Long title: Multiple imputation in the presence of an incomplete binary variable created from an underlying continuous variable

Anneke C Grobler[1,2],*, Katherine Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, 50 Flemington road, Parkville, Victoria 3052, Australia
2. Department of Paediatrics, The University of Melbourne, Parkville, Victoria, Australia

Abstract

Multiple imputation is used to handle missing at random (MAR) data. Despite warnings from statisticians, continuous variables are often recoded into binary variables. With MI it is important that the imputation and analysis models are compatible; variables should be imputed in the same form they appear in the analysis model. With an encoded binary variable more accurate imputations may be obtained by imputing the underlying continuous variable. We conducted a simulation study to explore how best to impute a binary variable that was created from an underlying continuous variable. We generated a completely observed continuous outcome associated with an incomplete binary covariate that is a categorised version of an underlying continuous covariate, and an auxiliary variable associated with the underlying continuous covariate. We simulated data with several sample sizes, and set 25% and 50% of data in the covariate to MAR dependent on the outcome and the auxiliary variable. We compared the performance of five different imputation methods: 1) imputation of the binary variable using logistic regression; 2) imputation of the continuous variable using linear regression, then categorising into the binary variable; 3&4) imputation of both the continuous and binary variables using fully conditional specification (FCS) and multivariate normal imputation (MVNI); 5) substantive-model compatible (SMC) FCS. Bias and standard errors were large when the continuous variable only was imputed. The other methods performed adequately. Imputation of both the binary and continuous variables using FCS often encountered mathematical difficulties. We recommend the SMC-FCS method as it performed best in our simulation studies.

Key words: Binary variable; Compatibility; Fully conditional specification; Multiple Imputation; Multivariate Normal Imputation

*Email: anneke.grobler@mcri.edu.au

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/bimj.201900011.

1. Introduction

Public health and medical research commonly involves the converting of continuous variables into dichotomous variables for analysis. For example, body mass index (BMI) is often categorised into normal weight and obese, and blood pressure into normotensive or hypertensive (Royston, et al., 2006).

Missing data is a common problem in all medical research. Multiple imputation is one principled technique that addresses the uncertainty inherent in data analysis where data are missing and assumed to be missing at random (MAR) (Harel and Zhou, 2007; Rubin, 1987), that is when the probability of dropout is related to the observed data, but not unobserved data (Rubin, 1976).

Multiple imputation starts by specifying a joint distribution for the variables in the analysis model and other variables that are associated with the incomplete variables, known as the imputation model. This model is then fitted to the observed data to obtain initial estimates of the model parameters. The parameters of the imputation model are then updated by drawing a sample from their posterior distribution and these models are used to impute the missing values. This process is repeated to create multiple imputed datasets, which are then analysed using the analysis model as for complete case analysis. Finally, the estimates from the analysis models are then combined across the multiple datasets using the formulae suggested by Rubin (1987).

Since 1994, it has been highlighted that multiple imputation inference requires congeniality in order to be valid (Meng, 1994). Compatibility means that the imputation and analysis models can both be derived from a well-defined joint model for all the variables involved. The analysis model cannot include variables or relationships between variables, such as interactions, that were not included in the imputation model (Carpenter and Kenward, 2013; Enders, 2010; Harel and Zhou, 2007; Raghunathan, 2016). For a large number of imputed datasets, compatibility of the imputation procedure and the analysis model ensures that inference on multiple imputation data approximates the maximum likelihood procedure. Simplistically, the compatibility requirement means that variables should be imputed in the same form as they appear in the analysis model.

It is reasonable to belief that continuous variables contain more information about the underlying construct than binary variables (Royston, et al., 2006) potentially resulting in more accurate imputations, improved model performance and statistical inference. In addition, imputation of binary variables directly e.g. using logistic regression can encounter mathematical difficulties problems during the multiple imputation process. These two facts combined raises the question regarding how best to impute a binary variable of interest when it is derived from an underlying continuous variable.

The aim of this research was to explore how best to impute a binary variable that has been created from an underlying continuous variable when the binary variable is required in the analysis model. We focus on the setting where there is missing data in a key exposure variable or risk factor and we are interested in estimating the relationship with a completely observed continuous outcome. We compare 1) imputation of the binary variable using logistic regression; 2) imputation of the underlying continuous variable using linear regression, then categorising into the binary variable; 3 and 4) imputation of both the continuous and binary variables

using fully conditional specification (FCS) and multivariate normal imputation (MVNI), respectively using a simulation study and a real data example.

2. Methods

2.1 Notation

Let X_i denote the values of a continuous covariate, and Y_i , a completely observed continuous outcome variable, for subject i $(i = 1, \dots, n)$. X_i^{bin} is a binary variable derived from X_i using a specific cut-off; if $X_i > a$ then $X_i^{bin} = 1$, else $X_i^{bin} = 0$. W_i, Z_i and V_i denote the values of an auxiliary variable, and a continuous and binary covariate, respectively.

