Statistica Neerlandica (2003) Vol. 57, nr. 1, pp. 36–45

A toolkit in SAS for the evaluation of multiple imputation methods

Jaap P.L. Brand*

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, USA

Stef van Buuren

Department of Statistics, TNO Prevention and Health, Leiden, The Netherlands

Karin Groothuis-Oudshoorn

Department of Statistics, TNO Prevention and Health, Leiden, The Netherlands

Edzard S. Gelsema†

Department of Medical Informatics, Rotterdam, The Netherlands

This paper outlines a strategy to validate multiple imputation methods. Rubin's criteria for proper multiple imputation are the point of departure. We describe a simulation method that yields insight into various aspects of bias and efficiency of the imputation process. We propose a new method for creating incomplete data under a general Missing At Random (MAR) mechanism. Software implementing the validation strategy is available as a SAS/IML module. The method is applied to investigate the behavior of polytomous regression imputation for categorical data.

Key Words and Phrases: multiple imputation, proper imputation, missing data mechanism, simulation.

1 Introduction

The occurrence of missing data is a pervasive problem in data analysis. Multiple imputation (MI) (cf., Rubin, 1987, 1996) is a method to reflect the additional variability in estimates due to missing values. With MI, for each missing data entry $m \geqslant 2$ values are imputed resulting in m completed data sets. These m completed data

*JaapBrand@ms.soph.uab.edu.

†Edzard Gelsema was Jaap Brand's supervisor when this work was performed at the Department of Medical Informatics, Erasmus University, Rotterdam as part of his PhD Research. In 2000 Edzard Gelsema died. That his name is included in the list of authors should be seen as a posthumous mark of honour.

sets are analyzed separately by the complete-data method of interest, and the m intermediate results of these analyses are pooled into one final result according to certain rules. Several methods exist for pooling the m completed data results (cf., Li *et al.* 1991a,b, Meng and Rubin, 1992).

The statistical properties of MI have been studied quite extensively. See for example Rubin (1987) and Schafer (1997). Nevertheless, there are situations in which one would like to evaluate the properties of MI in more detail. For example, one might be interested in the performance of MI under special missing data mechanisms (MDM), for special types of data, and for special types of imputation procedures. The ideas and tools described in this paper are meant to be useful for such an evaluation.

Our starting point is Rubin's concept of proper MI. Properness is a desirable property, since the results of an MI method can only be guaranteed to be valid if the imputations are generated by a proper method. On the other hand, it has been generally recognized that MI may also work well under improper methods (cf., Schafer, 1997). The tools presented in this paper also aid in studying how much improperness can be accepted before MI breaks down. Rubin's conditions for properness are formulated in a general Bayesian context with $m = \infty$. We develop a slightly simplified formulation that allows for empirical verification. Properness depends on the statistic of interest: a method can be proper for one type of outcome measure and improper for another one. We therefore study properness on a range of measures, called target statistics.

We first introduce terminology, outline the properness conditions according to Rubin, and present our simplification of them. We then describe our general evaluation strategy. One specific problem is how to generate missing data. We introduce a new method to generate missing data under a class of MAR mechanisms and apply the validation strategy to polytomous regression methods for imputing categorical data.

2 Methods

2.1 *Terminology*

As in Rubin (1987), let Q be the quantity of interest. Complete data statistics are represented by $(\hat{Q}, U)$, where $\hat{Q}$ is a point estimate of a population parameter of interest Q, and U is the estimated variance-covariance matrix of this estimate. Suppose that multiple imputation creates m completed data sets. Let $\hat{Q}_{*i}$ be a point estimate of Q computed from the ith ($i = 1,\ldots,m$) imputed data set, and let $\bar{U}_{*i}$ be an estimate of the variance-covariance matrix of $\hat{Q}_{*i}$. The pooled result is represented by the triple $(\bar{Q}_m,\ \bar{U}_m,\ B_m)$. Here $\bar{Q}_m$ is the average of the m completed data estimates $\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}$, the within imputation variance $\bar{U}_m$ is the average over the m completed data variance-covariance matrices $U_{*1},\ldots,U_{*m}$, and the between imputation variance B_m is the variance-covariance matrix of the m completed data estimates $\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}$

