Article

Multiple Imputation of Composite Covariates in Survival Studies

Lily Clements [1], Alan C. Kimber [2,*] and Stefanie Biedermann [3]

[1] School of Mathematical Sciences, University of Southampton; Lily.Clements@soton.ac.uk
[2] School of Mathematical Sciences, University of Southampton; A.C.Kimber@soton.ac.uk
[3] School of Mathematics and Statistics, The Open University; Stefanie.Biedermann@open.ac.uk
* Correspondence: A.C.Kimber@soton.ac.uk

Abstract: Missing covariate values are a common problem in survival studies, and the method of choice when handling such incomplete data is often multiple imputation. However, it is not obvious how this can be used most effectively when an incomplete covariate is a function of other covariates. For example, body mass index (BMI) is the ratio of weight and height-squared. In this situation, the following question arises: Should a *composite* covariate such as BMI be imputed directly, or is it advantageous to impute its *constituents*, weight and height, first and to construct BMI afterwards? We address this question through a carefully designed simulation study that compares various approaches to multiple imputation of composite covariates in a survival context. We discuss advantages and limitations of these approaches for various types of missingness and imputation models. Our results are a first step towards providing much needed guidance to practitioners for analysing their incomplete survival data effectively.

Keywords: multiple imputation; SMCFCS; FCS; composite covariate; survival analysis

1. Introduction

A problem faced by analysts in survival studies is the ubiquity of missing covariate values. If not handled appropriately, the effects can be wide ranging and the loss of data can lead to inefficiencies and introduce bias into analyses. A widely used approach to analyse incomplete data sets is multiple imputation (MI), introduced in [13]. However, there are situations where it is not obvious how this can be used most effectively.

A motivating data set for our research was supplied by NHS Blood and Transplant and is a rather typical routinely collected survival data set. It involves censored survival times for 7732 kidney transplant patients and contains information on 30 covariates thought to be potentially related to post-transplant survival. Full details are given in [11]. Whilst there are relatively few missing values for most of these covariates, one, Body Mass Index (BMI) of the kidney donor has over 60 per cent of values missing. BMI is defined as the ratio of an individual's weight in kilograms to the square of the height in metres:

$$BMI = weight(kg)/height(m)^2.$$

This is an example of a composite covariate: it is a function of two constituents, namely weight and height in this case. In [11] the authors investigate whether MI is an appropriate approach for dealing with such a high proportion of missing values in the context of survival analysis and show that MI outperforms listwise deletion of observations with at least one missing covariate (complete case analysis). However, they do not use the composite nature of BMI explicitly in their analysis.

There are several possible approaches to imputing a composite covariate. The two main ones are active imputation and passive imputation. In active imputation, also called "Just another variable", the composite covariate is imputed directly like any other variable. As a result, the functional relationship between the imputed composite covariate and its constituents is diminished. This is the approach used in [11]. In passive imputation, the constituents are imputed and the composite covariate is only then constructed. Hence

Citation: Clements, L.; Kimber, A.C.; Biedermann, S. Multiple Imputation of Composite Covariates in Survival Studies. *Preprints* **2022**, *1*, 0. https://doi.org/

the functional relationship is preserved in passive imputation. However, the relationship between the composite covariate and other variables in the data set, such as the outcome variable, can be underestimated since the other variables do not directly influence the imputation of the composite covariate.

To combat this issue with passive imputation, a modification called Substantive Model Compatible Fully Conditional Specification (SMCFCS) has been proposed; see, for example, [2], [6]. In SMCFCS the outcome variable is accounted for when imputing the composite covariate in order to preserve the relationship between the composite covariate and the outcome variable.

MI for composite covariates in a survival analysis context has not been widely explored in the literature, so our aim is to make a start to filling this gap. We will investigate the performance of active and passive imputation for survival data via a simulation study, the design for which has been informed by the motivating data set

In Section 2, we first introduce some background on missing data mechanisms and MI before describing the design of our simulation study and giving some criteria for comparing the performance of active and passive imputation. In particular, we introduce the variants of MI to be considered in our study and investigate how the missing data mechanism and the presence of further covariates - beyond the composite covariate and its constituents - will affect the performance of the different MI methods. The results are given in Section 3, and their implications are further discussed in Section 4.

2. Background and Methods

In this section, we first outline different missingness mechanisms since these may have an impact on the performance of MI. Following this is a brief overview of the general concept of MI. We then introduce more specific methods relating to MI, such as Fully Conditional Specification (FCS) and variants of active and passive imputation. Finally, we provide the design of our simulation study, with details of the process to generate the data, to set missing values, and the models to impute the missing values.

2.1. Missingness Mechanisms

Assume a $n \times p$ data set with n observations and p covariates. Denote this data set by $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ where $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})^{'}$ for the j^{th} covariate, $j = 1, ..., p$. For each x_{ij}, denote a missingness indicator, r_{ij}, where $r_{ij} = 1$ if x_{ij} is observed, and $r_{ij} = 0$ otherwise. These values build a missingness matrix $\boldsymbol{R} = (\boldsymbol{r}_1, ..., \boldsymbol{r}_p)$ where $\boldsymbol{r}_j = (r_{1j}, ..., r_{nj})^{'}$. $\boldsymbol{X}$ can be decomposed into a missing part, $\boldsymbol{X}_M$, and an observed part $\boldsymbol{X}_O$, where $\boldsymbol{X}_O$ represents the observed values in the data set $\boldsymbol{X}_O = (x_{ij}|r_{ij} = 1)$. Similarly, $\boldsymbol{X}_M = (x_{ij}|r_{ij} = 0)$ denotes the unobserved values.