2.2 Simulation study

We conducted a simulation study to evaluate the performance of five possible multiple imputation methods that could be used in this context. In all simulations we simulated 1000 datasets with three different sample sizes (1000, 500, and 200) chosen to represent realistic sample sizes used in the literature. The number of simulations performed was based on the desired accuracy of the regression coefficient using the formula given in Burton, et al. (2006). If the variance was approximately 1 then 1000 simulations are required to produce an estimate to within 1% accuracy of the true regression coefficient of 6 with a significance level of 0.05.

2.3 Simulation of complete data

The binary covariate, V_i was created, with half the sample being in the first category and half the sample in the second category, e.g. a variable such as sex. The continuous covariate, Z_i was created as a random variable from a normal distribution with a mean of 80 and a standard deviation of 10. The continuous variable of interest X_i was then generated from a linear equation of the binary covariate, V_i, the continuous covariate, Z_i, and a normally distributed error term, thus for any individual:

$$X_i = \alpha V_i + \beta Z_i + e_i$$

We set $\alpha = 2$, $\beta = 0.25$ and generated e_i from a Normal(0,2) distribution, which represent strong but potentially realistic relationships between these variables. An auxiliary variable, W_i, was simulated to be associated with X_i, using

$$W_i = X_i + e_i$$

where e_i was generated from a Normal(0,7) distribution, in order to create a larger spread of values. In this simulation study, we deliberately generated the covariates and auxiliary variable to be associated with the continuous variable X_i as we believe this would more likely be the true relationship in practice and to ensure that there truly is "more information" to be gained in the continuous version of X_i during multiple imputation.

Next the binary version , X_i^{bin}, was generated in two scenarios; in the first scenario 15% of the dataset was set to be in the extreme category (this can be thought of as diseased/obese/high blood pressure) by using a cut-off

of 25 (corresponding to the cut-off for overweight on BMI). In the second scenario 30% of the dataset was set to the extreme category, using a cut-off of 19.

Finally, Y_i was simulated as

$$Y_i = \alpha + \beta_1 X_i^{bin} + \beta_2 V_i + \beta_3 Z_i + e_i$$

where $\alpha = 60$, $\beta_1 = 6$, $\beta_2 = 2$, $\beta_3 = 0.3$ and e_i is simulated from a $\text{Normal}(0,2)$ distribution, where the β_1 and β_2 were selected to be reasonably large but realistic and α was selected to provide a reasonable range for Y_i. In the data generation model we set Y_i to depend on X_i^{bin} rather than on X_i because we wanted to ensure that we were fitting the correct analysis model.

2.4 Imposing missing data

All variables were completely observed, with the exception of X and X_{bin} where missingness was imposed on the data. Three different missing data mechanisms were simulated. The first was assuming that missing data was missing completely at random (MCAR). In this scenario 25% of X_i and the corresponding X_i^{bin} variables were set to missing randomly, regardless of the values of any of the variables.

In the second and third scenarios data were set MAR, where the probability of missingness in X_i was determined by a logistic regression model dependent on Y_i and W_i

$$\text{logit}(\pi_{miss}) = \gamma + \beta_1 Y_i + \beta_2 W_i$$

β_1 and β_2 was set to 0.693, corresponding to an odds ratio of 2, a reasonably strong, but potentially realistic association with missingness and γ was set to ensure approximately 25% and 50% missingness, respectively. If X_i was missing, X_i^{bin} was also set to missing.

2.5 Multiple imputation models evaluated

We compare the performance of five different multiple imputation models, which can all easily be implemented in standard software. All analyses were conducted in `Stata`, version 14.2:

i) Imputation of the binary variable of interest using logistic regression. Using this approach, "mi impute logit" was used to impute the binary variable directly. This approach involves first creating the binary variable for the non-missing observations and then imputing the binary variable. This imputation model is compatible with the analysis model.

ii) Imputation of the underlying continuous variable using linear regression, then deriving the incomplete binary variable using the pre-specified cut-point to the imputed continuous variable. This was implemented using "mi impute regress" in Stata. This imputation model is not compatible with the analysis model.

iii) Imputation of both the continuous and binary variables using FCS, also known as multiple imputation using chained equations (MICE). Imputing two derivations of the same variable as separate variables in this way is referred to as "Just Another Variable" (JAV). In this method the binary variable is imputed using a logistic regression model and the underlying continuous variable is imputed using a linear

regression model in a sequential fashion. The model cycles through the variables, imputing missing values in each variable in turn using a model for the distribution of that variable conditional on all the other variables in the imputation model. This was implemented using "mi impute chained" in `Stata`. This imputation model is compatible with the analysis model.

iv) Imputation of both the continuous and binary variables using JAV with multivariate normal imputation (MVNI) where all the variables in the imputation model are assumed to be jointly normally distributed. This was implemented using "mi impute mvn". Following MI, the imputed X_i^{bin} variable was not categorised, but the non-integer values were used in the analysis. This method could also have been evaluated by imputing both variables using linear regression in FCS. This imputation model is not compatible with the analysis model.

v) Imputation using substantive model compatible (SMC) FCS which imputes missing values using a modified version of FCS where each partially observed covariate is imputed from an imputation model that is compatible with the analysis model (Bartlett, et al., 2015). In this method missing X_i is imputed from a joint model, then X_i^{bin} is derived as $X_i^{bin} = 1$ if X_i is larger than the specified cut-off. Rejection sampling is then used to draw a value which is bounded by a quantity depending on the substantive model. This imputation model is compatible with the analysis model.

