**RESEARCH ARTICLE**

# Multiple imputation of incomplete multilevel data using Heckman selection models

Johanna Munoz*[1] | Matthias Egger[2] | Orestis Efthimiou[2] | Vincent Audigier[3] | Valentijn M. T. de Jong[1,4,6] | Thomas. P. A. Debray[1,5,6]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[2]University of Bern, Switzerland

[3]CNAM, Laboratoire CEDRIC-MSDMA, 2 rue Conte, 75003 Paris, France

[4]Data Analytics and Methods Task Force, European Medicines Agency, Amsterdam, The Netherlands

[5]Smart Data Analysis and Statistics, Utrecht, The Netherlands

[6]These authors contributed equally

**Correspondence**
* Johanna Munoz. Email:
j.munozavila@umcutrecht.nl

**Present Address**
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands

Missing data is a common problem in medical research, and is commonly addressed using multiple imputation. Although traditional imputation methods allow for valid statistical inference when data are missing at random (MAR), their implementation is problematic when the presence of missingness depends on unobserved information, i.e. the data are missing not at random (MNAR). Unfortunately, this MNAR situation is rather common, in observational studies, registries and other sources of real-world data. While several imputation methods have been proposed for addressing individual studies when data are MNAR, their application and validity in large datasets with clustering remains unclear. We therefore explored the consequence of MNAR data in hierarchical data in-depth, and proposed a novel multilevel imputation method for common missing patterns in clustered datasets. This method is based on the principles of Heckman selection models and adopts a two-stage meta-analysis approach to impute binary and continuous variables that may be outcomes or predictors and that are systematically or sporadically missing. After evaluating the proposed imputation model in simulated scenarios, we illustrate it use in a cross-sectional community survey to estimate the prevalence of malaria parasitemia in children aged 2-10 years in five subregions in Uganda.

**KEYWORDS:**
Heckman model; IPDMA; Missing not at random; Selection models;Multiple imputation;

## 1 | INTRODUCTION

Over the past few years, data sharing efforts have substantially increased, and researchers increasingly often have access to IPD from large combined datasets derived from electronic health records (EHR) or from multiple randomized or observable trials (i.e. in IPD meta-analysis, IPD-MA). For example, the clinical practice research datalink (CRPD)[1] is an electronic health record dataset in the UK, which has been used in a variety of medical research, such as the evaluation of health policy and drug efficacy. A recent example of an IPD-MA is the emerging risk factor collaboration,[2] where data were combined from approximately 1.1 million individuals across 104 observational studies to investigate associations of cardiovascular diseases with several predictors. Individuals in these large datasets tend to be clustered in centres, countries or studies, where they have been subject to similar healthcare processes. Moreover, clusters may also differ in participant eligibility criteria, follow-up length, predictor and outcome definitions, or in the quality of applied measurement methods. Hence, differences (i.e.,heterogeneity)

between clusters with respect to baseline covariates and outcomes, while the structure of the correlations between these variables is likely to be different across different clusters.

A usual problem is that such clustered datasets may contain many incomplete variables. For example, in registry data it is common that test results are not available for all patients, as the decision to test may be at the discretion of the primary care physician or because the patient refuses to undergo testing. It is also possible that variables are systematically missing across clusters. For instance, in an IPD-MA, studies may have collected information on different variables. Missing values may thus appear for all participants of a study in the combined dataset. The presence of missing data can lead to loss of statistical power, imbalance in cluster size, bias in parameter estimates and therefore to erroneous conclusions as the analysis could be based on an unrepresentative sample.

To address the presence of missing data, it is important to consider the proportion of missingness and the missing mechanism for each incomplete variable. Rubin (1976)[3] identified three missing mechanisms where the probability of missingness: 1) is constant (missing completely at random; MCAR), 2) depends on observed data only (missing at random; MAR), or 3) depends on unobserved information even after conditioning on all observable variables (missing not at random; MNAR). Traditional imputation methods are designed to address incomplete data sets where variables are MCAR or MAR. Their implementation is justified when there is not systematic difference between units with missing and with complete data or when the missingness of a variable is strongly related to variables measured in the study.

Although it is possible to formally rule out MCAR,[4] it is impossible to assess whether data are MAR or MNAR as the observable data are not enough to test the assumptions of both mechanisms.[5] For this reason, researchers often conveniently assume that data are MAR, or present results of a complete case analysis. In practice, however, unless missingness is artificially introduced by study design, e.g. when a test is only taken on patients with certain characteristics, missingness will often (at least partially) depend on unobserved information.

The missingness mechanism of a variable may be a mixture of MAR and MNAR, depending on the availability of data that may predict the missingness patterns. Therefore, the MAR mechanism may become more influential than MNAR when more information is recorded. For example, health professionals can decide which vital signs to measure on admission. When it is observed that low blood pressure (BP) measurements are not recorded, it could be assumed that this variable is MNAR, as the missingness of BP depends on its own value; however, BP could be considered as MAR variable if additional information is available to explain its missingness, such as the severity of the patient on admission.

Registries are notoriously prone to incomplete variables that are MNAR, due complex recording processes.[6] For example, laboratory tests are taken only in certain patients based on symptoms that are often incompletely recorded. Data from randomized trials may also suffer from MNAR, for example when study participants that experience unfavorable results drop out of the study. Also, heterogeneity of the primary objective or resources of the studies may result in variables relevant to explain the missing process not being recorded at all in some of the studies.

