MULTIPLE IMPUTATION TECHNIQUE: HANDLING MISSING DATA IN REAL WORLD HEALTH CARE RESEARCH

Rapeepong Suphanchaimat[1,2], Supon Limwattananon[1,3] and Weerasak Putthasri[1]

[1]International Health Policy Program (IHPP), Ministry of Public Health, Nonthaburi; [2]Banphai Hospital, Ban Phai; [3]Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen, Thailand

Abstract. The problem of missing data is ubiquitous in real world health care research. Its consequences include an introduction of bias and a loss of statistical power. Several methods to account for missing data are proposed. In recent literature, Multiple Imputation (MI) has become more widely used. This study therefore sought to exemplify and discuss the pros and cons of the MI application in relation to some conventional methods, which were Mean Substitution and Regression Imputation. Data from previously published article were used. The variables of interest were serum creatinine and age. Counterfactual dataset with missing data was generated and used to assess to what extent each method estimated the missing values close to the actual data. Under Missing At Random (MAR) assumption, the study presumed that age was related to the missingness of serum creatinine. Mean Substitution tended to produce biased estimate than other methods. Regression Imputation, although producing less biased estimate, did not account for data uncertainty. MI yielded mean and standard deviation estimates closest to the actual data, compared to other two methods. To sum up, MI seems to have advantages over Mean Substitution and Regression Imputation, as it can preserve data variability and also account for uncertainty of missing data. However, it is worth noting that, more importantly than which method is used, the researcher(s) should check thoroughly if the missingness mechanism is not missing not at random (MNAR); and such assessment needs firm background knowledge on the aims/objectives and the methodology of the study.

Keywords: age, introduction of bias, loss of statistical power, Mean Substitution, missingness mechanisms, Multiple Imputation, Regression Imputation, serum creatinine, Tenofovir

INTRODUCTION

Missing data are defined as records or observations that would be meaningful to the analysis and are supposed to be made, but for some reasons are not (Horton and Kleinman, 2007). It is almost a ubiquitous problem in most clinical and public health research. Because the primary goal of most analyses is to make valid inferences concerning a population of interest, missing data often undermine this goal in that they make the sample estimate different from the true population parameter.

Correspondence: Rapeepong Suphanchaimat, International Health Policy Program (IHPP), Ministry of Public Health, Nonthaburi 11000, Thailand.
Tel: +66 (0) 2590 2366; Fax: +66 (0) 2590 2385
E-mail: rapeepong@ihpp.thaigov.net

Table 1
Types of missing data.

Type	Characteristics
Missing completely at random (MCAR)	There is no systematic difference between the missing values and observed values; P(R=1 | X,Y) = P(R=1). For example, records of blood pressure were missed because of technical errors of sphygmomanometer.
Missing at random (MAR)	Any systematic difference between the missing values and the observed values can be explained by differences in observed data; P(R=1 | X,Y) = P(R=1 | X). For example, missing blood pressure might be lower than the observed ones because younger patients tended to be missed.
Missing not at random (MNAR)	This means the chance of seeing Y depends on Y, even after conditioning on X; equivalently, f(Y | X,R = 0) ≠ f(Y | X,R = 1). For example, patients with higher blood pressure were likely to miss the appointment due to high blood pressure caused headache.

Y, dependent variable; X, independent variable; P(R=1), probability function of being missing data; P(R=1 | X, Y), probability function of being missing data given X and Y; f(Y | X, R), distribution function of Y given X and R (Little and Rubin 2002; Sterne *et al*, 2009).

Aside from real world public health and epidemiological research, Fiero *et al* (2016) suggested that missing data are even present in most randomized controlled trials (RCTs). Hussain *et al* (2015) highlighted that missing data contributed not only the introduction of biased estimate but also the reduction of statistical power.

Despite the existing knowledge of potential biases derived from missing data, there is evidence that efforts to address missing data in practice remained suboptimal (Horton and Kleinman, 2007). Horton and Switzer (2005) found that of the 331 articles published in The New England Journal of Medicine during 18-month period, between 2004 and 2005, only 26 (8%) reported some way to manage missing data. Over half of the mentioned 26 papers applied an ad hoc imputation strategy (such as Mean Substitution).