For all five methods the imputation models included the outcome, the covariate and the auxiliary variables and we imputed 50 datasets. Note, the imputed values of X_i^{bin} and X_i will not necessarily be consistent with each other in the last two multivariable methods.

We also carried out a complete data analysis for each simulation scenario, before any data was set to missing. We also carried out a naïve complete case analysis for each scenario which included only cases who did not have missing data in X_i^{bin}.

The analysis model was an adjusted linear regression model for Y_i on X_i^{bin};

$$Y_i = \alpha + \beta_1 X_i^{bin} + \beta_2 V_i + \beta_3 Z_i + e_i$$

We compared inferences regarding the β_1 coefficient compared with the true value used to generate the data (i.e. 6). In our simulation study we investigated the bias in the β_1 coefficient estimation, the empirical standard error, the percentage error in the model based standard error using the empirical standard error as comparison and the coverage of the 95% confidence interval for β_1 calculated using Rubin's Rules.

2.6 Motivating example

The Longitudinal Study of Australian Children (LSAC) began in 2004 as a nationally-representative sample of 5107 0-1 year old Australian children followed every 2 years. The cross-sectional biophysical Child Heath CheckPoint module was nested between waves 6 and 7 of LSAC from February 2015 to March 2016 and assessed multiple physical health outcomes including BMI and sleep. Both LSAC (Edwards, 2014) and CheckPoint (Wake, et al., 2014) are described in more detail elsewhere.

There is strong evidence that short sleep duration is associated with obesity in both children and adults (Cappuccio, et al., 2008). We used the data collected in CheckPoint to investigate this association; with child BMI z-score as outcome and short sleep duration as predictor, adjusting for child sex and age.

BMI was calculated as kg/m^2 from measurements taken by trained staff and transformed into z-scores using CDC population normative data (Kuczmarski, et al., 2000). Objectively-measured sleep characteristics were collected using tri-axial, wrist-worn GENEActiv accelerometers worn on the child's non-dominant wrist for eight consecutive days. Sleep time characteristics were derived from raw accelerometer data, using self-reported records of bedtime and wake-time as a guide to locating sleep onset and offset. Data were processed using *Cobra* custom software. Sleep duration was calculated as the difference between sleep onset (the start of the first three consecutive minutes scored as sleep) and offset (the end of the last five consecutive minutes scored as sleep).

To illustrate the analysis when the binary category included 15% and 30%, respectively, short duration of sleep was defined using a cut-off at the 15th and 30th percentile of the observed data; 520 minutes (8.7 hours) and 545 minutes (9.1 hours), respectively. We applied the five methods tested in the simulation study to this dataset.

3. Results

3.1 Results of simulation study

[Table 1 about here]

For all scenarios, bias and percentage error in the standard error were large when the underlying continuous variable was imputed only, while the coverage of the 95% confidence interval was small. This was even the case when the data were MCAR (Table 1).

When imputation was conducted using logistic regression (Model i), imputed both continuous and binary variables using FCS (Model iii), or SMC FCS (Model v) all performed well. These were the scenarios where the imputation model was compatible with the analysis model. As expected these three methods performed better for larger sample sizes than for smaller sample sizes, with less missing data (25% vs 50%), and when more observations were in the diseased category (30% vs 15%). All three of these models had coverage of the 95% confidence interval close to 95%. In general SMC FCS performed better than the other approaches.
It is interesting to compare the performance of the two JAV methods. In most of the simulation studies the bias was larger when using MVNI than using FCS, while coverage was better for FCS than for MVNI. The exceptions were when the missingness was MCAR and the sample size was relatively small (n=200).

Imputing just the binary variable and imputing both variables using FCS had smaller bias than the complete case analysis in MAR settings, except when sample size was only 200. Imputing both variables using MVNI had similar or larger bias than the complete case analysis.

However, bias and coverage might not be the only considerations to take into account in practical settings when analysing real data. Some of the models experienced numerical difficulties during the imputation model fitting step such that `Stata` does not produce imputed values and parameter estimates due to prefect prediction of the binary exposure variable. As expected, this was more prevalent with smaller sample sizes, larger percentage of missing data and only occurred when 15% of the simulated subjects were in the diseased category. Table 2 summarises the percentage of models that had numerical difficulties for each of the imputation approaches in each of the simulation scenarios. In the most extreme scenario; small sample size (200), small number of subjects in the disease category (15%), and large amount of data missing (50%), 65.8% of the imputations had numerical difficulties with only two of the multiple imputation methods used. Of note, the SMC FCS models took a longer time to run than the other approaches, but did not experience numerical difficulties and could be implemented in all scenarios.