When $P(\boldsymbol{R}|\boldsymbol{X}) = P(\boldsymbol{R})$, the incomplete values are Missing Completely at Random (MCAR) [12]. Hence, the distribution is the same between the observed portion of a variable, and the unobserved portion of that same variable.

When $P(\boldsymbol{R}|\boldsymbol{X}) = P(\boldsymbol{R}|\boldsymbol{X}_O)$ the incomplete values are Missing at Random (MAR) [12]. Missing values in an incomplete variable may depend on the observed values of other variables.

2.2. Multiple Imputation

First introduced by [13], MI is a widely used approach to handle missing values. In MI, the incomplete observations are imputed M times using an imputation model, yielding M complete data sets. Following this, an estimate of a parameter of interest, Q, is calculated for each multiply imputed data set using a substantive, or analysis, model [6]. Denote the M estimates of Q by $\hat{Q}_m$, with corresponding estimated variances, $\hat{V}_{W_m}$, $m = 1, ..., M$. In the *Pooling* phase, the estimates of the parameter of interest are combined by a set of rules called "Rubin's Rules" [14]. A pooled estimate, $\bar{Q}_M$, for the parameter of interest is calculated as the average of the M estimates:

$$\bar{Q}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m. \tag{1}$$

The estimated total variance of the pooled estimator of Q is given by

$$\hat{V}_T = \bar{V}_W + \left(1 + \frac{1}{M}\right)\hat{V}_B, \tag{2}$$

where $\bar{V}_W$ denotes the within-imputation variance,

$$\bar{V}_W = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_{W_m}, \tag{3}$$

and $\hat{V}_B$ denotes the between-imputation variance:

$$\hat{V}_B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m - \bar{Q}_M)^2. \tag{4}$$

2.2.1. Fully Conditional Specification (FCS)

Multiple imputation can be facilitated by Fully Conditional Specification (FCS). The procedure for FCS is as follows [1]:

1. To obtain initial values, all incomplete values in a data set are replaced with a "placeholder", such as the mean for that variable.
2. Take one variable with placeholder values, x_j, and set the placeholder values back to missing.
3. Subset the data set to the complete case form.
4. Fit a regression model where the outcome variable is x_j. Choose which of the remaining $p-1$ variables in the data set to fit as covariates. This regression model is an *Imputation Model*, denoted by $f(\boldsymbol{X}_M|\boldsymbol{X}_O)$.
5. Impute missing values in x_j by using the estimated coefficients from the imputation model.
6. Repeat steps 2-5 for any other variable that contains placeholder values.
7. Repeat steps 2-6 until the estimate of the parameter of interest converges. This results in a complete data set.

This results in one of the M complete data sets. The FCS process is repeated M times resulting in M imputed values for each missing value.

Within FCS, approaches such as Bayesian Linear Regression (BLR) or Predictive Mean Matching (PMM) discussed further in [16] can be utilised. However, FCS has some shortcomings when it comes to passive imputation. As discussed, passive imputation preserves the functional relationship between the constituents and the composite covariate by constructing the composite covariate after imputing the constituents. However, in passive imputation other variables in the data set do not influence the imputed value of the composite covariate. As a result, the effect of the covariate is attenuated, so that, for example, for a positive coefficient the bias of the estimator is negative [19].

One approach to combat the issues with passive imputation is to apply a modified version of FCS called SMCFCS. In SMCFCS, each incomplete variable is imputed with an imputation model compatible with a user-specified substantive model. Examples of achieving this are given in [2]; for example, by restricting the parameter space of the imputation model. As a result all variables in the analysis model are accounted for when imputing the composite covariate. Hence the relationships between the composite covariate and the other variables in the analysis model are preserved. This is discussed further in [2] and [6].

2.3. Simulation Study Design

A simulation study is conducted to compare the performance of active and passive imputation for a ratio functional form. In the simulation study a complete data set is first generated with the values chosen in the generating process based on the underlying data and analysis model used in [11]. This generating process is given in more detail in Section 2.3.1. Following this, three different missingness mechanisms are imposed: MCAR and two MAR mechanisms, outlined further in Section 2.3.2. The missing values are subsequently imputed by MI, discussed in 2.3.3. Approaches applied to evaluate the performance of the imputation models are given in Section 2.3.4.

In addition, other factors are varied in the simulation study to investigate firstly how these different factors impact the imputation process, and further investigate how they impact active and passive imputation. These additional factors are the number of observations in each replicated data set ($N = 500, 1000$, and 2000), the percentage of observations that are censored (10%, 15%, and 20%), whether an auxiliary variable, Z, is present, whether FCS or SMCFCS is applied. Additionally, in the case of FCS, another factor altered is whether BLR or PMM is applied.

