Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Multiply imputing missing values arising by design in transplant survival data

Laura Pankhurst [1], **Robin Mitra** [2], **Alan Kimber** [3], and **Dave Collett** [1]

[1] Statistics and Clinical Studies, NHS Blood and Transplant

[2] Department of Mathematics and Statistics, Lancaster University *

[3] Mathematical Sciences, University of Southampton

Received zzz, revised zzz, accepted zzz

Key words: Missing data; Multiple imputation; Stepwise selection; Survival analysis; Transplant data

In this article we address a missing data problem that occurs in transplant survival studies. Recipients of organ transplants are followed up from transplantation and their survival times recorded, together with various explanatory variables. Due to differences in data collection procedures in different centres or over time, a particular explanatory variable (or set of variables) may only be recorded for certain recipients, which results in this variable being missing for a substantial number of records in the data. The variable may also turn out to be an important predictor of survival and so it is important to handle this missing-by-design problem appropriately. Consensus in the literature is to handle this problem with complete case analysis, as the missing data are assumed to arise under an appropriate missing at random mechanism that gives consistent estimates here. Specifically the missing values can reasonably be assumed not to be related to the survival time. In this article, we investigate the potential for multiple imputation to handle this problem in a relevant study on survival after kidney transplantation, and show that it comprehensively outperforms complete case analysis on a range of measures. This is a particularly important finding in the medical context as imputing large amounts of missing data is often viewed with scepticism.

1 Introduction

The presence of missing data is a common problem in many fields and can complicate important statistical analyses. Common ad-hoc strategies to handle this problem, while easy to apply, can often result in biased estimates and incorrect conclusions being made.

The focus of this article is on missing values in transplant survival data, where these can occur for many different reasons. In the motivating example for this paper, transplant survival times following kidney transplantation depend on covariates associated with the donor (such as age, cause of death), the transplant procedure (such as cold ischaemic time) and the recipient (such as primary disease, diabetes status). Models for the survival time are then used in the development of organ allocation schemes (Johnson *et al.*, 2010), to advance clinical practice and to inform patients of their likely prognosis (Watson *et al.*, 2012). In data of this kind, missing values are frequently encountered and can often occur by design. For example, changes in the transplant procedure, clinical practice, and post-transplant immunosuppression, may mean that a variable, previously not routinely measured, is subsequently found to be an important determinant of outcome. For example, the body mass index (weight in kg/height^2 in m^2) of a recipient was not recorded until

*Corresponding author: e-mail: r.mitra@lancaster.ac.uk

2003, when evidence from elsewhere suggested that it could be relevant to graft and patient survival. When this occurs, certain individuals, whose transplant takes place after this discovery was made, will have information recorded on this variable, while all other individuals will not have this information reported. Any such variable will then have a significant proportion of missing values which arise by design rather than through some random (unknown) mechanism. Similarly, there may be differences between transplant centres in factors that are recorded. For example, a measure of disease severity in potential liver transplant recipients, known as the UKELD (UK End stage Liver Disease) score (Barber *et al.*, 2011), is relevant to post-transplant survival, but one centre did not record this variable between 2001 and 2006. This has important consequences for an analysis of national data to model the dependence of survival on the UKELD score.

In each of these situations, there are systematically missing data across subsets of individuals in distinct periods of time. This is different from traditional missing data patterns present in data where missing values may be spread in a more random way throughout the data set or in a monotonely increasing pattern through time if missing data are due to drop out. Missing values essentially arise by design here, as a result of differences in data collection procedures across different time periods or locations, so that a distinct part of the data is unobserved. It is evident that a complete case analysis on such data could have a severe impact on inferences as a large proportion of the data would be discarded.

To handle this missing by design problem, we propose to multiply impute the missing values and evaluate the performance of this method. This approach fills in missing values from the predictive distribution arising from a statistical model fit to the complete data. This is done multiple times to generate multiple completed data sets. This will allow analysts to apply their usual complete data methods to the completed data sets and make appropriate inferences through some simple combining rules; see Schafer (1999) for a review of the approach. We assume analysts are interested in fitting a Cox model to their data and obtaining maximum (partial) likelihood estimators for the regression coefficients. For a complete data set these estimators would be consistent.

In particular, we explore the advantages of using multiple imputation (MI) over a complete case analysis (CC). In the medical community, while MI is becoming increasingly popular (Bartlett *et al.*, 2015b; Kenward and Carpenter, 2007) multiply imputing a large proportion of values for a variable is typically viewed with scepticism, and caution is recommended in the literature (White *et al.*, 2011). Furthermore, CC may often be viewed as an appropriate method with this type of missing data problem as the missing data mechanism may be assumed to be missing completely at random, and so estimates would be consistent. Typically this is a very strong assumption to make, and a missing at random assumption is more appropriate here as the missing data mechanism would depend on other variables in the study, perhaps deterministically, e.g. on transplant year or transplant centre. Nevertheless, this does not change the conclusion that CC would result in consistent regression coefficient estimates. Provided the probability of being a complete case depends on the covariates only (not the response) and these are conditioned on in the regression model, regression coefficients will be consistent, see Little and Rubin (2002, p. 43) and Bartlett *et al.* (2015a) for more details. However, we also consider the missing completely at random scenario later in this article and illustrate the benefits of using MI here as well.

It is often quite challenging to determine a plausible model to fit to the data due to the large number of variables present, typically measured on different scales; for example some may be measured on a binary scale while others may be measured on a continuous scale. For this reason we use an imputation approach based on chained equations (Van Buuren *et al.*, 1999) which can handle missing values arising in a variety of situations. This is a commonly used imputation method in the literature.

The article is organised as follows. Section 2 reviews the framework for multiple imputation of missing data and briefly describes how an imputation method based on chained equations works. Section 3 describes the motivating application and compares the performance of MI and CC to

handle the problem of missing data. Section 4 presents further comparisons between MI and CC based on simulations constructed from the complete case subsample taken from the data used in Section 3. Finally Section 5 presents some concluding remarks.

2 Multiple imputation inference

We assume we have n observations in our data set, each measured on p covariates. Denote the $n \times p$ covariate data set by $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, with x_{ij} corresponding to the jth covariate value for the ith individual, $i = 1, \ldots, n$, $j = 1, \ldots, p$. For each covariate value x_{ij} we also define a missing data indicator m_{ij}, where $m_{ij} = 1$ implies x_{ij} is missing and $m_{ij} = 0$ implies x_{ij} is observed. The covariate data set X can then be decomposed into its observed and missing parts denoted by $X_{obs} = \{x_{ij} : m_{ij} = 0\}$ and $X_{mis} = \{x_{ij} : m_{ij} = 1\}$ respectively. Let M denote the corresponding matrix of missing data indicators.

The literature on missing data typically makes some assumptions about the missing data mechanism, i.e. the process that produces the missing values. This is often expressed with a conditional distribution of the form:

$$p(M|X, \phi)$$

where ϕ are parameters that govern the missing data mechanism. If the above expression reduces to the following,

$$p(M|X_{obs}, \phi)$$

then the data are said to be missing at random (MAR). This is the most commonly used missing data assumption. If missing data are MAR and the parameters ϕ are distinct from the parameters characterising the model for the data, e.g. coefficients from a regression model fit to the data, then the missing data mechanism is said to be ignorable and is a necessary condition for MI to be a valid method to consider here. In this article we assume that the data are MAR and parameters for the missing data mechanism are distinct from parameters characterising the data model. A special case of MAR occurs if the missing mechanism can be expressed as,

$$p(M|\phi)$$

in which case the data are said to be missing completely at random (MCAR). See Rubin (1976) for more information.