[Table 2 about here]

The easiest imputation model to fit, with small bias, namely to impute the binary variable using logistic regression was also the model most likely to have numerical problems. This can be an issue in practice if it is not possible to obtain estimates. The method where a continuous variable is imputed never had numerical problems, but had unacceptably high bias and low coverage.

3.2 Results of motivating example

Of the 1874 children who attended the CheckPoint assessments, 1371 had sleep data collected, thus 27% has missing data. Two children were excluded from this analysis because they did not have BMI measured. We used logistic regression to investigate whether missingness depended on any variables. In an exploratory analysis of the data, the following variables were found to be associated with missingness: sampling stratum, age, socioeconomic status, maternal smoking status.

For the scenario where 15% of children were classified as having sleep of short duration, the complete case analysis gave a regression coefficient of 0.21 (95% CI: 0.06-0.35; p-value=0.006); meaning that children who had short sleep duration had a 0.21 higher BMI z-score than children who sleep adequately. In the analysis where the binary variable is imputed using logistic regression the regression coefficient was 0.22 (95% CI: 0.07-0.36; p-value=0.003). In the analysis where the continuous variable is imputed using linear regression the regression coefficient was 0.18 (95% CI: 0.04-0.32; p-value=0.012). The results from the FCS imputation and MVNI were 0.14 (95% CI: -0.002-0.29; p-value=0.05) and 0.21 (95% CI:0.06-0.36 ; p-value=0.006); respectively. The results from the SMC FCS imputation was 0.21 (95% CI: 0.07-0.36; p-value=0.004). The results are fairly similar with all multiple imputation methods and the complete case analysis, except for FCS and imputing the continuous variable using linear regression (Figure 1).

[Figure 1: Results of multiple imputation using five different methods for the CheckPoint example. Regression coefficients with 95% confidence intervals for the association of BMI z-score with short sleep duration]

For the scenario where 30% of children were classified as having sleep of short duration the complete case regression coefficient was 0.12 (95% CI: 0.004-0.23; p-value=0.042). In the analysis where the binary variable is imputed using logistic regression the regression coefficient was 0.13 (95% CI: 0.01-0.24; p-value=0.028). In the analysis where the continuous variable is imputed using linear regression the regression coefficient was 0.12 (95% CI: 0.01-0.23; p-value=0.037). The results from the FCS imputation and MVNI were 0.08 (95% CI: -0.02-0.19; p-value=0.130) and 0.12 (95% CI: 0.004-0.23; p-value=0.042); respectively. The results from the SMC FCS imputation was 0.12 (95% CI: 0.00-0.24; p-value=0.049). The results are fairly similar with all multiple imputation methods and the complete case, except for FCS which had a smaller effect size (Figure 2).

4. Discussion

Through simulation we compared various approaches to impute an incomplete binary variable of interest which is derived from an underlying continuous variable. We showed that imputing the continuous variable and calculating the binary variable after imputation leads to severely biased results. Our results suggest that it is much better to impute the binary variable directly, even if intuitively this means throwing away potentially useful data. Another approach is to impute both the continuous and binary versions of the variable, using either FCS or MVNI or better still to use SMC FCS where each partially observed variable is imputed from an imputation model that is compatible with the analysis model

This finding that it is important to impute the variable in the same form as required for analysis is not novel. Since 1994 statisticians have recommended that the imputation model and the analysis model should be compatible for results to be valid (Meng, 1994). More recently Xie and Meng expanded on the importance of congeniality between the imputation and analysis models for validity of results (Xie and Meng, 2017). They used a general estimating equation decomposition theorem to present multiple imputation inference as a combination of the true model (God's model), the imputer's knowledge and the analyst's knowledge. Our results investigate one specific example of an analysis and imputation model that are not compatible and reiterate the findings of others. Others have evaluated other instances of incompatible models, for example models including quadratic and interactions terms, and found similar results (Seaman, et al., 2012).

Although imputing the binary variable directly resulted in the lowest bias these models often had numerical difficulties such that parameter estimates were not obtained for the simulated dataset. This is not ideal in practice as it would mean that an alternative approach may well be needed anyway. The summaries of simulations in Table 1 only include the models that did not have numerical problems, potentially biasing the summaries towards these models. The simulation study we did is a very simple scenario. In real life, the multiple imputation models would be even more complex, with more incomplete variables, and thus be more likely to fail.

The imputed values of X_i^{bin} and X_i are not necessarily consistent with each other with either MVNI or FCS. Von Hippel (2009) claimed that the inconsistency did not matter for the estimation of the parameters in the

analysis model under the MAR assumption. However, Seaman et al (2012) argued that the stronger condition of MCAR is required for these JAV models to give consistent estimation of the parameters of the analysis model.