(cf., RUBIN, 1987, p. 76). The assumption that makes multiple imputation work is that $\bar{Q}_m$ is normally distributed as $\bar{Q}_m \sim N(\hat{Q}, B)$, where B is the variance-covariance matrix of $\bar{Q}_\infty$ for an infinite number of imputations. Rubin shows that B is an asymptotically efficient estimator of $(1 + m^{-1})B_m$ for finite m.

2.2 *Proper imputation*

Randomization-valid analysis under the frequentist perspective can be guaranteed only if both the complete-data inference is randomization-valid, and if the multiple imputation procedure is proper (cf., RUBIN, 1987, p. 119). Under a fixed MDM, a multiple imputation procedure with m $= \infty$ is proper for the set of complete-data statistics $(\hat{Q}, U)$ if the following conditions are satisfied.

First, $\bar{Q}_\infty$ is an unbiased estimate of $\hat{Q}$ and normally distributed with a variance-covariance matrix B under the underlying MDM:

$$\bar{Q}_\infty \sim N(\hat{Q}, B). \tag{1}$$

Second, the between imputation variance B_∞ estimated by MI is approximately equal to the variance-covariance matrix B of $\bar{Q}_\infty$ under the underlying MDM:

$$B_\infty \approx B. \tag{2}$$

Third, the complete data variance $\bar{U}_\infty$ estimated by MI is approximately equal to the complete data variance U:

$$\bar{U}_\infty \approx U. \tag{3}$$

In these equations, the symbol $\approx$ indicates equality in the sense of lower order variability (cf., RUBIN, 1987). Finally, the variance-covariance matrix B of $\bar{Q}_\infty$ is stable under repeated sampling:

$$B \sim (B_0 \lll U_0). \tag{4}$$

In equation (4) B_0 is the expected value of the variance-covariance B under repeated sampling and U_0 is the true variance-covariance matrix of $\hat{Q}$ under sampling. This condition means that the B is distributed around B_0 and the variance-covariance matrix of this distribution is of lower order variance than U_0 in the sense of RUBIN (1987). It is difficult to empirically verify Rubin's conditions directly for a given set of data. Not only is m finite in practice, it is also not clear which criterion should be used to distinguish proper from improper cases in equations (2)–(4). In the sequel, we use a slightly simplified set of conditions:

$$E[\bar{Q}_m] = \hat{Q}, \tag{5}$$

$$E[\bar{U}_m] = U, \tag{6}$$

$$Var(\bar{Q}_m) = \left(1 + m^{-1}\right)E[B_m], \tag{7}$$

$$P\left(\bar{Q}_m - \left(\sqrt{(1+m^{-1})B_m}\right)t_{m-1;0.975} \leqslant \hat{Q} \right.$$
$$\left. \leqslant \bar{Q}_m + \left(\sqrt{(1+m^{-1})B_m}\right)t_{m-1;0.975}\right) = 0.95. \tag{8}$$