To illustrate the different types of missing data mechanisms consider a scenario where we only have two variables measured for each unit i, x_{i1} and x_{i2} with x_{i1} fully observed and x_{i2} containing some missing values. Also suppose γ_0, γ_1 and γ_2 are some unknown (constant) parameter values. If the missing data mechanism can be expressed as:

$$p(m_{i2} = 1|x_{i2}, x_{i2}) = \frac{\exp(\gamma_0)}{1 + \exp(\gamma_0)}$$

then the missing data are MCAR, each unit has the same probability to be missing. If instead the mechanism is expressed as:

$$p(m_{i2} = 1|x_{i2}, x_{i2}) = \frac{\exp(\gamma_0 + \gamma_1 x_{i1})}{1 + \exp(\gamma_0 + \gamma_1 x_{i1})}$$

then the missing data are MAR, where units have potentially different probabilities to be missing but this only depends on x_{i1} which is fully observed and it is possible to estimate values of γ_0 and γ_1 for example through a logistic regression. Finally if the mechanism is expressed as:

$$p(m_{i2} = 1|x_{i2}, x_{i2}) = \frac{\exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2})}{1 + \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2})}$$

then the missing data are missing not at random (MNAR), the missing probability depends on x_{i2} which is not fully observed and so estimating the model parameters is not possible.

MI works by completing the missing values in the data set with draws from the posterior predictive distribution $p(X_{mis}|X_{obs})$. This is done repeatedly, r times, to create r completed data sets. We denote these data sets by $X_{com}^{(k)}, k = 1, \ldots, r$. We assume that analysts of the data are interested in making inferences about some estimand in the population, Q. This could be for example the mean of a variable, or the coefficient from a regression model. The analyst performs the same complete data inference that would have been performed in each of the completed data sets; specifically, point and variance estimates, q_k and u_k respectively, are obtained for Q in each of the imputed data sets $X_{com}^{(k)}$. These estimates are then combined by the analyst to obtain the following quantities:

$$\bar{q}_r = \frac{\sum_{k=1}^{r} q_k}{r} \tag{1}$$

$$\bar{u}_r = \frac{\sum_{k=1}^{r} u_k}{r} \tag{2}$$

$$b_r = \frac{\sum_{k=1}^{r} (q_k - \bar{q}_r)^2}{r-1} \tag{3}$$

The analyst can then use $\bar{q}_r$ as a point estimate for Q and estimate the variance by $T_r = \bar{u}_r + (1 + 1/r)b_r$. The estimate T_r incorporates the additional uncertainty due to the presence of imputed values in the completed data sets. Analysts can obtain confidence intervals using a t-distribution with degree of freedom ν, where

$$\nu = (r-1)\left(1 + \frac{r}{r+1}\frac{\bar{u}_r}{b_r}\right)^2.$$

See Rubin (1987) for more details.

2.1 Imputation using chained equations

It is often quite challenging to determine an expression for the posterior predictive distribution, $p(X_{mis}|X_{obs})$, in closed form. Typically the missing data pattern will be non-monotone, and the variables in the data set will often be recorded on different measurement scales. A popular imputation method in such situations is that based on chained equations (Van Buuren *et al.*, 1999). We give a brief description of how the approach works here.

The approach essentially imputes missing values in a sequential iterative process. Suppose all the variables in the data set, $x_1, \ldots, x_p$ contain some missing values. The chained equations scheme first fills in missing values in x_j by sampling from the marginal observed distribution of this incomplete variable. Once this initial stage has taken place, imputations are generated from a sequence of full conditional distributions $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p, \psi_j)$, $j = 1, \ldots, p$. Each predictive distribution is obtained from an appropriate regression model depending on the measurement scale of x_j. For example if x_j is binary then a logistic regression could be used, while if x_j is continuous then a normal regression model may be considered. Flat priors are used for the regression model parameters ψ_j. This is done in an iterative process, so at iteration t draw imputations for missing values in x_j to create an imputed $x_j^{(t)}$ from $p(x_j|x_1^{(t)}, \ldots, x_{j-1}^{(t)}, x_{j+1}^{(t-1)}, \ldots, x_p^{(t-1)}, \psi_j)$ so we condition on the most recent imputed values for missing covariate values in this regression model. This is done T times until it is assumed that the imputations have stabilised. Often T is not very large with $T = 10$ or 20 typically being deemed to be sufficient (Buuren and Groothuis-Oudshoorn, 2010). The imputed data set resulting from this last iteration yields one of the imputed data sets. This is then done r times to create r different imputed data sets.

We note that this imputation method does not guarantee that imputations are drawn from a proper posterior predictive distribution, essentially this is because the parameters ψ_j are updated from their conditional posterior distributions rather than updating the parameters from their posterior distribution based on a model for the joint distribution of the variables, and a set of full conditional distributions does not guarantee that there is a corresponding proper joint density. This is what distinguishes this approach from the more formal Gibbs sampling approach to generate imputations from a joint model. Nevertheless, this approach has been widely used and performs well; see the thorough empirical evaluation of Van Buuren (2007). We note that a theoretical justification of the chained equations approach has been developed (Liu *et al.*, 2013; Hughes *et al.*, 2014). In particular, Hughes *et al.* (2014) propose a non-informative margins condition that, if satisfied, guarantee draws from the chained equations approach correspond to draws from a joint model for the data. However, if the potential lack of a relevant joint density is of concern, the approach of Lee and Mitra (2016) may be used to guarantee that imputations are drawn from a proper posterior predictive distribution.

In the next section the chained equations approach is used to handle missing values arising in a study of survival times after kidney transplantation.

3 Imputation in data on kidney transplant survival

To investigate factors associated with transplant survival time, defined to be the earlier of graft failure or patient death, data were obtained from the UK Transplant Registry, held by NHS Blood and Transplant, on adult, first kidney only transplants between 1 January 2001 and 31 December 2008. During this period there were 7732 transplanted patients. A Cox regression model for the hazard of transplant failure is adopted. The donor factors considered for inclusion in the model are the recipient unit, age, gender, ethnicity, body mass index, type (whether live donor, deceased donor following circulatory death, or deceased donor following brain stem death), cytomegalovirus (CMV) status and cause of death. Transplant factors were year of transplant, waiting time, cold ischaemic time (the time from retrieval of organ to transplant in the recipient), HLA match grade and whether or not kidney is transplanted locally. Recipient factors were recipient transplant unit, age, gender, ethnicity, body mass index, CMV status, primary disease (diabetes, glomerulonephritis, HU-syndrome, polycystic kidneys, nephritis, other), the degree of sensitisation, serum creatinine and the ACORN index of deprivation.

