9-7-2006

Multiple imputation - Review of theory, implementation and software

Ofer Harel
University of Connecticut, oharel@stat.uconn.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

STATISTICS IN MEDICINE
Statist. Med. 2005; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Multiple imputation Review of theory, implementation and software

Ofer Harel[†], X.H. (Andrew) Zhou[‡§]

[†] *Department of Statistics, University of Connecticut, 215 Glenbrook Road Unit 4120 Storrs, CT 06269-4120, USA*
[‡] *HSR&D Center of Excellence, VA Puget Sound Health Care System, 1660 South Columbian Way, 1/424, Seattle, WA 98108 USA*
[§] *Department of Biostatistics, School of Public Health, University of Washington, F600 Health Sciences, Box 357232, Seattle, WA 98195-7232, USA*

SUMMARY

Missing data is a common complication in data analysis. In many medical settings missing data can cause difficulties in estimation, precision and inference. Multiple imputation (MI) [1] is a simulation based approach to deal with incomplete data. Although there are many different methods to deal with incomplete data, MI has become one of the leading methods. Since the late 80's we observed a constant increase in the use and publication of MI related research. This tutorial does not attempt to cover all the material concerning MI, but rather provides an overview and combines together the theory behind MI, the implementation of MI, and discusses increasing possibilities of the use of MI using commercial and free software. We illustrate some of the major points using an example from an Alzheimer disease (AD) study. In this AD study, while clinical data are available for all subjects, postmortem data are only available for the subset of those who died and underwent autopsy. Analysis of incomplete data requires making unverifiable assumptions. These assumptions are discussed in detail in the text. Relevant S-Plus code is provided. Copyright © 2005 John Wiley & Sons, Ltd.

1. INTRODUCTION

Incomplete data occupy a central place in public health and clinical research. Almost every researcher has to deal with incomplete data from time to time, and some have to deal with them on a regular basis. There are several methods to deal with missing data, including ad-hoc methods such as case deletion, and mean substitution and more principled methods such as maximum likelihood methods, multiple imputation (MI), or others. All these methods have been developed in order to allow the researcher to make statistically valid inferences on parameters under study, but not all of them do. Since most data analysis methods and software were developed to handle complete data (rectangular), even a small amount of missing data can

*Correspondence to: Ofer Harel, Department of Statistics, University of Connecticut, 215 Glenbrook Road Unit 4120 Storrs, CT 06269-4120, USA

Contract/grant sponsor: Grant ; contract/grant number: 0101010101

Received 01 January 2005
Revised 01 January 2005

cause great harm (bias, inefficiency etc.). Therefore there is a need to consider the importance of the missing data issue [2].

This paper is concentrated on one of the many methods to deal with incomplete data sets, multiple imputation. New development of MI methodology has proliferated in the literature [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13].

Although Missing at random (MAR) is a non-testable assumption, it has been pointed out in the literature that we can get very close to MAR if we include enough variables in the imputation models ([4, pp. 27–28], [12, 14, 15]). In addition, efficient estimation with non-ignorable missing data requires good prior knowledge about the missing data mechanism due to the fact that the data contain no information about which non-ignorable models would be appropriate, and because the results would usually be sensitive to the assumed non-ignorable model. For these reasons, in this paper we decided to focus on ignorable models. We will discuss possible consequences of violating the ignorability assumption.

In order to illustrate the missing data procedure, we are using data collected by the National Alzheimer Coordinating Center (NACC). Since 1984 the center has maintained a cumulative database on subjects from approximately 30 National Institute of Aging (NIA)-funded Alzheimer disease (AD) centers. In our example, we are using an observational study of AD. While clinical data are available for all subjects (N=34,874), postmortem data are only available for the subset of those who died and underwent autopsy (N1=1536). In this example we are going to investigate the two neuropathological (NP) diagnostic criteria, NIA/Reagan (D1) and the Khachaturian criteria, the Consortium to Establish a Registry for Alzheimer Disease CERAD (D2), with the clinical dementia diagnosis of Alzheimers (T). In this data set some of the subjects' true status was verified ($V = 1$) while others were not. The importance of this example is the fact that only 4.5% have complete data, but if we use only those who received an autopsy, we will lose the majority of the data. Using MI allows us to use all the available data.

Our tutorial has three objectives:

1. To review key theoretical ideas formulating the basis of MI (section 2) and its implementation (section 3).
2. To provide a limited software availability list and the main purpose of each package, and to provide simple code which the reader will be able to use with minor modifications.
3. To illustrate by example the implementation of MI to deal with categorical missing data.

The remainder of this paper is arranged in the following manner: Section 2 provides an overall review of MI and more information about the assumptions and theory behind it. Section 3 introduces the implementation issues of the imputation stage. This section discusses imputation techniques, and different data types that will require different imputation models. Section 4 summarizes different types of commercial and free software. Section 5 demonstrates the application of MI using a data example. Section 6 is a summary of the important issues presented in the paper.

2. THE THEORY BEHIND MULTIPLE IMPUTATION

2.1. Overall review

Multiple imputation (MI) was designed for complex surveys that are used to create public-use data sets to be shared by many users. It has been proven that MI is very valuable in many other settings as well. The goal of MI is to provide valid inference in difficult scenarios, in which the data is incomplete, different users are using different models and analyses, and when the reasons for missing values are not known.

Many users are using public-use data sets. These users differ in their statistical proficiency, their computing power, objectives, and scientific questions. Most users only have access to a complete-data methodology and software. Data-base constructors have additional information about the data that can help in modeling the missing values. Therefore, it would be preferred if the missing data modeling was done by the data constructors and not by the users. However, in many cases, that is not possible.

One of the basic objectives in MI is to enable the user to use complete-data procedures. Several ad-hoc, single imputation procedures can achieve this objective. For example, "complete-case analysis," "available-case analysis," mean substitution and several other procedures will allow the user to use a complete-data procedure. But in many cases these procedures do not yield statistically valid results for the scientific estimands. A scientific estimand is a quantity of interest, which can be calculated for the population in mind, and does not change its value according to the study design (sample size, design, non-response, etc.). In our inference, we are interested in unbiased (or approximately unbiased) estimates with a nominal confidence interval. The most obvious limitation of single imputation is the underlying assumption that the imputed value is the true value. This limitation leads to underestimation of the variance, which affects confidence intervals and statistical tests.

Multiple imputation is a general method that incorporates the uncertainty into the imputation process. From an inferential point of view, one of the main reasons to use MI is the fact that the data-collection information, both observed and unobserved, can be incorporated into the imputation. MI is comprised of three stages: imputation stage, in which the missing data are imputed; analysis stage, in which each complete data set is analyzed using a complete-data technique; and the last stage, in which the results from the analysis are combined in order to yield a final result that combines the uncertainty in the data and the uncertainty due to missing values.

One has to remember that there might be different inputs for the imputation and analysis stages. Therefore, in the case in which the imputer has additional resources over the analyst (i.e. more variables to use), there is a case of uncongeniality. The inferential uncongeniality usually implies superiority of the MI inference in term of validity and efficiency. The quality of imputation is crucial, i.e. the imputation model should be as general and as objective as possible. Creating additional imputations will yield better results, and on occasion, will allow one to choose only a subset of the imputed data sets and not use all the imputed data sets in the same analysis. The other scenario, in which the analyst has a richer model, is not recommended (see [7, 12]).

Next, we are going to delve deeply into each of the steps.

2.2. Assumptions

2.2.1. Ignorability Let M be a set of random indicator variables that partitions Y_{com} into Y_{obs} and Y_{mis}. In general, M can be regarded as an array of the same size as Y_{com} containing 0 in every position where the corresponding element of Y_{com} is observed and 1 in every position where the element of Y_{com} is missing. We will refer to M as the missing indicator(s) or the "missingness." Based on the work of Rubin [16]; see also [2, 17, 18], missing data can be often categorized as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). The key distinction is whether the cause of the missingness is related directly to levels of the missing variable(s) (MNAR), or whether the missingness is due to other variables that are either irrelevant (MCAR), or measured and included in the statistical model (MAR).