The MNAR mechanism is considered as non-ignorable, in this case it is necessary to specify information about the missingness process in addition to assumptions about the observed data. A popular approach to address MNAR is selection models such as the one proposed by Heckman (1976).[7] Briefly, the Heckman selection model corrects for selection bias by estimating two linked equations: a outcome equation, where the missing variable is associated with predictors, and a selection equation, that accounts for the inclusion of observations in the sample. An important feature of the Heckman selection model is that it does not assume data to be MNAR, so that it can also be used when data are MCAR or MAR. It therefore offers an appealing solution to incomplete data sets when the missingness mechanism is not precisely known.

Over the past few years, several extensions and adaptations to the Heckman selection model have been proposed for multiple imputation. Among them, Galimard et al.(2016)[8] implemented a chained equations imputation method for continuous variables, which was extended to binary and categorical variables by employing copula estimates.[9] Also, Ogundimu and Collins (2019)[10] proposed a chained equations imputation method that is less dependent on normality assumptions.

In clustered data sets, multilevel imputation methods are required to properly propagate uncertainty within and across clusters.[11] However, to our knowledge, existing multilevel imputation methods mainly focus on situations where data are MAR, and do not adopt Heckman selection models. Although Hammon and Zinn (2020)[12] recently proposed an extension that allows for the inclusion of random intercept effects, it can only be used for binary missing variables and assumes that the effect of explanatory variables on the missingness mechanisms is common across clusters.

Therefore, the aim of this work is to develop a multilevel imputation method for continuous and binary variables that are both sporadically and systematically MNAR, and that this can be applied for an incomplete outcome or multiple incomplete predictors in the data sets.

In section 2 we provide an introduction to the Heckman model and its estimation, and we extend it to a hierarchical setting. In section 3 we define the main steps of the proposed imputation method. In section 4 we provide the settings and results of a simulation study to evaluate the performance of the proposed imputation method. In section 5 we illustrate the method using the survey information collected in different sub-districts in Uganda to estimate the prevalence of malaria in children. Finally in section 6 we summarize the results, outline limitations and propose future extension of our method.

## 2 | THE HECKMAN MODEL

The Heckman selection model was initially proposed as a method to correct for selection bias, in which individuals are not randomly selected from the population, leading to inconsistent estimates and erroneous conclusions.[7]

Selection bias occurs when the inclusion of an observation into the sample is influenced by unobserved variables (e.g., the respondent's level of trust toward healthcare entities may cause them to self-select out of the study or refuse to sign consent for a test), which in turn either influence the outcome of interest (e.g., the result of a blood test), or are related to other unobserved variables that influence the outcome of interest.[13]
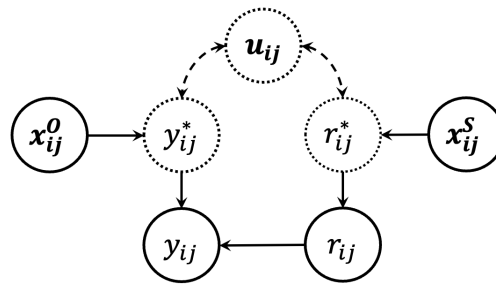


**FIGURE 1** DAG Heckman selection model: here the nodes (dotted lines = latent variables, continuous lines = observed variables) describe the relationship between $y^*_{ij}$ the latent response and $r^*_{ij}$ the latent selection variables of a $j$th unit in a $i$th cluster, that are dependable on $x^O_{ij}$ and $x^S_{ij}$ sets of predictors and are correlated through $u_{ij}$ unobservable variables.

This is visualized in Figure 1, where for the $j$th individual or unit within the $i$th cluster, there is $y^*_{ij}$, a latent outcome variable, and $r^*_{ij}$, a latent selection variable, which are correlated through $u_{ij}$ an unobserved or unrecorded variable, with $i \in [1, 2, ..., N]$ and $j \in [1, 2, ..., n_i]$. Here, both latent variables are related to the sets of predictor covariates $x^O_{ij}$ and $x^S_{ij}$. From $r^*_{ij}$ one can derive $r_{ij} = I(r^*_{ij} >= 0)$ a selection indicator of $y^*_{ij}$ into the sample, and with this in turn, one can define $y_{ij} = y^*_{ij}, \forall r_{ij} = 1$ the observable outcome variable.

Denoting $\boldsymbol{y}^*_i = (y^*_{i1}, y^*_{i2}, ..., y^*_{in_i})^T$ and $\boldsymbol{r}^*_i = (r^*_{i1}, r^*_{i2}, ..., r^*_{in_i})^T$ the vectors of latent outcomes and latent selections in the cluster $i$, then Heckman's model is defined by two main equations: the outcome equation (1), which describes the relation between the latent outcome ($y^*_{ij}$) and a set of covariates ($X^O_i = (x^O_{i1}, x^O_{i2}, ..., x^O_{in_i})^T$), and the selection equation (2) which models the likelihood that the outcome is observed in the sample as a function of another set of covariates ($X^S_i = (x^S_{i1}, x^S_{i2}, ..., x^S_{in_i})^T$).

$$y^*_i = X^O_i \beta^O_i + \epsilon^O_i \tag{1}$$
$$r^*_i = X^S_i \beta^S_i + \epsilon^S_i \tag{2}$$

Here $\beta^O_i$ and $\beta^S_i$ are $p \times 1$ and $q \times 1$ coefficient parameter vectors and $\epsilon^O_i = (\epsilon^O_{i1}, \epsilon^O_{i2}, ..., \epsilon^O_{in_i})^T$ and $\epsilon^S_i = (\epsilon^S_{i1}, \epsilon^S_{i2}, ..., \epsilon^S_{in_i})^T$ are the residual terms vectors for the outcome and selection equations, respectively.