The simulation is repeated for 1000 replications in line with the sample size calculation given in [5]. The substantive model fitted is an exponential AFT model, and so the parameter of interest, Q, is the true coefficient of the composite covariate in the substantive model. The simulation study is conducted in R. The simulation for FCS is performed using the MICE package [15], and the SMCFCS simulation is performed using the smcfcs package [3].

2.3.1. Generating the Data

The variables are generated to follow a structure similar to that of variables in the motivating data set analysed in [11]. Information from a data set on survival after a cardiothoracic transplant is also consulted to generate the variables to provide additional or supporting information.

Two constituents, U_1, U_2, are generated to follow a structure similar to that of weight in kg and height in cm, respectively. To account for the skewed distribution in the weight variable, $U_1 \sim Gumbel$ with location and scale parameters 64 and 14 respectively. The height variable, U_2, is generated from a linear regression model with both U_1 and $\log(U_1)$ as predictors to account for non-linearity in the relationship. An error term is also given to reflect the distribution of the height variable, $\epsilon \sim N(0, 8.6^2)$:

$$U_2 = -36.0 - 0.36U_1 + 54.0\log(U_1) + \epsilon.$$

These coefficient values are chosen since they are the estimated coefficients when fitting this linear model in the kidney data set with recipient height and weight. Then the composite covariate, X_3, is generated to be like BMI, hence $X_3 = \frac{U_1}{(U_2/100)^2}$.

Two further covariates, X_1 and X_2, having different correlations with X_3 are created. X_1 takes a distribution similar to recipient age and is generated from a linear regression model involving both constituents and the composite variable as predictors to maintain a relationship between X_1 and X_3,

$$X_1 = 3.2 - 0.12U_1 + 0.14U_2 + 1.18X_3 + \epsilon.$$

The coefficient values are the estimated coefficients calculated when fitting this linear model with the underlying variables in the motivating data set and $\epsilon \sim N(0, 13^2)$ since recipient age is roughly normally distributed, with $\sigma = 13$ to equal the standard deviation of recipient age in the motivating data set. The model to generate X_1 values can result in age values less than 20 which is out of the range of the motivating data sets. Therefore, any X_1 values less than 20 are then re-generated by $X_1 \sim U(20, 100)$. As a result, $cor(X_1, X_3) \approx 0.3$.

The second covariate is based on donor age. We generate $X_2 \sim N(40, 100)$, but if $X_2 < 20$ then we re-generate X_2 by $X_2 \sim U(20, 45)$. X_2 has virtually no relationship with X_3. As a result, $cor(X_2, X_3) \approx 0$. The parameter values are chosen so that X_2 reflects donor age in the motivating data set.

An auxiliary variable, Z, was generated to have a correlation of approximately 0.5 to the composite covariate in order to assess the effect of auxiliary variables on the performance of the MI variants. Hence Z is based on a variable 'waist measurement' from the US National Health and Nutrition Examination Survey investigated in [18].

$$Z = -8.8 + 0.21U_2 + 2X_3 + \epsilon,$$

where $\epsilon \sim N(0, 256)$. σ is slightly inflated from the standard deviation of waist measurement in the motivating data set to decrease the correlation between X_3 and Z to approximately 0.5. This approach can result in values that are too small to be realistic. Hence if $Z < 40$, Z is re-generated by $Z \sim U(40, 150)$ resulting in $cor(Z, X_3) \approx 0.5$.

Finally, survival time and a censoring indicator are produced. Survival time is generated by

$$time = \exp(6 - 0.02X_1 - 0.02X_2 + 0.05X_3 + \epsilon).$$

The chosen coefficient values are influenced by the estimated coefficients in the transplant data sets. Additionally, $\epsilon \sim$ Gumbel(0, 1). Hence the survival time is exponentially distributed. To have approximately 15% of observations censored, any observation with a survival time above 500 is right censored at 500. Censoring percentages of 10% and 20% are achieved analogously by varying the survival time where censoring takes place. A censoring indicator is then introduced.

2.3.2. Generating Missing Values

We chose to generate approximately 30% missingness in the composite covariate X_3 and three different missingness mechanisms are investigated, denoted by MCAR, MAR1 and MAR2. For MCAR, values of X_3 are set to missing independently with probability 0.3. In the MAR1 structure, X_3 is set to missing with probability 0.5 when X_1 is smaller than its median, otherwise the probability that X_3 is missing is 0.1. In the MAR2 structure, X_3 is missing for the smallest 30% of the X_1 values.

We consider three different situations for missingness in the constituents, i.e. scenarios where only height is observed, only weight is observed, and where both constituents are missing. To incorporate this, a dummy variable, W_1, is randomly generated for each row of the generated data with a missing value of X_3 such that

$$P(W_1 = 1) = 1/3$$
$$P(W_1 = 2) = 1/3$$
$$P(W_1 = 3) = 1/3.$$

When $W_1 = 1$, only the corresponding value of U_1 is set as missing. When $W_1 = 2$, only the corresponding value of U_2 is set as missing. When $W_1 = 3$, both U_1, U_2 values are set as missing.

2.3.3. Applying Multiple Imputation

MI is applied with $M = 30$. The choice of M is so that M is roughly equivalent to the percent of missing values in X_3, as recommended in [19].