Before delving into methods for handling missing data, it is important to understand mechanisms why data are missing. In modern research, missingness mechanisms are often described as falling into one of these three categories: 1) Missing Completely At Random (MCAR), 2) Missing At Random (MAR), and 3) Missing Not At Random (MNAR) (Table 1) (Little and Rubin, 2002; Sterne *et al*, 2009).

When MCAR exists, the analysis validity is not affected much. The only real penalty in failing to account for missing data is a loss of statistical power. MNAR is a more damaging situation, as it can be addressed only by a modification of study design. Given that MNAR exists, sensitivity analysis is recommended (Heraud-Bousquet *et al*, 2012). More realistic settings are MAR. If MAR is assumed, based on epidemiological and

clinical background knowledge, then unbiased and more statistically powerful analyses, relative to the analyses only on observed cases, can be done by including individuals with incomplete data (Sterne *et al*, 2009).

Although the best method for handling missing data is to prevent the problem by well planning of the study design and collecting the data more carefully, one almost always faces problems of missing data to some extent (Kang, 2013). In recent years, there have been many statistical methods developed to resolve any potential bias and a loss of power in the analysis of data with missing data. The methods ranged from fairly simple to advanced ones.

Examples of common approaches, although becoming less acceptable, are Complete Case Analysis (analysis only based on fully observed data), Adding Dummy Presenting Missing Data, and Mean Substitution (replacing missing data with mean value of observed variables). More advanced techniques are Regression Imputation (RI) (replacing missing data with predicted values derived from regression analysis done on fully observed data) and Multiple Imputation (MI) (replacing missing data with a 'set' of plausible values containing natural variability and uncertainty of the right values).

With the development of novel statistical software that can reduce calculating time, MI is recommended as useful method in producing unbiased estimates in novel clinical and public health study with missing data (given MAR is hold) (Dong and Peng, 2013; Dziura *et al*, 2013).

Although the above approach is theoretically accepted in the research arena, quite a few studies explored the application of MI in a practical setting. This study therefore aimed to exemplify, and to discuss the pros and cons of an application of MI, relative to some common methods for handling missing data, namely Mean Substitution and Regression Imputation.

MATERIALS AND METHODS

Data were retrieved from recent retrospective cohort study previously published in a peer reviewed domestic journal (Petchkum and Suphanchaimat, 2016). The study was exercised at the outpatient HIV clinic at Somdejprajaotaksin-Maharaj Provincial Hospital in Tak Province, with an aim to assess the association of Tenofovir (TDF) on nephrotoxicity. The total volume of participating individuals with complete collection of key variables, namely, age, sex, history of taking TDF, and baseline serum creatinine was 343. In addition to these variables, serum creatinine information at the 12th-month of a follow-up period was required. However, 15 patients did not show up at the hospital for the 12th-month appointment, making up the amount of missing data as 4.4%.

This study applied Complete Case Analysis and concluded that TDF created a risk of nephrotoxicity. However, one might argue that the estimate was subject to bias, or at least suffered from a loss of statistical power. Given that the study included only OPD HIV cases without renal insufficiency at the outset, it was unlikely that the missingness mechanism was MNAR (in other word, it was very unlikely that patients with high creatinine tended not to show up). Conversely, there is a study that suggested that more patients in the working-age group showed up at facilities than the older ones (Limwattananon *et al*, 2012); and this notion was likely in such setting, as the clinic was operated only during official hours when

Table 2
Key mathematical details.

