Research Article

Statistics in Medicine

Received 14 November 2014, Accepted 20 January 2016 Published online 18 February 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6902

A multiple imputation approach for MNAR mechanisms compatible with Heckman's model

Jacques-Emmanuel Galimard,[a,b*†] Sylvie Chevret,[a,b,c] Camelia Protopopescu[d] and Matthieu Resche-Rigon[a,b,c]

Standard implementations of multiple imputation (MI) approaches provide unbiased inferences based on an assumption of underlying missing at random (MAR) mechanisms. However, in the presence of missing data generated by missing not at random (MNAR) mechanisms, MI is not satisfactory. Originating in an econometric statistical context, Heckman's model, also called the sample selection method, deals with selected samples using two joined linear equations, termed the selection equation and the outcome equation. It has been successfully applied to MNAR outcomes. Nevertheless, such a method only addresses missing outcomes, and this is a strong limitation in clinical epidemiology settings, where covariates are also often missing.

We propose to extend the validity of MI to some MNAR mechanisms through the use of the Heckman's model as imputation model and a two-step estimation process. This approach will provide a solution that can be used in an MI by chained equation framework to impute missing (either outcomes or covariates) data resulting either from a MAR or an MNAR mechanism when the MNAR mechanism is compatible with a Heckman's model. The approach is illustrated on a real dataset from a randomised trial in patients with seasonal influenza. **Copyright © 2016 John Wiley & Sons, Ltd.**

Keywords: missing data; missing not at random (MNAR); multiple imputation: chained equation; Heckman's model; sample selection

1. Introduction

Although unwanted, missing data occur in almost all clinical and epidemiological research. Missing data are usually classified following the Little and Rubin framework [1, 2]: either missing completely at random (MCAR), missing at random (MAR) when the probability of data being missing does not depend on unobserved data conditionally on the observed data or missing not at random (MNAR) when the probability of being missing still depends on unobserved data. Inadequate handling of missing data in statistical analyses can lead to biased and/or inefficient estimates of the parameters of interest such as means, regression coefficients or standard errors, resulting in potential incorrect conclusions. However, if methods to handle MCAR or MAR data in clinical epidemiology have been widely described and studied, methods adapted for MNAR mechanisms are less studied [1, 3]. Nevertheless, in certain clinical contexts such as adherence studies, one cannot reasonably exclude MNAR mechanisms, given that missing data on adherence are most likely related to the adherence itself, requiring methods appropriate for MNAR data.

Given that MAR and MNAR are not fully verifiable from the data, underlying assumptions can be evaluated by creating departures from those assumptions [3, 4]. Thus, most of the time, such assumptions are addressed by performing such sensitivity analysis. In contrast, specific methods to address MNAR

[a]*INSERM U1153, Statistic and Epidemiologic Research Center Sorbonne Paris Cité (CRESS), ECSTRA Team, Saint-Louis Hospital, 75010 Paris, France*
[b]*Paris Diderot University, Paris 7, SPC, Paris, France*
[c]*SBIM, Saint-Louis Hospital, APHP, Paris, France*
[d]*The BIVIR Study Group, INSERM UMR 912 (SESSTIM), 13006, Marseille, France*
**Correspondence to: Jacques-Emmanuel Galimard, SBIM, Hôpital Saint-Louis, 1 avenue Claude Vellefaux, 75010 Paris, France.*
†*E-mail: jacques-emmanuel.galimard@inserm.fr*

missing data in the scientific literature are rare. This is largely because estimation is more difficult when data are MNAR than when data are MAR or MCAR. Indeed, it implies that in practice, it is necessary to either (i) describe the statistical relationship between the chance of seeing a variable and its (unseen) value or (ii) describe the difference in the distribution of patients with and without missing observations [5]. Then, the key point of such approaches consists in modelling the missing mechanism. However, in clinical epidemiology, such mechanisms are rarely precisely known, and researchers prefer to include explanatory variables in the analyses as much as possible to make more plausible a MAR assumption and bet on the fact that relatively little information remains in the unseen data [5, 6].

Several approaches have been proposed to specifically address MNAR data, such as pattern-mixture models [7–10], protective estimation [11] or sample selection method [3]. The latter approach, also known as the Heckman's model, has been popular primarily in econometric statistics [12, 13]. It was originally proposed in 1976 by James J. Heckman to address selected samples using two joined equations, termed the selection equation and the outcome equation. This method has been successfully applied to missing outcome issues, where the sample without missing data is considered a selected sample. In practice, the two equations are joined by their error terms through a bivariate normal distribution. Originally developed for continuous outcomes based on a probit model for the selection equation and a linear model for the outcome equation, it has been extended to other types of outcomes, such as binary outcomes, using various general linear models as the outcome equation [14, 15]. Heckman proposed a two-step estimator. The first step consists of estimating parameters of the selection model, while the second step, using a correction term obtained from the first step estimates, allows the calculation of unbiased estimates of the outcome equation parameters. A constraint of the Heckman's model implies the inclusion of different sets of covariates, which may or not be nested in the selection equation and the outcome equation, to avoid collinearity issues [16]. Under this so-called exclusion-restriction rule, valid estimates are obtained [16]. Nevertheless, this approach deals only with missing outcomes, while missing data on covariates are handled by excluding subjects with those missing observations from the analysis. Of course, such a complete case analysis is not satisfactory. It leads, at least, to a loss of information in the case of MCAR mechanisms, or to potentially biased estimates in the case of MAR mechanisms with regard to these covariates.

In the presence of missing data that are not MNAR, several approaches have been proposed. A non-exhaustive list of methods includes direct likelihood methods [3], weighted estimating equations [17] and multiple imputation (MI) [18–20]. All these approaches, usually considered to be valid under MAR, differ in terms of models, assumptions and difficulties in implementation [3]. In the presence of complex patterns of missing data, MI appears to be the method of choice, notably because of its flexibility and ease of implementation due to the widespread development of statistical packages dedicated to its implementation [6]. The key concept of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing data. Therefore, MI is a Bayesian procedure: Given a joint prior distribution for the observed data and a specific data model, we obtain a posterior distribution of the missing values given the observed data [6]. Then, random draws are made from this posterior distribution to impute missing values. To reflect the uncertainty in plausible values, several multiple datasets are created and analysed individually but identically to obtain a set of parameter estimates. Finally, the estimates are combined to obtain the final estimates, variances and confidence intervals using Rubin's rules [18]. In the presence of several variables with missing values, imputed datasets could be obtained either by joint modelling or by chained equations. Joint modelling involves specifying a multivariate distribution of the missing data and usually relies on Markov Chain Monte Carlo methods to draw imputed values [21]. MIs by chained equations (MICE) is a simple, flexible and practical method to generate final imputations based on a sequential approach. This approach involves specifying a separate imputation model for each incomplete variable, given all other variables, and repeatedly imputing the variables in an iterated sequence [6]. After, generally, 5 to 20 iterations, the process is considered to give the correct posterior distribution of the missing data. As with MI in general, it is crucial that the imputation model is consistent (or congenial) with the model of interest that will subsequently be fitted to the imputed datasets [2, 22].