Consider a Bayesian joint model for the complete data and the missingness,

$$P(Y_{com}, M, \theta, \phi) \;=\; P(Y_{com}|\theta)\, P(M|Y_{com}, \phi)\, P(\theta, \phi), \tag{1}$$

where ϕ represents the unknown parameters of the conditional distribution of M given Y_{com}. Ignorability requires two conditions. The first is that the joint prior distribution for θ and ϕ must factor into independent priors,

$$P(\theta, \phi) \;=\; P(\theta)\, P(\phi),$$

in which case θ and ϕ are said to be "distinct." In this case, the change of one parameter will not cause a change in the other parameter. The other condition is missing at random (MAR) in which the missingness distribution depends only on the observed data, and all the missingness information is contain in the observed part of the data. It states that

$$P(M|Y_{com}, \phi) \;=\; P(M|Y_{obs}, \phi)$$

at the actual values for M and Y_{obs} realized in the current data and for all ϕ [16]. It is easy to show that under MAR and distinctness, the predictive distribution for Y_{mis} given Y_{obs} and M,

$$\begin{aligned} P(Y_{mis}|Y_{obs}, M) \;&=\; P^{-1}(Y_{obs}, M) \\ &\quad \times \iint P(Y_{com}|\theta)\, P(M|Y_{com}, \phi)\, P(\theta, \phi)\, d\phi\, d\theta, \end{aligned} \tag{2}$$

is equal to its distribution $P(Y_{mis}|Y_{obs})$ given Y_{obs} alone. Rubin [16] argues that ignorability is the weakest general condition under which the distribution of M does not need to be taken into account when making likelihood-based or Bayesian inferences about θ. Without ignorability, multiple imputations would have to be drawn from (2).

2.2.2. Congeniality The validity of MI also rests on the relationship between Q and the parameters of (1). Let Q be the measure of interest, and U its variance. If Q is a function of θ, and if $\hat{Q}$ and U are approximately a complete-data posterior mean and variance,

$$\hat{Q} \;\approx\; E(Q|Y_{obs}, Y_{mis}), \tag{3}$$

$$U \;\approx\; V(Q|Y_{obs}, Y_{mis}), \tag{4}$$

then MI yields an approximate Bayesian inference for Q. Meng [13] calls this setting "congenial." In other words, in the case in which the imputation model and the analysis model are the same we will get congenial results, it is not always true if the models are different. MI yields valid inferences not only in congenial settings but in certain uncongenial ones as well—where the imputer's model (1) is more general (i.e. makes fewer assumptions) than the complete-data estimation method, or when the imputer's model makes additional assumptions that are well-founded. The properties of MI when the imputer's and analyst's models differ are also discussed by [3], [7] and [12].

2.2.3. Proper Imputations The validity of MI rests on the properties of the distribution from which the missing data $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$ were drawn. Rubin [1] describes conditions under which the approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu$$

holds, in a frequentist sense, over repetitions of the sampling non-response and imputation processes; imputations for which these conditions hold are said to be "proper." Rubin's definition, like many frequentist criteria, is useful for evaluating the properties of a given method but provides little guidance for one seeking to create such a method in practice. For this reason, Rubin recommends that imputations be created through a Bayesian process.

2.2.4. Implications of Violating the MAR assumption Although the distinction between Missing at random (MAR) and Missing not at random (MNAR) is based on a non-testable assumption, it has been pointed out in the literature that by including enough variables in the imputation model the MAR assumption becomes more plausible ([4, pp. 27–28], [12, 14, 15]). In addition, efficient estimation with non-ignorable missing data requires good prior knowledge about the missing data mechanism due to the fact that the data contain no information about which non-ignorable models would be appropriate, and because the results would usually be sensitive to the assumed non-ignorable model.

When the ignorability assumption does not hold, one needs to draw the imputation from the posterior distribution of the missing data given the observed and the missingness $P(Y_{mis}|Y_{obs}, M)$. In order to accomplish this task, one needs to model the joint distribution of the complete data and the missingness. There are several ways to construct non-ignorable models based on different factorizations of the joint distribution of the complete data and the missingness mechanism.

Selection models, which first appeared in the econometrics literature [19, 20] combine a model for the distribution of the complete data with a conditional model for the missingness given the data such that $P(Y, M|X) = P(Y|X)P(M|Y, X)$. The results from these models tend to be highly sensitive to departure from the assumptions about the shape of the complete-data population [2, Chap. 11]. For more information about likelihood and bayes estimation using selection models refer to [21], and [22].

Pattern-mixture models is another way to construct non-ignorable models. Little [23] coined the term for this alternative for selection models. Using this model, one first models the marginal distribution of the missingness, and then models the conditional distribution of the complete data given the missingness patterns. The population of the complete-data becomes a mixture of distributions, weighted by the probabilities of the missingness pattern. In this case, we factor the joint distribution as follows $P(Y, M|X) = P(M|X)P(Y|M, X)$. This requires

the researcher to specify a model for the missingness (in many cases it will be a bernoulli distribution or logistic regression). Then the data will be modeled such that each missing data pattern might behave differently.

Frailty models is another way to construct non-ignorable models. In this case random effects are are used to affect the dependence between the responses Y and the missing mechanism M. These models are also known as shared parameter models [24, 25]. In this cases the factorization of the joint model is $f(Y, M|X) = \int f(Y|X;\beta)f(M|X;\beta)dF(\beta|X)$.

2.3. Imputation

In MI [1], we first impute m independent versions of the missing data from the posterior predictive distribution $P(Y_{mis}|Y_{obs}, M)$ under a joint model for the complete data $Y_{com} = (Y_{obs}, Y_{mis})$ and M, where (Y_{obs}, Y_{mis}) is the observed and missing parts of the data, and M is the set of missingness indicators for Y_{com}. The missingness indicators are random variables that separate the complete data into those two parts. In special cases, we assume ignorability which allows the model of M to drop out; therefore, under the ignorability assumption, we can impute the missing values from $P(Y_{mis}|Y_{obs})$.

2.3.1. Drawing imputations In practice, MI are usually created by Bayesian rather than frequentist arguments. That is, they are typically drawn from a posterior predictive distribution for the missing data given the observed data. Let $P(Y_{com}|\theta)$ denote a model for the complete data with unknown parameter θ. The posterior predictive distribution for Y_{mis} is

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)\, P(\theta|Y_{obs})\, d\theta, \tag{5}$$

where

$$P(\theta|Y_{obs}) \propto P(\theta) \int P(Y_{obs}, Y_{mis}|\theta)\, dY_{mis} \tag{6}$$

is the observed-data posterior distribution for θ and $P(\theta)$ is the prior distribution. The right-hand-side of (5) suggests that MI may be drawn by repeating this two-step process for $j = 1, \ldots, m$: first, draw $\theta^{(j)}$ from $P(\theta|Y_{obs})$, given by (6); then draw $Y_{mis}^{(j)}$ from $P(Y_{mis}|Y_{obs}, \theta^{(j)})$. Rubin [1] demonstrates the method in examples where (6) is relatively simple. In more complex situations, special computational techniques such as Markov Chain Monte Carlo may be needed [4]. Software for generating MI is now available from a variety of commercial and non-commercial sources.

2.4. Analysis

Imputing the data results in m complete data sets. Using common complete-data methods we gather the estimates $(\hat{Q}^1, \hat{Q}^2, \ldots, \hat{Q}^m)$ and squared standard errors $(U^1, U^2, \ldots, U^m)$. This analysis will be equivalent to the analysis that would have been done if we had complete data. Only in this case, it will be done m times. Estimates we might consider are regressions coefficients, odds ratio, etc.