- For each imputation, the parameter of interest, θ, is estimated and its standard error is recorded, for instance, θ = E(Y), the average value of Y. Let θ'_m and Var (θ'_m) denote the estimate of θ and its variance from m^{th} imputation.
- Supposed M denotes the number of imputations, the estimate of all imputations of θ is the average of the estimates from all imputed datasets: $\theta'_{MI} = (\sum_{m=1}^{M} \theta'_M)/M$.
- The within-imputation variance is given by $\sigma_w^2 = (\sum_{m=1}^{M} Var(\theta'_M)/M$. This quantifies uncertainty due to a finite sample.
- The between-imputation variance is given by $\sigma_b^2 = (\sum_{m=1}^{M} (\theta'_M - \theta'_{MI})^2/M$. This quantifies uncertainty due to missing data.
- The overall uncertainty in the estimate θ'_{M1} is given by $Var(\theta'_{MI}) = \sigma_w^2 + (1+1/M)\sigma_b^2$

some patients of working age might find difficulty to leave their job. In such circumstances, it was possible that the MAR assumption held. Therefore, the variables of interest in this article were 'Age' and '12^{th}-Month Serum Creatinine.'

Before estimating the values of missing data by different methods, it might be better to go back the fully observed cases and assess that, if some data were missed from the fully observed dataset, which method could best handle the missing data. Therefore, the analysis was done by the following steps.

In the first step, the relationship between having missing data and age was determined. In this step, logistic regression was applied (age as predictor variable, and being missing data as dependent variable with 1 if a record was missed, and 0 if otherwise).

In the second step, going back to complete cases (328 observations), using the information (logistic regression coefficient) from the first step to generate a new binary variable indicating whether or not the 12^{th} month serum creatinine information of each individual was likely to be missed (coded as 1 if the data was likely to be missed and 0 if otherwise). This step was done by inverse logit function. Note that the statistical software randomly generates the data based on the coefficient applied, and each round of generation might produce a slightly different result. Thus, the results shown in this article were just one of many simulations.

In the third step, each record with the binary variable '1' produced in Step 2 was treated as missing for its 12^{th}-month serum creatinine—as if counterfactual dataset was created.

In the fourth step, different methods, namely mean substitution, were used to estimate the value of 12^{th}-month serum creatinine, which was treated as missing in the third step.

In the fifth step, the estimated values were compared against the actual 12^{th}-month serum creatinine value to determine the most appropriate value estimating method.

Finally, returning to the entire dataset, the least biased method was then applied to estimate the value of 12^{th}-month creatinine of the 15 missing records.

STATA 12® (Stata Corporation: College Station, TX) was used for the analysis. Key mathematical details of the estimation appear in Table 2. The aforementioned steps are displayed in Fig 1.