Despite all these available approaches, it remains true that, in the presence of data with several types of suspected missingness mechanisms, one lacks a general framework able to address MNAR and MAR missing data in the same estimation process. Accordingly, we propose to extend the MI approach to MNAR missing data when the MNAR mechanism is compatible with a Heckman's model. More specifically, we propose using an imputation model based on the Heckman's model with a two-step estimator and, hence, to consider a MICE procedure to simultaneously address several types of missingness

mechanisms in the same dataset and impute both MNAR and MAR missing data in the outcome or the covariates when the MNAR mechanism is compatible with a Heckman's model.

The paper is organised as follows. Section 2 presents the BIVIR data, the motivating example evaluating the efficacy of a bitherapy for seasonal influenza on a general-population basis, exhibiting a non-negligible amount of missing data both for secondary outcomes and for covariates. In Section 3, we describe the Heckman's model and propose an imputation model. We then make explicit the MICE process. In Section 4, the performance of the developed imputation model is evaluated using a simulation study. Section 5 illustrates this two-step approach using the BIVIR data. Finally, Section 6 provides some discussion.

2. Motivating example

The BIVIR study was a three-arm parallel randomised clinical trial aiming to assess the efficacy of the oseltamivir–zanamivir combination relative to each monotherapy in patients with seasonal influenza. It was conducted by 145 general practitioners throughout France during the 2008–2009 seasonal influenza epidemic and included 541 patients. The primary analysis of the trial showed that the oseltamivir–zanamivir combination appeared less effective than oseltamivir monotherapy and not significantly more effective than zanamivir monotherapy on the proportion of patients with nasal influenza reverse transcription-PCR below 200 copies [23]. We focused our work on a secondary objective of the study, which was to assess the impact of initial body temperature, sick leave status and tobacco status at inclusion on the persistence of flu symptoms at 48 h after randomisation. This outcome was measured using a severity score, defined as a weighted sum (ranging from 0 to 78) of 13 intensity symptoms on day 2 [24]. Unfortunately, severity score information based on self-report was missing for 127 (23%) patients. In view of the self-report method of data collection, the MNAR missing data mechanism was highly likely. Indeed, patients with very serious symptoms or, at the opposite extreme, those without symptoms might be more likely not to record data on their severity symptoms.

In such circumstances, in which there are missing data on the outcome, possibly MNAR, one could propose to directly apply a Heckman's model with a probit selection model and a linear outcome model including risk factors. Of course, one should keep in mind the exclusion-restriction criteria, that is, at least one covariate is included in the selection equation and not in the outcome equation. As it has been recommended that this variable should be known to be unlinked directly to the outcome, thus, we introduced the gender in selection model [16].

In addition to the missing data on the outcome, the initial body temperature was also missing for 182 (34%) patients. We supposed that the mechanism of missingness was MAR. The methods that we present below are notably dedicated to handle such MNAR missing data in the outcome and MAR missing data in the predictor. Moreover, we will extend our approaches to handle such MNAR and/or MAR missing data whatever their model position, either in the outcome or in the predictor.

3. Methods

Let Y_i be a continuous variable of interest (or outcome) and X_i a p row vector of covariates for individual $i = 1, \ldots, n$. Let us adopt the following linear regression model

$$Y_i = X_i\beta + \varepsilon_i \tag{1}$$

as the analysis equation (also called the outcome equation in the Heckman's framework), with β a p-vector of fixed effects and independent $\varepsilon_i \sim N(0, \sigma^2)$. Let us suppose an MNAR mechanism underlying missing values of Y. We propose an MI model for MNAR missing data adapted from the Heckman's model. We will then describe the use of the new imputation model in a chained equations framework.

3.1. Heckman's model

As previously described, Heckman's model, also called sample selection method, deals with non-random samples. Van Buuren and Greene have described the sample selection method available to handle non-ignorable missing data (MNAR) [25, 26]. Let us suppose missing data only in Y. Let R_{yi} be the indicator

of Y_i missingness (equal to 1 if Y_i is observed and 0 if Y_i is missing). Heckman's model introduces a selection equation that represents the non-random sampling of the missingness process:

$$P\left(R_{yi}=1|X_i^s\right)=\Phi\left(X_i^s\beta^s\right) \tag{2}$$

where X_i^s is a q-row vector of variables potentially associated with the missingness, β^s is a q-vector of coefficients and Φ is the standard normal cumulative distribution function. The key point of the Heckman's model is that the latent variable formulation of the selection equation $R_{yi}^*=X_i^s\beta^s+\varepsilon_i^s$ is linked to the outcome equation through the error terms ε_i^s, with $R_{yi}=1$ for individual i if $R_{yi}^*>0$ and $R_{yi}=0$ otherwise. In other words, Equations 1 and 2 are joined by a bivariate normal distribution of their error terms ε and ε^s as follow:

$$\begin{array}{l} R_{yi}^*=X_i^s\beta^s+\varepsilon_i^s \\ Y_i=X_i\beta+\varepsilon_i \end{array}, \quad \text{with} \quad \begin{pmatrix}\varepsilon^s\\ \varepsilon\end{pmatrix}\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{\varepsilon^s}^2 & \rho\sigma_{\varepsilon^s}\sigma_\varepsilon\\ \rho\sigma_{\varepsilon^s}\sigma_\varepsilon & \sigma_\varepsilon^2\end{pmatrix}\right)$$

where ρ is the correlation coefficient between ε and ε^s and $\sigma_{\varepsilon^s}=1$ as Equation 2 is a probit model [26]. Because of the bivariate normal distribution, $E[Ry^*|Y]=X^s\beta^s+\rho\frac{\sigma_{\varepsilon^s}}{\sigma_\varepsilon}(Y-X\beta)$. When ρ equals 0, selection and outcome equations are not correlated, $E[Ry^*|Y]$ does not depend on Y and the mechanism is MAR. When ρ is not equal to 0, $E[Ry^*|Y]$ still depends on Y, and the mechanism is MNAR. The more ρ increases, the more the MNAR mechanism is important. Heckman expressed the conditional mean and the conditional variance of Y, given X and X^s, as follows [26–28]:

$$E\left(Y_i\,|X_i,X_i^s,R_{yi}=1\right)=X_i\beta+\rho\sigma_\varepsilon\lambda_i \tag{3}$$

$$Var\left(Y_i\,|X_i,X_i^s,R_{yi}=1\right)=\sigma_\varepsilon^2\left(1-\rho^2\delta_i\right)\quad\text{with}\quad\delta_i=\lambda_i\left(\lambda_i+X_i^s\beta^s\right) \tag{4}$$

where

$$\lambda_i=\frac{\phi\left(X_i^s\beta^s\right)}{\Phi\left(X_i^s\beta^s\right)}$$

is the inverse Mills ratio (IMR), ϕ is the standard normal density and Φ is the standard normal cumulative distribution function. Then each individual in the study sample receives an individual value of the IMR corresponding to the error from the probit equation explaining the selection [27]. Indeed, the IMR is sometimes called a 'control function' – literally a function that controls for selection bias [27, 29].