The JAV models seem to be a good compromise between models that do not have numerical problems and models that have low bias. Of the two JAV models the model fitted using FCS had lower bias than the model using MVNI; and the same bias as the logistic regression imputation model. However, the JAV model using FCS fits a logistic regression model for the binary outcome and thus similar numerical problems were experienced as when just the binary variable was imputed. This FCS model therefore does not appear to provide any advantage over simply imputing the binary variable. In contrast, the MVNI approach to impute both the binary and the continuous variable had larger bias and few numerical problems. The SMC FCS model overcame both these problems by providing a model with low bias and good coverage without having numerical problems. The only disadvantage of the SMC FCS model was the long computation time.

Our example illustrated the effect of non-compatible imputation and analysis models in a very simple scenario. In practice the analysis model can be much more complicated, for example including quadratic terms, transformed variables (for example log transformed variables) or interactions (Von Hippel, 2009). An exploration of multiple imputation in these more complex scenarios was conducted by Seaman et al (2012) where they evaluated multiple imputation with three methods; imputing, then transforming, using predictive mean matching and JAV models; when non-linear effects (quadratic effects) or interactions were present. They also came to the conclusion that "Given the current state of available software, JAV is the best of a set of imperfect methods for linear regression with quadratic or interaction effect" (Seaman, et al., 2012). Mitani et al (2015) discussed multiple imputation in the context of an analysis model with multi-level categorical interaction effects by comparing joint models and FCS models through a simulation study. They investigated both imputing the interaction term as JAV or imputing the main effects and then deriving the interactions (passive imputation). They compared the joint modelling approach to FCS using the JAV methods and evaluated an improved passive imputation approach under FCS. The improved passive imputation using FCS was superior to the other approaches investigated. Tilling also showed that imputation models that did not include interactions in the imputation model that were present in the analysis model resulted in biased estimates (Tilling, et al., 2016). Bartlett, et al. (2015) suggested the SMC FCS method which utilised FCS where each imputation model is compatible with the analysis model. In a series of simulation studies they showed this method to give consistent estimates, as was also the case in the simulation study presented here.

One caution is that the results of this simulation study should not be read as promoting the arbitrary dichotomising of continuous variables for analysis. This dichotomisation can arbitrarily dichotomise a relationship which is often on a continuum, and leads to a loss of information, with an accompanying loss of statistical power and the need for larger sample sizes compared to analysis of the continuous variable (Fedorov, et al., 2009), as well as increasing the risk of false positive results (Royston, et al., 2006). Others have written extensively on the loss of information that occurs when continuous variables are analysed as binary information (Altman and Royston, 2006; MacCallum, et al., 2002; Naggara, et al., 2011; Royston, et al., 2006). In our

simulated example, the data were generated so that the binary version of the variable of interest was associated with the outcome and thus the analysis model considered was the correct analysis models. In most real life situations that is not the case and the continuous variable would be a better choice to fit. In practice, the assumption that the correct analysis model includes the categorised version of a continuous variable rather than the continuous version should be critically investigated in each application and might not hold in the majority of cases. Not only providing advantages in terms of power and statistical efficiency, but also enabling the use of linear regression models during the multiple imputation step, thus removing the problems encountered when models had numerical problems. The researcher should consiwe only cvder the nature of the relationship between the variables and only analyse a binary variable when appropriate and there is evidence of a threshold effect.

Our simple simulation study had several limitations. We only consider a contrived, extremely simple example with missingness in one variable only to ensure that we could tease out the effects of the different imputation approaches without getting entangled with the complexity of real data scenarios. We did limited simulations of a small number of scenarios. The goal was to compare, in a controlled environment, what the effect of different simulation strategies would be. If approaches do not perform well in this simple scenario, it will only perform worse in more complicated scenarios. Further investigation, with different scenarios, and different forms of non-compatible models is needed. We acknowledge that in practice situations would be more complicated than in this example, although that generally exacerbates any issues with the MI process.

We recommend the message to researchers doing multiple imputation should be that having a multiple imputation model with the continuous variable is severely biased with low coverage and should not be done. The first step should be to transform the variable to a binary variable and attempt to do the multiple imputation with a logistic regression model using this binary version of the variable. An alternative, particularly if faced with numerical difficulties due to perfect prediction, it to use SMC FCS which was found to have acceptably low bias and good coverage.

Conflict of Interest
The authors have declared no conflict of interest

References

Altman, D.G. and Royston, P. (2006). The cost of dichotomising continuous variables. *British Medical Journal* **332**(7549), 1080-1080.

