_{mis}, \theta|Y_{obs})d\theta$$

Simulating Y_{com}^{rep} is based on the complete-data model as in (2). However, the exact algebraic forms of the model may not be of the primary interest to practitioners. We suggest using existing imputation packages to simulate Y_{com}^{rep} automatically. Suppose $Y = (Y_1, Y_2)$, where Y_1 includes incomplete variables for which the imputation model is of assessment interest. If Y_2 is complete, we can first create a duplicate dataset in which Y_2 is retained but the incomplete Y_1 is totally left out. Then we can concatenate the original and duplicate sets together as $(Y_{1,obs}, Y_{1,mis}, Y_2, Y_{1,obs}^{rep}, Y_{1,mis}^{rep}, Y_2)$ where $Y_{1,mis}$, $Y_{1,obs}^{rep}$, and $Y_{1,mis}^{rep}$ are unobserved. Independent imputations for the concatenated set produce $Y_{1,mis}^{l}$, $Y_{1,obs}^{rep,l}$, and $Y_{1,mis}^{rep,l}$, $(l = 1, \ldots, L)$, and can be easily implemented using existing imputation code for the original dataset. If Y_2 contains missing data as well, they can be "filled in" by multiple imputation prior to applying PPC to Y_1. Suppose Y_2 consists of $Y_{2,obs}$ and $Y_{2,mis}$, then

$$\begin{aligned} P_{B,com} &= P(Q(Y_{1,obs}^{rep}, Y_{1,mis}^{rep}) > Q(Y_{1,obs}, Y_{1,mis})|Y_{1,obs}, Y_{2,obs}) \qquad (3) \\ &= \int P(Q(Y_{1,obs}^{rep}, Y_{1,mis}^{rep}) > Q(Y_{1,obs}, Y_{1,mis})|Y_{1,obs}, Y_{2,obs}, Y_{2,mis}) \\ &\quad P(Y_{2,mis}|Y_{1,obs}, Y_{2,obs})dY_{2,mis} \\ &\approx \frac{1}{L}\sum_{l=1}^{L} P(Q(Y_{1,obs}^{rep}, Y_{1,mis}^{rep}) > Q(Y_{1,obs}, Y_{1,mis})|Y_{1,obs}, Y_{2,obs}, Y_{2,mis}^{l}) \end{aligned}$$

where $Y_{2,mis}^{l} \sim P(Y_{2,mis}|Y_{1,obs}, Y_{2,obs})$.

The calculation of $P_{B,com}$ in (3) also fits naturally into the SRMI framework, where the imputation process is decomposed into conditional imputations of partitions of the data.

4.2.4 Approximating the Bayesian P-value based on a modest number of imputations

Typical PPC procedures calculate P_B based on a large number of replicates, perhaps in the scale of thousands; see the examples of Gelman et al. (2004, Chap. 6). In principle, this could be done for a multiple imputation analysis, but it might be computationally intensive and requires storage of extensive simulated data, especially a problem for imputation of large-scale survey data. We propose to approximate P_B based on a modest number of imputations, using the same normal approximation typically used in multiple imputation (Rubin 1987, Chap. 3, Sec. 3).

We assume the normal approximation holds for the posterior predictive distribution of scalar Q, $[Q(Y^{rep})|Y] \dot{\sim} N(\mu_Q, \sigma_Q^2)$. For example, the normal assumption might hold well for generalized linear model regression parameters with large sample size. In some cases, a transformation of the quantities of interest improves the normality of Q, as with variances. The key idea is to estimate this approximation $[Q(Y^{rep})|Y]$ from L drawn values of $Q(Y^{rep,l})$. If $[Q(Y^{rep,l})|Y] \sim iidN(\mu_Q, \sigma_Q^2)$, then it is not difficult to show that $[Q(Y^{rep})|\{Q(Y^{rep,l})\}] \sim t_{L-1}(\frac{\sum_{l=1}^{L} Q(Y^{rep,l})}{L}, (1+\frac{1}{L})\frac{\sum(Q(Y^{rep,l}) - \frac{\sum_{l=1}^{L} Q(Y^{rep,l})}{L})^2}{L-1})$, and the P-value can be approximated based on the cumulative distribution function.