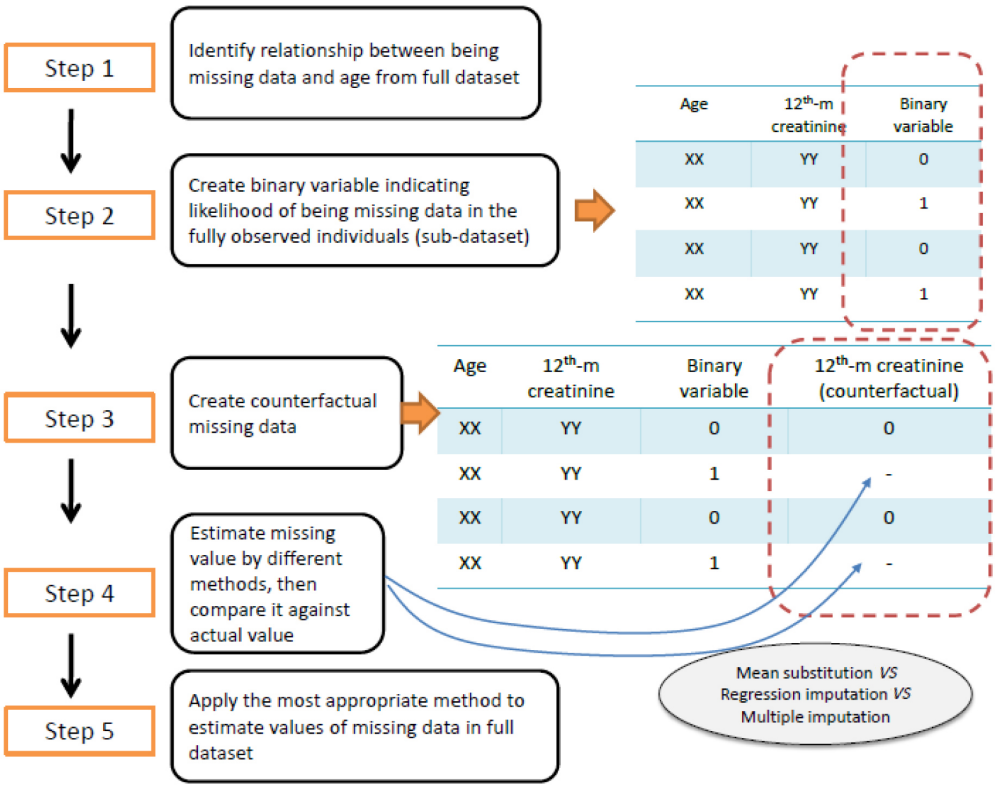

Fig 1–Analytical framework.

The research was approved and conducted according to the ethical standards of Somdejprajaotaksin-Maharaj Hospital.

RESULTS

At the outset, the total 343 observations were explored. The mean 12th-month serum creatinine (in mg/dl) was 0.846 (calculated from 328 observations), and mean age (in years) of the participants was 35.184. By dividing the dataset into sub-groups, namely, fully observed sub-group and missing data sub-group, it was found that the mean age of patients with fully observed 12th-month serum creatinine was 35.332, while the mean age of individuals with missing creatinine was 31.933. The logistic regression of being missing creatinine on age was performed. The negative coefficient of -0.048 reflected that the likelihood of having missing data tended to decline by additional age-year. Although the coefficient did not show strong statistical significance, it somehow exhibited the potential link between missing data and age, and therefore MAR was assumed (Table 3 and Table 4).

By focusing on fully observed sample, the coefficients in Table 4 were used to create counterfactual data based on inverse logit function. As a result, 23 missing variables were created. The counterfactual 12th-month serum creatinine amongst 305 observations had a mean of 0.851 and standard deviation of 0.238. In the next step, 3 different methods were applied to

Table 3
Age and 12th-month serum creatinine of the participants.

Variable	*n*	Mean	Standard deviation
12th-month serum creatinine (mg/dl)	328	0.846	0.236
Age (years)	343	35.184	8.422
Age in patients with fully observed creatinine (years)	328	35.322	8.402
Age in patients with missing creatinine (years)	15	31.933	8.481

Table 4
Logistic regression between being missing creatinine (Y) and age (X).

Variable	Coefficient	Standard error	*p*-value	95% Confidence interval
Age	-0.048	0.032	0.127	-0.110, 0.014
Constant term	-1.469	1.046	0.160	-3.518, 0.580

Table 5
Regression analysis of 12th-month serum creatinine (Y) on age (X)-R^2=0.063.

Variable	Coefficient	Standard error	*p*-value	95% Confidence interval
Age	0.007	0.002	<0.001	0.004, 0.010
Constant term	0.599	0.574	<0.001	0.486, 0.712

estimate value of these 23 missing data as per the following details.

For Method 1, mean substitution, all 23 missing values were replaced by the mean value of 0.851. Hence, this method yielded the mean creatinine of 328 observations of 0.851, and the standard deviation declined to 0.229 due to a decrease of data variability.

For Method 2, regression imputation, 12th-month serum creatinine was regressed on age. The results yielded regression coefficient of 0.007 with statistical significance, implying that serum creatinine in the elderly tended to be higher than the younger subjects (Table 5). In the next step, predicted value of missing creatinine was imputed (based on the aforementioned regression coefficient), making a final estimate of mean creatinine of 0.849, with standard deviation of 0.230.

For Method 3, MI was executed. The predicted values, so-called, imputes were substituted for the missing values, contributing to a full dataset, namely, the 'imputed dataset'. Each imputation was performed based on a regression of 12th-month serum creatinine of age, and this process was performed multiple times. While it is recommended to perform at least 50 datasets, this study increased the number of imputed datasets to 200 with an aim to have a more consistent estimate. Examples of imputed datasets are demon-

Table 6
Examples of imputed datasets.