Generally the same variables can be used on the matrix of predictor variables $X^O_i$ and $X^S_i$. However, to avoid multicollinearity problems,[14] it is recommended to include in $X^S_i$ at least one variable that is not included in the outcome model.[15] This variable is commonly known as an exclusion restriction variable (ERV), and should only be associated with the selection in the sample $r^*_{ij}$ (relevance condition) but not with the actual observation $y^*_{ij}$ (exclusion condition).[16] The ERV meets by definition the relevance and exclusion conditions in order to provide independent information about the selection process and to facilitate the estimation of the Heckman model.

In the presence of selection bias, the aforementioned outcome equation will yield biased estimates of $\beta^O_i$ if no efforts are made to adjust for the non-representativeness of the observed $X^O_i$ and $y_i$ values. For this reason, the Heckman model aims to jointly

estimate the outcome and selection equation by defining a relation between their respective error distributions. For instance, Heckman's original model[7] assumes that residual terms have a bivariate normal distribution (BVN),

$$\begin{pmatrix} \epsilon_{ij}^O \\ \epsilon_{ij}^S \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i \\ \rho_i \sigma_i & 1 \end{pmatrix} \right)$$

where $\sigma_i$ corresponds to the variance of the error in the outcome equation and $\rho_i$ to the correlation between the error terms of the outcome and selection equations in the $i$-th cluster. As in a probit model, this model assumes a unit variance for the error term of the selection equation. The unit variance has no consequence on the observable values of $r_{ij} = \{0, 1\}$, since they only depend on the sign of $r_{ij}^*$ and not on its scale.

The interpretation of $\rho_i$ is fairly straightforward. When $\rho_i = 0$, the participation does not affect the outcome model and missing data can be considered MCAR (if data are missing completely at random) or MAR (if missingness is already explained by $x_{ij}^O$). Conversely, when $\rho_i \neq 0$, this suggests that data are MNAR.

## 2.1 | Heckman model estimation

Under the BVN assumption, the parameters of the Heckman model coefficients can be estimated using the two-step Heckman method (H2S)[7] or the full information maximum likelihood method (FIML).[17] However, both methods may lead to inconsistent estimators when the true underlying distribution is not the BVN.[16] To overcome this problem, other approaches have been proposed that relax the distribution assumptions, among them copula models.[18] The copula approach uses a function, known as a copula, that joins the marginal distribution of the error terms of the selection and outcome equation which are specified separately. The Heckman model can be expressed with the following bivariate normal copula:

$$F(r_i^*, y_i^*) = \Phi \left( r_i^* - X_i^S \beta_i^S, \frac{y_i^* - X_i^O \beta_i^O}{\sigma_i}, \rho_i \right),$$

where $\Phi$ is the bivariate standard normal cumulative distribution function. Thus, to estimate the parameters of the Heckman method, it is sufficient to specify the marginal distributions of the error terms, and link them with a suitable copula function. In our imputation method we estimate the Heckman model using the copula method, as in real life data generating mechanism may deviate a BVN distribution. This facilitates the Heckman model estimation and makes it more robust to deviations of distributional assumptions.

## 2.2 | Hierarchical model

The Heckman model can be extended to hierarchical settings, i.e., in cases when individuals or sampling units are nested within clusters, as is the case in EHR or IPD. In this case, sample units from the same cluster are expected to share some characteristics (e.g., the distribution of variables and correlations between them, relationships between exposure and outcomes, missing processes).

Since most statistical analyses assume that the sample units are independent of each other, more complex hierarchical models dealing with nested data are required. This hierarchical complexity must not only be taken into account in the analysis model, but also when dealing with missing data, thus requiring imputation models that are congenial to the analysis model, i.e. that make the same assumptions about the data.

Different procedures can be adopted to combine information between clusters; however, in our imputation method we opted for the two-stage approach that is often used in meta-analyses.[19] This is because such an approach is computationally less intensive and could potentially generate fewer convergence problems in the estimation of the Heckman hierarchical model compared to other approaches.

Briefly, in a first stage we estimate $\theta_i = \{\beta_i^O, \beta_i^S, \sigma_i, \rho_i\}$ the cluster specific parameters of the Heckman model. That is, we estimate separately for each of the $N$ clusters, the parameters of the two equations in each study, the outcome model and the selection model. In the second stage, we combine all $\theta_i$ by using a random effects multivariate meta-analysis model.

In a random effects model, $\theta_i$ parameters are assumed to be drawn independently and identically from a latent multivariate normal distribution of parameters with a population mean $\Theta$ and a population variance $\psi$.[20] Thus, we can write $\theta_i = \Theta + b_i$ using $b_i \sim N(0, \psi)$ random effects to allow for between-study heterogeneity in observed data relationships, and between-study heterogeneity in missing patterns.

# 3 | USING THE HECKMAN MODEL TO IMPUTE MISSING DATA

We follow a similar approach proposed by Resche-Rigon and White (2018)[21] for multilevel imputation of data. Briefly, their method was developed to impute variables from a hierarchical structure (i.e., when there are units grouped within a cluster). This method involves estimating an outcome equation (describing the relationships of the observed data to the missing variable) separately in each cluster, after which the parameter estimates of that equation are pooled using random effects meta-analysis.

With this method, values can be imputed in very common scenarios in IPD, e.g., sporadic and systematic missing patterns. In particular, when the response variable is systematically missing within a group, i.e., when $y_{ij}$ are totally missing within a cluster, the imputation values are drawn from a (generalized) linear model conditional on $\Theta$ the marginal population parameters, i.e., those estimated after pooling the cluster-specific parameters $\theta_i$. On the other hand, when the variable is sporadically missing within the cluster, i.e., there are some observed $y_{ij}$ within the cluster, the imputation model is conditional on the shrunk-cluster parameters, i.e., those coming from the shrinkage of $\theta_i$ towards $\Theta$.