Two different FCS procedures are evaluated. Firstly, FCS-BLR is investigated since it is a quick, commonly used approach that is designed for use with continuous variables, such as X_3. In addition, BLR is the approach applied in [11]. Also, FCS-PMM is investigated since PMM, unlike BLR, does not involve any underlying normality assumptions. Moreover, studies have found PMM to enhance the imputation procedure; see [10], [19]. In addition to FCS, SMCFCS is applied for passive imputation models with BLR because users of MI

often rely on readily available software and SMCFCS PMM does not currently fulfil this criterion.

Four imputation models are investigated, two of them being active and two passive. The presence of the outcome variables as predictors in the imputation model enables the values in the outcome variable to influence the imputation of the incomplete variable. This preserves the relationship between the outcome variables and incomplete variable. Hence, in all these imputation models, survival time and censoring indicator (called 'status' below) are used as predictors in order to avoid incompatibility issues [8]. Additionally, X_3 is not a predictor of either constituent in the imputation models in order to avoid circularity ([17]). The four imputation models are:

1. Active imputation without constituents present as predictors (AWO).

$$X_3 \sim X_1 + X_2 + time + status$$

2. Active imputation with constituents present as predictors (APA).

$$U_1 \sim U_2 + X_1 + X_2 + time + status$$
$$U_2 \sim U_1 + X_1 + X_2 + time + status$$
$$X_3 \sim U_1 + U_2 + X_1 + X_2 + time + status.$$

3. Standard Passive Imputation (PNP).

$$U_1 \sim U_2 + X_1 + X_2 + time + status$$
$$U_2 \sim U_1 + X_1 + X_2 + time + status$$
$$X_3 = \frac{U_1}{(U_2/100)^2}.$$

4. Log-Passive Imputation (LNP). In this imputation model, the constituents are first log-transformed before imputation takes place; $U_1^* = \ln(U_1), U_2^* = \ln(U_2)$:

$$U_1^* \sim U_2^* + X_1 + X_2 + time + status$$
$$U_2^* \sim U_1^* + X_1 + X_2 + time + status$$
$$X_3 = \exp(U_1^* - 2 \times (U_2^* - ln(100))).$$

Z is an additional predictor in the imputation models given when an auxiliary variable is present.

The substantive model fitted to the imputed data sets is an exponential AFT model:

$$surv(time, status) \sim X_1 + X_2 + X_3.$$

In each replication, the $M = 30$ estimated coefficients are pooled by Rubin's Rules, giving $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.

2.3.4. Comparing Imputation Models

To compare the imputation models, the pooled estimated coefficients of X_3, $\bar{Q}_M$ defined in (1), can be compared to the value of the true coefficient, $Q = \beta_3 = 0.05$. Since there are 1000 replications, the mean of the pooled estimates across all replications, $\hat{Q}$, can be calculated. One approach to evaluate the performance of the imputation models is to estimate the percentage bias (PB),

$$\text{PB} = |\frac{\hat{Q} - Q}{Q}| \times 100. \quad (5)$$

Another approach is to estimate the coverage rate (CR). The CR is the proportion of replications where the true value of Q is in the 95% Wald-type confidence interval. Ideally the CR should be close to 95%.

To distinguish between imputation models that perform well, the average width of the confidence intervals is calculated [16]. The mean average width (AW) is then calculated from all 1000 replications.

Finally, the between-imputation variances and within-imputation variances defined in (3) and (4), respectively, can be used to help identify underlying problems in the MI procedure ([7]). Two examples of this are Fraction of Missing Information (FMI) and Relative Increase in Variance (RIV), where

$$\text{FMI} = \frac{\hat{V}_B + \frac{\hat{V}_B}{M}}{\hat{V}_T}, \tag{6}$$

$$\text{RIV} = \frac{\text{FMI}}{1 - \text{FMI}} = \frac{\hat{V}_B + \frac{\hat{V}_B}{M}}{\bar{V}_W}, \tag{7}$$

where the total variance, $\hat{V}_T$, has been defined in (2).

From (6), FMI lies between 0 and 1 and indicates the proportion of the total variance in the estimated coefficients that is attributable to missing values in the associated variable. A large FMI value indicates that the missing values in the variable are causing a large proportion of the variability in the estimated coefficients.

Definition (7) shows that a large RIV value indicates that $\hat{V}_B$ is large relative to $\bar{V}_W$. Higher RIV values indicate either poor predictors in the imputation model for the associated variable or that a large proportion of the associated variable is missing and thus imputed [20]. FMI and RIV are calculated for each replication and averaged over the 1000 replications.

3. Results

Table 1 shows the main results of the simulation study when $N = 2000$ and 15% of observations are censored. Results from other values of N and proportions of censoring are given in the supplementary material ($N = 500, 1000, 2000$; proportions of censoring $10\%, 15\%, 20\%$). For each imputation method, each missingness structure with and without the presence of an auxiliary variable, the PB, CR and AW estimates are given. Table 2 contains the corresponding FIV and RIV results when $N = 2000$ and 15% of observations are censored. We first outline the general trends in the results under the missingness and auxiliary variable conditions and then discuss the overall performance of the imputation methods. When commenting on individual numbers, these are taken from Tables 1 and 2, while the overall results on the different imputation methods are supported by Tables S1-S8 in the supplementary material.