2.5. Combining

Rubin [1] develops the rules for combining the estimates and their standard errors. First, we are going to introduce the results for scalar quantity, then we will discuss the rules for multidimensional estimates.

2.5.1. Scalar estimates Rubin's rules proceed as follows: Let $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ denote the scalar estimate for Q that would be used if complete data were available, and let $U = U(Y_{obs}, Y_{mis})$ denote its variance estimate. We must assume that with complete data, tests and intervals based on the normal approximation

$$(\hat{Q} - Q)/\sqrt{U} \sim N(0, 1) \tag{7}$$

would be appropriate. With incomplete data set, we have random versions or imputed $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$ from which we calculate the imputed-data estimates $\hat{Q}^{(j)} = \hat{Q}(Y_{obs}, Y_{mis}^{(j)})$ and their estimated variances $U^{(j)} = U(Y_{obs}, Y_{mis}^{(j)})$, $j = 1, \ldots, m$. The overall estimate of Q is $\bar{Q} = m^{-1} \sum \hat{Q}^{(j)}$. To obtain a standard error for $\bar{Q}$, we calculate the between-imputation variance $B = (m-1)^{-1} \sum (\hat{Q}^{(j)} - \bar{Q})^2$ and $\bar{U} = m^{-1} \sum U^{(j)}$ the within-imputation variance. The estimated total variance is $T = (1 + m^{-1})B + \bar{U}$, and tests and confidence intervals are based on a Student's t approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu, \tag{8}$$

with degrees of freedom $\nu = (m-1)\left[\frac{T}{(1+m^{-1})B}\right]^2$.

2.5.2. Multidimensional estimates Let $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ be the estimate for Q, a $k \times 1$ vector of unknown parameters, and $U = U(Y_{obs}, Y_{mis})$ denote its covariance matrix. We must assume that with complete data, tests and intervals based on the normal approximation $(\hat{Q} - Q), \sim N_k(0, U)$, and that p-values are based on the multivariate Wald test.

The multivariate extension for the combining rules is: $\bar{Q} = m^{-1} \sum \hat{Q}^{(t)}$ for the overall estimate. The estimated total variance is $T = (1 + m^{-1})B + \bar{U}$, where the between-imputation variance is $B = (m-1)^{-1} \sum (\hat{Q}^{(t)} - \bar{Q})(\hat{Q}^{(t)} - \bar{Q})^T$ and the within-imputation variance is $\bar{U} = m^{-1} \sum U^{(t)}$.

For small m, the between-imputation covariance matrix is very noisy, and many times not fully ranked. This might cause problems in finding the reference distribution for the test statistic. Li et al. [26] get around this complication by assuming that the between- and within-imputation matrices are proportional to each other. This assumption implies that the rates of missing information for all the components of Q are the same. In that case, the variance estimate is $\tilde{T} = (1 - s_1)\bar{U}$, where $s_1 = \frac{(1+\frac{1}{m})tr(BU^{-1})}{k}$. It follows that the test statistic is

$$G_1 = \frac{(\bar{Q} - Q_0)^T \tilde{T}^{-1} (\bar{Q} - Q_0)}{k},$$

distributed F_{k,ν_1}, where the p-value for testing, $Q = Q_0$, is $p = P(F_{k,\nu_1} \geq G_1)$. The degrees of freedom for the test are $\nu_1 = 4 + (t-4)[1 + (1 - 2t^{-1})s_1^{-1}]^2$ when $t = k(m-1)$ is greater then 4, and $\nu_1 = t(1 + k^{-1})(1 + s_1^{-1})^2/2$ when t is less then or equal to four.

In many scenarios the researcher is interested in the p-values themselves. In this case, one might ask if it is possible to combine the p-values from the m imputed data sets in order to

get a final p-value taking into considerations the variability in the data and the fact that the data are incomplete. Li et al. [27] answer this exact question. Consider the m complete data Wald statistics: $g_W^{(t)} = (Q^{(t)} - Q_0)^T (U^{(t)})^{-1} (Q^{(t)} - Q_0)$ where $t = 1, 2, \ldots, m$. In this case the statistic is

$$G_2 = \frac{\bar{g}_W k^{-1} - (m+1)(m-1)^{-1} s_2}{1 + s_2},$$

with $\bar{l}_W = \frac{1}{m}\Sigma_{t=1}^m l_W^{(t)}$, the average of the statistics, and $s_2 = (1 + m^{-1})[\frac{1}{m-1}\Sigma_{t=1}^m (\sqrt{l_W^{(t)}} - \sqrt{\bar{l}_W})^2]$. The combined p-value for testing $Q = Q_0$ is: $p = P(F_{k,\nu_2} \geq G_2)$ with degrees of freedom $\nu_2 = k^{-3/m}(m-1)(1 + s_2^{-1})^2$. This procedure was developed partly by theoretical argument and partly via simulations. It was developed for $m = 3$ and it is best used with this choice. Li et al. [27] suggest using this method as a guide for the p-values. The p-value found using this method should be interpreted as a range of p-values from $0.5p$ to $2p$. For example, if we have found the p-value to be 0.03 one should look at it as if the p-value is in the interval $(0.015, 0.06)$.

Additional combining rules were established by Meng and Rubin [28] for combining likelihood ratio test (LRT) statistics, for testing $Q = Q(\xi)$ where ξ is a k dimensional parameter vector. In complete data cases, the LRT statistic will be $g_L = 2[l(\hat{\xi}|Y_{obs}, Y_{mis}) - l(\hat{\xi}_0|Y_{obs}, Y_{mis})]$ which is asymptotically chi-square under the null. Let $g_L^{(t)}$ be the LRT statistics from the tth imputation, where $(\hat{\xi}^{(t)}, \hat{\xi}_0^{(t)})$ are the maximizers of the likelihoods. Also, let $\bar{g}_L = \frac{1}{m}\Sigma_{t=1}^m g_L^{(t)}$, and $\bar{\xi} = \frac{1}{m}\Sigma_{t=1}^m \hat{\xi}^{(t)}$, $\bar{\xi}_0 = \frac{1}{m}\Sigma_{t=1}^m \hat{\xi}_0^{(t)}$. Finally let the average of LRT statistics evaluated at $\bar{\xi}, \bar{\xi}_0$ be $\tilde{g}_L = \frac{1}{m}\Sigma_{t=1}^m g_L(\bar{\xi}, \bar{\xi}_0|Y_{obs}, Y_{mis}^{(t)})$. The test statistic proposed by Meng and Rubin [28] is

$$G_3 = \frac{\tilde{g}_L}{k(1 + s_3)},$$

where $s_3 = \frac{m+1}{k(m-1)}(\bar{g}_L - \tilde{g}_L)$. The p-value that follows is $p = P(F_{k,\nu_3} \geq G_3)$ with degrees of freedom $\nu_3 = 4 + (t-4)[1 + (1 - 2t^{-1})s_3^{-1}]^2$ when $t > 4$ and $\nu_3 = t(1 + k^{-1})(1 + s_3^{-1})^2/2$ otherwise.

Little and Rubin [2] suggest that the first combination rules (result with G_1), which are the multivariate analogues of the scalar rules are the most accurate combining rules. For small m, the assumption of proportionality between the between- and within-variance components is required. When the complete data analysis does not produce a complete-data variance-covariance matrix, but only produce p-values. The two combining rules applies (G_2 and G_3). The procedure which is simpler to use is (G_2), but the more precise (asymptotically as precise as G_1), is G_3.