Our imputation approach differs crucially from the previous approach as here we estimate two correlated equations (instead of a single outcome equation) in each cluster, thus obtaining $\theta_i = \{\beta_i^O, \beta_i^S, \sigma_i, \rho_i\}$ parameters from both equations which are then pooled into a $\Theta$ parameter set at the marginal or population level. Our imputation method was created for imputing datasets with a single missing outcome or covariate variable, but it can also be used to impute multiple incomplete variables in a data set.

In the following subsections, we explain the method when the incomplete variable is the outcome variable but this method can also be used to impute any incomplete MNAR predictor variable in the dataset. In the latter case, in the outcome equation (1), the incomplete predictor variable must be specified as the dependent variable and $X_i^O$ as the set of covariates associated with it. On the other hand, in the selection model(2), $r_i^*$ corresponds to the latent variable of selection of the incomplete predictor variable together with its $X_i^S$ associated set of predictors.

## 3.1 | Univariate imputation

Given an outcome variable $y = (y_1, y_2, ..., y_N)^T$, that consists of $y_{ij}^{miss}$ missing and $y_{ij}^{obs}$ observable values, we generate independent draws from the posterior predictive distribution for the missing data, $y_{ij}^{miss}$, given the observable data information $y_{ij}^{obs}$.

$$p(y_{ij}^{miss}|y_{ij}^{obs}) = \int_\theta p(y_{ij}^{miss}|\theta, y_{ij}^{obs})p(\theta|y_{ij}^{obs})\mathrm{d}\theta$$

Here we implicitly assume vague prior distributions for each of the parameters included in the parameter vector $\theta$. Because the integration can be performed computationally by sampling from the posterior predictive distribution $p(\theta|y_{ij}^{obs})$, our imputation method can be carried out in the following two steps:

1. Draw a $\theta$ parameter vector, $\theta^*$, from $p(\theta|y_{ij}^{obs})$, their posterior distribution.

2. Draw $y_{ij}^{miss}$ from $p(y_{ij}^{miss}|\theta^*)$, their predictive distribution for a given $\theta^*$ vector.

Below we describe each step in depth:

### 3.1.1 | Draw the $\theta^*$ parameter vector

**Fit $p(y_{ij}^{obs}|\theta_i)$, the heckman selection model at cluster level**

Initially, we use the copula method to estimate the set of cluster-specific parameters, $\widehat{\theta}_i = \{\widehat{\beta_i^O}, \widehat{\beta_i^S}, \widehat{\sigma}_i, \widehat{\rho}_i\}$, using all $j$ units with observable measurements $y_{ij}^{obs}$ within each cluster $i$. The Heckman model is estimated with the **gjrm** function of the GJRM R package,[22] under the bivariate model with the nonrandom sample selection (BSS) specification, from which we obtain not only the parameters' point estimates $\widehat{\theta}_i$, but also their corresponding $\widehat{S(\theta_i)}$ within-cluster variance-covariance matrix.

**Fit a meta-analysis model**

In this step, we pool the parameters $\widehat{\theta}_i$ with a random effects meta-analysis model using only the clusters with observable information, i.e., those with no systematically missing outcome. In particular, we pool the $p$ coefficients of the $\beta^O$ in the outcome equation and estimate a multivariate random effects meta-analysis model with them. Similarly we combine all $q$ coefficient parameters of the $\beta^S$ in the selection equation.

We also perform a univariate random effects meta_analysis on $\sigma'$, the log-transformed parameter of $\sigma$, and another on $\rho'$, the fisher-transformed parameter of $\rho$.

The meta-analysis model is performed with the **mixmeta** function of the R package mixmeta[23], which allows the use of maximum likelihood (ML), restricted maximum likelihood (REML), and moments estimation methods. For the simulation and illustrative study, we use the restricted REML estimation method, which is recommended as it has a good balance between unbiasedness and efficiency.[24]

**Draw the marginal parameters $\Theta$**

From the meta-analysis model, we obtain the marginal estimates $\widehat{\Theta}$ and the between-cluster variance matrix $\widehat{\psi}$, i.e. variance-covariance matrix of the random effects, with their corresponding variance-covariance matrices $\widehat{S_\Theta}$ and $\widehat{S_\psi}$, which are used to draw the $\Theta^*$ and $\psi^*$ parameters, from their posterior distribution as follows:[25]

$$\Theta^* \sim N(\widehat{\Theta}, \widehat{S_\Theta})$$
$$\psi^* \sim N(\widehat{\psi}, \widehat{S_\psi})$$

**Draw the cluster parameters $\theta_i^*$**

We draw the shrunked-cluster-parameters $\theta_i^*$ for each $i$ cluster from the following posterior distribution conditional on $\Theta^*$ and $\psi^*$.

$$\theta_i^* \sim N\left( \left(\psi^{*-1} + \widehat{S_{\theta_i}}^{-1}\right)^{-1} \left(\psi^{*-1}\Theta^* + \widehat{S_{\theta_i}}^{-1}\widehat{\theta_i}\right), \left(\psi^{*-1} + \widehat{S_{\theta_i}}^{-1}\right)^{-1} \right)$$

As can be seen, the mean and variance of the posterior distribution is a combination between the estimated marginal and cluster-specific parameters. Here the weights on the cluster-specific parameters $\widehat{\theta_i}$ and the marginal parameters $\Theta^*$ are inversely proportional to the within cluster variance $\widehat{S_{\theta_i}}$ and between clusters variance $\psi^*$. For example, when $\widehat{S_{\theta_i}} < \psi^*$ the mean of the conditional distribution gives more weight to the estimated cluster-specific parameter. Conversely, when $\widehat{S_{\theta_i}} > \psi^*$, more weight is given to the estimated marginal parameters. Therefore, in case of a cluster with systematic missingness, it is as if the within-cluster variance is infinite ($\widehat{S_{\theta_i}} \to \infty$), so all the weight is assigned to the parameters estimated at the marginal level. That is, when there is no information in a cluster, we rely entirely on the marginal information.