In the presence of incomplete data, we can construct a t distribution with the mean as the average of the $\{Q(Y_{com}^{rep,l}) - Q(Y_{obs}, Y_{mis}^{l})\}$, and the variance as $1 + \frac{1}{L}$ times the between-replicate variance of the differences, with $df = L - 1$. The corresponding P-value can be approximated by the probability of being nonnegative under the t distribution.

5. Release and Use of Multiply Imputed Datasets

The CanCORS SCC releases the multiply imputed patient survey data (5 imputations) as well as the original ones to the PDCR sites for analysis. The imputations are updated as raw data are updated periodically. In addition, analysts pose their questions and inputs regarding imputations via an analytic discussion forum. These feedback are considered and incorporated into the timely updates of the imputations.

6. Example

We selected a subset of CanCORS patient survey data to illustrate our methods. The substantive interest of this study is to identify patient characteristics and preferences that are associated with their decision to enroll for hospice care, which in general includes a broad array of palliative and support services for individuals with terminal illness. The cohort consists of all advanced lung cancer patients from CanCORS ($n = 2261$). Study variables are patients' hospice decision and variables that might be associated with it. Table 3 describes the sample. Although the missingness proportion for each variable is generally low, the complete-case analysis discards around 16% of the sample and is certainly suboptimal.

We applied three different multiple imputation (MI) methods as follows:

(1) MI based on multivariate normality among all variables, implemented using SAS PROC MI. This method is obviously inappropriate because it ignores the discreteness of nominal variables and treats their codes as ordinal.

(2) MI based on a general location model (Olken and Tate 1961) for variables of mixed type, implemented using the "mix" library in R. The general location model treats pdcr, race, marital_status, insurance, and english as nominal and the remaining variables as continuous; both the loglinear and conditional normal models include only the main effects.

(3) SRMI implemented using IVEware; it treats comorbidity, income, marital_status, agegroup as continuous variables and the others as categorical variables.

In all three methods, we rounded continuous imputed values to the nearest integer.

In addition to the descriptive statistics of all variables in the cohort, a major analysis of interest is to identify potential predictors for doctor's early discussion (within 4 months of diagnosis) of hospice use with patient (mddishsp). We ran a logistic regression for the outcome mddishsp; the predictors include all other variables in the cohort except hospice and comorbidity. Multiple imputation analyses from different methods yielded very similar results and they identified the same set of significant predictors at 0.05 level. This is not surprising because the proportions of incomplete data are quite low. The regression results show that early discussion of hospice is more likely for (1) Whites compared to Hispanics; (2) married than for divorced/seprated/never married; (3) those aged 81+ years than those 55 and under; (4) those who did not receive chemotherapy; (5) those who had not had a heart attack; (6) those who have depression or diabetes; (7) those who died within 1 year of diagnosis; (8) those at UCLA compared to those at VA.

Table 4: Number of $P_{B,com} < 0.05$ or $P_{B,com} > 0.95$

Methods	Mean	Std.	Logistic Coef.	Logistic SE.
Normal	11	14	7	31
Mix	11	13	3	33
SRMI	6	6	1	17

PPCs comparing the parameter estimates between the original and simulated completed data show evidence of lack of fit for each method. Table 4 lists the number of extreme P-values if the test quantities Q are chosen to be the means and standard deviations of incomplete variables, as well as regression coefficients and their associated standard errors in the logistic model. SRMI models appear to produce a better fit for the data than the other two methods because it has the fewest extreme P-values. On the other hand, imputation models that are shown not to fit, judged by the posterior P-value, might not lead to practically invalid inferences if the missingness proportion is low, as suggested by the similarity of the inferences across three methods in this case.