3. IMPUTATION IMPLEMENTATION

3.1. General principles

Drawing from the posterior distribution in the imputation stage of MI is the most complicated task. There are several algorithms that accomplish this function; some have been implemented into readily available software and some were not. The methods that have been implemented into readily available software are: data augmentation [4], sampling importance/resampling

(SIR; [1]), and Hot-deck ([2]). Bayesian approximation is one method that has not yet been implemented into readily available software.

3.2. Imputation

Some of the following procedures will result in proper imputation while others might not. In most cases, we would prefer to achieve proper imputations, but many times that is hard to accomplish. It has been shown that in the case of improper MI valid or approximately valid results can be archived. Therefore, there are times that non-proper procedure will be used.

3.2.1. Data Augmentation Data Augmentation (DA) [29] is a method based on Markov Chain Monte Carlo (MCMC) methodology. MCMC creates draws from probability distributions (f). Since these distributions are either hard to find, or do not have a closed form, MCMC is used to generate a sequence $\{X^{(1)}, X^{(2)}, \ldots, X^{(t)}, \ldots\}$ such that each X depends in some way on the previous one, and where the stationary distribution ($X^{(t)}$ as $t \to \infty$) is a draw from the target distribution function f.

Data augmentation is closely related to Gibbs sampling [30]. In DA one will partition the random vector x into two parts, such that $x = (y, z)$. In this case the joint distribution $P(x)$ is hard to simulate from, but it is easy to simulate from $P(y|z) = g(y|z)$ and from $P(z|y) = h(z|y)$. In most of the incomplete-data scenarios, the observed-data posterior distribution $P(\theta|Y_{obs})$ is intractable but after the data augmentation, the complete-data posterior distribution $P(\theta|Y_{obs}, Y_{mis})$ becomes much easier to handle.

The iterative procedure is as follows: for a current guess of the parameter $\theta^{(t)}$, draw values to replace the missing values from

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)}).$$

Then given $Y_{mis}^{(t+1)}$ we would draw new value for θ from the complete-data posterior distribution

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}).$$

Repeating the iterative procedure until stationary state will produce draws from $P(\theta, Y_{mis}|Y_{obs})$, which implies that at the end of the procedure one can draw both $P(Y_{mis}|Y_{obs})$ and $P(\theta|Y_{obs})$.

3.2.2. Hot-Deck Hot-deck refers to the computer cards matching of available donors for a non-respondent [2]. Most hot-deck procedures replace the missing values with values from similar responding units in the sample. The procedure in which we find the donors to provide information (imputed values) for incomplete records is different according to the particular technique used.

The matching process uses filter variables. Records are considered a match if they have the same values on the filter variables. Other hot-deck imputation methods use a distance function matching, or nearest neighbor imputation, in which non-respondents get the values of their closest neighbor.

3.2.3. Bayesian Approximation Bayesian approximation, or the approximate Bayesian Bootstrap, was described by Rubin and Schenker [31] and Rubin [1, pp. 123–124], and used in [32] and [33]. A single set of imputations is created by following steps:

- Choose your "most important" variable to impute, or the variable with the most amount of missing values.
- Bootstrap a sample from the complete data; fit a model predicting the missing variable using all other variables, and additional information about the model (if applicable).
- Temporarily fill all missing values and apply the previous stage to incomplete cases in order to compute predicted means.
- Match each incomplete case with m complete cases based on the distance of the predicted means.
- Randomly choose one of the matched complete cases and use this case to impute all variables of the incomplete case.

Repeating these steps m times will result in m sets of imputations.

3.2.4. Sampling Importance/Resampling This is a non-iterative procedure that allows the imputer to create draws from the posterior distribution on Y_{mis} in the restricted circumstances in which the number of imputations is limited and the rates of missing information is modest. The Sampling Importance/Resampling (SIR) [34] has an additional advantage, in that it can be applied even when the imputation task is intractable. To use this method, one needs a good approximation of the joint posterior distribution for (Y_{mis}, θ), for example, $\tilde{P}(Y_{mis}, \theta) = \tilde{P}(\theta|X, Y_{obs})\tilde{P}(Y_{mis}|X, Y_{obs}, \theta))$. In addition, one needs to have the importance ratios

$$Ratio(Y_{mis}, \theta) \propto \frac{P(Y|X, \theta)P(\theta)}{\tilde{P}(Y_{mis}, \theta|X, Y_{obs})},$$

for all possible (Y_{mis}, θ) at the observed (X, Y_{obs}), where $\tilde{P}(Y_{mis}, \theta|X, Y_{obs})$ is the joint posterior distribution approximation, and X is all observed covariates. When $\tilde{P}(Y_{mis}, \theta|X, Y_{obs}) = P(Y_{mis}, \theta|X, Y_{obs})$, we can say that the imputation task is tractable and only the distribution of θ should be approximated and the importance ratios do not depend on Y_{mis}.

Follow the next three steps for generating SIR imputations:

- Draw M values of (Y_{mis}, θ) from the approximate joint posterior distribution, where M is large relative to m.
- Calculate the importance ratios for each draw, $Ratio_1, \ldots, Ratio_M$.
- Draw m values of Y_{mis} with probability proportional to $Ratio_1, \ldots, Ratio_M$from the values in the first stage.

3.3. Specific data types

There are many ways to implement the imputation stage. The imputer will have to choose the most appropriate method and impute accordingly. The types of variables have an important role in the imputation procedure. The most common imputation model was designed for continuous variables under the assumption of multivariate normality. In some scenarios [1, pp. 166] the model can be determined with a close form and DA is not needed. In other scenarios a MCMC procedure is needed. To date there are implementations for the data conjoin categorical variables model, the mixed model for categorical and continuous variables, and the model for multi-level data. More detailed information can be found in [4].

3.3.1. Multivariate normal [4, 5] The most common analysis model for continuous multivariate data is a joint normal model. It is true that in practice the multivariate normal assumption does not always hold, but in many cases the normal model will still be very useful, even thought the data are not normal. There is no well grounded theory about the rules in which the normality assumption is crucial. But deviating from the assumption can still result in reasonable answers. If the amount of missing information is low, even if there is a great deviance from the assumption, the data imputed do not have much influence on the final results. Sometimes, transformations can be applied to make the data more normal.

The most common method to impute a univariate Y using a collection of predictors X is to use the normal linear regression model. In this case $Y_i \sim N(X_i\beta, \sigma^2)$, using very straight forward random generation from normal and chi-square distributions, one can impute the values of the missing $Y's$ in close form.

In the multivariate form, the Sweep operator [35] is needed. Many times in order to get a conjugate family, the normal inverted Wishart is used.

3.3.2. Categorical [4, 5] Both in the biomedical and social research fields, the use of categorical variables is common. In many cases it is possible to use the multivariate normal methods for the categorical data, but there are imputation procedures that apply to this type of data specifically. The imputation model will follow a multinomial model, both saturated, and log-linear model.

The advantage of the saturated multinomial model is that it allows three-way and higher associations between variables (while the multivariate normal allows only two-way interactions). The downfall of the saturated multinomial model is in the presence of many variables with many levels. In that case, the observed data cannot support such complexity, and a reduced model is necessary. The log-linear model does exactly that, which allows us flexible models for many types of data.

3.3.3. Mixed continuous and categorical [4] In most theoretical settings, the type of the variables does not matter. But usually, the models discussed are assumed to be either continuous or categorical, not many models join these two types together. In practice, however, we often see these two types of variables (continuous and categorical) as a part of one data set.

The general location model introduced by Olkin and Tate [36] combines the two types of variables by modelling the marginal distribution of the categorical part, and the conditional distribution of the continuous variables given the categorical ones. More specifically, consider that $X_1, X_2, \ldots, X_p$ denote a set of continuous variables and X represents the continuous part of the data. In addition, let $Z_1, Z_2, \ldots, Z_q$ denote a set of categorical variables where Z represents the categorical part of the data. The general location model will be defined by the marginal distribution of Z and the conditional distribution of X given Z.