## 3.1.2 | Draw $y_{ij}^{miss}$ observation

Having $\theta_i^*$, the shrunk-cluster parameters vector for each cluster, we back-transform $\sigma^*$ and $\rho^*$ to the original scale. Then $y_{ij}^{miss}$, the missing values, can be drawn from $p(y_{ij}^{miss}|\theta_i^*)$, their predictive distribution given $\theta_i^*$, as follows:

**Continuous missing variable**

The imputed value of $y_{ij}^{miss}$ can be drawn from the conditional expectation of $y_{ij}$ on unobserved measurements:[26]

$$\mu = E[y_{ij}|r_{ij} = 0, \beta_i^{O*}, \beta_i^{S*}, \rho_i^*, \sigma_i^*]$$
$$\mu = x_{ij}^O\beta_i^{O*} + \rho_i^*\sigma_i^* \frac{-\phi(x_{ij}^S\beta_i^{S*})}{\Phi(-x_{ij}^S\beta_i^{S*})}$$
$$y_{ij}^{miss} \sim N(\mu, \sigma_i^{*2})$$

where $\phi(.)$ is the probability density function and $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

**Binary missing variable**

When $y_{ij}^{miss}$ is a binary variable, the imputed value is drawn from a Bernoulli distribution with a proportion parameter $p_{ij}^*$ given by $P[y_{ij} = 1|r_{ij} = 0]$, that is the conditional probability that $y_{ij} = 1$ given that the measure is unobservable ($r_{ij} = 0$). The $p_{ij}^*$

is obtained from a bivariate probit model, as follows: [26]

$$p_{ij}^* = P[y_{ij} = 1 | r_{ij} = 0, \boldsymbol{\beta}_i^{O*}, \boldsymbol{\beta}_i^{S*}, \rho_i^*]$$

$$p_{ij}^* = \frac{\Phi_2(\boldsymbol{x}_{ij}^O \boldsymbol{\beta}_i^{O*}, -\boldsymbol{x}_{ij}^S \boldsymbol{\beta}_i^{S*}, -\rho_i)}{\Phi(-\boldsymbol{x}_{ij}^S \boldsymbol{\beta}_i^{S*})}$$

$$y_{ij}^{miss} \sim Ber(p_{ij}^*)$$

where $\Phi_2(.)$ corresponds to the bivariate normal cumulative distribution function.

## 3.2 | Multivariate imputation

When there are simultaneous missing variables in a dataset, our imputation method can be extended in a Gibbs sampler procedure. Particularly, our imputation method has been implemented according to the structure of the MICE R package, [27] that allows imputing multiple incomplete predictors and covariates in a given dataset.

Briefly, MICE (multiple imputation of chained equations) was built under the fully conditional specification framework, where for each incomplete variable a conditional imputation model is specified based on another variables in the dataset. This process is carried out iteratively, so that in each iteration the missing values of an incomplete variable are drawn from the conditional distribution based on the updated variables in the previous iteration.

Our imputation model can then be used in the imputation of any incomplete variable in a dataset following an MNAR mechanism, even if it is an outcome, predictor or auxiliary variable of the main analysis model. Furthermore, as implemented according to the MICE framework, the imputation method could be used simultaneously with other imputation methods available in MICE or in add-in packages such as MICEMD [28], to impute datasets with multiple incomplete variables that differ in type and missing mechanism.

# 4 | SIMULATION STUDY

## 4.1 | Aim

We designed a simulation study aimed to compare the performance of alternative methods for imputing a single missing outcome variable in a hierarchical dataset, where the missingness follows a MNAR mechanism. In our scenarios we considered systematically missingness, i.e. the case some of the studies did not report hte variable at all.

## 4.2 | Data-generation mechanism

We generated the data from a Heckman selection model with bivariate normal distribution error terms. For simplicity we started from "basic scenario" where the database collected information from $N = 10$ clusters of $n_i = 1000$ individuals. However, we altered both $N$ and $n_i$ in sensitivity analyses (Table 1).

For each dataset, we generated $X_{1i}$ a treatment indicator variable from a Bernoulli distribution with a probability of treatment on each cluster equal to 0.6. Next, we simulated the mean of two continuous covariates from a multivariate normal distribution $\mu_h \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.2 & 0.015 \\ 0.015 & 0.2 \end{pmatrix}))$, with $h = \{2, 3\}$. We then simulated for each cluster a baseline covariate $X_{2i} \sim N(\mu_2, 1)$ and a exclusion restriction variable $X_{3i} \sim N(\mu_3, 0.5)$.

Here, we considered $X_{1i}$ and $X_{2i}$ as predictors in the outcome equation, i.e., $\boldsymbol{X}_i^O = [1, \boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}]$. For the selection equation we included both variables and the $X_{3i}$ exclusion restriction variable, $\boldsymbol{X}_i^S = [1, \boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}, \boldsymbol{X}_{3i}]$. Then in case of a missing continuous variable, we calculate the latent variables $\boldsymbol{y}_i^*$ and $\boldsymbol{r}_i^*$ as follows:

$$\boldsymbol{y}_i^* = \boldsymbol{\beta}_i^O \boldsymbol{X}_i^O + \boldsymbol{\epsilon}_i^O$$
$$\boldsymbol{r}_i^* = \boldsymbol{\beta}_i^S \boldsymbol{X}_i^O + \boldsymbol{\epsilon}_i^S$$