7. Future Research

In an ongoing imputation project for multimode, multiwave survey data from a multisite observational study of cancer care, we used SRMI to impute missing data with complicated patterns, and PPC to assess the adequacy of the imputation model.

Many unsolved methodological and empirical questions arise from the current work. We classify potential research topics into the following three main categories:

(1) Topics related to SRMI: (a) Enhancing the current SRMI software to allow more data types, such as ordinal variables; (b) Developing improved rules for model selection; (c) Incorporating informative prior distributions into the model to enhance predictive power and reduce reliance on variable selection, especially for variables with sparse observations; (d) Investigating the effect of model incoherence on imputation inferences; (e) Developing formal convergence criteria for the Gibbs-like chain of sequential imputation.

Table 3: Cohort sample information

Variable	Label and Classification	Missingness frequency
comorbidity	0=none, 1=mild, 2=moderate, 3=severe	0
mddishsp	1=hospice discussed, 0=no	0.57%
income	1= <20k, 2= 20-40k, 3=40-60k, 4=>60k	0.62%
gender	0=male, 1=female	0.22%
race	1=white, 2=black, 3=hispanic, 4=asian, 5=other	0.18%
english	1=yes, 0=no	0
education	1= less than high school, 2=high school/some college, 3=college degree or more	1.95%
marital_status	1= married/live with partner, 2=widowed, 3=divorced/separated, 4=never married	6.50%
mi	1=heart attack, 0=no	0.62%
chf	1=heart failure, 0=no	0.62%
stroke	1=stroke, 0=no	0.62%
lung_disease	1=lung disease, 0=no	0.62%
diabetes	1=diabetes, 0=no	0.62%
depression	1=depression, 0=no	0.62%
chemotherapy	1=chemo, 0=no	0.35%
insurance	1=medicare, 2=medicaid, 3=private, 4=other	7.39%
hospice	1=hospice used, 0=no	1.24%
agegroup	1=21-55 yrs, 2=56-60, 3=61-65, 4=66-70, 5=71-75, 6=76=80, 7=81+	2.43%
deceased	1=deceased within 1 yr of dx, 0=no	0
pdcr site/code	10=CRN, 20=NCCC, 30=UAB, 40=UCLA, 50=Iowa, 70=VA	0

(2) Topics related to PPC: (a) We note that in addition to the posterior P-values, PPC results might also provide other insights into the imputation inferences. For example, denote by $\hat{\beta}_{obs,mis}$ and $\hat{\beta}^{rep}_{obs,mis}$ as the average estimates from the original and replicated completed data across simulations, respectively. Also note that $\hat{\beta}^{rep}_{obs,mis}$ is the multiple imputation estimate under the model. Since $\hat{\beta}_{obs,mis}$ is obtained from data produced from a mixture of the true model (for the observed part) and the imputation model (for the imputed part), we might conjecture that $\hat{\beta}_{obs,mis} \approx (1 - f_{mis})\beta_{com} + f_{mis}\hat{\beta}^{rep}_{obs,mis}$, where f_{mis} stands for the fraction of missing information of the variable (Rubin 1987). Hence $\beta_{com} \approx \frac{\hat{\beta}_{obs,mis} - f_{mis}\hat{\beta}^{rep}_{obs,mis}}{1-f_{mis}}$, and the bias of $\hat{\beta}_{obs,mis} \approx \frac{f_{mis}}{1-f_{mis}}(\hat{\beta}^{rep}_{obs,mis} - \hat{\beta}_{obs,mis})$. We can also obtain the approximate credible interval of the bias based on the posterior simulations.

Others topics include (b) Investigating the effect of model uncongeniality (Meng 1995) on assessment results; (c) Devising procedures to incorporate/calibrate imputation model based on assessment results; (d) Exploring the strategy of using test quantities of observed data, $Q(Y^{rep}_{obs})$; (e) Devising practical procedures to implement PPC within the SRMI framework.