Recipient body mass index was missing in 4937 patients (64%). This quantity was not recorded at all until 2003, and in subsequent years it was missing for at least 35% of transplant recipients. Missing values in this variable are thus strongly related to transplant year and transplant unit. The data also contained missing values in 10 of the remaining variables. In particular, CMV was missing for 15% of recipients, serum creatinine in 11%, donor body mass index in 5%, while the ACORN index, donor CMV status, HSP status, cold ischaemic time, recipient and donor ethnicity were missing in less than 3% of patients. Unlike recipient BMI, missing values in other variables are not strongly related to the year of transplant or transplant unit. All missing values were assumed to be MAR.

MI using 30 replicates was performed using chained equations, with some restrictions placed on the values of imputed variables to ensure that imputed values of cold ischaemic time and serum creatine were positive, while values of BMI were greater than 8.0. We also considered results based on imputations without these restrictions but results were not qualitatively different and are not presented here. Default choices in the MICE package in R were used to impute the categorical variables. For binary variables (or equivalently variables with only two categories) logistic regression imputation models were used, and for categorical variables with more than two categories multinomial logistic regression models were used. For continuous variables, missing values were imputed using normal linear regression models to be consistent with the other fully parametric

models used for imputing missing categorical values. In each imputation model, all other variables were included as covariates as main effects only, including the outcome variable, transplant survival time, and the corresponding censoring indicator. We note that other approaches could have been used to impute missing values in the continuous variables such as predictive mean matching. In results not presented here, we also imputed missing continuous variables using predictive mean matching instead and found inferences were similar to those obtained here.

We note that there are alternative approaches that instead include a function of the survival time as a covariate in the imputation model. White and Royston (2009) suggest instead including the Nelson Aalen estimator of the cumulative hazard and Bartlett *et al.* (2015b) propose an alternative approach based on rejection sampling whereby each imputed value is sampled from a proposal distribution with a rejection rule used to determine whether this value would be accepted or not. Advantages of the Bartlett *et al.* (2015b) approach have been noted by Keogh *et al.* (2018), in particular robustness to model mis-specification. The way we have included survival times in the imputation model was deemed to be the simplest option to consider and would be easily compatible with the MICE package in R. This was a key factor in our decision when the objective is to compare performance with complete case analysis, which is argued for its simplicity. Thus, we did not want to greatly complicate the procedure for producing imputations, but we note that there is the potential to further optimise the performance of the multiple imputation approach using approaches suggested in the above literature.

Standard methods of variable selection are impractical with 30 separate data sets, and so the method of Wood *et al.* (2008) is used. Here, all 30 imputed data sets are stacked and variable selection performed on the 7732×30 observations. To determine the significance of a covariate adjusted for the inflated number of observations, each observation is weighted by $1/30$. This leads to a model with 11 covariates. A Cox regression model with these variables was fitted to each imputed dataset and the estimates combined to give the values shown in Table 1. For comparison, this table also shows the estimates obtained from CC using these variables, although this is based on just 28% of the data. Note that certain point and interval estimates could not be obtained using CC due to the greatly reduced sample size and the impact of censoring. Some additional metadata is presented in Table 2 in the appendix; this includes sample sizes for each continuous variable and levels of each categorical variable in the original data, complete case data and multiply imputed data (presented as a range across imputations where appropriate). Table 2 also includes the proportion of missing values present in each variable in the original data.

covariate	number of patients	complete cases		multiple imputation	
		estimate	95% CI	estimate	95% CI
Primary renal disease					
2	564	0.064	(-0.476, 0.604)	0.471	**(0.238, 0.704)**
3	1026	-0.936	(-1.501, -0.371)	-0.575	**(-0.809, -0.340)**
4	640	0.060	(-0.480, 0.600)	0.134	**(-0.106, 0.374)**
5	1326	0.118	(-0.371, 0.607)	0.073	**(-0.140, 0.286)**
6	930	0.072	(-0.446, 0.591)	0.161	**(-0.063, 0.384)**
7	2866	0.011	(-0.470, 0.491)	0.048	**(-0.156, 0.253)**
Recipient unit					
2	505	-0.429	(-1.502, 0.644)	0.074	**(-0.231, 0.379)**
3	633	-0.332	(-0.830, 0.165)	-0.127	**(-0.426, 0.172)**
4	173	0.110	(-0.565, 0.784)	0.221	**(-0.141, 0.584)**
5	153	0.142	(-0.478, 0.762)	-0.096	**(-0.501, 0.309)**
6	224	-0.271	(-0.859, 0.317)	-0.389	**(-0.765, -0.014)**

Continued on next page

Table 1 – *Continued from previous page*

covariate	number of patients	complete cases estimate	complete cases 95% CI	multiple imputation estimate	multiple imputation 95% CI
7	425	-0.083	(-0.617, 0.450)	-0.328	**(-0.659, 0.004)**
8	282	1.516	(-0.546, 3.577)	-0.335	**(-0.745, 0.074)**
9	319	-0.342	(-1.011, 0.328)	-0.001	**(-0.338, 0.335)**
10	408	-14.987	($-\infty$, ∞)	-0.349	**(-0.729, 0.031)**
11	348	-0.423	(-1.027, 0.182)	-0.466	**(-0.816, -0.116)**
12	241	0.050	(-0.539, 0.640)	0.090	**(-0.244, 0.424)**
13	361	-0.271	(-1.012, 0.470)	-0.434	**(-0.769, -0.099)**
14	396	0.058	(-0.456, 0.572)	0.065	**(-0.249, 0.378)**
15	236	-0.025	(-0.577, 0.526)	0.030	**(-0.324, 0.385)**
16	491	-0.308	(-0.963, 0.346)	-0.134	**(-0.450, 0.182)**
17	104	-0.066	(-0.722, 0.590)	-0.232	**(-0.692, 0.227)**
18	295	0.010	(-0.719, 0.739)	-0.170	**(-0.510, 0.170)**
19	603	-0.396	(-1.205, 0.413)	-0.192	**(-0.496, 0.112)**
20	293	-0.390	(-0.940, 0.160)	-0.023	**(-0.356, 0.310)**
21	270	-0.376	(-0.990, 0.237)	-0.138	**(-0.480, 0.205)**
22	376	0.409	(-0.241, 1.058)	-0.113	**(-0.440, 0.214)**
23	370	-0.455	(-0.997, 0.086)	-0.178	**(-0.509, 0.153)**
ACORN index					
Urban prosperity	727	0.139	(-0.264, 0.541)	0.141	**(-0.038, 0.320)**
Comfortably off	1952	0.085	(-0.187, 0.357)	0.118	**(-0.008, 0.245)**
Moderate means	1277	0.369	(0.078, 0.661)	0.318	**(0.178, 0.458)**
Hard pressed	1844	0.358	(0.085, 0.630)	0.344	**(0.216, 0.471)**
Transplant year					
2002	952	NA	NA	0.039	(-0.103, 0.181)
2003	931	NA	NA	-0.175	(-0.337, -0.013)
2004	1038	-0.023	(-0.571, 0.526)	-0.021	**(-0.177, 0.135)**
2005	915	-0.237	(-0.806, 0.332)	-0.235	**(-0.411, -0.059)**
2006	939	-0.064	(-0.638, 0.509)	-0.259	**(-0.444, -0.074)**
2007	914	0.149	(-0.425, 0.723)	-0.018	**(-0.202, 0.166)**
2008	1073	-0.048	(-0.637, 0.540)	-0.150	**(-0.344, 0.043)**
Recipient sex - female	2940	0.203	(0.002, 0.403)	0.159	**(0.068, 0.249)**
Serum creatinine	6866	0.006	(0.005, 0.007)	0.006	**(0.006, 0.006)**
Donor age	7732	0.005	(-0.002, 0.013))	0.004	**(0.000, 0.007)**
Recipient age	7732	0.026	(0.018, 0.034)	0.027	**(0.024, 0.031)**
Recipient BMI	2795	-0.017	(-0.037, 0.003)	-0.024	**(-0.042, -0.007)**
Donor BMI	7315	0.014	(-0.004, 0.031)	0.010	**(0.000, 0.019)**
Donor CMV					
Status positive	3721	0.314	(0.118, 0.509)	0.135	**(0.047, 0.224)**