To obtain unbiased estimates of β, the outcome equation parameters (Equation 1), Heckman proposed a two-step estimator directly inspired by Equation 3. It can be summarised as follows:

(1) Estimate parameters $\hat{\beta}^s$ of the selection Equation 2 by maximum likelihood (ML).
(2) Estimates of the previous probit model are then used to construct consistent estimates of the IMR term. Compute for each observation i in the observed sample $\widehat{\lambda}_i=\frac{\phi\left(X_i^s\hat{\beta}^s\right)}{\Phi\left(X_i^s\hat{\beta}^s\right)}$.
(3) Estimate $\widehat{\beta}$ and $\widehat{\beta}_\lambda$ from the following linear equation:

$$Y_i=X_i\beta+\widehat{\lambda}_i\beta_\lambda+\eta_i,\ \eta\sim N\left(0,\sigma_\eta^2\right) \tag{5}$$

with β_λ a scalar coefficient associated with the IMR.

We then obtain unbiased estimates of β. However, in this two-step procedure, the error term η_i is heteroscedastic, $Var\left(\eta_i|X_i,R_{yi}=1,\hat{\beta}_i^s\right)=\sigma_\varepsilon^2\left(1-\rho^2\delta_i\right)$. Hence, Heckman applied a corrector term to obtain valid estimates of variances based on a diagonal matrix of $\left(1-\rho^2\delta_i\right)$ and adjusted to take into account for the first step estimation [12, 26, 28].

If the two-step procedure was recommended for computational reasons, it is also known to generate a slightly higher standard error than the one-step ML estimator. Indeed, in the econometric literature, the two-step method is considered more rapid but less efficient than the one-step ML estimator [27, 28]. As

computational considerations are crucial in developing efficient MI, we choose the two-step estimator. Moreover, this estimator offers an easy way to develop an imputation model by incorporating the IMR as a predictor in Equation 5.

3.2. The imputation model using Heckman's model

Let us still suppose an MNAR missing mechanism on Y. To develop the imputation model, the main idea is to use the conditional expectation of Y as unobserved to impute Y_{miss}. Following Equation 3, we obtain the conditional expectation [26]

$$E\left(Y_i \,|X_i, X_i^s, R_{yi} = 0\right) = X_i\beta + \frac{-\phi\left(X_i^s\beta^s\right)}{1-\Phi\left(X_i^s\beta^s\right)}\rho\sigma_\epsilon \tag{6}$$

We then propose to use the corresponding linear model as the imputation model to draw missing Y_i for individual i:

$$Y_i = X_i\beta + \frac{-\phi\left(X_i^s\beta^s\right)}{1-\Phi\left(X_i^s\beta^s\right)}\beta_{\lambda i} + \eta, \;\; \eta \sim N\left(0, \sigma_\eta^2\right) \tag{7}$$

Note that, because of the symmetrical definition of Equations 6 and 3, β and β_λ should be equal in the two equations, so estimates obtained with Equation 5 could be used to impute Y_i. Motivated by the approximate proper imputation algorithm of a quantitative variable with a normal distribution, using a linear model described by Carpenter and Kenward [5], the main steps for one imputation become as follows:

(1) Obtain $\widehat{\beta^s}$ estimates from the selection model (Equation 2).
(2) Compute $\widehat{\lambda}_i$ for each observation.
(3) Obtain $\hat{\beta}$, $\hat{\beta}_\lambda$ and $\hat{\sigma}_\eta$ using the Heckman two-step estimator (Equation 5).
(4) Draw $\left(\sigma_\eta^{2*}, \beta^*, \beta_\lambda^*\right)$ using approximate proper imputation for the linear model adding the Heckman variance correction as detailed in Appendix A [5, 25].
(5) Draw η^* from $N\left(0, \sigma_\eta^{2*}\right)$.
(6) For each missing Y, impute Y^* using the imputation model

$$Y_i^* = X_i\beta^* + \frac{-\phi(\widehat{X_i^s\beta^s})}{1-\Phi(\widehat{X_i^s\beta^s})}\beta_{\lambda i}^* + \eta^*$$

We then obtain an imputation model for MNAR data when the Heckman's model is valid. Moreover, one should insist, at that point, on the fact that the final analysis model remains the model originally planned, that is, model of Equation 1. Indeed, the process for generating missing Y, recognising the bias introduced by the selection, precludes the use of a modified analysis model.

From a computational point of view, we modified the function proposed in the *mice* R-package by van Buuren [30] to impute quantitative variables using Heckman's model with the variance correction (the computational R code is available in the Supporting Information).

3.3. Multiple imputation by chained equations using Heckman's imputation model

The first aim of this work was to develop a MICE procedure to impute missing outcomes with an MNAR mechanism and missing predictors with MAR mechanisms. Let us suppose a simple model with three continuous variables, Y, X_1 and X_2, where Y is the outcome of interest with MNAR missing data, X_1 is a first predictor without missing data and X_2 is a second predictor with MAR missing data depending on X_1. Let us define R_y and R_2 as the missing indicators of Y and X_2, respectively. Figure 1 represents the causal diagram of the relationships between these variables. The full arrows symbolise causal relationships, whereas indirect associations are represented by dashed lines.

The diagram of the Figure 1 shows that we must condition by R_y to obtain only direct paths between X_2 and (Y, X_1) and then unbiased dependence between X_2 and (Y, X_1) [31]. In the MICE procedure, we should define an imputation model for each variable with a missing value, that is, Y and X_2. The model

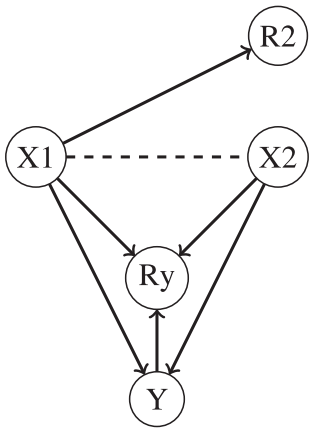

Figure 1. Causal diagram of the relationships between the studied variables.

for Y is the Heckman imputation model proposed previously with X_1 and X_2 in the outcome equation and a potential other covariate in the selection model for the exclusion-restriction criteria. For X_2, the conditional mean of X_2 depends also on R_y. We then propose to include R_y in the imputation model of X_2, as previously described by van Buuren in a non-ignorable missing data situation [25]. Therefore, we will use a linear imputation model of the form $X_2 \sim X_1 + Y + R_y$. The diagram of the Figure 1 does not prove that R_y is a linear and additive predictor for X_2. Nevertheless, first-order approximations show that at least a linear and additive term for R_y should be present in the equation. The complete MICE procedure will consist of sequentially imputing Y and X_2 using the two imputation models. As recommended, the number of iterations should be at least 10 [6].