Here we assumed that all coefficient parameters varied across studies, by including cluster-specific random effects as:

$$\beta_{hi}^O = \beta_h^O + b_{hi}^O$$
$$\beta_{hi}^S = \beta_h^S + b_{hi}^S$$

**TABLE 1** Data generation scenarios

| Scenario | Incomplete variable | $\rho$ | $N; n_i$ | Missing process |
|---|---|---|---|---|
| Base | Continuous | 0.6 | 10;1000 | Heckman(BVN) |
| M(N)AR | Continuous, Binary | MAR:0, MNAR:0.3,0.6,0.9 | 10;1000 | Heckman(BVN) |
| Size and cluster number | Continuous | 0.6 | 10;50, 10;100, 10;1000, 50;1000, 100;1000 | Heckman (BVN) |
| Distribution deviations | Continuous | 0.6 | 10;1000 | Heckman(BVN), Heckman(t-skew), Explicit |

We fixed coefficients $\beta_h^O = \{0.3, 1, 1\}$ and $\beta_h^S = \{-0.8, 1.3, -0.7, 1.2\}$ in order to get around 40% of sporadically missing values on the response $y_{ij}$ in the entire data set. Additionally, we ensured that the $y_{ij}^*$ observations were systematically missing in 20% of the clusters included in the data set, by removing the outcome values in the 20% of the clusters.

We assumed that random effects were independent within equations ($b_{h0}^O \perp\!\!\!\perp b_{h1}^O \perp\!\!\!\perp b_{h2}^O$ and $b_{h0}^S \perp\!\!\!\perp b_{h1}^S \perp\!\!\!\perp b_{h2}^S$), but were linked between both selection and outcome equations through a bivariate normal distributed as:

$$\begin{pmatrix} b_h^O \\ b_h^S \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_{bh}^2 \begin{pmatrix} 1 & \rho * 0.4 \\ \rho * 0.4 & 1 \end{pmatrix} \right)$$

with the parameters $\sigma_{b0}^2 = \sigma_{b1}^2 = \sigma_{b2}^2 = 0.4$. We considered that the correlation parameter of the random effects between equations is 40% of the value of the assumed correlation parameter between error terms $\rho$. In addition, we included a random effect on the exclusion restriction variable given by $b_3 \sim N(0, 0.2)$ assuming that the intracluster variation in the exclusion restriction effect is lower than the variation on other coefficient parameters effects. The $\rho$ parameter was given different values depending on the simulated missing mechanism (Table 1).

As regards the error terms, they were bivariate normal distributed as:

$$\begin{pmatrix} \epsilon_i^O \\ \epsilon_i^S \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho\sigma_i \\ \rho\sigma_i & 1 \end{pmatrix} \right)$$

whose $\sigma_i^2$ is variable across clusters and distributed as $log(\sigma_i) \sim N(0, 0.05)$.

## 4.3 | Adittional scenarios

In addition to the basic scenario described above, we explored additional data generating mechanisms. Specifically we investigated the performance of the imputation methods under the following scenarios:

- **M(N)AR scenarios:** We explored a missing MAR mechanism ($\rho = 0$), a scenario when data followed a MNAR mechanism with a low ($\rho = 0.3$), intermediate ($\rho = 0.6$) and strong correlation ($\rho = 0.9$) between $y^*$ and $r^*$. We also explored a scenario when the missing variable is binary. Therefore we simulated a $y_i$ binary incomplete variable, by keeping similar parameters to the ones used in the simulation of a missing continuous variable, but here we defined the observable binary variable as:

$$r_i = I(r*_i > 0)$$
$$y_i = I(y_i^* > 0) \forall r_{ij}^* > 0$$

- **Influence of sample size and cluster number**: we explored different configurations regarding the number of patients per cluster $n_i = \{50, 100, 1000\}$ and the number of clusters $N = \{10, 50, 100\}$.

- **Violation of distributional assumptions:** Here, we aimed to investigate how the imputation models behave in settings with departures from the bivariate normal distribution, performed two sensitivity analyses.

   1. **Skewed-t:** We drew error terms from a bivariate skewed student-t distribution using the same location parameter and covariance matrix of the normal distributed settings, with 4 degrees of freedom and an $\alpha = \{-2, 6\}$ parameter which regulates the asymmetry of the density.

   2. **Explicit:** In addition we simulated an explicit missingness process, where error terms of the selection and outcome equations were independently normal distributed and the selection of observations depended on the value of the outcome variable, as $ry_i^* = 0.3y_i^* + \epsilon_i^S$. These settings assure that the percentage of missingness on the outcome variable is around 60% in all the evaluated scenarios.

   3. **Normal:** As a reference, we provide the results from the basic scenario of the main simulation study, where the error terms were Bivariate normal distributed.

## 4.4 | Estimands

The estimands were the parameter coefficients of the outcome equation $\beta^O = \{\beta_0^O, \beta_1^O, \beta_2^O\}$, with special emphasis on the treatment effect parameter $\beta_1^0$. We also report the estimated variance of the random effects and residual errors $\sigma_{b0}^2, \sigma_{b1}^2, \sigma_{b2}^2, \sigma_e^2$.

## 4.5 | Analysis Methods

For each scenario we simulated 500 datasets over which we evaluated the following imputation methods:

- **Complete case analysis (CCA):** We removed all patients with missing observations.

- **1l.Heckman:** Multiple imputation based on the Heckman model without no study specification, following the imputation method proposed by Galimard et al.(2016).[8]

- **2l.MAR**: Multiple imputation assuming MAR for hierarchical datasets, we used the multilevel imputation model ( 2l.2stage.norm and 2l.2stage.bin) from the micemd R package,[28] which are described in Audigier et al. (2018) paper.[11]

- **2l.Heckman**: The proposed imputation method based on the Heckman model for hierarchical datasets.