In these equations $E[.]$ and $Var[.]$ are the expectation and variance taken under repeatedly generation of incomplete data sets by the underlying MDM and subsequent application of multiple imputation. Equations (5) and (6), which are simplifications of the equations (1) and (3), require that $\bar{Q}_m$ and $\bar{U}_m$ are unbiased estimates of the complete data statistics $\hat{Q}$ and U, respectively. Equation (7), which simplifies condition (2), requires that $(1+m^{-1})B_m$ is an unbiased estimate of the variance of $\bar{Q}_m$. Finally, equation (8) states that the 95% confidence interval given by $\bar{Q}_m \pm (\sqrt{(1+m^{-1})B_m})t_{m-1;0.975}$, where $t_{m-1;0.975}$ is the 0.975 quantile of the Student-t distribution with $m-1$ degrees of freedom, has an actual coverage of at least 95%. This interval is based on the assumption that $\bar{Q}_m$ is normally distributed with a mean equal to $\hat{Q}$ and a variance equal to $(1+m^{-1})B_m$, where the estimate B_m of B has the same distribution as $(\chi^2_{m-1}/(m-1))B$, with χ^2_m a χ^2 random variable with $m-1$ degrees of freedom. Note that this is an interval for the complete statistic $\hat{Q}$ and not for the usual population parameter of interest Q. The observed variability in $\bar{Q}_m$ reflects only the uncertainty due to the missing information. We note that the obtained coverage in (8) should be interpreted with care. An actual coverage of 95% does not automatically imply properness, since a bias of $\bar{Q}_m$ with respect to $\hat{Q}$ in combination with an overestimation of the between imputation variance B may result in an actual coverage of 95% or more. Thus, the actual coverage measure needs to be supplemented by other information. In particular, we need to know whether $\bar{Q}_m$ is approximately unbiased with respect to $\hat{Q}$. Finally, equation (4) is a minor technical condition that primarily depends on the sample size, which will not be considered of major importance.

2.3 *Validation strategy*

Except for trivial cases, evaluation of properness of MI is analytically intractable, and is therefore best done via simulation. We have written SAS/IML software that assists in the actual calculations. Several aspects can be varied: the complete data set, the MDM, the amount of missing information, the imputation method, and so on. Properness can be established for one or more target statistics that measure aspects of univariate or bivariate distributions of the variables. Table 1 provides an overview of the target statistics that are currently implemented in the software.

The log odds-ratio is used if one or both variables are binary, otherwise Cramer's C is taken.

Figure 1 depicts the flow of the validation protocol. For a given complete data set, complete data target statistics $(\hat{Q}, U)$ are computed. The method generates N incomplete data sets applying a user-specified MDM to the complete data. The

Table 1. Target statistics implemented.

Uni- or bivariate	Type of measurement	Target statistic
Univariate	Continuous	Mean, quantiles
Univariate	Categorical	Proportions
Bivariate	2 continuous	Pearson correlation
Bivariate	1 continuous, 1 categorical	Conditional means
Bivariate	2 categorical	Log odds-ratio, Cramer's C

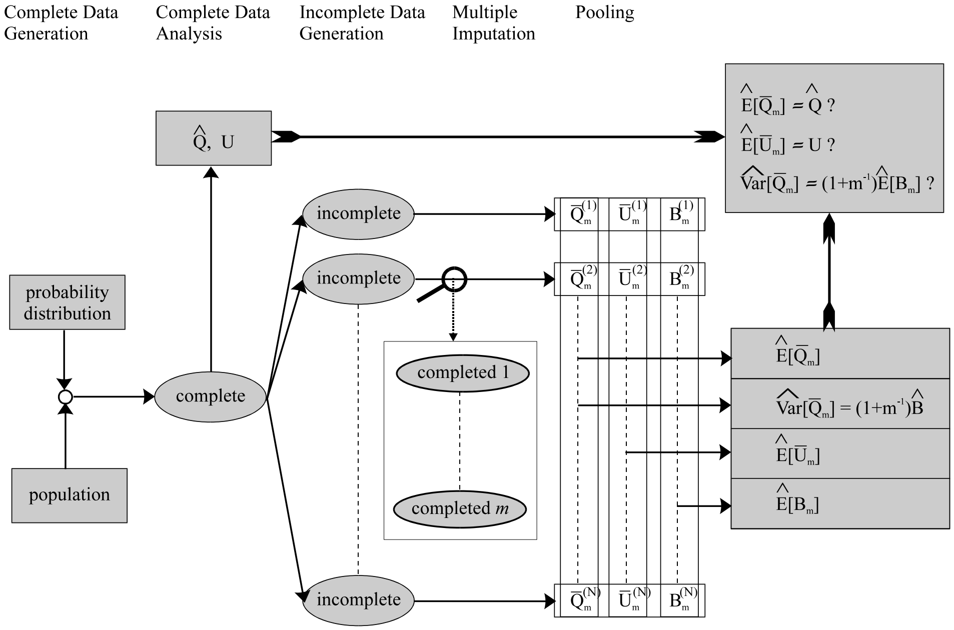
Complete Data Generation
Complete Data Analysis
Incomplete Data Generation
Multiple Imputation
Pooling
$\hat{Q}$, U
$\hat{E}[\bar{Q}_m] = \hat{Q}$?
$\hat{E}[\bar{U}_m] = U$?
$\widehat{Var}[\bar{Q}_m] = (1+m^{-1})\hat{E}[B_m]$?
probability distribution
population
complete
incomplete
incomplete
incomplete
completed 1
completed *m*
$\bar{Q}_m^{(1)}$ $\bar{U}_m^{(1)}$ $B_m^{(1)}$
$\bar{Q}_m^{(2)}$ $\bar{U}_m^{(2)}$ $B_m^{(2)}$
$\bar{Q}_m^{(N)}$ $\bar{U}_m^{(N)}$ $B_m^{(N)}$
$\hat{E}[\bar{Q}_m]$
$\widehat{Var}[\bar{Q}_m] = (1+m^{-1})\hat{B}$
$\hat{E}[\bar{U}_m]$
$\hat{E}[B_m]$