Continued on next page

Table 1 – *Continued from previous page*

covariate	number of patients	complete cases estimate	complete cases 95% CI	multiple imputation estimate	multiple imputation 95% CI

Table 1: Coefficient estimates and 95% confidence intervals from a Cox proportional hazards regression using multiple imputation and complete case analysis respectively. For each coefficient the interval in bold is the shorter one. Primary renal disease categories: 1 - Glomerulonephritis (absorbed into the intercept), 2 - Pyelonephritis/Interstitial Nephritis, 3 - Miscellaneous, 4 - Polycystic kidneys, 5 - Hypertension/Renovascular Disease, 6 - Diabetes, 7 - Not Reported

We see that in general the point estimates from the multiply imputed data and the complete case analysis are similar, but the 95% confidence intervals tend to be narrower when using MI. There is thus some indication that using MI has some advantages over using CC. In particular the 95% confidence interval for the coefficient of recipient BMI (rbmi), which contains a large proportion of values missing by design, includes zero under CC but does not include zero in the multiply imputed data. Similarly, confidence intervals for coefficients on donor age and donor BMI include zero in the complete case data but not in the multiply imputed data. Note also that variables with no missing data, such as donor age and recipient age, tend to have estimates with much higher precision under MI. We investigated sensitivity of the results to different numbers of imputations, ranging from 10 to 50 in increments of 10. We found that results were robust to the number of imputations with the subset selected from the stepwise regression always remaining the same. For the remainder of the article we focus on inferences obtained for the mid value of 30 imputations as that allows a decent number of imputations to be considered while also not placing undue computational burden on the simulations in the next section.

Despite this, we do not know the true coefficient values in the above model, and as such it is not possible to determine which approach (MI or CC) performs better. To gain further insight into the performance of both methods, in the next section we design a simulation involving the complete cases that can better compare both approaches.

4 Simulation involving the complete cases

4.1 Comparison of multiple imputation and complete cases based on the analysis model in Section 3

We first remove the incomplete cases from the observed data to obtain the complete case subsample. This results in a data set comprising 2131 transplanted patients. As this sample size is much smaller than in the original data, we resample the rows of this data set with replacement (i.e. bootstrap rows of the data) to increase the sample size to 4000. Henceforth, in this section we refer to this data set as the complete data set. We then re-introduce the patterns of missing data present in the original data back into the fully observed complete data set. This corresponds to a non-parametric model-free approach for introducing missing values. To illustrate this consider a simple example, suppose a data set has three variables, x_1, x_2 and x_3. In the data set, 60% of units are fully observed, 15% of units are missing in x_2 only, 15% are missing in x_3 only and 10% are missing in both x_2 and x_3. The approach would introduce missing values by randomly assigning each unit to be fully observed with probability 0.6, missing in x_2 only with

probability 0.15, missing in x_3 only with probability 0.15 and missing in both x_2 and x_3 only with probability 0.1. Doing so allows us to avoid specifying a parametric model for the missing data mechanism and is thus more robust to this type of mis-specification. We note that this approach of simulating missing values only preserves the missing data patterns, and we cannot make any statements about the missing data mechanism. The approach has been used previously to good effect to design simulation studies, e.g. by Mitra and Reiter (2011). We can then perform similar analyses to those performed in the previous section, handling the missing values using either MI or CC. We repeat this process 250 times and summarise the results in various ways to compare the performance between using MI and CC.

Treating the coefficient estimates from the fully observed complete data set as the true parameter values, we can obtain a measure of mean squared error (MSE) for the estimates found using MI and complete case analysis respectively. Figure 1 plots the MSEs for each regression coefficient obtained using MI against the corresponding MSE obtained from a complete case analysis. Note that the plot of MSEs is on the log scale to aid clarity. There were two covariates for which it was not possible to obtain estimates of the regression coefficients when using CC, and so these were not included in the plots. Specifically, the coefficient of the effect of the level of recipient unit 10 could not be estimated in the original data due to censoring, and the coefficient of the effect of recipient unit 8 could not be estimated because there were only two observations in this group, one of which was censored, so inevitably there would have been some simulated data sets containing no observable survival times in this group. Out of the remaining 42 coefficient estimates, when using MI all 42 estimates had a smaller MSE than when using CC. This can be seen from Figure 1 where all points in both plots are below the line. We also see that gains for MI are slightly more pronounced for coefficients corresponding to categorical variables. This could be due to the reduced sample sizes in each level of a category resulting from a complete case analysis (as seen from Table 2) which inflate variances.

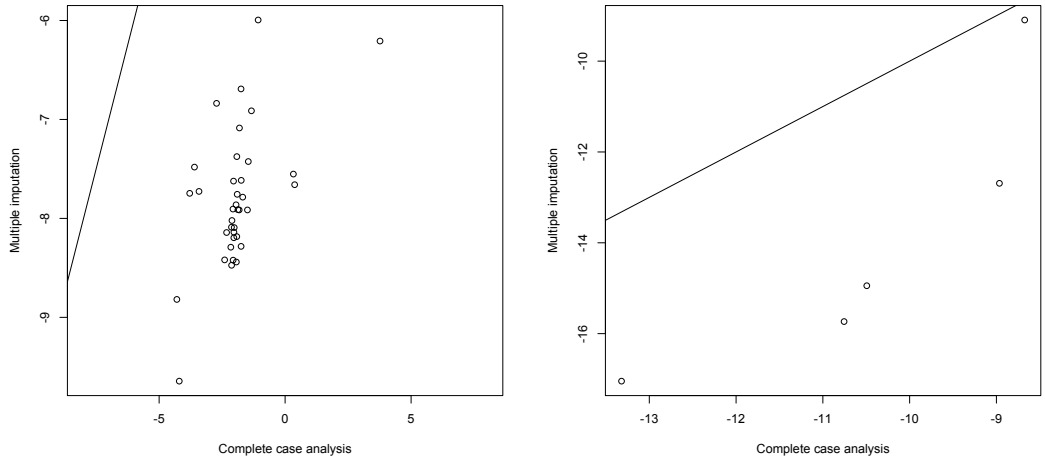

Figure 1 Plots of log(MSE) in coefficient estimates for categorical covariates (left) and continuous covariates (right) obtained from using MI against CC in the simulation involving the complete case subsample. Points below the $\log y = \log x$ line (included) indicate a larger MSE for a complete case analysis. Note that the plot is on the log scale.