## 4.6 | Estimating procedures

After the imputation procedure, we estimated the following (generalized) mixed linear effect model using the *lmer()* function from the *lme4* R package.[29] $y_i = \beta_i^O X_i^O + \epsilon_i^O$ In case of missing binary variable, we used the same matrix of predictors but on a binary model estimated with the *glmer()* function from the *lme4* R package. Then, we pooled the estimates of the $\beta_i^O$ and the variance of the random effect and residual errors of the multiple imputed datasets according to Rubin's rule,[30] over which we calculated the performance measures on the estimands.

To calculate the coverage of the parameter coefficients' \95% confidence intervals (CI), we estimate CI with the Wald method. Although it is possible to obtain CI through the profile or bootstrap method for the variances of the random effects, we prefer not to estimate them (and hence the coverage) for the random effects parameters due to computational time.

## 4.7 | Performance measures

We calculated the following measures, usually employed to evaluate imputation methods,[31] according to the formulas provided in Morris et al.(2019):[32]

- **Bias:** Bias on the coefficient and random effect parameters.
- **Coverage:** Coverage of the 95% confidence intervals for the coefficient parameters.
- **Width**: Average width of the confidence interval on the coefficient parameters.
- **RMSE:** Root mean squared error of the coefficient and random effect parameters.

In addition, in the appendix table we reported the empirical standard errors (EmpSE), Monte Carlo standard errors (ModSE) on the coefficient parameters, average processing time (time in seconds) and the percentage of datasets where the imputation method converged (run), i.e., the imputation method generated an output.

## 4.8 | Software

For the simulation study and illustrative examples we used R version 4.0.4 in a linux environment.[33]

The Heckman 2L imputation method is available in the MICEMD R package[28] (as **mice.2l.heckman()**) and also on the github repository https://github.com/johamunoz/Statsmed_Heckman where you can also find all the codes accompanying this paper and a toy example that explains how to implement the method in mice.

## 4.9 | Results from the simulation study

### 4.9.1 | Descriptive statistics of the simulated datasets

We generated data sets of 10 clusters of 1000 patients each in each scenario. For example, for the base scenario (Table 1) in which the error terms followed a normal distribution, out of the 500 datasets generated, we obtained that on average 60.8% of the Y response was missing, with the lowest missing percentage being 26.84% and the maximum being 73.67%. At the cluster level, looking only at the sporadic missing clusters, we found an average of 35.25% missing values, but there were clusters with no missing data at all and up to 98% missing data in the outcome. Processing times differ in the imputation methods evaluated, the 1l. Heckman takes an average of 10 seconds, the 2l.MAR takes 1.5 and the 2l. Heckman takes about 18 seconds. It should be noted that the time of the last two approaches depends strongly on the number of clusters included and the sample size in the data set. (See Annex excel file)

### 4.9.2 | Results M(N)AR scenarios

**Continuous incomplete variable**

Figure 2 shows the results of simulations where the missing variable was continuous. In the MAR scenario, i.e. when $\rho = 0$, all imputation methods provided similar unbiased estimates of the coefficient parameters, but as rho increased, i.e. the mechanism became MNAR, the estimates for the complete case analysis and the MAR-IPD imputation method became biased.

As it is expected both Heckman-based imputation (1l. and 2l.) models provided less biased estimates of the coefficient parameters in MNAR scenarios, but the random effects estimates on the 1l.heckman were far away from the true values as no cluster information was considered at all.

Regarding coverage, the 2l.Heckman imputation method provided the best coverage values across all the $\rho$ scenarios with a coverage level close to the nominal 95% interval. Even though 2l.Heckman was not properly a randomization-valid method[31], as it was not unbiased and had a coverage above 95% across all the $\rho$ values, it led to better results in terms of bias and coverage compared to the evaluated methods.

The 2l.Heckman method, also resulted in better estimates in terms of RMSE than the other methods evaluated, via a better trade-off between bias and variance.

In particular, our method provided an advantage over the 1l.Heckman method in systematic missingness scenarios, by adding a cluster variable in the imputation model, a feature lacking from 1l.Heckman. This can be seen in the estimates of the random effects parameters of the 1l.Heckman method, which were more biased than those of the other methods in which cluster information was included, i.e. 2l.MAR and 2l.Heckman.

Regarding the width of the 95% CI of the estimated parameters, we observed that using the 2l.Heckman model we obtained larger CIs than when using other evaluated methods, which generally increased as $\rho$ moved away from zero. In particular, we observed in $\beta_2$ that the CI length was larger in the MAR scenario. By checking the simulation results further, we noted that the high variability came from one simulation in which the linear model estimation had singularity problems (not shown here).