ID	Actual creatinine	Counterfactual creatinine	Im_1	Im_2	Im_3	Im…	Im_199	Im_200
1	0.87	MISS	0.79	0.47	0.63	…	0.96	1.25
2	1.20	1.20	1.20	1.20	1.20	…	1.20	1.20
3	1.03	MISS	0.93	0.47	1.19	…	0.85	1.06
4	0.75	0.75	0.75	0.75	0.75	…	0.75	0.75
5	0.80	0.80	0.80	0.80	0.80	…	0.80	0.80
…	…	…	…	…	…	…	…	…
50	0.50	MISS	0.62	1.08	0.88	…	0.89	0.66
…	…	…	…	…	…	…	…	…
328	0.60	0.60	0.60	0.60	0.60	…	0.60	0.60

Im_*n*, Imputation round n^{th}…, Data not shown.
Missing creatinine in the counterfactual datasets were generated based on inverse logit function.

Table 7
Summary of the estimates of 12^{th}-month serum creatinine by different methods.

Data	Method	Number	Mean	Standard deviation
Actual data	Full sample	328	0.846	0.236
	Complete case analysis	305	0.851	0.238
Counterfactual data	Mean substitution	328	0.851	0.229
	Regression imputation	328	0.849	0.230
	Multiple imputation	328	0.849	0.238

Parameters (mean and standard deviation) obtained by multiple imputation were grand mean of parameter across all imputed datasets, not a single value (See Table 2 for mathematical notes).

strated in Table 6. The mathematical note of MI is shown in Table 2.

Overall, the imputed datasets led to the mean of 12^{th}-month creatinine of 0.849, with a variance of estimated mean of 0.000184 (taking into account both within- and between-imputation variances). The (mean of) standard deviation of all 200 imputed datasets was 0.238. The summary of the creatinine estimates by all methods above is presented in Table 7.

It is apparent that regression imputation and MI produced estimates closer to the actual value than mean substitution. Yet, MI was more likely to preserve data variability (as reflected by a closer standard deviation to the true standard deviation) than regression imputation. As a result, MI was selected and was used to handle missing data in the full dataset (N=343). This led to the final estimates of 12^{th}-month creatinine mean of 0.845 (with variance of estimated mean of 0.000169), and the standard deviation of the entire dataset of 0.236.

DISCUSSION

This study suggested, as described by recent literature, that MI has many advantages over single imputation methods

(such as Mean Substitution and Regression Imputation) for handling missing data (Sinharay *et al*, 2001; Azur *et al*, 2011; Kang, 2013). As presented in Table 7, the main disadvantage of Mean Substitution is that it adds no new information but only increases sample size, and this contributes to underestimate of the errors. Besides, should missing values not be strictly random, mean substitution will be subject to inconsistent bias (Malhotra, 1987).

For Regression Imputation, it has a key advantage over complete case analysis for its capability in retaining a great deal of data and in avoiding the alteration of the distribution shape of the data. Yet, similar to Mean Substitution, Regression Imputation does not account for uncertainty in the estimates and standard error of estimate tends to be reduced (Kang, 2013).

In contrast, MI produced mean estimate closest to the actual mean of the entire dataset. Moreover, it is still able to restore the natural variability of the missing values as it incorporates the uncertainty due to missing data (as noticed in Table 7 that mean standard deviation of imputed datasets was 0.238, very close to the standard deviation of actual data of 0.236). Kang (2013) underlined that the restoration of the natural variability of the missing data is achieved by replacing the missing data with the imputed values, which are predicted using the variables correlated with the missing data (In this scenario, it was age variable).

One may notice that there was little difference of estimates between methods in this study, and therefore a simple complete case analysis might be acceptable. The potential explanation for the trivial difference in the study results is a small volume of missing data. Graham (2009) also suggested that in circumstances where the number of missing data is less than 5%, the missingness can be assumed random, and hence bias from complete case analysis is ignorable.