As expected most of the contribution to the MSE, in both MI and CC, is due to variance with bias being relatively small. This is to be expected as we assume missingness is primarily MAR

and so bias should be small. To illustrate this Figure 2 presents the ratio of variance to MSE for each coefficient arising from MI plotted against the corresponding ratio from CC. We see that almost all points are greater than 0.8 and points with lower ratios tend to be similar for both analysis methods indicating that there is little difference, as expected.

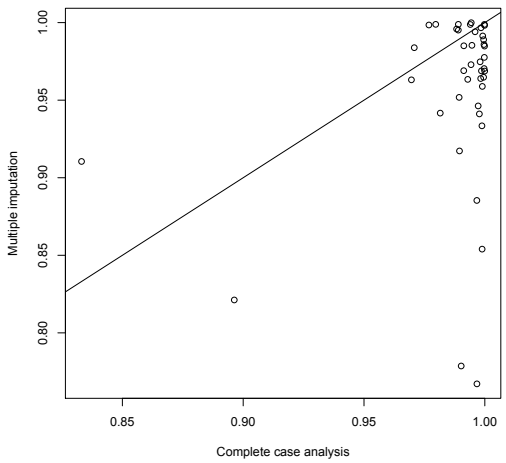

Figure 2 Plot of the ratio of variance to MSE for MI against CC. The $y = x$ line is included

We also compare performance of interval estimates between MI and CC. Specifically we compare the average 95% confidence interval width of each regression coefficient obtained from MI and a CC, where the averaging is over the 250 replications. As before, the two average confidence interval lengths, corresponding to recipient unit levels 8 and 10 could not be computed when performing CC and are not considered in the comparison. Of the remaining 42 average lengths, 41 were smaller when using MI. This indicates that MI has the ability to obtain more precise estimates with shorter confidence interval lengths in general than using CC. Figure 3 plots the average lengths for each coefficient from using MI against the average length from CC, with the y = x line included. Points below the line indicate improved performance for MI over CC. All but one point are below the line, confirming that all the average lengths are smaller when using MI. There were several average lengths for CC that were infinite and as such cannot be included on the plot. These also corresponded to various levels of recipient unit where most of the survival times were censored. The simulation results in this section indicate that MI has the potential to improve performance in making point and interval estimates of regression coefficients compared with CC.

4.2 Simulation involving stepwise model selection

We also compare performance between MI and CC when performing stepwise regression. We again use the method of Wood *et al.* (2008) to perform stepwise regression with the multiply imputed data sets. Specifically we fit a Cox proportional hazards regression to the (fully observed) complete data, and perform backwards elimination stepwise regression to the full model, where the final model is selected with the smallest AIC. There are 14 variables retained which include donor age, BMI and CMV; recipient age, sex, ethnicity, serum creatinine, and primary disease; as well as the transplant factor whether or not the kidney was transplanted locally.

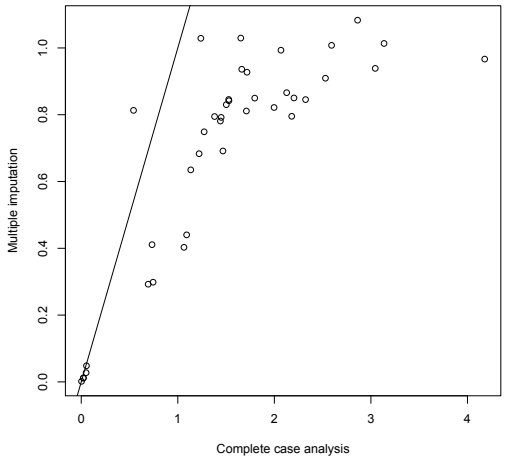

Figure 3 Plots of average 95% confidence interval lengths of coefficient estimates obtained from using MI and CC in the simulation involving the complete cases. Points below the $y = x$ line indicate a larger average length for CC.

We then run the same simulation as performed above, repeatedly introducing missing values into the complete case data and dealing with the missing values using MI or CC, but now we apply stepwise backwards elimination after fitting a Cox proportional hazards regression to the full data. We then compare performance between MI and CC as follows:

- The proportion of covariates in the final model from the original complete data, selected in the final model obtained from the incomplete data.
- The proportion of covariates selected in the final model obtained from the incomplete data that are not present in the final model from the complete data.

Figure 4 presents boxplots of these two measures. In the left plot, higher plots indicate better performance and in the right plot lower plots indicate better performance. We see that MI outperforms CC in both.

4.3 Sensitivity analysis

In this section we consider how sensitive the results from the previous section are to specification of the missing data mechanism and imputation model. To address these we consider firstly introducing missing values using a known MAR mechanism, and secondly mis-specifying the imputation model used to impute missing recipient BMI values. Results similar to those presented in the previous section are also presented for each of these scenarios.

4.3.1 MAR mechanism

The results from Sections 4.1 and 4.2 are not obtained from simulated data generated using an explicit missing data mechanism, rather the simulated data was generated to preserve the distribution of the missing data patterns present in the original data. To complement this, we also perform a simulation study where we know the exact mechanism that creates the missing data. Specifically, after subsetting on the complete cases, we introduce missing values into the

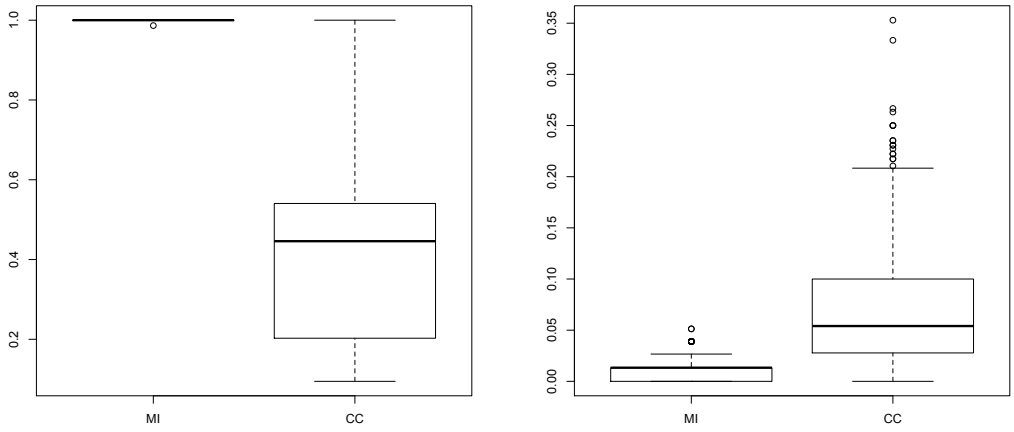

Figure 4 Proportion of correct covariates included (left) and proportion of covariates selected that should not be included (right) across the 250 replications

recipient BMI variable using an MAR mechanism. We do this by fitting a main effects logistic regression model to the missing data indicators corresponding to whether each recipient BMI value was observed or not, using all other variables in the data as covariates. Having created the complete data set of 4000 units as detailed above, we then use the coefficient estimates to estimate missing probabilities for each unit and introduce missing values repeatedly, taking draws each time from this distribution. We find that solely doing this with the completed data results in a smaller proportion of missing values in the recipient BMI variable than what occurred in the original data, with only approximately 24% of values missing approximately. This is because certain features of the data would have inevitably changed after subsetting on the complete cases, e.g. early transplant years would have been omitted as they all contained missing BMI values. To allow a more meaningful comparison to be made we introduce extra missing data into this variable using a MCAR mechanism where each unit has the same probability of 0.4 to be missing. This allows the proportion of missing values in this variable to be approximately equal to the proportion missing in the real data (0.6385). No other missing values are introduced into the complete case subsample, so all other variables remain fully observed. This section thus investigates the performance of MI in a scenario where the missing values arise from a known MAR mechanism constructed to incorporate features estimated from the original data. We note that this is a specific MAR mechanism which depends only on covariate values obtained from the complete case subsample and other MAR mechanisms could be considered. Under this mechanism the estimators of interest obtained from CC and MI should be consistent.