The effect of sample size, within the range considered, on PB and CR appears to be negligible. A higher censoring proportion slightly reduces CR and increases AW, with little effect on PB. The relative performance of the imputation methods compared with each other is not affected by either N or the proportion of censoring.

Table 1. PB, CR, and AW for the estimated coefficients of the composite covariate in a exponential AFT substantive model when $N = 2000$ and 15% of observations are censored.

			No Auxiliary Variables			One Auxiliary Variable		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MCAR	FCS-BLR	AWO	1.44	95.1	0.0239	0.88	96.2	0.0231
		APA	1.28	96.3	0.0230	0.88	96.6	0.0224
		PNP	3.12	95.1	0.0228	2.76	95.5	0.0223
		LNP	0.26	95.7	0.0231	0.08	96.7	0.0226
	FCS-PMM	AWO	1.92	94.3	0.0238	1.82	95.9	0.0229
		APA	1.14	96.0	0.0232	0.58	96.7	0.0222
		PNP	0.86	95.6	0.0232	0.24	96.8	0.0221
		LNP	0.60	95.8	0.0231	0.16	96.8	0.0224
	SMCFCS-BLR	PNP	0.06	96.5	0.0228	0.02	96.2	0.0222
		LNP	0.54	96.4	0.0230	0.98	96.5	0.0224
MAR1	FCS-BLR	AWO	1.66	95.1	0.0231	0.28	95.8	0.0224
		APA	0.96	96.2	0.0223	0.20	96.1	0.0219
		PNP	1.08	96.7	0.0222	1.84	95.6	0.0217
		LNP	2.56	95.3	0.0224	1.74	96.0	0.0221
	FCS-PMM	AWO	2.46	95.0	0.0232	1.54	96.2	0.0222
		APA	4.80	93.1	0.0226	2.96	95.1	0.0217
		PNP	4.46	93.4	0.0226	2.46	95.9	0.0216
		LNP	2.90	95.3	0.0225	1.94	96.1	0.0219
	SMCFCS-BLR	PNP	0.12	96.8	0.0221	1.06	96.1	0.0216
		LNP	0.58	96.2	0.0223	0.34	96.0	0.0218
MAR2	FCS-BLR	AWO	4.46	93.3	0.0224	1.32	94.8	0.0219
		APA	2.56	95.1	0.0218	0.76	94.8	0.0215
		PNP	0.32	95.2	0.0217	1.34	94.5	0.0213
		LNP	4.64	92.4	0.0220	2.84	94.6	0.0217
	FCS-PMM	AWO	0.64	76.7	0.0231	3.20	93.7	0.0216
		APA	7.12	88.4	0.0223	4.40	93.3	0.0213
		PNP	6.88	88.8	0.0222	3.88	92.9	0.0212
		LNP	5.24	91.2	0.0221	3.26	93.8	0.0216
	SMCFCS-BLR	PNP	0.74	94.5	0.0212	2.44	94.2	0.0210
		LNP	0.10	95.4	0.0218	0.80	95.1	0.0214

Table 2. Mean FMI and mean RIV values for $\hat{\beta}_3$ over the 1000 replications for all imputation models.

			No Auxiliary Variables		One Auxiliary Variable	
			FMI	**RIV**	**FMI**	**RIV**
MCAR	FCS-BLR	AWO	0.304	0.436	0.256	0.344
		APA	0.244	0.322	0.209	0.263
		PNP	0.259	0.348	0.223	0.286
		LNP	0.224	0.288	0.193	0.238
	FCS-PMM	AWO	0.281	0.390	0.245	0.324
		APA	0.229	0.296	0.197	0.244
		PNP	0.230	0.297	0.200	0.249
		LNP	0.225	0.289	0.195	0.241
	SMCFCS-BLR	PNP	0.242	0.326	0.204	0.260
		LNP	0.216	0.280	0.179	0.222
MAR1	FCS-BLR	AWO	0.296	0.420	0.250	0.333
		APA	0.231	0.300	0.204	0.255
		PNP	0.249	0.331	0.218	0.278
		LNP	0.202	0.252	0.175	0.211
	FCS-PMM	AWO	0.247	0.328	0.210	0.264
		APA	0.208	0.262	0.179	0.217
		PNP	0.207	0.260	0.180	0.218
		LNP	0.199	0.248	0.174	0.210
	SMCFCS-BLR	PNP	0.232	0.309	0.197	0.249
		LNP	0.191	0.240	0.158	0.190
MAR2	FCS-BLR	AWO	0.286	0.400	0.243	0.320
		APA	0.223	0.286	0.196	0.243
		PNP	0.240	0.314	0.210	0.265
		LNP	0.185	0.226	0.159	0.188
	FCS-PMM	AWO	0.248	0.344	0.178	0.215
		APA	0.188	0.231	0.161	0.191
		PNP	0.188	0.230	0.161	0.191
		LNP	0.178	0.215	0.154	0.181
	SMCFCS-BLR	PNP	0.222	0.289	0.189	0.236
		LNP	0.169	0.206	0.139	0.163