(3) Topics specific to CanCORS or similar datasets: (a) Combining information from multiple sources. Future CanCORS datasets will include patient data collected from medical records and administrative databases. We will need to adopt an imputation method that incorporate correlations of information from different sources; (b) Comparing the performance of nonresponse weighting and multiple imputation approaches to the block nonresponses; (c) Developing imputation procedures for missing data in scale questions (group of items that are commonly combined into a single scale-score before analysis, such as the SF-12 physical and mental functioning scales).

REFERENCES

Abayomi, K., Gelman, A.E., and Levy, M. (2007) “Diagnostics for multivariate imputations”, unpublished manuscript.

Ayanian, J.Z., Chrischilles, E.A., Fletcher, R.H., *et al.* (2003) ”Understanding cancer treatment and outcomes: the Cancer Care Outcomes Research and Surveillance Consortium”, *Journal of Clinical Oncology*, **22**, 2292-2296.

Barnard, J. and Meng, X.L. (1999), “Applications of multiple imputation in medical studies: from AIDS to NHANES”, *Statistical Methods in Medical Research*, **8**, 17-36.

Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004), *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A.E., Mechelen, I.V., Verbeke, G., Heitjan, D.F., and Meulders, M. (2005), “Multiple imputation for model checking: completed-data plots with missing and latent data”, *Biometrics*, **61**, 74-85.

Gelman, A.E., Meng, X.L., and Stern, H.S. (1996) “Posterior predictive assessment of model fitness via realized discrepancies (with discussion)”, *Statistical Sinica*, **6**, 733-807.

Gelman, A.E., King, G., Liu, C. (1998) “Not asked and not answered: multiple imputation for multiple surveys”, *Journal of the American Statistical Association*, **93**, 846-857.

Little, R. J. A. and Rubin D. B. (2002), *Statistical Analysis of Missing Data*, New York: Wiley.

Meng, X.L. (1994), “Posterior predictive P-values”, *Annals of Statistics*, **22**, 1142-1160.

Meng, X.L. (1995), “Multiple imputation with uncongenial sources of input (with discussion)”, *Statistical Science*, **10**, 538-573.

Schenker, N., Raghunathan, T.E., Chiu, P.L., Makuc, D.M., Zhang, G., and Cohen, A.J. (2006), “Multiple imputation for missing income data in the National Health Interview Survey”, *Journal of the American Statistical Association*, **101**, 924-933.

Schenker, N., Treiman, D.J., and Weidman, L. (1993) "Analyses of public use decennial census data with multiply imputed industry and occupation codes", *Applied Statistics*, **42**, 545-556.

Olkin, I. and Tate, R.F. (1961), "Multivariate correlation models with mixed discrete and continuous variables", *Annals of Mathematical Statistics*, **32**, 448-465.

Raghunathan, T. E., Lepkowski, J. M., VanHoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey Methodology*, **27**, 85-95.

Raghunathan, T.E. and Siscovick, D.S. (1996), "A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertnsives", *Applied Statistics*, **45**, 335-352.

Tu, X.M., Meng, X.L., and Pagano, M. (1993), "The AIDS epidemic: estimating the survival distribution after AIDS diagnosis from surveillance data", *Journal of the American Statistical Association*, **88**, 26-36.

Rubin, D.B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician", *Annals of Statistics*, **12**, 1151-1172.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D.B. (1996), "Multiple imputation after 18+ years (with discussion)", *Journal of the American Statistical Association*, **91**, 473-489.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

van Buuren S (2007). "Multiple imputation of discrete and continuous data by fully conditional specification", *Statistical Methods in Medical Research*, **16**, 219-242.

van Buuren S, Boshuizen, H.C., Knook, D.L. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis", *Statistics in Medicine*, **18**, 681-694

van Buuren S, Brand J.P.L., Groothuis-Oudshoorn, K., Rubin D.B. (2006). "Fully conditional specification in multivariate imputation", *Journal of Statistical Computation and Simulation*, **76**, 1049-1064.