We perform similar analysis to that performed in Sections 4.1 and 4.2. Specifically, using the complete case subsample, we fit a Cox regression of survival time on the same 11 covariates that were included in the regression model from Section 3. We then introduce missing values into the recipient BMI variable using the MAR mechanism above and obtain coefficient estimates from fitting the Cox regression (including the same 11 covariates) to the incomplete data, where we use both MI and CC to handle the problem of missing data. We repeat this process 250 times, creating 250 incomplete data sets, and compare the results from both MI and CC to the estimates obtained from the regression model fit prior to introducing missing values. For similar reasons described in the previous section, it was not possible to obtain estimates for one coefficient.

Figures 5 and 6 summarise the results and show similar gains for MI over CC as seen in Section 4.1. Namely we see that MI results in estimates with a smaller (proxy) measure of mean squared error for all coefficients, with again more pronounced gains in the estimation of coefficients corresponding to categorical variables. In addition MI results in smaller average confidence interval lengths for all but two coefficients.

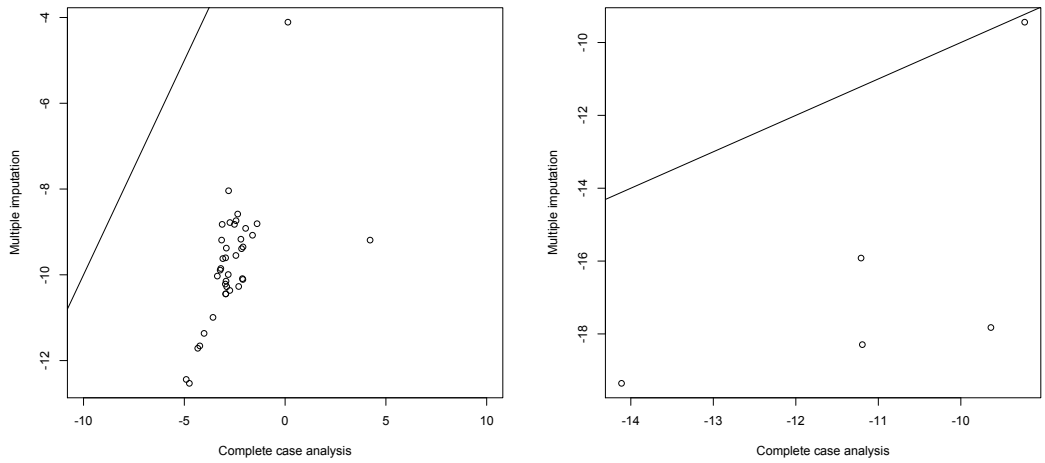

Figure 5 Plots of log(MSE) in coefficient estimates for categorical covariates (left) and continuous covariates (right) obtained from using MI against CC in the simulation involving the complete case subsample. Points below the $\log y = \log x$ line indicate a larger MSE for CC. Note that the plot is on the log scale.

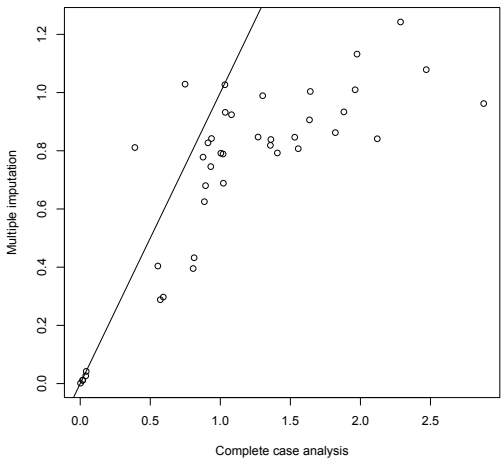

Figure 6 Plots of average 95% confidence interval lengths in coefficient estimates obtained from using MI and CC in the simulation involving the complete cases. Points below the $y = x$ line indicate a larger average length for CC.

Similar to Section 4.2, we also perform a stepwise backwards elimination model selection procedure when fitting the Cox model to the complete data. For each incomplete data set, we also perform the same stepwise procedure, using both MI and CC to deal with the missing data. An equivalent figure to that of Figure 4 is presented here (Figure 7) that summarises the results. We see similar results to those observed in Section 4.2. On average MI results in inclusion of a higher proportion of covariates that were originally included in the complete data model obtained through stepwise selection, as compared to CC. MI also results in inclusion of fewer covariates that were not originally included in the complete data model, again as compared to CC.

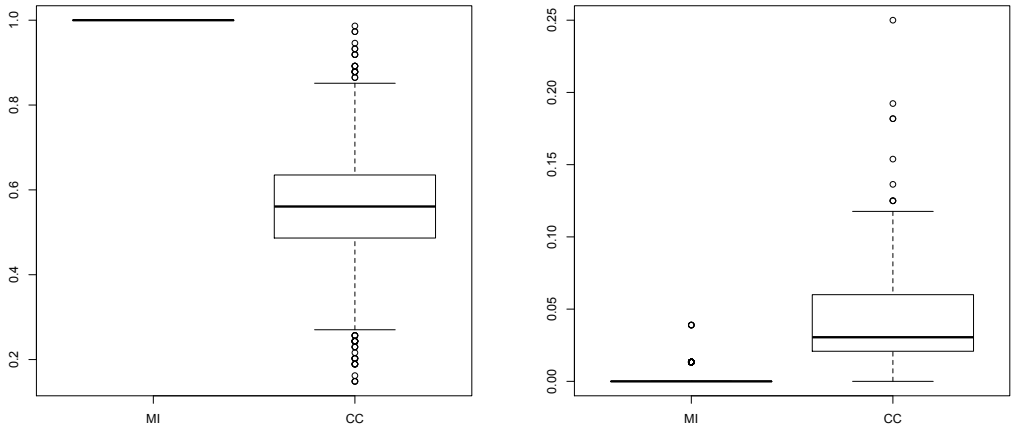

Figure 7 Proportion of correct covariates included (left) and proportion of covariates selected that should not be included (right) across the 250 replications

4.3.2 Imputation model mis-specification

We now consider the effect of mis-specifying the imputation model. Clearly there are many ways that the model could be mis-specified. We decide to consider the effect of leaving out a potentially highly predictive variable from the imputation of recipient BMI. We fit a main effects linear model to recipient BMI with all other variables in the data as covariates, and examine the most significant covariates. We find that recipient age is one of the most highly significant predictors with a p-value of 1.37×10^{-12}. Also intuition suggests that this should be an important predictor of BMI. Thus we choose to omit this variable when imputing missing recipient BMI. A complete data set and missing patterns are generated using the method described in Section 4.1. As before results are obtained over 250 replications. Similar comparisons and plots are produced as in Section 4.3.1.