In practice, for each cell of the theoretical contingency table, a multivariate normal model will be assigned. When there are too many cells, a restricted model will be used by using a log-linear model for the cell probabilities and a linear model for the within-cell means.

3.3.4. Panel data – Multi level data [37] When dealing with longitudinal data and/or clustered data, it is common to use the linear or generalized mixed effects models [38]. These models allow for the response to be missing but when some covariates are missing imputation

models can be the answer. The multivariate model is comprised of a conditional linear mixed model for each covariate, with fixed effects for all other covariates.

In this case our model can be represented as: $y_i = X_i\beta + Z_i b_i + \epsilon_i$ where X, Z are known covariates, β is the coefficient for the fixed terms and b is the coefficient for the random terms. Let $\epsilon_i \sim N(0, \Sigma)$, and $b \sim N(0, \Phi)$. Also, let $\theta = (\beta, \Sigma, \Phi)$, then the imputation procedure will follow drawing from these distributions:

$$b^{(t+1)} \sim P(b|Y_{obs}, Y_{mis}^{(t)}, \theta^{(t)}), \quad (9)$$

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t)}), \quad (10)$$

$$y_{mis}^{(t+1)} \sim P(y_{mis}|Y_{obs}, \theta^{(t+1)}), \quad (11)$$

4. SOFTWARE

There are many statistical packages that support multiple imputation. The packages can be separated into freeware and commercial packages. Some of the packages are more general while others can be considered specific packages. We are not trying to give a detailed comparison of the packages, but just hint at the available opportunities.

Most statistical software has attempted to deal with incomplete data. Some commercial packages capable of producing MI include:

- S-PLUS [39] – The missing data library uses both EM [40], and DA [29], where DA is usually used to generate the imputations. The library supports different models for multivariate normal (for continuous variables, see section 3.3.1), categorical variables (see section 3.3.2), and the conditional Gaussian for imputation involving both continuous and categorical variables (see section 3.3.3).
- SAS [41] – The MI and MIANALYZE procedures in SAS use regression methods and propensity scores for monotone missing data, and DA (MCMC) for arbitrary missing data. The MCMC statement supports only models for continuous variables (section 3.3.1).
- SOLAS [42] – This package uses predictive mean model and propensity scores to impute continuous variables (section 3.3.1). It uses discriminant models to impute binary and categorical variables (section 3.3.2).
- LISREL [43] – This is specialized software for structural equation models (SEM). This package uses maximum likelihood techniques and is not much used for MI.
- STATA [44] – Hot-deck imputation was implemented in 1999 by Mander and Clayton. Recently MI was implemented using the multiple imputation by chain equations (MICE) described by Van-Buuren and Oudshoorn [45]. MICE uses mainly regression imputation for the univariate imputation and regression switching [45, 46] for the multivariate imputation. The regression switching algorithm is an algorithm of the same type as the Gibbs sampler. (Look at MICE freeware for more information).

There are several stand alone freeware, and other functions and libraries that can be used in order to implement MI. Some of these freeware programs are:

- Stand alone NORM – [47] A stand alone Windows package that uses EM and DA for imputing under the multivariate normal assumption (section 3.3.1), which can be downloaded from the web site, http://www.stat.psu.edu/ jls/misoftwa.html.

- Norm, Cat, Mix – [4] Libraries for S-plus (4.0) for multivariate normal (section 3.3.1), categorical (section 3.3.2), and mix continuous and categorical data (section 3.3.3), which can be downloaded from the web site, http://www.stat.psu.edu/ jls/misoftwa.html.
  The following libraries have been written to allow them to be used using R. These routines use DA for the imputation stage.
- PAN – [37, 48] S-plus (4.0) library for panel data (section 3.3.4). This is the only routine that will use DA for multilevel type data and can be downloaded from the web site, http://www.stat.psu.edu/ jls/misoftwa.html.
- MICE – [45] Multiple Imputation by Chain Equations (MICE) is a library for S-plus or R. This package uses predictive mean matching and regression methods, logistic and polytomous regressions, and discriminant analysis for monotone missing data, and use MCMC for non-monotone data (sections 3.3.1–3.3.2). This package can be downloaded from the web site, http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm.
- IVEWARE – [49] Imputation and Variance Estimation software performs single or multiple imputations of missing values using the Sequential Regression Imputation Method described in [50]. This program works on idea similar to MICE, but allows normal linear regression, binary logistic regression, multinomial logit, and Poisson regression (sections 3.3.1–3.3.3). This package can be downloaded from the web site, http://www.isr.umich.edu/src/smp/ive/.
- EMCOV – [51] Uses the EM algorithm to obtain maximum likelihood estimates of means and covariances in the presence of missing data. It can also impute missing values. This package can be downloaded from the web site, http://methcenter.psu.edu/downloads/EMCOV.html.
- AMELIA – [52, 53] A software for Windows or Gauss, it works mainly with EM, but allows the use of MCMC as well. It allows the imputation of continuous and categorical variables (without the use of MCMC) can be downloaded from the web site, http://gking.harvard.edu/stats.shtml.
- WinLTA – [54, 55, 56] A stand alone software designed for latent class analysis. The new version will allow MI using DA and can be downloaded from the web site, http://methcenter.psu.edu/downloads/winlta.html.

All packages that implement MI will allow combining the results according to Rubin's rules. Some of these programs are reviewed and compared by Horton and Lipsitz [57].

5. EXAMPLE
Assessment of verification bias

5.1. Background

The National Alzheimer Coordinating Center (NACC) maintains a cumulative database on subjects from approximately 30 NIA-funded AD centers, since 1984. Our example comes from the NACC database, an observational study on Alzheimer disease (AD). While clinical data are available for all subjects (N=34,874), postmortem data are only available for the subset of those who died and underwent autopsy (N1=1536).

The NACC data arose not from a single standardized sampling of a population but from heterogeneous sampling strategies. Subjects may have been selected and enrolled in these

Table I. NACC Data

	CERAD	NIT	$T=1$	$T=0$	Sum
$V=1$	$D_2=1$	$D_1=1$	1028	112	1140
		$D_1=0$	59	27	86
	$D_2=0$	$D_1=1$	15	3	18
		$D_1=0$	149	143	292
$V=0$			27245	6093	33338
Sum			28496	6378	34874

centers for various reasons. The clinical diagnostic criteria are not standardized over centers, and although all centers are using the same criteria, it is open to a local (center) interpretation. The neuropathological (NP) diagnostic criteria are not the same criteria over time or across centers. The Khachaturian criteria and the Consortium to Establish a Registry for Alzheimer Disease (CERAD) criteria for the neuropathological assessment of Alzheimer disease (AD) emphasize senile or neuritic plaques, age, and clinical history. A newer scheme stressing topographic staging of neurofibrillary changes in addition to neuritic plaques has been proposed by the National Institute on Aging (NIA)-Reagan Institute Consensus Conference. This scheme assigns cases to high, intermediate, or low likelihood categories that the dementia is due to AD.

We illustrate how the use of different gold standards can affect the sensitivities and specificities of a single diagnostic test. In the data set we have a single AD diagnostic test, but due to many reasons we have two NP diagnostic criteria, considered as gold standards.

5.2. Settings

In this example we are going to investigate the two NP diagnostic criteria, NIA/Reagan (D1) and CERAD (D2), with the clinical measure being "Was the primary clinical dementia diagnosis Alzheimers at the last measurement" (T). It follows that some of the subjects' true status was verified ($V=1$) while others were not. The data is summarized in Table I

We can separate the data into two parts. First, when T, D_1 and D_2 are all observed ($V=1$), we can call it the observed data (part A). Second, when T is observed but $D_1\&D_2$ are missing ($V=0$), we refer to this as the missing part (B) (Table II). Let x_{ijk} be the count of subjects such that $D_1=i$, $D_2=j$, and $T=k$. Each complete data count x_{ijk} is a sum of two parts, $x_{ijk}=x^A_{ijk}+x^B_{ijk}$. Although x^A_{ijk} is totally observed, x^B_{ijk} is not, instead we observe only the marginal total $x^B_{++k}=x^B_{10k}+x^B_{01k}+x^B_{00k}+x^B_{11k}$. The observed data, $Y_{obs}=\{x^A_{ijk}, x^B_{++k} : i,j,k=0,1\}$, is represented in Table II.