4. Simulation study

The objective of the simulation study was to evaluate the performance of our approach under various scenarios. We generated N independent datasets in several settings defined by the type of missingness mechanism for the outcome and the presence or absence of MAR missing data for the predictors. We then analysed the data with our proposed approaches and with several standard approaches, namely, complete case analysis, MI and Heckman's original model.

4.1. Data-generating process

Let us generate three independent and identically normally distributed covariates, X_1, X_2 and X_3, where $X_j \sim N(0, 0.3^2)$, and a continuous outcome, Y. The first two covariates, X_1 and X_2, are associated with Y through $(\beta_0, \beta_1, \beta_2)$, a vector of coefficients in the following linear model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The missing indicator R_y of Y was generated according to the following algorithm: if $\beta_0^s + \beta_1^s X_1 + \beta_2^s X_2 + \beta_3^s X_3 + \varepsilon^s < 0$, then $R_y = 0$; otherwise, $R_y = 1$. Thus, R_y depended on X_1, X_2 and X_3 through the vector $\left(\beta_0^s, \beta_1^s, \beta_2^s, \beta_3^s\right)$. We included X_3 as a third independent covariate in the generating model of missingness to fulfil the exclusion-restriction rule of Heckman's model.

According to the Heckman's model, the error terms ε and ε^s were generated using a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon^s \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where ρ is the correlation between the two error terms. If ρ equals 0, the missing data mechanism is a MAR mechanism. If ρ is not zero, then the data are MNAR given that the probability of missing data is linked to the three explanatory variables (X_1, X_2 and X_3) but also to Y (via ε and ε^s). The more ρ increases, the more the MNAR mechanism is important. In the outcome equation, $(\beta_0, \beta_1, \beta_2)$ was fixed at $(0, 1, 1)$. In the missingness equation, $(\beta_0^s, \beta_1^s, \beta_2^s, \beta_3^s)$ was fixed at $(0.54, 1, -0.5, 1)$ to assure approximately 30% of missing data. Different correlations ρ between error terms were considered to ensure different missingness mechanism: 0 (MAR), 0.3 (MNAR) and 0.6 (MNAR).

To evaluate the performance of methods with an MNAR mechanism that differed from a Heckman's model, we simulated an MNAR mechanism according to Ganjali's proposal [32]. We named this simulation 'non-Heckman'. We generated R_y using a Bernoulli distribution with parameter $P(R_y = 1) = \Phi(0.75 + Y)$ for each observation.

Finally, we generated missing values on X_2 using a simple MAR mechanism (depending on X_1) to evaluate the performance of the MICE approach: for $X_1 > 0.157$, X_2 was considered missing using a Bernoulli distribution with parameter 54%, and for $X_1 \leq 0.157$, X_2 was considered missing using a Bernoulli distribution with parameter 20%. We finally obtained approximately 30% missing data on X_2.

A total of $N = 1000$ independent datasets of size $n = 2000$ were generated for each setting. Our simulation datasets were obtained with R statistical software, version 2.15.2 and the package *mvtnorm* [33, 34].

4.2. Analysis methods

The analysis outcome model was a linear model on Y including X_1 and X_2 as predictors:

$$Y \sim X_1 + X_2 \tag{8}$$

The data were first analysed before any data deletion as a benchmark for the MI procedure. The incomplete data were then analysed using the following methods in the presence of only missing Y:

- The complete case analysis (*CCA*) method consisted of estimating the parameters of the outcome model based only on complete observations.
- The Heckman's model consisted of using a two-step estimator (*HE*2) with a probit regression selection model adjusted on X_1, X_2 and X_3 to obtain the estimated IMR. The parameters of the outcome equation were estimated using a linear regression including the IMR in the predictors, and the variance correction was applied.
- Standard MI (MI_{norm}) for Y consisted of applying a linear regression imputation model and an approximate proper imputation algorithm with X_1, X_2 and X_3 in the imputation model. The incomplete data were imputed $m = 10$ times, and the final estimates were obtained using Rubin's rules applied to model 8.
- The MI approach proposed for Y used a Heckman's model as described in the preceding text (*MIHE*). X_1, X_2 and X_3 were included in the imputation model. The incomplete data were imputed $m = 10$ times, and the final estimates were obtained using Rubin's rules applied to model 8.

For the dataset with missing data for X_2 and Y, *HE*2 was considered. The *HE*2 procedure was applied on a complete case basis for X_2. Moreover, the two MICE procedures MI_{norm} and *MIHE* were applied. To impute the missing X_2, we used a linear regression imputation model and an approximate proper imputation algorithm. To verify the merit of using R_y in the imputation model for X_2, we applied our method with and without R_y in the imputation model. Ten iterations of the chained equation process were considered, and the number of imputations was $m = 10$.

In each data-generating setting, the performance of each method was assessed by computing the empirical mean of the parameter estimates, the relative bias (Rbias), the root mean square of estimated standard errors (SE_{cal}), the empirical Monte Carlo standard errors (SE_{emp}), the root mean square error (RMSE), the coverage of nominal 95% confidence intervals (Cover) of β_1 and the percentage of the 2000 observations used by each method (% used).

The analysis was performed with the R statistical software, version 2.15.2, including the following packages: *mice* and *sampleSelection* [28, 30, 33].

4.3. Results

The results of the first simulation study, which considered a possible MNAR missing mechanism only for the outcome, are shown in Table I. With a MAR mechanism ($\rho = 0$), all methods appeared to give unbiased estimates of β_1. Only approximately 70% of the observations were considered in the *CCA*. The standard errors observed using *HE*2 and *MIHE* were greater than the *CCA* and MI_{norm} standard errors. For the Heckman's model and *MIHE*, the coverage was slightly less than 95%.

With an MNAR mechanism with $\rho = 0.3$ and $\rho = 0.6$, *CCA* and MI_{norm} were biased, the bias increasing with ρ from approximately 13% to approximately 28%. *HE*2 and *MIHE* appeared to be unbiased. The

Table I. Simulation results for $\beta_1 = 1$ estimates with $\rho = 0$ representing a MAR mechanism, $\rho = 0.3$ and 0.6 representing an MNAR mechanism.