Fig. 1. A schematic overview of the validation protocol of multiple imputation.

number of replications, N, is typically chosen in the range of 200 to 1,000. Each incomplete data set is completed and analyzed by MI, resulting in solutions $(\bar{Q}_m^{(1)}, \bar{U}_m^{(1)}, B_m^{(1)}), \ldots, (\bar{Q}_m^{(N)}, \bar{U}_m^{(N)}, B_m^{(N)})$. Estimates for $E[\bar{Q}_m]$, $E[\bar{U}_m]$ and $E[B_m]$ are obtained by taking averages over the replications. $Var[\bar{Q}_m] = (1 + m^{-1})B$ is estimated by $V\hat{a}r(\bar{Q}_m)$ as the observed variance between all $\bar{Q}_m^{(i)}$, $i = 1, \ldots, N$. The coverage according to equation (8) is calculated as the percentage of confidence intervals including $\hat{Q}$.

If the imputation method is proper, we expect $\hat{E}[\bar{Q}_m] \approx \hat{Q}$, $\hat{E}[\bar{U}_m] \approx U$, $V\hat{a}r(\bar{Q}_m) = (1 + m^{-1})\hat{E}[B_m]$, and an actual coverage rate close to 95%. With $N = 500$, the rate's standard error due to simulation is approximately 1%.

Some combinations of validation criteria and target statistics are uninformative. For example, for proportions the complete data variance U is a function of $\hat{Q}$ and of the sample size n. For correlations, this variance depends on n only. The validation statistics U and $\bar{U}_m$ are therefore not reported for proportions and correlations. Confidence intervals for each target statistic were determined by the usual methods. The sampling distribution of the Cramer's C measure is intricate, and very skew for population values near -1 or $+1$. Coverage rates for Cramer's C are therefore not computed.

2.4 *A class of MAR MDM's*

This section describes the class of Missing at Random (MAR) MDM's that we implemented in the validation software. An MDM is called MAR if the probability that a data entry is missing depends on the observed data and is independent of the unobserved data. If, in addition, this probability is also independent of the observed data, then we speak of Missing Completely at Random (MCAR). MCAR is thus a special case of MAR. The important thing about MAR is that all information about the missing data is contained in the observed data, but structured in a way that complicates the analysis.