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006) .The design of simulation studies in medical statistics. *Statistics in Medicine* **25**(24):4279-4292

Bartlett, J.W., Seaman, S.R., White, I.R., Carpenter, J.R., Alzheimer's Disease Neuroimaging Initiative. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* **24**(4):462-87

Cappuccio, F., Taggart, F. Kandala, N-B., Currie, A., Peile, E. *et al.* (2008). Meta-Analysis of Short Sleep Duration and Obesity in Children and Adults. Sleep **31**, 619-26

Carpenter, J. and Kenward, M.G. (2013). *Multiple imputation and its application.* John Wiley and Sons, Chichester (United Kingdom)

Edwards, B. (2014). Growing Up in Australia: The Longitudinal Study of Australian Children Entering adolescence and becoming a young adult. *Family Matters* **95**, 5-14

Enders, C.K. (2010). *Applied Missing Data Analysis*. The Guilford Press, New York

Fedorov, V., Mannino, F. and Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics* **8**(1), 50-61

Harel, O. and Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine* **26**(16), 3057-3077

Kuczmarski, R.J., Ogden, C.L., Grummer-Strawn, L.M*., et al.* (2000). CDC growth charts: United States. *Adv Data* **314**, 1-27

MacCallum, R.C., Zhang, S., Preacher, K.J. and Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods* **7**(1),19-40

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–558

Mitani, A.A., Kurian, A.W., Das, A.K., Desai, M. (2015). Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects. *Statistical Methodology* **27,** 82-99

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., Altman, D.G. (2011). Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. *American Journal of Neuroradiology* **32**(3), 437-440

Raghunathan, T. (2016). *Missing Data Analysis in Practice*. CRC Press, Boca Raton, FL

Royston, P., Altman, D.G. and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* **25**(1), 127-141

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

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**(3), 581-592

Seaman, S.R., Bartlett, J.W. and White, I.R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology* **12**, 46-46

Tilling, K., Williamson, E.J., Spratt, M., Sterne, J.A.C. and Carpenter, J.R. (2016). Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *Journal of clinical epidemiology* **80**, 107-115

Von Hippel, P.T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology* **39**(1), 265-291

Wake, M., Clifford, S., York, E., Mensah, F., Burgner, D, Davies, S. *et al.* (2014). Introducing Growing Up in Australia's Child Health CheckPoint. *Family Matters* **95**, 15-23

Xie, X. and Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica* **27**, 1485-1594

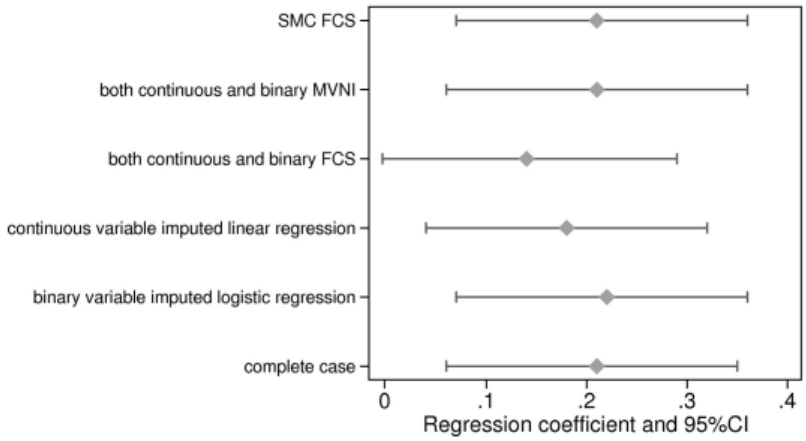

Figure 1: Results of multiple imputation using five different methods for the CheckPoint example with 15% of sample in the diseased category . Regression coefficients with 95% confidence intervals for the association of BMI z-score with short sleep duration

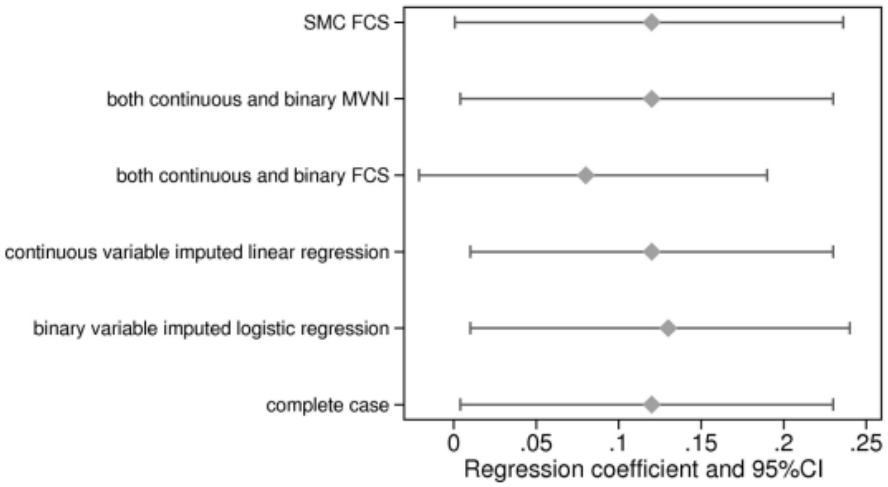

Figure 2: Results of multiple imputation using five different methods for the CheckPoint example with 30% of sample in the diseased category . Regression coefficients with 95% confidence intervals for the association of BMI z-score with short sleep duration