Methods	ρ	mean	Rbias (%)	SE_{cal}	SE_{emp}	RMSE	Cover (%)	% used
Before deletion	0	0.9981	−0.2	0.0746	0.0757	0.0757	94.7	100
	0.3	1.0028	0.3	0.0746	0.0724	0.0725	95.7	100
	0.6	0.9996	0.0	0.0747	0.0756	0.0756	95.4	100
	Non-Heckman	1.0014	0.1	0.0747	0.0757	0.0757	94.4	100
CCA	0	0.9974	−0.3	0.0917	0.0932	0.0932	95.2	69
	0.3	0.8663	−13.4	0.0898	0.0883	0.1602	68.4	69
	0.6	0.7281	−27.2	0.0840	0.0871	0.2855	10.3	69
	Non-Heckman	0.7546	−24.5	0.0788	0.0779	0.2575	11.6	70
*HE*2	0	0.9987	−0.1	0.1271	0.1324	0.1324	94.3	100
	0.3	1.0017	0.2	0.1258	0.1231	0.1231	95.7	100
	0.6	1.0047	0.5	0.1230	0.1256	0.1257	95.1	100
	Non-Heckman	0.9412	−5.9	0.4692	0.4301	0.4341	98.1	100
MI_{norm}	0	0.9969	−0.3	0.0934	0.0936	0.0937	94.3	100
	0.3	0.8611	−13.9	0.0916	0.0899	0.1655	67.3	100
	0.6	0.7160	−28.4	0.0856	0.0881	0.2974	8.8	100
	Non-Heckman	0.7552	−24.5	0.0804	0.0786	0.2571	13.3	100
MIHE	0	0.9987	−0.1	0.1314	0.1374	0.1374	93.0	100
	0.3	1.0031	0.3	0.1304	0.1274	0.1274	94.5	100
	0.6	1.0035	0.4	0.1275	0.1290	0.1290	93.5	100
	Non-Heckman	0.9366	−6.3	0.4357	0.4438	0.4483	90.8	100

Non-Heckman, corresponds to a MNAR mechanism for which the probability of Y missingness is equal to $\Phi(0.75 + Y)$. Means, relative bias (Rbias), root mean square of estimated standard errors (Se Calc.), observed standard errors (Se Emp.), root mean square error (RMSE), coverage (Cover) and percentage of the 2000 observations used by each method (% used).
MAR, missing at random; MNAR, missing not at random.

standard errors observed using *HE*2 and *MIHE* seemed similar, and the coverage of *MIHE* was still slightly less than 95%.

With a 'non-Heckman' MNAR mechanism, *CCA* and MI_{norm} were biased approximately 25%. *HE*2 and *MIHE* were the least biased methods showing biases about 6%. The coverages were high, 98%, for *HE*2 and decreased to 91% for *MIHE*. The standard errors increased dramatically with both approaches. Thereby, while *MIHE* is the least biased, its RMSE (0.4483) is increased as compared with the MI_{norm} RMSE (0.2571). However, because of the bias, confidence intervals by MI_{norm} exhibited substantial undercoverage. The results for the β_2 estimates were similar and are given in the Supporting Information.

Table II presents the results for the estimation of β_1 in the presence of missing outcome data and in the presence of MAR missing data on X_2 depending on X_1. With a MAR mechanism for the outcome ($\rho = 0$), all methods were unbiased. Biases obtained with MI_{norm} with R_y were smaller than those observed with MI_{norm} without R_y. With an MNAR mechanism for the outcome ($\rho = 0.3$ or $\rho = 0.6$), MI_{norm} were biased, while the Heckman's model and *MIHE* with R_y appeared not biased. *MIHE* without R_y appeared to be slightly biased (approximately 2%). *MIHE* with *Ry* provided lower standard errors than *HE*2 linked to the fact that *HE*2 used only approximately 70% of the observations. As previously observed, the coverage of *MIHE* was only slightly lower than 95%.

With a 'non-Heckman' MNAR mechanism, biases were similar to those obtained without missing data on X_2 except for *MIHE* without R_y, which showed biases of approximately 43%.

Table III presents the results for the estimation of β_2 in the presence of missing outcome data and in the presence of MAR missing data on X_2 depending on X_1. The results were similar to those observed with β_1 except that the improvement in terms of standard error was smaller for *MIHE* with R_y approach.

Table II. Simulation results for $\beta_1 = 1$ estimates with $\rho = 0$ representing a MAR mechanism, $\rho = 0.3$ and 0.6 representing an MNAR mechanism in the presence of MAR missing data on X_2.

Methods	ρ	mean	Rbias (%)	SE_{cal}	SE_{emp}	RMSE	Cover (%)	% used
Before deletion	0	0.9989	−0.1	0.0746	0.0745	0.0745	94.8	100
	0.3	1.0025	0.3	0.0746	0.0771	0.0771	93.3	100
	0.6	0.9974	−0.3	0.0747	0.0760	0.0760	93.5	100
	Non-Heckman	1.0014	0.1	0.0746	0.0728	0.0728	95.8	100
*HE*2	0	0.9981	−0.2	0.1698	0.1696	0.1696	95.6	69
	0.3	0.9969	−0.3	0.1690	0.1641	0.1641	96.2	69
	0.6	0.9974	−0.3	0.1641	0.1693	0.1693	94.9	69
	Non-Heckman	0.9332	−6.7	0.6732	0.5722	0.5761	99.0	71
MI_{norm}	0	0.9874	−1.3	0.0967	0.0965	0.0973	93.8	100
without *Ry*	0.3	0.8485	−15.2	0.0955	0.1013	0.1822	63.3	100
	0.6	0.7051	−29.5	0.0904	0.0898	0.3083	10.5	100
	Non-Heckman	0.7671	−23.3	0.0822	0.0843	0.2477	20.0	100
MI_{norm}	0	0.9930	−0.7	0.0970	0.0955	0.0958	94.9	100
with *Ry*	0.3	0.8555	−14.5	0.0956	0.1003	0.1759	65.4	100
	0.6	0.7129	−28.7	0.0901	0.0895	0.3007	12.4	100
	Non-Heckman	0.7578	−24.2	0.0822	0.0833	0.2561	17.2	100
MIHE	0	0.9801	−2.0	0.1350	0.1308	0.1323	94.9	100
without *Ry*	0.3	0.9851	−1.5	0.1357	0.1325	0.1333	94.8	100
	0.6	0.9704	−3.0	0.1325	0.1293	0.1326	94.2	100
	Non-Heckman	0.5665	−43.4	0.5999	0.5852	0.7283	84.0	100
MIHE	0	0.9923	−0.8	0.1354	0.1324	0.1326	94.4	100
with *Ry*	0.3	0.9991	−0.1	0.1340	0.1350	0.1350	94.5	100
	0.6	0.9910	−0.9	0.1327	0.1318	0.1321	93.9	100
	Non-Heckman	0.9438	−5.6	0.5090	0.3905	0.3945	96.6	100

Non-Heckman, corresponds to an MNAR mechanism for which the probability of Y missingness is equal to $\Phi(0.75 + Y)$. Means, relative bias (Rbias), root mean square of estimated standard errors (Se Calc.), observed standard errors (Se Emp.), root mean square error (RMSE), coverage (Cover) and percentage of the 2000 observations used by each method (% used).
MAR, missing at random; MNAR, missing not at random.

5. Application to BIVIR data

The equation developed for the linear analysis of the severity score included *sick leave*, *temperature* and *tobacco status* as predictors. Among the 127 patients with missing outcomes, 48 were also missing *temperature* values. A total of 261 observations had at least one missing value for the *severity score* and/or for the *temperature*.