We start with a complete data matrix Z with n rows and q columns. Let z denote a row from Z. A missing data pattern (MDP) for z is a vector $r = (r_1,\ldots,r_q)$, where $r_i = 1$ if z_i is observed and $r_i = 0$ if z_i is missing ($i = 1,\ldots,q$). There exist 2^q possible MDP's for vector z. In the sequel we exclude the pattern $(1,\ldots,1)$ (i.e., all observations observed) and $(0,\ldots,0)$ (i.e., all observations missing) from the set of possible MDP's. The MDM is constructed in two steps. First, an MDP is assigned to each row z. Next, within the group of rows with identical MDP's, a certain proportion of the rows is made incomplete, where the probability may depend on the observed part of the row.

To be more specific, let parameter α be the expected fraction of incomplete cases, i.e., the units with at least one missing observation. Denote by $f(r) \in [0,1]$ the probability that z is a candidate for MDP r. The probabilities $f(r)$ are specified by the user and should sum up to 1 over the different MDP's. Note that $f(r) = 0$ implies that r is not a candidate MDP. Each case is assigned a specific MDP by taking a random draw from the distribution specified by $\alpha f(r)$. The expected number of incomplete cases for MDP r is thus equal to $\alpha n f(r)$.

Within the group of all rows with MDP r, missing data are created as follows. Choose $a(r) = (a_1(r),\ldots,a_q(r))$ as a vector of weights with q the number of MDP's considered, and define the linear combination $s(r) = \sum_{j=1}^{q} a_j(r) r_j z_j$. Note that $s(r)$ is, by construction, a linear combination of the observed data only. In addition, choose $k + 1$ quantile breakpoints $0 = \theta_0(r) < \theta_1(r) < \cdots < \theta_k(r) = 1$ and a vector $g(r) = (1, g_2(r),\ldots,g_k(r))$ with the relative risk of being missing with respect to the first (reference) quantile group. The cases with MDP r are then split into k groups $C_i(r)$, $i = 1,\ldots,k$, according to the quantiles of the linear combination $s(r)$ defined by the quantile levels $\theta_i(r)$, $i = 0,\ldots,k$. Next, the entries of row z are made incomplete with probability $P(z \quad \textit{incomplete} | z \in C_i(r)) = \alpha g_i(r) / (\sum_{j=1}^{k} (\theta_j(r) - \theta_{j-1}(r)) g_j(r))$. It is not difficult to show that this procedure generates an incomplete data matrix with an expected fraction of incomplete cases α.

The procedure is defined such that the data are MAR. The vector with weights $g(r)$ specifies how much the mechanism deviates from MCAR. When all the weights equal 1, the MDM is MCAR. We can specify $a(r)$ such that $s(r)$ will be highly correlated with the observed data. This provides a means to create substantial differences between the complete and incomplete cases.

3 Application

We now demonstrate the validation method for the case of one incomplete categorical variable. Using the method of Section 2.4, we define four different MDM's. The imputation method is based on polytomous regression. Our goal is to study how well imputation by polytomous regression performs under each of the mechanisms.

Data are from the Mammography Experience Study, as published in HOSMER and LEMESHOW (2000). This data set contains six responses from a survey of 412 women on knowledge, attitude and behavior towards mammography. The variable 'Mammographic experience' (ME) with response categories: $0 =$ never, $1 =$ during the past year, $2 =$ over one year ago is modeled by polytomous regression for ordinal data. This analysis results in estimated response probabilities per category. We first replace the original ME-variable by random draws based on these probabilities. This ensures that the polytomous regression model fits the data, and that complete-data issues will not affect the results.

Simulations are done using $N = 500$ replications. Every replication starts by generating 50% of missing data in the data under four different MAR MDM's, called MCAR, MARRIGHT, MARTAIL and MARMID. MDM's MCAR deletes observations in ME in a random fashion, MARRIGHT creates more missing values on the right side of the distribution, MARTAIL deletes more cases from both tails, while MARMID introduces more nonresponse in the center of the distribution.

The three category proportions of ME and the three averages of PB conditional on the categories of ME are considered as the six target statistics. The complete data statistic of interest $\hat{Q}$ and its variance U are estimated from the complete data with $n = 412$. For each replication, $i = 1,\ldots,500$, missing data are generated according to the specified mechanism, and multiple imputation with $m = 5$ is applied to the incomplete data.