5.3. Imputation

For the imputation stage we use DA in order to draw the missing values from their posterior distribution. As we have categorical data, we can regard the data as coming from a multinomial distribution, which allows us to use S-plus for the imputation (See Appendix for code). In

Table II. Data

	Gold Standard		Diagnostic test	
			$T=1$	$T=0$
$V=1$	$D_2=1$	$D_1=1$	x^A_{111}	x^A_{110}
		$D_1=0$	x^A_{101}	x^A_{100}
	$D_2=0$	$D_1=1$	x^A_{011}	x^A_{010}
		$D_1=0$	x^A_{001}	x^A_{000}
$V=0$			x^B_{++1}	x^B_{++0}

order to proceed with the data augmentation algorithm, let us choose the parameters for the prior Dirichlet distribution to be $\alpha = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)$, resulting in Jeffreys (flat/non-informative) prior. Therefore, our predictive distributions are as follows:

$$\begin{aligned}
(x^B_{ij1}, x^B_{ij1})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{++1}, (\theta_{111}/\theta_{++1}, \theta_{101}/\theta_{++1}, \theta_{011}/\theta_{++1}, \theta_{001}/\theta_{++1})),\\
(x^B_{ij0}, x^B_{ij0})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{++0}, (\theta_{110}/\theta_{++0}, \theta_{100}/\theta_{++0}, \theta_{010}/\theta_{++0}, \theta_{000}/\theta_{++0})),\\
\theta &\sim D(\alpha),\\
\theta|Y &\sim D(x_{111}+0.5, x_{101}+0.5, x_{011}+0.5, x_{001}+0.5,\\
&\qquad x_{110}+0.5, x_{100}+0.5, x_{010}+0.5, x_{000}+0.5),
\end{aligned}$$

where $x_{ijk} = x^A_{ijk} + x^B_{ijk}$ $i,j,k = 0,1$, and t is the number of iteration. Using S-plus 6.2 [39] we use MI ($m = 10$) to impute the missing values.

5.4. Analysis

When both diagnostic tests and true status are available for all subjects, estimation of the sensitivity and specificity confidence intervals is equivalent to estimating the confidence interval of a conditional binomial proportion (sensitivity and specificity for each gold standard separately). This estimation is not trivial due to the skewed nature of the binomial distribution, especially when the proportion is close to 0 or 1. Consider a random variable $X \sim Bin(n,p)$, the standard interval for p is the Wald interval in which $\hat{p} \pm \kappa\sqrt{n\hat{p}\hat{q}}$, where $\hat{p} = X/n$, $\hat{q} = 1-\hat{p}$ and κ is the $(1-\alpha/2)$ percentile of the standard normal distribution. When using MI, one can use several complete-data methods to estimate the sensitivities and specificities and use the different results as sensitivity analysis. Harel and Zhou [58] performed a simulation study comparing six different complete-data procedures and showed that one method was superior to all other complete-data methods. This method is:

Logit (Rubin) interval: Rubin and Schenker [59] suggested the use of the confidence interval for $\theta = \theta(p) = logit(p) = log(\frac{p}{1-p})$ under a normal approximation. Using a Bayesian argument with the Jeffreys prior distribution, we can show that the distribution of θ is approximately normal. Therefore, if $\hat{\theta}_X$ is the estimate of θ, it follows that $(\theta - \hat{\theta}_X) \sim N(0, V_X)$ where $-V_X^{-1}$ is the second derivative of the log posterior of θ evaluated at $\hat{\theta}_X$. It follows that $\hat{\theta}_X = logit(\tilde{p})$ where $\tilde{p} = \frac{X+1/2}{n+1}$, with $V_X = [(n+1)\tilde{p}(1-\tilde{p})]^{-1}$. Hence, the $100(1-\alpha)\%$ confidence interval

is

$$logit^{-1}\{logit(\tilde{p}) \pm \frac{\kappa}{\sqrt{(n+1)\tilde{p}(1-\tilde{p})}}\}. \quad (12)$$

For incomplete data sets arising from verification bias, we can not use estimation methods for binomial proportions to derive the sensitivity, specificity and their confidence intervals anymore. Begg and Greenes [60] proposed a bias correction method for estimating the sensitivity and specificity. One shortcoming of using this method occurs in the case in which there are two diagnostics tests, where one needs to estimate the sensitivity and specificity for each diagnostic test without using any information from the other test. In this case we will use the settings in Table III.

Begg-Greenes ($B\&G$) interval: Begg and Greenes [60] proposed a moment-based method to estimate sensitivity and specificity under the ignorability assumption for the verification process. If we follow the notation of Table IIIa , where we sum over one of the gold standard and consider only the other one. It follows that $B\&G$'s sensitivity estimate is

$$\hat{\pi}_{1BG} = \frac{(x^A_{11}n_1)/(x^A_{11}+x^A_{01})}{(x^A_{11}n_1)/(x^A_{11}+x^A_{01}) + (x^A_{10}n_2)/(x^A_{10}+x^A_{00})},$$

with variance

$$\hat{var}(\hat{\pi}_{1BG}) = (\hat{\pi}_{1BG}(1-\hat{\pi}_{1BG}))^2(\frac{n}{n_1n_2} + \frac{x^A_{01}}{x^A_{11}(x^A_{11}+x^A_{01})} + \frac{x^A_{00}}{x^A_{10}(x^A_{10}+x^A_{00})}),$$

and $B\&G$'s specificity estimate is

$$\hat{\pi}_{2BG} = \frac{(x^A_{00}n_2)/(x^A_{10}+x^A_{00})}{(x^A_{01}n_1)/(x^A_{11}+x^A_{01}) + (x^A_{00}n_2)/(x^A_{10}+x^A_{00})},$$

with variance

$$\hat{var}(\hat{\pi}_{2BG}) = (\hat{\pi}_{2BG}(1-\hat{\pi}_{2BG}))^2(\frac{n}{n_1n_2} + \frac{x^A_{11}}{x^A_{01}(x^A_{11}+x^A_{01})} + \frac{x^A_{10}}{x^A_{00}(x^A_{10}+x^A_{00})}).$$

Using this information, the $100(1-\alpha)\%$ confidence intervals for sensitivity and specificity will be

$$\hat{\pi}_{1BG} \pm \kappa\sqrt{\hat{var}(\hat{\pi}_{1BG})}, \quad (13)$$

$$\hat{\pi}_{2BG} \pm \kappa\sqrt{\hat{var}(\hat{\pi}_{2BG})}, \quad (14)$$

respectively.

5.5. Combining results

Using Rubin's rules, we combine the results of the $m = 10$ imputations to get a final answer that takes into consideration both the variability in the data, and the variability introduced due to the fact that the data were incomplete. The results are summarized in Tables IV and V. Table IV lays out the results for the sensitivities of the logit-based MI procedure and the $B\&G$ method. The results are the estimate (sensitivity), its standard error (SE), the upper and lower bound of the confidence interval, and the confidence interval length. The results are given for both NP diagnostic criteria, NIT/Regan (D_1) and CERAD (D_2). Table V lays out similar results for the specificities.