Figures 8 and 9 summarise the results and still show gains for MI over CC as seen in Section 4.1. Namely we see that MI results in estimates with a smaller (proxy) measure of mean squared error for all coefficients, with more pronounced gains in the estimation of coefficients corresponding to categorical variables. In addition MI results in smaller average confidence interval lengths for all but one coefficient. It is encouraging still to see these gains even when the imputation model has been partly mis-specified by omitting an important predictor.

Similar to Section 4.2, we also perform a stepwise backwards elimination model selection procedure when fitting the Cox model to the complete data set. For each incomplete data set, we also perform the same stepwise procedure, using both MI and CC to deal with the missing data.

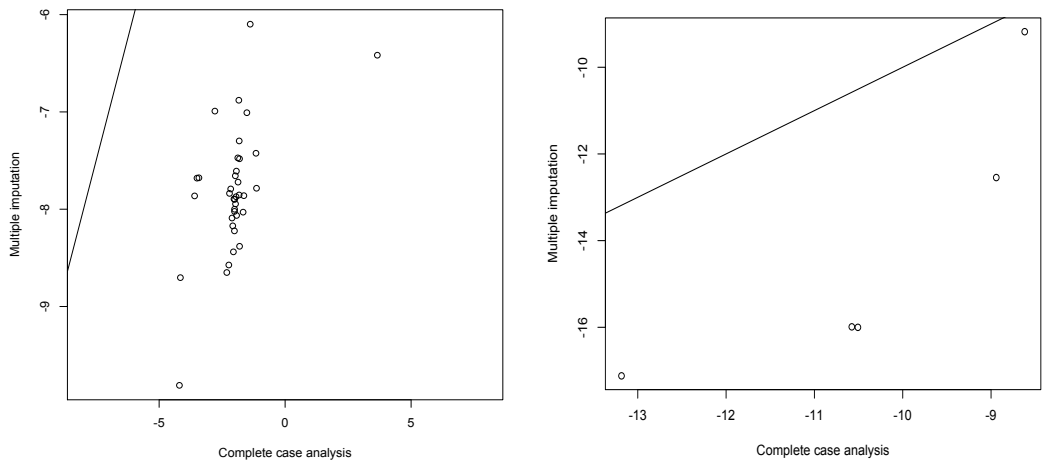

Figure 8 Plots of log(MSE) in coefficient estimates for categorical covariates (left) and continuous covariates (right) obtained from using MI against CC in the simulation involving the complete case subsample. Points below the $\log y = \log x$ line indicate a larger MSE for CC. Note that the plot is on the log scale.

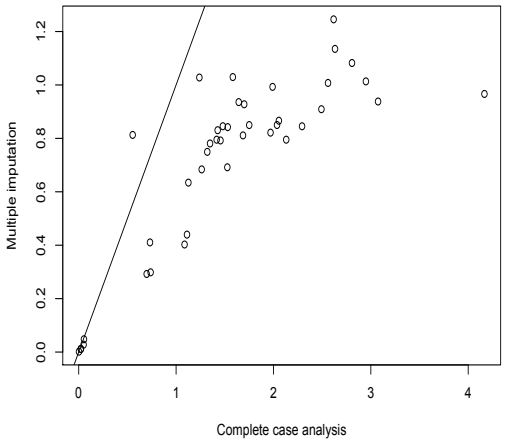

Figure 9 Plots of average 95% confidence interval lengths in coefficient estimates obtained from using MI and CC in the simulation involving the complete cases. Points below the $y = x$ line indicate a larger average length for CC.

An equivalent figure to that of Figure 4 is presented here (Figure 10) that summarises the results. We see the same general trend that was observed in Section 4.2. On average MI results in inclusion of a higher proportion of covariates that were originally included in the complete data model obtained through stepwise selection, as compared to CC. MI also results in inclusion of fewer covariates that were not originally included in the complete data model, again as compared

to CC. The gains here are not as pronounced as with the equivalent plots in previous sections, but this is to be expected as the imputation model has been partly mis-specified.

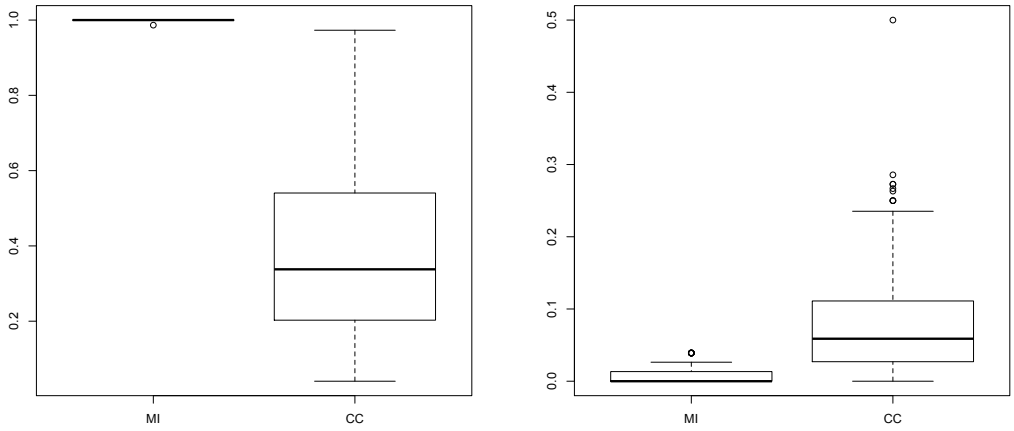

Figure 10 Proportion of correct covariates included (left) and proportion of covariates selected that should not be included (right) across the 250 replications

To summarise, the simulation results indicate that MI has the potential to offer significant gains over CC when the missing data arise by design or are MAR. In this paper we have only considered missing data mechanisms that are relevant to our application, and not other mechanisms such as informative mechanisms. Nevertheless, this finding does highlight an important, and perhaps less well known, issue that imputing missing values that arise through a systematic process, or a MAR mechanism, offers significant advantages over complete case analysis, even when the proportion of missing values is quite substantial and when the imputation model is partly mis-specified.

5 Conclusions

Missing data arise naturally in a variety of practical settings in medical studies. In particular, for routinely collected health data, such as the transplant survival data used to motivate this paper, lack of awareness that a specific covariate is important until part-way through the prospective data collection process can lead to a missing by design issue amongst the covariates (though not for the response variable). It is fair to say that such missingness by design is likely to fall under the heading of MAR so that CC might yield consistent estimates of regression coefficients. However, the difficulty here is that typically the incomplete observations still contain considerable information despite, say, one covariate out of several being missing. Thus CC is likely to be highly inefficient, even if it is consistent. In such circumstances MI gives an attractive solution that is relatively easy to apply.

In the analysis of the transplant survival data it is clear that the regression coefficient estimates are broadly similar whether one uses CC or MI. However, the variability of these estimates is considerably less when using MI. The simulation study indicates that, in a setting like the motivating data example, MI yields estimates with smaller MSE and also has benefits in terms of variable selection.