The important caveat of using MI (and also other methods) is the presumption that the missingness mechanism is MAR, not MNAR. In this example, it is quite straight forward as the outcome of interest (serum creatinine) is associated with age, as supported by background clinical knowledge. Yet, in a setting where several variables are collected and the relationship between variables is more complex, it demands for meticulous checking for the underlying MAR assumption. In this regard, Carpenter *et al* (2007) recommended the weighting approach after MI in order to impute estimates under MNAR assumption (where the weights depend on the assumed degree of departure from MAR). Although, this study did not aim to investigate the weighting approach, the method is worth mentioning, as it becomes more and more accepted in the current public health and clinical research fields (Heraud-Bousquet *et al*, 2012).

One of the limitations of this article was that there are several unmentioned methods for managing missing data, for instance, Expectation Maximization (EM), last observation carried forward (LOCF) and maximum likelihood (ML). All of them have pros and cons and should be selected with caution.

Besides, as there was only one variable with missing data, the analysis in this case was quite straightforward. For more than one missing variable, particularly in case of an assumed joint distribution between variables, more complex analysis tool is needed. One of the suggested tools is multivariate imputation by chained

equations (MICE) where a battery of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data (van Buuren, 2007; Sepulveda *et al*, 2014).

Another important limitation is the unavailability of patients' characteristics data aside from age. For instance, domiciles and insurance status may affect the missingness (such as a patient living far from the facility might be less willing to have a follow-up visit). Should such a circumstance occur, it may imply that the missingness is linked to unobserved variables, and this means a violation of MAR assumption.

In conclusion, this article suggested an application of several methods in managing missing data in the real world health care research. MI seems to have advantages over Mean Substitution and Regression Imputation as it can preserve data variability and also account for uncertainty of missing data. However, it is worth noting that, more importantly than which method is used, the researcher(s) should check thoroughly if the missingness mechanism is MCAR or MAR, and such assessment needs firm background knowledge on the aims/objectives and the methodology of the study.

ACKNOWLEDGEMENTS

The authors would like to express our sincere thanks to Dr Porkaew Petchkum, the lead researcher of the TDF-nephrotoxicity project for data sharing conditioning upon academic purpose. Support from the IHPP staff was much appreciated.

REFERENCES

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; 20: 40-9.

Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res* 2007; 16: 259-75.

Dong Y, Peng C-YJ. Principled missing data methods for researchers. *SpringerPlus* 2013; 2: 222.

Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med* 2013; 86: 343-58.

Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016; 17: 1-10.

Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol* 2009 60: 549-76.

Heraud-Bousquet V, Larsen C, Carpenter J, Desenclos J-C, Le Strat Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Med Res Methodol* 2012; 12: 1-11.

Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am Statist* 2007; 61: 79-90.

Horton NJ, Switzer SS. Statistical methods in the journal. *N Engl J Med* 2005; 353: 1977-9.

Hussain JA, White IR, Langan D, *et al*. Missing data in randomized controlled trials testing palliative interventions pose a significant risk of bias and loss of power: a systematic review and meta-analyses. *J Clin Epidemiol* 2015; 74: 57-65.

Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol* 2013; 64: 402-6.

Limwattananon S, Neelsen S, Tangcharoensathien V, van Doorslaer E, O'Donnell O.

What does Universal Coverage do?: the impact on health care utilization and expenditure in Thailand. Nonthaburi: IHPP Ministry of Public Health, 2012.

Little R, Rubin D. Statistical analysis with missing data. 2nd ed. New York: Wiley, 2002.

Malhotra NK. Analyzing marketing research data with incomplete information on the dependent variable. *J Mark Res* 1987; 24: 74-84.

Petchkum P, Suphanchaimat R. Incidence and associated risk factors of nephrotoxicity due to Tenofovir in HIV-infected patients. *J Health Sci* 2016; 25: 1-12.

Sepulveda N, Manjurano A, Drakeley C, Clark TG. On the performance of multiple imputation based on chained equations in tackling missing data of the African alpha3.7 -globin deletion in a malaria association study. *Ann Hum Genet* 2014; 78: 277-89.

Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychol Methods* 2001; 6: 317-29.

Sterne JAC, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; 338: b2393.

van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007; 16: 219-42.