Table III. Data summary for one diagnostic test

a. aggregated data

		$T=1$	$T=0$
$V=1$	$D=1$	x^A_{11}	x^A_{10}
	$D=0$	x^A_{01}	x^A_{00}
$V=0$		x^B_{+1}	x^B_{+0}
Total		n_1	n_2

b. complete data

	$T=1$	$T=0$
$D=1$	x_{11}	x_{10}
$D=0$	x_{01}	x_{00}
Total	n_1	n_2

Table IV. Sensitivity - results

	Methods	Sensitivity	SE	upper	lower	dif
NIA-Regan	Logit-MI	0.8300	0.0092	0.8472	0.8113	0.0359
	B&G	0.9023	0.0065	0.9151	0.8894	0.0256
CERAD	Logit-MI	0.8677	0.0086	0.8836	0.8499	0.0338
	B&G	0.8884	0.0062	0.9006	0.8762	0.0245

Table V. Specificity - results

	Methods	Specificity	SE	upper	lower	dif
NIA-Regan	Logit-MI	0.5949	0.0261	0.6450	0.5426	0.1024
	B&G	0.4454	0.0200	0.4845	0.4062	0.0783
CERAD	Logit-MI	0.5070	0.0350	0.5756	0.4382	0.1373
	B&G	0.4666	0.0233	0.5123	0.4208	0.0915

5.6. Inference

In Harel and Zhou [58], we compared the performance of the logit-based MI and the *B&G* method and found that the logit-based MI outperformed the *B&G* method. Since the sample size in this example is quite large ($N = 34,874$), and we have proper MI, we can assume that the MI results are more appropriate than the *B&G* results. When looking at the first NP diagnostic criteria (D_1), we find that the *B&G* method is overestimating the sensitivity, while underestimating the specificity. On the other hand, when looking at the second NP diagnostic criteria (D_2), although the *B&G* method is still overestimating the sensitivity, and underestimating the specificity, the differences are much smaller. Overall, it seems that when using the NIA/REGAN NP diagnostic criteria (D_1), the diagnostic test (T) is less sensitive but more specific compared to the case in which one uses the CERAD NP diagnostic criteria (D_2). One must remember that the *B&G* method calculated the sensitivity and specificity for each diagnostic test separately, averaging over the other test.

6. SUMMARY

Missing data is a common and major complication in data analysis. The scope of the problem is very broad, and can cause inconvenience, bias, reduction of precision and loss of information. Many researchers in various medical and public health fields are becoming more aware of the need to deal with incomplete data sets in a methodical way. The volume of the research regarding missing data and multiple imputation increases constantly, but still there is a need for many more researchers to delve into missing data, multiple imputation methods and applications research.

There are three main objectives to this manuscript: The first objective is to review some key theoretical ideas forming the basis of MI and its implementation. The second one is to provide a limited software availability list detailing the main purpose of each package, to provide simple code which the reader will be able to use with minor modifications. The third one is to illustrate by example the practical implementations of MI, dealing with categorical missing data. As the space is limited and the amount of information is extensive, the information in this tutorial is introductory by nature. Some topics were introduced briefly, while other topics were not covered at all. Some of the topics we did not cover (and still have a great potential for interesting methodology) include using MI in large data sets, both in number of observations and in the number of variables, nonparametric and semi-parametric imputation, imputation evaluation and more. More researchers need to put their time and effort in the advancing of MI methodology.

Acknowledgements

We thank three referees for invaluable comments. We also thank Tom Koepsell and Walter A. Kukull for their helpful insight on verification bias in the NACC database. This work was supported in part by the grants: R01HL62567, AHRQ R01HS013105, and U01 AG16976.

Professor Xiao-Hua Zhou is presently a core investigator and senior biostatistician at the Northwest *HSR&D* Center of Excellence within the VA Puget Sound Health Care System. The views expressed in this article are those of the author and do not necessarily represent the views of the Department of Veteran Affairs.

APPENDIX
S-plus code

```
# Attach the S+MissingData library
library(missing)

# Define the functions that are going to be used throughout the program
logit_function(x){
    y_log(x/(1-x))
    y}
invlogit_function(x){
    y_exp(x)/(1+exp(x))
```

```
    y}
mi.inference_function(est,std.err,confidence=.95){
  qstar_est[[1]]
  for(i in 2:length(est)){qstar_cbind(qstar,est[[i]])}
  qbar_apply(qstar,1,mean)
  u_std.err[[1]]
  for(i in 2:length(std.err)){u_cbind(u,std.err[[i]])}
  dimnames(u)[[1]]_dimnames(qstar)[[1]]
  u_u^2
  ubar_apply(u,1,mean)
  bm_apply(qstar,1,var)
  m_dim(qstar)[2]
  tm_ubar+((1+(1/m))*bm)
  rem_(1+(1/m))*bm/ubar
  nu_(m-1)*(1+(1/rem))**2
  alpha_1-(1-confidence)/2
  low_qbar-qt(alpha,nu)*sqrt(tm)
  up_qbar+qt(alpha,nu)*sqrt(tm)
  pval_2*(1-pt(abs(qbar/sqrt(tm)),nu))
  fminf_(rem+2/(nu+3))/(rem+1)
  result_c(qbar,sqrt(tm),nu,pval,low,up,rem,fminf)
  result}

# Set the data
D1_rep(c("Positive", "Negative", "NA"),6)
D2_rep(rep(c("Positive", "Negative", "NA"),each=3),2)
Test_c(rep("Positive", 9), rep("Negative", 9))
count_c(1028,59,0,15,149,0,0,0,27245,112,27,0,3,143,0,0,0,6093)
data.grouped_data.frame(D1=factor(D1), D2=factor(D2), Test=factor(Test), count=count)
# set m - # of imputations
m_10
ka_qnorm(0.975,0,1)
# Generate m imputations under the saturated loglinear model
pre_preLoglin(data=data.grouped, margin= count~D1:D2:Test)
set.seed(1231)
data.imp_impLoglin(pre, margins = ~D1:D2:Test, prior=0.5,
    nimpute=m, control=list(niter=100))
# extract the imputed data
for (j in 1:m){
    tmp_miSubscript(data.imp,j)
    imptable_tmp[,4]
# sensitivities calculation -- Find the sensitivities for the 2 gold standards
                             For all the analysis methods you want to compare.
    x1[j]_imptable[8]+imptable[6]
    x2[j]_imptable[8]+imptable[7]
    n1[j]_imptable[5]+imptable[6]+imptable[7]+imptable[8]
```

```
    n2[j]_n1[j]
    Se1AC[j]_(x1[j]+0.5*ka^2)/(n1[j]+ka^2)
    Se2AC[j]_(x2[j]+0.5*ka^2)/(n2[j]+ka^2)
    Vse1AC[j]_Se1AC[j]*(1-Se1AC[j])/(n1[j]+ka^2)
    Vse2AC[j]_Se2AC[j]*(1-Se2AC[j])/(n2[j]+ka^2)
    Se1R[j]_logit((x1[j]+0.5)/(n1[j]+1))
    Se2R[j]_logit((x2[j]+0.5)/(n2[j]+1))
    Vse1R[j]_1/((n1[j]+1)*((x1[j]+0.5)/(n1[j]+1))*(1-((x1[j]+0.5)/(n1[j]+1))))
    Vse2R[j]_1/((n2[j]+1)*((x2[j]+0.5)/(n2[j]+1))*(1-((x2[j]+0.5)/(n2[j]+1))))
    Se1ZL[j]_(x1[j]+0.5)/(n1[j]+1)
    Se2ZL[j]_(x2[j]+0.5)/(n2[j]+1)
    Vse1ZL[j]_Se1ZL[j]*(1-Se1ZL[j])/n1[j]
    Vse2ZL[j]_Se2ZL[j]*(1-Se2ZL[j])/n2[j]
    Se1MI[j]_(x1[j]+0.5)/(n1[j]+1)
    Se2MI[j]_(x2[j]+0.5)/(n2[j]+1)
    Vse1MI[j]_Se1ZL[j]*(1-Se1ZL[j])/n1[j]
    Vse2MI[j]_Se2ZL[j]*(1-Se2ZL[j])/n2[j]
# Specificities calculations -- Find the specificities for the 2 gold standards
                                For all the analysis methods you want to compare.
similar to sensitivities
}

# Combine the estimates and SE's for all imputations -- Diagnostic test #1 (NIA/REGAN)
Se1AC.comb_mi.inference(Se1AC, sqrt(Vse1AC))
Sp1AC.comb_mi.inference(Sp1AC, sqrt(Vsp1AC))
Se1R.comb_mi.inference(Se1R, sqrt(Vse1R))
Sp1R.comb_mi.inference(Sp1R, sqrt(Vsp1R))
Se1ZL.comb_mi.inference(Se1ZL, sqrt(Vse1ZL))
Sp1ZL.comb_mi.inference(Sp1ZL, sqrt(Vsp1ZL))
Se1MI.comb_mi.inference(Se1MI, sqrt(Vse1MI))
Sp1MI.comb_mi.inference(Sp1MI, sqrt(Vsp1MI))
x1.dat_cbind(mean(x1),mean(n1),mean(x1/n1),mean(y1),mean(n3),mean(y1/n3))
#
Similarly for diagnostic test #2 (CERAD)
```