We note that our simulation studies have been constrained by the size of the complete case subsample. An alternative approach might be to consider generating a synthetic simulated data set

as an alternative to what has been proposed. Using parametric modelling assumptions would result in sensitivity to model mis-specification, so the challenge would be to determine an appropriate model that would be robust to this and produce a faithful representation of the original data. We note that there is a substantial body of literature in generating synthetic data to protect data confidentiality with strong ties to multiple imputation that could be considered here. An example would be to use classification and regression trees to generate a synthetic version of the data (Reiter, 2005). Careful thought would need to be given to how to deal with the missing values present in the original data. This would be an interesting area of future investigation.

We have not considered the problem of MNAR in this article. Given the primary reasons behind why missing values occur here we do not think this would be an issue here. However, in other situations MI may not necessarily outperform CC. In these situations it may be of interest additionally to consider approaches that use inverse probability weighting and doubly robust strategies (Carpenter *et al.*, 2006).

We are also aware of the causal implications of including the survival time (outcome) variable in imputation models for missing covariate values. This is something that can divide opinion. On the one hand omitting the outcome variable avoids the risk of distorting the causal path and has been a strategy employed in some situations (Mitra and Reiter, 2011; D'Agostino and Rubin, 2000). However, the important information included in the outcome variable could lead to a much better predictive distribution for missing covariate values, and it is typical practice in survival studies to include the survival time in imputation models (White and Royston, 2009), hence our choice to include survival times in imputation models.

As the variable subject to missing by design in the application was BMI, which is a variable derived from a person's height and weight, there is more than one way the imputation process could have been implemented, which leads to an interesting avenue for future work. For example, rather than imputing the derived variable, BMI here, directly, we could instead impute the variables used to construct the derived variable, here height and weight, and then construct the derived variable based on the imputed data. Evidence in the literature suggests the two approaches might lead to different results (Morris *et al.*, 2014) and it would be interesting to explore this further in the context of missingness by design.

Acknowledgements

The authors undertook this work under NIHR grant RMOFS2012/03. The authors are grateful to all the transplant centres in the UK who contributed data on which this article is based. The ACORN data was supplied by CACI Ltd. They also thank Ben Hopson for helping in the initial stages of the R programming.

References

Barber, K., Madden, S., Allen, J., Collett, D., Neuberger, J., Gimson, A., *et al.* (2011). Elective liver transplant list mortality: development of a united kingdom end-stage liver disease score. *Transplantation* **92**, 4, 469–476.

Bartlett, J. W., Harel, O., and Carpenter, J. R. (2015a). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology* **182**, 8, 730–736.

Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2015b). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research* **24**, 4, 462–487.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 1–68.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**, 3, 571–584.

D'Agostino, Jr., R. B. and Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association* **95**, 451, 749–759.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology* **14**, 1, 28.

Johnson, R. J., Fuggle, S. V., O'neill, J., Start, S., Bradley, J. A., Forsythe, J. L., Rudge, C. J., of NHS Blood, K. A. G., Transplant, *et al.* (2010). Factors influencing outcome after deceased heart beating donor kidney transplantation in the united kingdom: an evidence base for a new national kidney allocation policy. *Transplantation* **89**, 4, 379–386.

Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research* **16**, 3, 199–218.

Keogh, R. H., Seaman, S. R., Bartlett, J. W., and Wood, A. M. (2018). Multiple imputation of missing data in nested case-control and case-cohort studies. *Biometrics* .

Lee, M. C. and Mitra, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational Statistics & Data Analysis* **95**, 24–38.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing data (2nd edition)*. Wiley-Interscience.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika* **101**, 1, 155–173.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine* **30**, 627–641.

Morris, T. P., White, I. R., Royston, P., Seaman, S. R., and Wood, A. M. (2014). Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine* **33**, 1, 88–104.

Reiter, J. P. (2005). Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of Official Statistics* **21**, 3, 441–462.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 1, 3–15.

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

Van Buuren, S., Boshuizen, H., and Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 6, 681–694.

Watson, C. J., Johnson, R. J., Birch, R., Collett, D., and Bradley, J. A. (2012). A simplified donor risk index for predicting outcome after deceased donor kidney transplantation. *Transplantation* **93**, 3, 314–318.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine* **28**, 15, 1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 4, 377–399.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine* **27**, 17, 3227–3246.

Appendix - Additional metadata

Table 2 below presents some additional metadata of interest. This includes sample sizes for each continuous variable and level of a categorical variable in the original data, complete case data and multiply imputed data (presented as a range across imputations where appropriate). Table 2 also includes the proportion of missing values present in each variable in the original data.

covariate	number of patients			proportion of missing data
	original	CC	MI	
Primary renal disease				
1	380	94	380	0
2	564	177	564	0
3	1026	301	1026	0
4	640	204	640	0
5	1326	398	1326	0
6	930	260	930	0
7	2866	697	2866	0
Recipient unit				
1	226	112	226	0
2	505	23	505	0
3	633	257	633	0
4	173	50	173	0
5	153	72	153	0
6	224	105	224	0
7	425	148	425	0
8	282	2	282	0
9	319	82	319	0
10	408	7	408	0
11	348	123	348	0
12	241	80	241	0
13	361	66	361	0
14	396	164	396	0
15	236	117	236	0
16	491	92	491	0
17	104	60	104	0
18	295	34	295	0
19	603	45	603	0
20	293	153	293	0
21	270	109	270	0
22	376	40	376	0
23	370	190	370	0
ACORN index				
Wealthy achievers	1724	545	1753 –1771	0.0269
Urban prosperity	727	167	751 – 764	0.0269
Comfortably off	1952	524	1985 – 2017	0.0269
Moderate means	1277	374	1299 – 1322	0.0269
Hard pressed	1844	521	1883 – 1903	0.0269

Continued on next page

Table 2 – *Continued from previous page*

covariate	number of patients			proportion of missing data
	original	CC	MI	
Transplant year				
2001	970	0	970	0
2002	952	0	952	0
2003	931	47	931	0
2004	1038	403	1038	0
2005	915	359	915	0
2006	939	379	939	0
2007	914	414	914	0
2008	1073	529	1073	0
Recipient sex				
male	4792	1336	4792	0
female	2940	795	2940	0
Serum creatinine	6866	2131	7732	0.112
Donor age	7732	2131	7732	0
Recipient age	7732	2131	7732	0
Recipient BMI	2795	2131	7732	0.6385
Donor BMI	7315	2131	7732	0.0539
Donor CMV Status				
Negative	3855	1106	3923 – 3947	0.02018
Positive	3721	1025	3785 – 3809	0.02018

Table 2: Numbers of patients corresponding to each covariate included in the model based on the full data, complete case data and after multiple imputation respectively. For categorical variables, numbers relate to each level of the variable. The final column present the proportion of values missing in each variable. Primary renal disease categories: 1 - Glomerulonephritis, 2 - Pyelonephritis/Interstitial Nephritis, 3 - Miscellaneous, 4 - Polycystic kidneys, 5 - Hypertension/Renovascular Disease, 6 - Diabetes, 7 - Not Reported