Results are displayed in Table IV. The *CCA* estimates included only 280 cases, that is, 52% of the entire dataset. Observations with missing *temperature* were ignored in the *HE*2 analysis; that is, only 66% of the observations were retained in the analysis. The Heckman selection model used for the analysis included *sick leave*, *temperature*, *tobacco status* and *gender* as predictors. *Gender* was chosen to ensure that the exclusion-restriction criteria were satisfied. Under that model, a MAR mechanism was ruled out for *severity score* by testing $\hat{\rho} = 0$ (p-value ≤ 0.001) [26].

The *MIHE* method considered all observations. The Heckman imputation model for *severity score* was the model used for the *HE*2 approach. In terms of the missing mechanism for *temperature*, we introduced a second Heckman's model with all other variables as predictors of the selection equation to test a potential MAR mechanism. The MAR mechanism was not ruled out (p-value $= 0.29$). We then considered a MAR mechanism to be plausible. *Temperature* was imputed using a linear model and an approximate proper imputation model with the missing indicator of *severity score* in the predictors. The MICE procedure was applied with $m = 10$ imputed datasets and 10 iterations.

Finally, a standard MICE approach supposing only the MAR mechanism was applied (MI_{norm}). A linear model and an approximate proper imputation model were used for *severity score* as well

Table III. Simulation results for $\beta_2 = 1$ estimates with $\rho = 0$ representing a MAR mechanism, $\rho = 0.3$ and 0.6 representing an MNAR mechanism in the presence of MAR missing data on X_2.

Methods	ρ	mean	Rbias (%)	SE_{cal}	SE_{emp}	RMSE	Cover (%)	% used
Before deletion	0	0.9997	0.0	0.0746	0.0761	0.0761	94.7	100
	0.3	0.9972	−0.3	0.0746	0.0763	0.0764	94.8	100
	0.6	1.0039	0.4	0.0747	0.0730	0.0731	95.7	100
	Non-Heckman	0.9954	−0.5	0.0746	0.0749	0.0750	94.5	100
*HE*2	0	0.9936	−0.6	0.1239	0.1235	0.1237	95.6	69
	0.3	0.9952	−0.5	0.1225	0.1246	0.1247	95.3	69
	0.6	1.0033	0.3	0.1194	0.1148	0.1148	95.5	69
	Non-Heckman	0.9187	−8.1	0.6530	0.5452	0.5512	98.6	71
MI_{norm} without *Ry*	0	0.9921	−0.8	0.1105	0.1125	0.1128	94.0	100
	0.3	1.0689	6.9	0.1074	0.1095	0.1294	88.1	100
	0.6	1.1567	15.7	0.0999	0.0997	0.1857	62.0	100
	Non-Heckman	0.7475	−25.3	0.0962	0.0967	0.2704	25.4	100
MI_{norm} with *Ry*	0	0.9931	−0.7	0.1100	0.1121	0.1123	93.4	100
	0.3	1.0690	6.9	0.1077	0.1096	0.1295	88.6	100
	0.6	1.1584	15.8	0.0995	0.0981	0.1863	61.7	100
	Non-Heckman	0.7490	−25.1	0.0965	0.0954	0.2685	26.3	100
MIHE without *Ry*	0	0.9900	−1.0	0.1229	0.1214	0.1218	94.8	100
	0.3	0.9784	−2.2	0.1237	0.1270	0.1288	93.5	100
	0.6	0.9751	−2.5	0.1237	0.1205	0.1230	94.0	100
	Non-Heckman	0.4886	−51.1	0.6402	0.6057	0.7927	83.6	100
MIHE with *Ry*	0	0.9932	−0.7	0.1210	0.1178	0.1180	94.2	100
	0.3	1.0013	0.1	0.1193	0.1172	0.1172	95.0	100
	0.6	1.0328	3.3	0.1144	0.1084	0.1133	94.4	100
	Non-Heckman	0.9477	−5.2	0.5108	0.3747	0.3783	96.3	100

Non-Heckman, corresponds to an MNAR mechanism for which the probability of Y missingness is equal to $\Phi(0.75 + Y)$. Means, relative bias (Rbias), root mean square of estimated standard errors (Se Calc.), observed standard errors (Se Emp.), root mean square error (RMSE), coverage (Cover) and percentage of the 2000 observations used by each method (% used).
MAR, missing at random; MNAR, missing not at random.

Table IV. Estimation of the predictive value of temperature, sick leave and tobacco for Y.

	CCA $n = 280$		*HE2* $n = 358$		*MIHE* $n = 541$		MI_{norm} $n = 541$	
Variable	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Sick leave	4.96	1.59	6.35	2.26	4.90	1.59	4.25	1.46
Temperature	1.59	1.27	0.48	1.80	0.90	1.47	1.36	1.16
Tobacco	−0.49	2.39	−2.54	3.33	−2.55	2.27	−1.53	1.96

as for *temperature*. $m = 10$ imputed datasets were obtained after 10 iterations of the chained equation procedure.

The estimated coefficients differed according to the estimation method. Depending on the values of the variables, the *MIHE* coefficients (4.90, 0.90 and −2.55) were closer to the *HE*2 coefficients (6.35, 0.48 and −2.54), whereas the MI_{norm} coefficients (4.25, 1.36 and −1.53) were closer to the *CCA* coefficients (4.96, 1.59 and −0.49) except for *tobacco* (−0.49 and −1.53 respectively). The standard errors obtained with *MIHE* (1.59, 1.47 and 2.27) were lower than the *HE*2 standard errors (2.26, 1.80 and 3.33). As expected, the *MIHE* and MI_{norm} standard errors were lower than the *HE*2 and *CCA* standard errors, respectively, because of a larger sample size.

6. Discussion

To our knowledge, no satisfactory approach has been proposed to handle datasets with MNAR mechanism in outcomes in combination with MAR in predictors. As it is not possible to check the MAR from the observed data alone [6, 35], several modelling strategies based on imputation models or pattern-mixture models with sensitivity analyses are used to evaluate the robustness of the MAR mechanism against MNAR mechanism assumptions [18, 25]. Notably, the delta-adjustment method acts by systematically varying the value of an increment (δ) to the imputed values [25, 36]. Nevertheless, such an approach does not correct a potential bias, and the assumption of ignorability, that is, the possibility that MAR holds, essentially states 'a belief that data are sufficient to correct the missing data' [25]. Conversely, a belief that an MNAR mechanism holds, that is, that data are not sufficient to correct the missing data unless some additional hypotheses, should conduct to the use of an adapted method of analysis. Otherwise, selection models are alternate approaches [3], such as the Heckman's model [12, 13] that introduces a correlation term between the selection equation and the outcome equation, but those approaches cannot handle MAR in predictors.

In this paper, we introduce an MNAR imputation model derived from Heckman's model in a chained equation process to impute missing values of both covariates and outcomes via an iterative process.