Table 2 reports eight validation statistics: three about $\hat{Q}$, two about U, two about B, and one coverage estimate. An imputation method is considered proper if $\hat{E}[\bar{Q}_m]$ is close to $\hat{Q}$, if $\hat{E}[\bar{U}_m]$ is close to U, if $\hat{E}[B_m]$ is close $V\hat{a}r[\bar{Q}_m]$, and if the coverage coefficient is close to 95%. The column labeled $\hat{E}[\hat{Q}_{cc}]$ indicates the expected value obtained by complete case analysis (listwise deletion). Comparing $\hat{E}[\hat{Q}_{cc}]$ with $\hat{Q}$, we see that complete case analysis is severely biased under MARRIGHT, MARTAIL and MARMID. Multiple imputation, however, succeeds to "repair the damage". $\hat{E}[\bar{Q}_m]$ is often close to $\hat{Q}$. For example, under MARRIGHT the proportion

Table 2. Simulation results of the validation of the method for imputing categorical data by polytomous regression under several MDM's.

		Validation statistics							
Mechanism	Target statistic	$\hat{Q}$	$\hat{E}[\hat{Q}_{cc}]$	$\hat{E}[\bar{Q}_m]$	U	$\hat{E}[\bar{U}_m]$	$V\hat{a}r[\bar{Q}_m]$	$\hat{E}[B_m]$	95% cvg.
MCAR	P(ME = 0)	0.50	0.50	0.49			0.00	0.00	96.4
	P(ME = 1)	0.32	0.32	0.33			0.00	0.00	95.2
	P(ME = 2)	0.18	0.18	0.18			0.00	0.00	97.6
	E(PB|ME = 0)	8.13	8.12	8.11	0.02	0.02	0.01	0.01	95.2
	E(PB|ME = 1)	6.79	6.80	6.82	0.02	0.02	0.02	0.02	95.6
	E(PB|ME = 2)	7.34	7.34	7.37	0.06	0.06	0.05	0.05	94.8
MARRIGHT	P(ME = 0)	0.50	0.62	0.50			0.00	0.00	96.0
	P(ME = 1)	0.32	0.21	0.32			0.00	0.00	97.8
	P(ME = 2)	0.18	0.16	0.18			0.00	0.00	97.0
	E(PB|ME = 0)	8.13	8.55	8.07	0.02	0.02	0.01	0.01	93.8
	E(PB|ME = 1)	6.79	7.67	6.84	0.02	0.02	0.01	0.02	99.0
	E(PB|ME = 2)	7.34	8.19	7.47	0.06	0.06	0.05	0.05	95.8
MARTAIL	P(ME = 0)	0.50	0.44	0.49			0.00	0.00	92.8
	P(ME = 1)	0.32	0.38	0.33			0.00	0.00	94.8
	P(ME = 2)	0.18	0.18	0.18			0.00	0.00	96.2
	E(PB|ME = 0)	8.13	7.76	8.14	0.02	0.02	0.01	0.01	95.2
	E(PB|ME = 1)	6.79	6.45	6.83	0.02	0.02	0.02	0.02	95.6
	E(PB|ME = 2)	7.34	6.81	7.29	0.06	0.06	0.06	0.05	92.6
MARMID	P(ME = 0)	0.50	0.61	0.52			0.00	0.00	97.2
	P(ME = 1)	0.32	0.22	0.31			0.00	0.00	98.2
	P(ME = 2)	0.18	0.17	0.17			0.00	0.00	96.6
	E(PB|ME = 0)	8.13	8.59	7.99	0.02	0.02	0.01	0.02	94.2
	E(PB|ME = 1)	6.79	7.81	6.85	0.02	0.03	0.02	0.04	99.0
	E(PB|ME = 2)	7.34	8.34	7.62	0.06	0.06	0.06	0.09	93.8

P(ME = 0) is estimated by complete case analysis as 0.62, while the true proportion is 0.50. Thus, complete case analysis is biased, but multiple imputation is not. Apart from the trivial MCAR case, this finding is consistent across all mechanisms. Table 2 also shows that the within (U) and between (B) components of the variances are very close to the ideal values. MARMID appears to be the most difficult case. The between imputation variances $E[B_m]$ may be slightly too large, e.g., 0.09 instead of 0.06, but the actual coverages are generally very good. All in all, the results clearly demonstrate that imputing categorical data by polytomous regression is well behaved under the studied MDM's.