Table 1: Bias, empirical standard error, relative % error in the standard error and coverage of the 95% confidence interval for the various simulated missing data scenarios using a complete case analysis and each of the four multiple imputation methods

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
Sample size 1000 **MCAR; 25% missing**					**Sample size 1000** **MCAR; 25% missing**			
Complete case analysis	0.00	0.27	-2.48	94.7	0.00	0.20	1.02	95.8
Impute binary variable only using logistic regression	-0.01	0.24	-1.22	94.6	-0.00	0.18	2.16	95.6
Impute continuous variable only using linear regression	-0.82	0.25	26.24	19.8	-0.70	0.19	31.48	11.4
Impute both continuous and binary variable using FCS	-0.01	0.24	-0.96	94.8	-0.00	0.18	2.19	95.3
Impute both continuous and binary variable using MVNI	-0.00	0.25	-1.89	95.0	-0.01	0.19	2.09	95.3

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
Substantive model compatible (SMC) FCS	-0.00	0.23	1.16	96.1	0.01	0.18	-1.40	94.0
MAR; 25% missing					**MAR; 25% missing**			
Complete case analysis	-0.16	0.30	-3.19	91.1	-0.14	0.20	1.60	89.9
Impute binary variable only using logistic regression	-0.01	0.24	0.82	95.5	-0.00	0.18	2.70	95.9
Impute continuous variable only using linear regression	-1.02	0.25	33.25	7.2	-0.63	0.18	29.12	19.5
Impute both continuous and binary variable using FCS	-0.01	0.24	1.41	95.2	-0.00	0.18	2.69	96.0
Impute both continuous and binary variable using MVNI	0.23	0.29	-7.06	83.7	-0.10	0.18	3.55	92.2
Substantive model compatible (SMC) FCS	-0.00	0.23	6.50	96.4	0.01	0.18	-1.01	94.4
MAR; 50%					**MAR; 50% missing**			

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
missing								
Complete case analysis	-0.20	0.38	-1.69	91.8	-0.20	0.25	-3.82	86.7
Impute binary variable only using logistic regression	-0.06	0.25	38.01	97.4	-0.02	0.18	5.45	96.2
Impute continuous variable only using linear regression	-1.95	0.30	42.03	0	-1.21	0.21	37.13	0.1
Impute both continuous and binary variable using FCS	-0.06	0.25	40.11	97.0	-0.02	0.18	3.28	96.3
Impute both continuous and binary variable using MVNI	0.45	0.39	-10.97	71.2	-0.18	0.21	2.30	85.9
Substantive model compatible (SMC) FCS	0.00	0.25	4.57	96.3	0.01	0.18	-0.93	94.4
Sample size 500 **MCAR; 25% missing**					**Sample size 500** **MCAR; 25% missing**			
Complete case	0.03	0.37	-0.68	95.0	-0.02	0.30	-2.83	94.5

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
analysis								
Impute binary variable only using logistic regression	-0.02	0.31	24.00	97.5	-0.03	0.26	2.08	95.6
Impute continuous variable only using linear regression	-0.80	0.35	26.99	57.1	-0.72	0.27	28.29	43.1
Impute both continuous and binary variable using FCS	-0.02	0.31	21.62	96.9	-0.03	0.26	1.19	95.9
Impute both continuous and binary variable using MVNI	0.00	0.34	1.80	96.1	-0.04	0.28	-1.99	93.8
Substantive model compatible (SMC) FCS	-0.00	0.33	-0.28	95.4	0.003	0.25	0.63	94.7
MAR; 25% missing					**MAR; 25% missing**			
Complete case analysis	-0.17	0.41	-0.24	93.8	-0.15	0.29	-0.41	91.5
Impute binary variable only using logistic	-0.05	0.30	34.67	97.8	-0.03	0.26	0.24	95.4

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
regression								
Impute continuous variable only using linear regression	-1.02	0.34	38.75	38.6	-0.65	0.26	28.92	50.8
Impute both continuous and binary variable using FCS	-0.05	0.30	32.64	97.8	-0.03	0.26	0.33	95.4
Impute both continuous and binary variable using MVNI	0.22	0.39	-1.44	90.9	-0.13	0.26	2.13	92.6
Substantive model compatible (SMC) FCS	-0.00	0.32	0.85	95.7	0.004	0.25	1.11	94.2
MAR; 50% missing					**MAR; 50% missing**			
Complete case analysis	-0.20	0.53	1.70	93.8	-0.21	0.35	-1.45	89.9
Impute binary variable only using logistic regression	-0.28	0.36	117.20	98.3	-0.06	0.26	15.96	96.7
Impute continuous variable only using linear	-1.96	0.41	45.72	2	-1.24	0.29	40.89	6.1