3.1. MCAR

MCAR is the simplest of the missingness structures and so serves as a baseline condition. All imputation methods have biases that are small in magnitude (PB $\leq$ 3.12 when there is no auxiliary variable and PB $\leq$ 2.76 when an auxiliary variable is present). All CR-values are close to the nominal value of 95 per cent. To put this in context, using the Wilson score method [4], CR $<$ 93.7 leads to a 95 per cent confidence interval for the true value of CR that does not contain the nominal value of 95 per cent. So, any such values supply evidence of underestimation. When an auxiliary variable is present PB is smaller, CR is higher and AW is lower than when there is no auxiliary variable. The only exceptions are SMCFCS-LNP in the case of PB and SMCFCS-PNP for CR.

3.2. MAR1

Firstly, comparing the results for the auxiliary variable conditions, PB is lower when the auxiliary variable is present than when it is absent (with the exception of SMCFCS-PNP), the CR is higher (with the exception of FCS-BLR-APA and the SMCFCS methods) and AW is lower. For FCS-PMM-APA and FCS-PMM-PNP with no auxiliary variable the CR-values are low, but the presence of an auxiliary variable brings these up to a better level. Comparison of MCAR and MAR1 results shows that PB is higher for MAR1 (except SMCFCS-LNP) and AW is lower for MAR1. The pattern for CR is less clear.

3.3. MAR2

Leaving aside FCS-PMM-AWO for the moment, PB is lower when the auxiliary variable is present (except for the SMCFCS methods). The main pattern in CR is that methods that have low CR-values with no auxiliary variable, have higher CR with an auxiliary variable: FCS-BLR-AWO, FCS-BLR-LNP and all the FCS-PMM methods. FCS-PMM-AWO has a particularly poor CR for the case of no auxiliary variable and a borderline CR with an auxiliary variable. Comparison of MCAR and MAR2 results shows the main feature to be that CR is lower for MAR2 than for MCAR (except FCS-BLR-PNP) and AW is also lower for MAR2 than MCAR.

3.4. Imputation methods

First consider the FCS-BLR methods. AWO has low CR under MAR2 (no auxiliary variable) and has uniformly highest AW across all conditions. LNP has a relatively high PB (4.46) and low CR under MAR2 (no auxiliary variable). APA and PNP have broadly satisfactory results. The FCS-PMM methods are less satisfactory here. AWO has a particularly poor CR under MAR2 (no auxiliary variable), APA and PNP have several low CR-values and high PB-values, while LNP has low CR under MAR2. The SMCFCS methods have relatively low PB, satisfactory CR and reasonably low AW under all conditions. A review of the FMI and RIV values in Table 2 shows that of the four methods with the best overall performance SMCFCS-LNP has the lowest values across the board.

3.5. FCS-PMM

As discussed in Section 3.4, FCS-PMM revealed several weaknesses under the MAR schemes, with large values of PB compared with FCS-BLR and SMCFCS-BLR and evidence of, sometimes severe, undercoverage. Investigating this further, we found that a reason for these issues may be incompatability of the way PMM imputes missing values with certain MAR mechanisms: Under PMM, the missing values of X_3 are imputed to follow the distribution of the observed X_3 values. However, under the MAR schemes used in this simulation study, the observed X_3 values are negatively skewed since smaller X_3 values are more likely to be missing; see Figure 1.

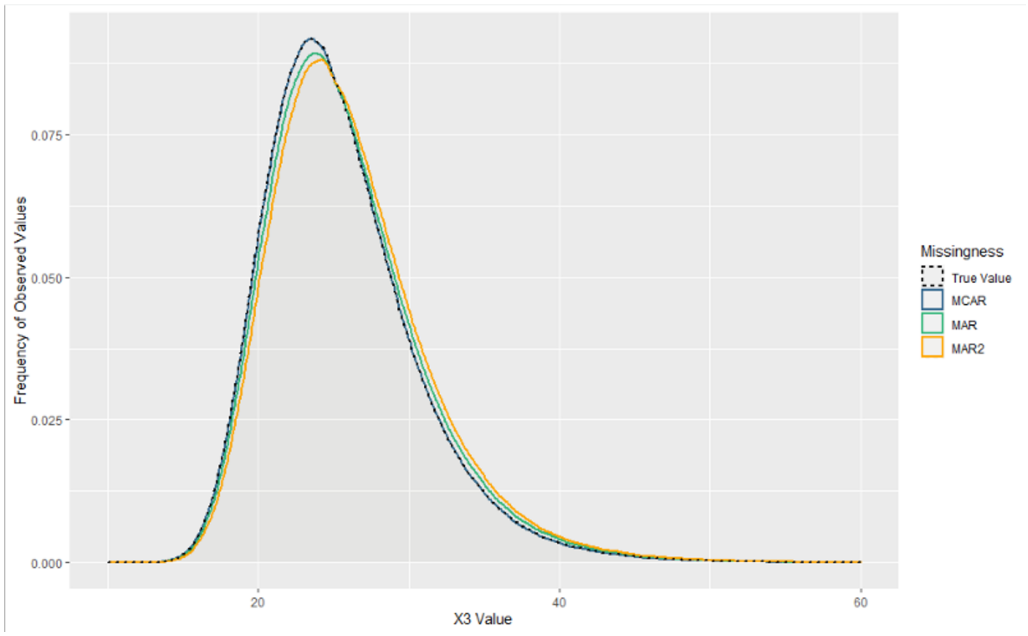

Figure 1. Kernel density plot of observed X_3 values for the different missingness structures, and for X_3 when no missing values are generated.