REFERENCES

1. D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* J. Wiley & Sons, New York, 1987.
2. R.J.A Little and D.B. Rubin. *Statistical Analysis with Missing Data.* J. Wiley & Sons, New York, 1987.
3. D.B. Rubin. Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91:473–489, 1996.
4. J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.
5. P.D. Allison. *Missing data.* Sage Publications Inc, 2002.
6. J.L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15, 1999.
7. J.L. Schafer. Multiple imputation in multivariate problems where the imputer's and analyst's models differ. *Statistica Neerlandica*, 2003. In press.

8. N. Schenker and A.H. Welsh. Asymptotic results for multiple imputation. *The Annals of Statistics*, 16:1550–1566, 1988.
9. J.M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000.
10. J. Barnard and D.B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86:948–955, 1999.
11. J.K. Kim. Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, 32(2):766–783, 2004.
12. L.M Collins, J.L Schafer, and C.M Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6:330–351, 2001.
13. X.L. Meng. Multiple imputation inference with uncongenial sources of input (with discussion). *Statistical Science*, 10:538–573, 1994.
14. H. Demirtas and J.L. Schafer. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22:2553–2575, 2003.
15. Donald B. Rubin, Hal S. Stern, and Vasja Vehovar. Handling "Don't know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90:822–828, 1995.
16. D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
17. P. Diggle and M.G. Kenward. Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43(1):49–93, 1994.
18. P. Diggle, K.-Y. Liang, and S.L. Zeger. *Analysis of longitudinal data*. Oxford University Press, 1994.
19. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492, 1976.
20. Takeshi Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24:3–61, 1984.
21. J.W. Hogan and N.M. Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16:259–272, 1997.
22. J.W. Hogan, J. Roy, and C. Korkontzelou. Tutorial in biostatistics handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455–1497, 2004.
23. Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.
24. Dean Follmann and Margaret Wu. An approximate generalized linear model with random effects for informative missing data (Corr: 97V53 p384). *Biometrics*, 51:151–168, 1995.
25. Paul S. Albert and Dean A. Follmann. Modeling repeated count data subject to informative dropout. *Biometrics*, 56(3):667–677, 2000.
26. K.H. Li, T.E. Raghunathan, and D.B. Rubin. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86:1065–1073, 1991.
27. K-H Li, X.L. Meng, T.E. Raghunathan, and D.B. Rubin. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1:65–92, 1991.
28. X.L. Meng and D.B. Rubin. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79:103–111, 1992.
29. M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.
30. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
31. D.B. Rubin and N. Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–374, 1986.
32. D.F. Heitjan and R.J.A. Little. Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40:13–29, 1991.
33. X.H. Zhou, G.J. Eckert, and W.M. Tierney. Multiple imputation in public health research. *Statistics in Medicine*, 20(9-1):1541–1549, 2001.
34. D.B. Rubin. Comment – a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fraction of missing information are modest: The sir algorithm. *Journal of American Statistical Association*, 82(398):543–546, 1987.
35. A.E. Beaton. The use of special matrix operations in statistical calculus. In *Research Bulletin RB–64–51*, Princton, NJ, 1964. Educational Testing Service.
36. I. Olkin and R.F. Tate. Multivariate correlation models with mixed discrete and continuous variables (Corr: V36 p343). *The Annals of Mathematical Statistics*, 32:448–465, 1961.
37. J.L. Schafer. *Multiple imputation with PAN*, 2000. Software.
38. N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.
39. J. Schimert, J.L. Schafer, T. Hesterberg, C. Fraley, and D. Clarkson. *Analyzing Missing Values in S-PLUS*. Insightful Corp., Seattle, WA, 2001.

40. A.P. Dempster, N.M. Laird, and D.B Rubin. Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.
41. Y. C. Yuan. Multiple imputation for missing data: Concepts and new developments. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, 2000. Paper 267.
42. SOLAS. *SOLAS*. Statistical Solutions Inc., Cork, Ireland, 2001. software.
43. K.G. Jreskog and D. Srbom. *LISREL 8: User's Reference Guide*. Scientific Software International, Chicago, 1996. software.
44. StataCorp. *Stata Statistical Software: Release 8*. StataCorp LP, College Station, TX, 2003. Software.
45. S. van Buuren and C.G.M. Oudshoorn. Flexible multivariate imputation by mice. *Leiden: TNO Preventie en Gezondheid*, TNO/VGZ/PG 99.054, 1999.
46. S. van Buuren, H.C. Boshuizen, and D.L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694, 1999.
47. J.L. Schafer. *NORM: Multiple imputation of incomplete multivariate data under a normal model, Version 2*. Department of Statistics, The Pennsylvania State University, University Park, PA, 1999.
48. J.L. Schafer. Imputation of missing covariates under a multivariate linear mixed model. Technical report, The Pennsylvania State University, 1997.
49. T.E. Raghunathan, P.W. Solenberger, and J. Van Hoewyk. *IVEware: Imputation and Variance Estimation Software Installation instruction and User Guide*. University of Michigan, Ann Arbor, MI, 2000. Software.
50. T.E. Raghunathan, J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, 2001.
51. J.W. Graham and S.M. Hofer. *EMCOV.EXE User's Guide (Computer program and manual)*. University of Southen California, Department of Prevention Research, Alhambra, CA, 1993. Software.
52. J. Honaker, A. Joseph, G. King, K. Schve, and N. Singh. *Amelia: A Program fo missing data (Windows version)*. Harvard University, Cambridge, MA, 2001. Software.
53. G. King, J. Honaker, A. Joseph, and K. Scheve. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95:49–69, 2001.
54. S.T. Lanza, L.M. Collins, J.L. Schafer, and B.P. Flaherty. Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods*, In press.
55. L.M Collins, B.P. Flaherty, S.L. Hyatt, and J.L. Schafer. *WinLTA user's guide part 1*. The Methodology Center, Penn State University, 1999. Software.
56. L.M Collins, S.L. Lanza, and J.L. Schafer. *WinLTA user's guide for data augmentation*. The Methodology Center, Penn State University, 2001. Software.
57. N.J. Horton and S.R. Lipsitz. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254, 2001.
58. O. Harel and X.H. Zhou. Multiple imputation for correcting for verification bias. *Statistics in Medicine*, 2006. In Press; http://www3.interscience.wiley.com/cgi-bin/fulltext/112315271/PDFSTART.
59. D.B. Rubin and N. Schenker. Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology*, pages 131–144, 1987.
60. C.B. Begg and R.A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39:207–215, 1983.