In our simulations, under MNAR mechanisms compatible with the Heckman's model, only *MIHE* and *HE*2 were unbiased. Thus, if the model is valid, our approach gives unbiased estimates, and the more the MNAR phenomenon is important, the more the differences with conventional approaches as *CCA* and MI_{norm} are important. Finally, compared with standard *MI*, biases were almost eliminated, variances only slightly increase and coverages were close to 95%.

To thoroughly evaluate our approach in a MICE procedure, we simulated MAR missing data for other covariates. Our proposed approach gives the results with the smallest biases and with smaller standard errors when MICE included *Ry* in other imputation models (for more detail, see the Supporting Information). Results were similar with an MCAR missing data for other covariates and are given in the Supporting Information. Although not simulated, the extension of our procedure to covariates MNAR implies first, to define a selection model for each covariate with a suspected MNAR mechanism and second, to include all missing data indicators in the iterative application of the MICE procedure.

There remains the potential problem of the compatibility between the full conditional distribution and the global joint distribution of the multivariate missing data. Theoretically, the conditional distribution should be derived from the joint probability distribution [25, 37]. It is a crucial point because misspecification of the models used to generate imputed values may result in biased parameter estimates and reduced efficiency [22]. If recent studies have identified compatibility in simple cases [38], the efficiency of the MICE approach is usually validated only by simulation studies, and the results appear robust when the compatibility assumption is not met [25]. Moreover, certain authors have argued that having a valid joint distribution is 'less important than incorporating information from other variables and unique features of the dataset' [39]. Then, even if a simulation study is never sufficiently complete, we believe that our approach of MI using a Heckman's model and its use in a MICE process including a missing data indicator in the other imputation models is valid and could be useful if the MNAR mechanism is compatible with the Heckman's model. However, under a 'non-Heckman' scenario, that is, a scenario not fully compatible with a Heckman's model, our simulation showed that the RMSE of estimates is increased; nevertheless, the *MIHE* estimates are almost unbiased and still provides almost perfect coverage. This is not the case for MI_{norm}, where because of the bias, coverage decreases severely.

Performances of the Heckman's model rely on the strict respect of the exclusion-restriction criteria [16, 28]. The Heckman's model has been particularly used in the case of missing self-reported outcome [40] or refusal bias in HIV epidemiological studies [41–44]. Indeed, in these studies, an MNAR mechanism is considered as patients are highly suspected to modify their probability to indicate their HIV experience or to realise an HIV test according to their HIV status. Hence, the use of a Heckman's model allows to correct the selection bias potentially introduced by the MNAR mechanism [42]. Moreover, it has been used by Ganjali as a benchmark for comparing the performances of the MI and direct likelihood approaches in the presence of MNAR mechanisms [32].

In accordance with our motivating example, we focused on a continuous outcome. However, a Heckman's model can be used with binary outcomes, based on a bivariate probit estimator [15, 26, 45]. This method has only been presented using a one-step ML estimator of a bivariate probit model. The development of an imputation model using parameters of the Heckman ML estimator would represent a possibility for extending the imputation model, adapted for MNAR missing data, to a binary variable.

Finally, we proposed a simple way to handle missing data with an MNAR mechanism under a Heckman assumption in the MICE procedure. This framework allows us to handle a very large category of missing data whatever the quality of the variables (quantitative or qualitative) and their place in the analysis model (outcome or covariate). Our code is available in the Supporting Information and could easily be run using the MICE package [30].

Appendix A

Detailed *MIHE* imputation algorithm:

(1) Obtain $\widehat{\beta^s}$ estimates from the selection model (Equation 2).
(2) Compute $\widehat{\lambda_i}$ for each observation.
(3) Obtain $\hat{\beta}$, $\hat{\beta}_\lambda$ and $\hat{\sigma}_\eta$ using the Heckman two-step estimator (Equation 5).
(4) Draw $\left(\sigma_\eta^{2*}, \beta^*, \beta_\lambda^*\right)$ using approximate proper imputation for the linear model adding the Heckman variance correction [5]:
   - Draw a random variable g^* following $g \sim \chi_\nu^2$ with ν the number of degrees of freedom of Equation 7.
   - Calculate $\sigma_\eta^{*2} = \frac{\hat{\sigma}_\eta^2}{g^*}$.

     Nevertheless, the real distribution of η is not homoscedastic because [26]

$$Var\left(\eta_i | X_i, R_{yi} = 1, \hat{\beta}_i^s\right) = \sigma_\varepsilon^2\left(1 - \rho^2\delta_i\right)$$

     Then we calculated

$$\sigma_\varepsilon^{*2} = mean\left(\sigma_\eta^{*2} / \left(1 - \rho^2\delta_i\right)\right).$$

   - Draw q independent $N(0, 1)$ variables in vector z^*, where q is the dimension of V_c.
   - Calculate $V_c^{\frac{1}{2}}$ by Cholesky decomposition.
   - Calculate $\left(\beta^*, \beta_\lambda^*\right) = (\hat{\beta}, \hat{\beta}_\lambda) + z^*\sigma_\varepsilon^{*2}V_c^{\frac{1}{2}}$.
(5) Draw η^* from $N\left(0, \sigma_\eta^{2*}\right)$.
(6) For each missing Y, impute Y^* using the imputation model

$$Y_i^* = X_i\beta^* + \frac{-\phi(\widehat{X_i^s\beta^s})}{1 - \Phi(\widehat{X_i^s\beta^s})}\beta_{\lambda i}^* + \eta^*$$

Appendix B

Bivir Study Group

Scientific Committee: *Steering Committee:* Leport C (principal investigator), Andreoletti L, Blanchon T, Carrat F, Duval X, Guimfack A, Lina L, Loubière S, Mentré F, Mosnier A, Tibi A, Tubach F, van der Werf S. *Clinical study manager:* Charlois - Ou C. *Other members:* Bricaire F, Cohen JM, Flahault A, Moatti JP, Vogel JY. *Invited members:* Eid Z (GSK), Peurichard C (GSK), Pecking M (Roche), Dantin S (Roche), Gysembergh-Houal A (AP-HP).

Independent data-monitoring committee: Chêne G, Hannoun C, Vittecoq D.

Monitoring and statistical analysis: Boucherit S, Dornic Q, Quintin C, Vincent C., and Atlanstat, Studypharm clinical research organisations.