This in turn affects the relationship between X_3 and the outcome variable as shown in Figure 2. While this relationship is preserved under MCAR missingness, there are clear differences in the relationship between 'survival time' and, respectively, the observed and the missing values of X_3 under MAR2. Imputing the missing values from the distribution of the observed values may thus result in estimated coefficients further from the true value of β_3 and thus increased PB and undercoverage.

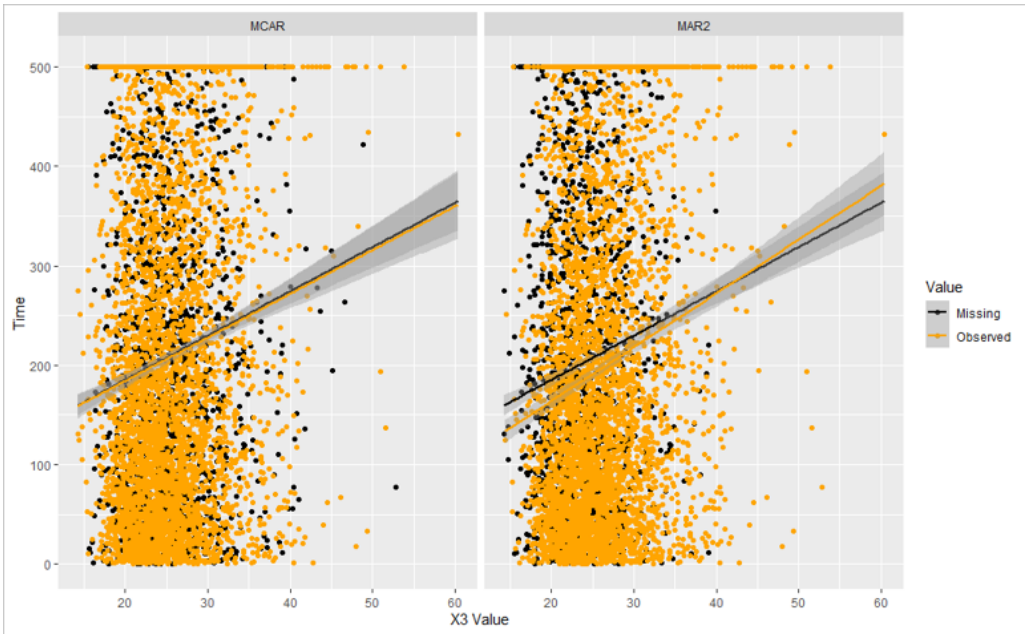

Figure 2. The relationship between X_3 and survival time, split by whether X_3 is missing or observed. These plots are displayed for an MCAR missingness structure and an MAR2 missingness structure. The values plotted are a random sample of three generated data sets of 2000 values each.

4. Discussion

In this paper we have undertaken a simulation study within a survival analysis context to investigate various aspects of MI. Whilst the study is modest in scale and in no way definitive, a number of aspects are worthy of attention.

First, it is rare to know for sure what the missingness mechanism is in a real application and clearly just because an imputation method performs well under an MCAR structure, it need not do so under a MAR structure. In practice it may be that a chosen MI approach with good MCAR properties may perform less well if the missingness (unbeknown to the analyst) is MAR. So, anything that can help the performance of such an MI method is to be welcomed. In our study, at least under the MAR1 structure, the presence of an auxiliary variable can be useful. For example, FCS-PMM-APA with no auxiliary variable has a low CR but its CR is reasonable in this case when an auxiliary variable is present. Conversely, in our study if an MI method performs well with no auxiliary variable (for example, SMCFCS-BLR-LNP under MAR2), it retains a good performance in the presence of an auxiliary variable. So, the indication is that if an appropriate auxiliary variable is available, it is worth considering incorporating it into the MI approach.

Secondly, we are interested in this study, in simple terms, about whether active or passive MI is preferable in a survival analysis context with BMI. In our study active imputation methods have performed well (FCS-BLR-APA) and poorly (FCS-PMM-AWO). Likewise passive imputation methods have performed well (FCS-BLR-PNP) and poorly (FCS-PMM-PNP). So the general question "which is better, active or passive imputation?" is too simplistic. Rather, it is important also to bring in other factors, such as whether to use BLR or PMM in further studies or in practice.

Thirdly, the idea of logging a ratio before undertaking imputation seems like an obvious thing to do. Typically, ratios like BMI are positive and positively skewed. Basic

statistics indicates that taking logs of such variables may make them less skewed. But in our study pre-imputation logging is not always beneficial; for example, FCS-PMM-LNP did not perform well with MAR2 and no auxiliary variable. On the other hand, SMCFCS-BLR-LNP was arguably the best performing approach in our study. So, it is important to bring in other factors, such as whether to use FCS or SMCFCS in further studies or in practice.