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
regression								
Impute both continuous and binary variable using FCS	-0.29	0.35	122.56	98.0	-0.06	0.26	14.28	96.1
Impute both continuous and binary variable using MVNI	0.42	0.52	-3.14	85.3	-0.22	0.29	3.24	89.4
Substantive model compatible (SMC) FCS	-0.01	0.32	1.04	95.6	0.001	0.26	0.86	94.2
Sample size 200 **MCAR; 25% missing**					**Sample size 200** **MCAR; 25% missing**			
Complete case analysis	0.01	0.58	2.92	95.7	-0.02	0.47	-0.30	94.4
Impute binary variable only using logistic regression	-0.38	0.50	92.72	97.5	-0.12	0.41	37.31	96.3
Impute continuous variable only using linear regression	-0.84	0.53	33.65	85.5	-0.73	0.42	29.21	79.4
Impute both continuous	-0.36	0.49	94.12	97.5	-0.13	0.40	37.31	96.2

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
and binary variable using FCS								
Impute both continuous and binary variable using MVNI	-0.04	0.56	2.09	95.1	-0.06	0.43	1.33	95.8
Substantive model compatible (SMC) FCS	0.01	0.53	0.30	94.7	-0.02	0.40	1.11	95.8
MAR; 25% missing					**MAR; 25% missing**			
Complete case analysis	-0.16	0.64	3.44	94.7	-0.16	0.46	-0.72	93.3
Impute binary variable only using logistic regression	-0.45	0.51	79.12	96.7	-0.12	0.41	43.86	96.5
Impute continuous variable only using linear regression	-1.08	0.56	38.28	78.2	-0.67	0.43	26.34	81.9
Impute both continuous and binary variable using FCS	-0.45	0.51	81.07	96.5	-0.11	0.39	43.04	96.5
Impute both continuous	0.19	0.64	-0.20	93.9	-0.16	0.42	1.23	94

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
and binary variable using MVNI								
Substantive model compatible (SMC) FCS	0.00	0.52	1.64	94.7	-0.02	0.41	-0.49	95.2
MAR; 50% missing					**MAR; 50% missing**			
Complete case analysis	-0.20	0.90	0.53	94.5	-0.20	0.54	2.21	93.6
Impute binary variable only using logistic regression	-1.14	0.61	111.07	94.2	-0.27	0.46	92.41	97.6
Impute continuous variable only using linear regression	-2.04	0.68	43.34	41.4	-1.29	0.49	35.11	50.8
Impute both continuous and binary variable using FCS	-1.13	0.61	111.94	94.5	-0.27	0.44	95.02	97.4
Impute both continuous and binary variable using MVNI	0.31	0.90	-0.94	92.4	-0.30	0.47	3.31	91.8
Substantive model compatible	-0.01	0.55	0.65	94.7	-0.03	0.42	-0.87	95.1

	Proportion data in disease category: 15%				Proportion data in disease category: 30%			
	Bias	**Empirical SE**	**Relative % error in SE**	**Coverage**	**Bias**	**Empirical SE**	**Relative % error in SE**	**Coverage**
(SMC) FCS								

Coverage =Coverage of nominal 95% confidence interval, SD = Standard deviation; SE = standard error; MCAR = missing completely at random; MAR = missing at random; FCS using fully conditional specification; MVNI = multivariate normal imputation; Relative % error in SE is defined as the percentage error in the model based standard error using the empirical SE as comparison

Table 2: Percentage of simulated datasets that encountered numerical problems during the multiple imputation step

	Proportion data in disease category: 15%			Proportion data in disease category: 30%
MCAR; 25% missing	**Sample size 1000**	**Sample size 500**	**Sample size 200**	**All sample sizes**
Impute binary variable only using logistic regression	0	1.1	35.3	0
Impute continuous variable only using linear regression	0	0	0	0
Impute both continuous and binary variable using FCS	0	1.1	35.3	0
Impute both continuous and binary variable using MVNI	0	0	0	0
Substantive model compatible (SMC) FCS	0	0	0	0
MAR; 25% missing			0	
Impute binary variable only using logistic regression	0	2.6	43.5	0
Impute continuous variable only using linear regression	0	0	0	0
Impute both continuous and binary variable using FCS	0	2.6	43.5	0
Impute both continuous and binary variable using MVNI	0	0	0	0
Substantive model compatible (SMC) FCS	0	0	0	0
MAR; 50% missing			0	
Impute binary variable only using logistic regression	0.8	19.2	65.8	0
Impute continuous variable only using linear regression	0	0	0	0
Impute both continuous and binary variable using FCS	0.8	19.2	65.8	0
Impute both continuous and binary variable using MVNI	0	0	0	0

	Proportion data in disease category: 15%			Proportion data in disease category: 30%
Substantive model compatible (SMC) FCS	0	0	0	0

MCAR = missing completely at random; MAR = missing at random; FCS using fully conditional specification; MVNI = multivariate normal imputation