Clinical investigators: Alea JR, Aleonard JL, Arditti L, Baranes C, Beaujard J, Beaurain C, Behar M, Beignot-Devalmont P, Biquet D, Blanchard M, Blot E, Bodin X, Bouaniche H, Boulet L, Bourgeois O, Bretillon F, Breton N, Broyer F, Buffler P, Camper E, Carissimo P, Carrera J, Causse P, Cayet JP, Cayron P, Cazard C, Chaix C, Chazerans D, Cheftel JA, Codron G, Cooren G, Coutrey L, Crappier JJ, Daugenet C, Dauzat C, Defreyn F, Delamare G, Delsart D, Demure P, Desmarchelier P, Domenech A, Dubois D Laroy G, Dubrana E, Dumond P, Dumont A, Durel G, Ellé P, Evellin F, Eyraud P, Fhal G, Fournillou JC, Galesne Herceg G, Gastan G, Geoffray B, Giagnorio P, Goguel J, Granger JF, Guenee P, Haushalter B,

Huber C, Hureau JP, Jacob L, Jami A, Jordan E, Jourde P, Journet L, Julien D, Jusserand JT, Korsec P, Laforest G, Lalanne G, Le Duff N, Le Guen-Naas A, Le Hir A, Lebois S, Leclerc S, Leclere V, Lejay D, Lejoly JM, Lemoine C, Lepine C, Lepoutre B, Leprince P, Lhoumeau P, Lognos B, Lustig G, Mannessier B, Marlier M, Marmor P, Martocq G, Massot J, Meme B, Mercier P, Mercier G, Mesnier PL, Meyrand G, Mongin G, Montavont J, Morlon P, Pantea D, Parisot J, Partouche H, Pertusa MC, Petot A, Peyrol Y, Piketty B, Poignant G, Pradere H, Rabaud D, Rachine L, Ragon B, Rambaud J, Richard P, Rigaill P, Robinson D, Rosenberg S, Ruetsch M, Sacareau D, Saint Lannes M, Saugues M, Sauvage P, Schaupp T, Schmitt C, Sellam A, Severin JF, Simian B, Specht L, Szmuckler I, Tetaud D, Trehou P, Triantaphylides JC, Triot P, Uge P, Urbain F, Urbina JC, Vailler P, Vallez V, Varnier H, Venot N, Verhun R, Vogel JY, Zanuttini-Vogt C, Zeline V.

Acknowledgements

We thank the BIVIR Study Group for the permission to use their data. We wish to thank the two reviewers and the editor for all their constructive and helpful comments that have clearly improved our manuscript.

References

1. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*, Vol. 539. Wiley: New York, 1987.
2. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002; **7**(2):147–177.
3. Molenberghs G, Kenward M. *Missing Data in Clinical Studies*, Vol. 61. John Wiley & Sons: Wiley: Chichester, 2007.
4. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007; **16**(3):259–275.
5. Carpenter JR, Kenward MG. *Missing Data in Randomised Controlled Trials a Practical Guide*. Birmingham: National Institute for Health Research: Birmingham, 2008.
6. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 2011; **30**(4):377–399.
7. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**(421):125–134.
8. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**(431):1112–1121.
9. Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* 1977; **72**(359):538–543.
10. Glynn RJ, Laird NM, Rubin DB. Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-selected Samples*. Springer, 1986; 115–142.
11. Brown CH. Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 1990; **46**(1):143–155.
12. Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, Number 4*. NBER, 1976; 475 –492.
13. Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979; **47**(1):153–161.
14. Dubin JA, Rivers D. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research* 1989; **18**(2-3):360–390.
15. Marra G, Radice R. A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics* 2013; **7**:1432–1455.
16. Puhani P. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* 2000; **14**(1):53 –68.
17. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**(429):122–129.
18. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 2009.
19. Schafer JL. Multiple imputation: a primer. *Statistical methods in medical research* 1999; **8**(1):3–15.
20. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**(3):445–464.
21. Schafer JL. *Analysis of Incomplete Multivariate Data*. CRC Press: CRC Press: Boca Raton, 1997.
22. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**(4):538–558.
23. Duval X, van der Werf S, Blanchon T, Mosnier A, Bouscambert-Duchamp M, Tibi A, Enouf V, Charlois-Ou C, Vincent C, Andreoletti L, Tubach F, Lina B, Mentré F, Leport C. Efficacy of oseltamivir-zanamivir combination compared to each monotherapy for seasonal influenza: a randomized placebo-controlled trial. *PLoS Med* 2010-11; **7**(11):1–12. e1000362.
24. Treanor JJ, Hayden FG, Vrooman PS, Barbarash R, Bettis R, Riff D, Singh S, Kinnersley N, Ward P, Mills RG. Efficacy and safety of the oral neuraminidase inhibitor oseltamivir in treating acute influenza: a randomized controlled trial. *Jama* 2000; **283**(8):1016–1024.
25. van Buuren S. *Flexible Imputation of Missing Data*. CRC Press: Boca Raton, 2012.
26. Greene WH. *Econometric Analysis: International Edition*, 7th Ed. Pearson: Edinburgh, 2011.
27. Vella F. Estimating models with sample selection bias: a survey. *The Journal of Human Resources* 1998; **33**(1):127–169.
28. Toomet O, Henningsen A. Sample selection models in R: Package sampleSelection. *Journal of Statistical Software* 2008; **27**(7):1–23.

29. Sales AE, Plomondon ME, Magid DJ, Spertus JA, Rumsfeld JS. Assessing response bias from missing quality of life data: the Heckman method. *Health and quality of life outcomes* 2004; **2**(1):1–10.
30. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *Journal of Statistical Software* 2011; **45**(3):1–67.
31. Greenland S, Pearl J, Robins J.M. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**(1):37–48.
32. Ganjali M, Ranji H. A comparison of several algorithms and models for analyzing multivariate normal data with missing responses. *Applications and Applied Mathematics* 2008-06; **3**:55–68.
33. R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2013.
34. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. *mvtnorm: Multivariate Normal and T Distributions*, 2014. R package version 0.9-9997.
35. Kaambwa B, Bryan S, Billingham L. Do the methods used to analyse missing data really matter? An examination of data from an observational study of intermediate care patients. *BMC Research Notes* 2012; **5**(1):1–12. 330.
36. Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics* 2013-11; **12**(6):337–347.
37. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*. Springer, 1996; 75–88.
38. Chen HY. Compatibility of conditionally specified models. *Statistics & Probability Letters* 2010; **80**(7):670–677.
39. Gelman A. Parameterization and Bayesian modeling. *Journal of the American Statistical Association* 2004; **99**(466): 537–545.
40. Protopopescu C, Raffi F, Roux P, Reynes J, Dellamonica P, Spire B, Leport C, Carrieri MP. Factors associated with non-adherence to long-term highly active antiretroviral therapy: a 10 year follow-up analysis with correction for the bias induced by missing data. *Journal of Antimicrobial Chemotherapy* 2009-01; **64**(3):599–606.
41. Brnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology (Cambridge, Mass.)* 2011-01; **22**(1):27–35.
42. McGovern M.E, Marra G, Radice R, Canning D, Newell ML, Bärnighausen T. Adjusting HIV prevalence estimates for non-participation: an application to demographic surveillance. *Journal of the International AIDS Society* 2015; **18**(1):1–11.
43. Reniers G, Araya T, Berhane Y, Davey G, Sanders EJ. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health* 2009; **9**:163.
44. Clark SJ, Houle B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. *PloS One* 2014; **9**(11):e112563.
45. Marra G, Radice R. *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling*, 2013. R package version 3.2-8.