Fourthly, [2] noted that more investigation of SMCFCS in a censored data context is needed. We concur with this given the results in the present study where SMCFCS-BLR-LNP in particular performed well in all scenarios. In addition, for BLR-PNP, SMCFCS showed improved performance compared with FCS.

The main limitation of our study is the focus on the exponential model as the substantive model, combined with a Type I censoring scheme albeit for different sample sizes and censoring percentages. While we have found several interesting results for these scenarios with different missingness mechanisms and in the presence/absence of an auxiliary variable, it would be interesting to see if our conclusions generalise to further widely used survival models, such as the Weibull model or Cox's proportional hazards model, with a variety of commonly encountered censoring schemes. Further avenues of interest that could be explored in future research include investigating different composite covariates and different percentages of missingness in the composite covariate and its constituents. Realistic simulation scenarios could be based on survival studies beyond those on organ transplantation, to broaden the appeal, and to increase the benefits, to practitioners in the area of survival studies.

Author Contributions: "Conceptualization, L.C., A.C.K. and S.B.; methodology, L.C., A.C.K. and S.B.; software, L.C.; validation, L.C., A.C.K. and S.B.; formal analysis, L.C.; investigation, L.C.; writing—original draft preparation, L.C.; writing—review and editing, A.C.K and S.B.; supervision, A.C.K. and S.B. All authors have read and agreed to the published version of the manuscript."

Funding: Lily Clements's research was funded by an EPSRC PhD studentship at the University of Southampton.

Data Availability Statement: Data generated in the simulation study can be found at https://github.com/lilyclements/mice-data
The kidney transplant data we used to motivate our simulation scenarios can be found at https://www.odt.nhs.uk/statistics-and-reports/access-data/ and the National Health and Nutrition Examination Survey can be found at https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=1999.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BMI	Body Mass Index
MI	Multiple Imputation
FCS	Fully Conditional Specification
SMCFCS	Substantive Model Compatible Fully Conditional Specification
MCAR	Missing Completely at Random
MAR	Missing at Random
BLR	Bayesian Linear Regression
PMM	Predictive Mean Matching
MAR1	First MAR structure used in the simulation study
MAR2	Stricter MAR structure used in the simulation study
AWO	Active Imputation when the constituents are not predictors
APA	Active Imputation when the constituents are predictors
PNP	Standard Passive Imputation
LNP	Passive Imputation when the constituents are first log-transformed
PB	Percentage Bias
CR	Coverage Rate
AW	Average Width
FMI	Fraction of Missing Information
RIV	Relative Increase of Variance

References

1. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*. **2011** *20(1)*, 40–49.
2. Bartlett JW, Morris TP. Multiple imputation of covariates by substantive-model compatible fully conditional specification. *The Stata Journal*. **2015**, *15(2)*, 437–456.
3. Bartlett J, Keogh R, Bonneville E, Thorn Ekstrøm, C. Package 'smcfcs'. Available online: https://github.com/jwb133/smcfcs (accessed on February 2021).
4. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* **2001**, *16*, 101–133.
5. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* **2006**, *25(24)*, 4279–4292.
6. Carpenter J, Kenward, M. *Multiple Imputation and its Applications*; Publisher: Wiley and Sons, **2012**.
7. Enders CK. *Applied Missing Data Analysis*; Publisher: Guilford press, **2010**.
8. von Hippel, PT. How to impute interactions, squares, and other transformed variables. *Sociological Methodology* **2009**, *39(1)*, 265–291.
9. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*; Publisher: Pearson, **2014**.
10. Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine*. **2014** *33(1)*, 88–104.
11. Pankhurst L, Mitra R, Kimber AC, Collett D., Multiply imputing missing values arising by design in transplant survival data. *Biometrical Journal* **2020**, *62*, 1192–1207.
12. Rubin DB, Inference and missing data. *Biometrika* **1976**, *63(3)*, 581–592.
13. Rubin DB. Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In Proceedings of the Survey Research Methods Section of the American Statistical Association 1978. *1*, 20–34.
14. Rubin DB. Multiple imputation for survey nonresponse. **1987**.
15. van Buuren, S. Package 'mice'. Available online: https://github.com/cran/mice (accessed on January 2020).
16. van Buuren, S. *Flexible Imputation of Missing Data*; Publisher: Chapman and Hall/CRC, **2018**.
17. van Buuren, S. MICE: Passive Imputation and Post-Processing. Available online: https://www.gerkovink.com/miceVignettes/Passive_Post_processing/Passive_imputation_post_processing.html (accessed on 11 March 2019).
18. Wagstaff DA, Kranz S and Harel O. A preliminary study of active compared with passive imputation of missing body mass index values among non-Hispanic white youths. *The American Journal of Clinical Nutrition*. **2009**, *89(4)*, 1025–1030.
19. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*. **2011**, *30(4)*, 377–399.
20. Eddings, W. A note on how to perform Multiple-Imputation diagnostics in Stata. Available online: http://www.stata.com/users/ymarchenko/midiagnote.pdf (accessed on 20 May 2020).