4 Conclusion

This paper describes a practical strategy for investigating the statistical properties of a given imputation method on a given data set. The method provides insight into the bias of the complete data target statistic $\hat{Q}$ under multiple imputation, the bias of the complete data within-imputation variability U, the bias of the between-imputation variability B, and the coverage of the 95% confidence interval under repeated imputation.

The validation strategy has been implemented in a SAS/IML module. It was used extensively to calibrate imputation methods in a project to build a missing data engine (cf., Brand *et al.*, 1994; van Buuren *et al.*, 1994; Brand, 1999). The software contains imputation methods and pooling methods, and is based on the same Gibbs sampling algorithm that is used in the S-Plus MICE library (van Buuren *et al.* 1999, 2000). One might thus also apply the software for the sole purpose of creating imputations under the Gibbs sampler. The properties of the Gibbs sampling imputation algorithm are published elsewhere.

We introduced a new method for creating missing entries in multivariate data according to a known MDM. We have limited the presentation to the class of MAR MDM's, but it is straightforward to generalize the method to include non-MAR MDM's. In fact, the software contains a routine to create MDM's that are not MAR.

Finally, this paper describes a method to study methods. We expect that in the coming years we will see many new and specialized imputation methods. We hope that our toolbox may assist in evaluating the properties of these methods.

Acknowledgements

This work was performed while the first author was at the Department of Medical Informatics, Erasmus University, Rotterdam and the Department of Statistics, TNO-PG, Leiden. The software is available from www.multiple-imputation.com.

References

Brand, J. P. L., S. van Buuren, E. M. van Mulligen, T. Timmers and E. S. Gelsema (1994), Multiple imputation as a missing data machine, in: J. G. Ozbolt (ed.), *Proceedings of the eighteenth annual symposium on computer applications in medical care (SCAMC)*, Hanley & Belfus, Inc., Philadelphia, 303–307.

Brand, J. P. L. (1999), *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*, Dissertation, Rotterdam, Erasmus University Rotterdam.

van Buuren, S., E. M. van Mulligen and J.P.L. Brand (1994), Routine multiple imputation in statistical databases, in: J. C. French and H. Hinterberger (eds), *Proceedings of the seventh international working conference on scientific and statistical database management*, IEEE Computer Society Press, Los Alamitos, 74–78.

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

van Buuren, S. and C. C. M. Oudshoorn (2000), *Multivariate imputation by chained equations:MICE V1.0 User's Manual. Report PG/VGZ/00.038*. Leiden: TNO Prevention and Health.Hosmer, D. W. and S. Lemeshow (1989), *Applied logistic regression*, Wiley, New York.

LI, K. H., T. E. RAGHUNATHAN and D. B. RUBIN (1991a), 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.

LI, K. H., T. E. RAGHUNATHAN and D. B. RUBIN (1991b), Significance levels from repeated p-values with multiply imputed data, *Statistica Sinica* **1**, 65–92.

MENG, X. L. and D.B. RUBIN (1992), Performing likelihood ratio tests with multiply imputed data sets, *Biometrika* **79**, 103–111.

RUBIN, D. B. (1987), *Multiple imputation for nonresponse in surveys*, Wiley, New York.

RUBIN, D. B. (1996), Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489.

SCHAFER, J. (1997), *Analysis of incomplete multivariate data*, Chapman and Hall, London.

Received: March 2002, Revised: August 2002.