**Bivariate incomplete variable**

In the case of a missing bivariate variable, we also found (Figure 3) that the 2l.Heckman method provided the least unbiased results on the coefficient parameters. However, we observed more unbiased estimates and a larger 95th percentile amplitude in the estimates of the variance of the random effects, especially in the $\sigma_{b2}$ compared to the estimates obtained in the continuous incomplete outcome scenario.
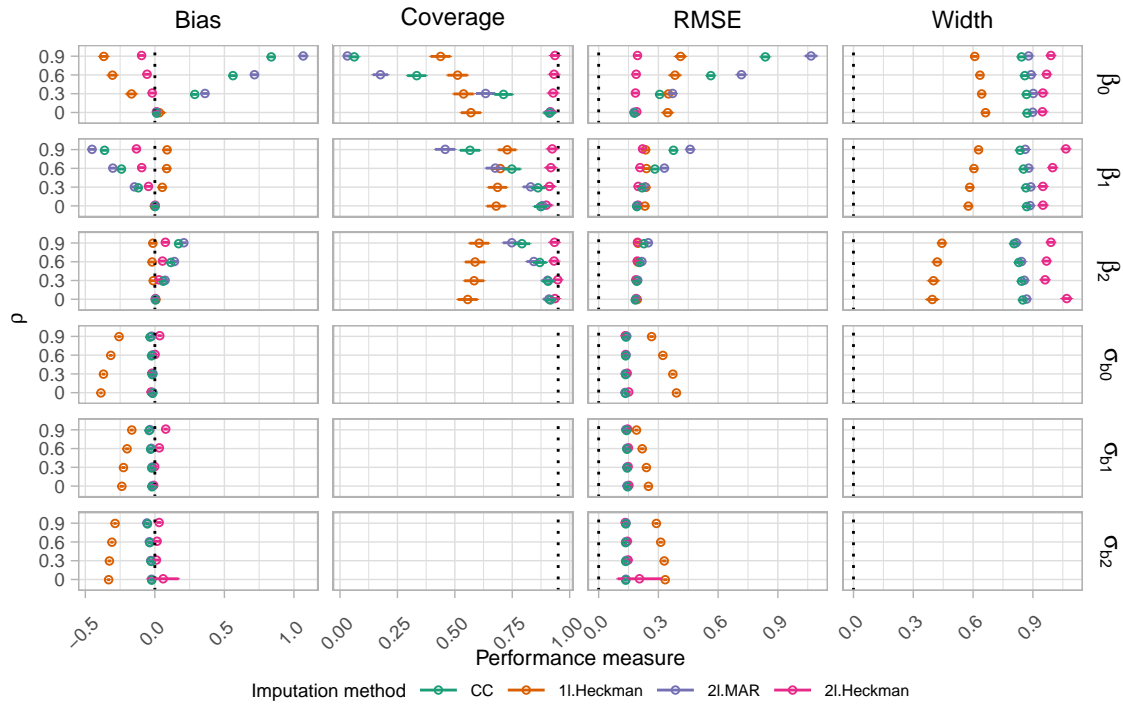
**FIGURE 2** Continuous incomplete variable by varing $\rho$, where dashed lines depict the target performance criteria value
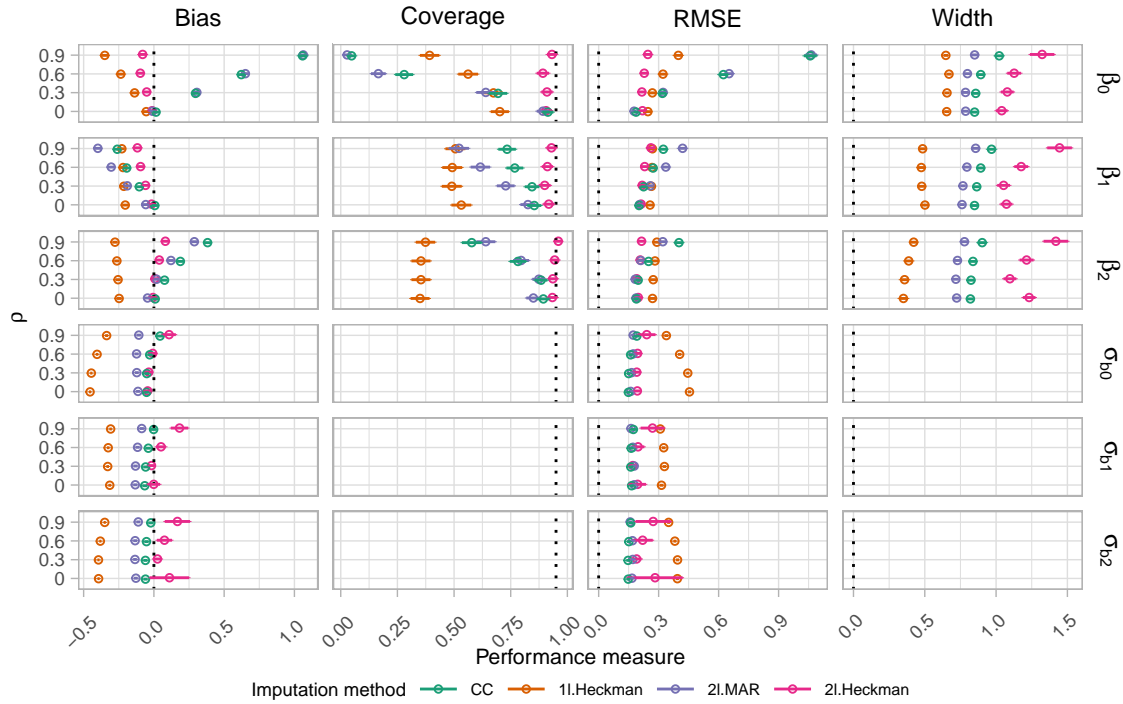


**FIGURE 3** Binary incomplete variable by varing $\rho$, where dashed lines depict the target performance criteria value

### 4.9.3 | Sensitivity analysis: number of clusters and sample size of clusters

We evaluated how robust our method was to variations in the number of clusters and also in the sample size of cluster (Figure 4).

By increasing N, the number of clusters, from 10 to 100, we observed that the bias was not affected but the width of the 95% CI of the estimates decreased (larger precision) and hence the RMSE decreased. On the other hand, by reducing the number of units per cluster (from $n_i$=1000 to $n_i$=100) the precision decreased for all coefficients, the bias on coefficient estimates were not drastically affected but the bias of random effects did.

When we reduced the sample size to 50 patients per study, the bias and RMSE of the $\sigma_b2$ were drastically affected (not shown here but in the Appendix). This could be explained in part due to the scarce information on certain clusters, which affects directly the estimation of the Heckman model on those clusters.
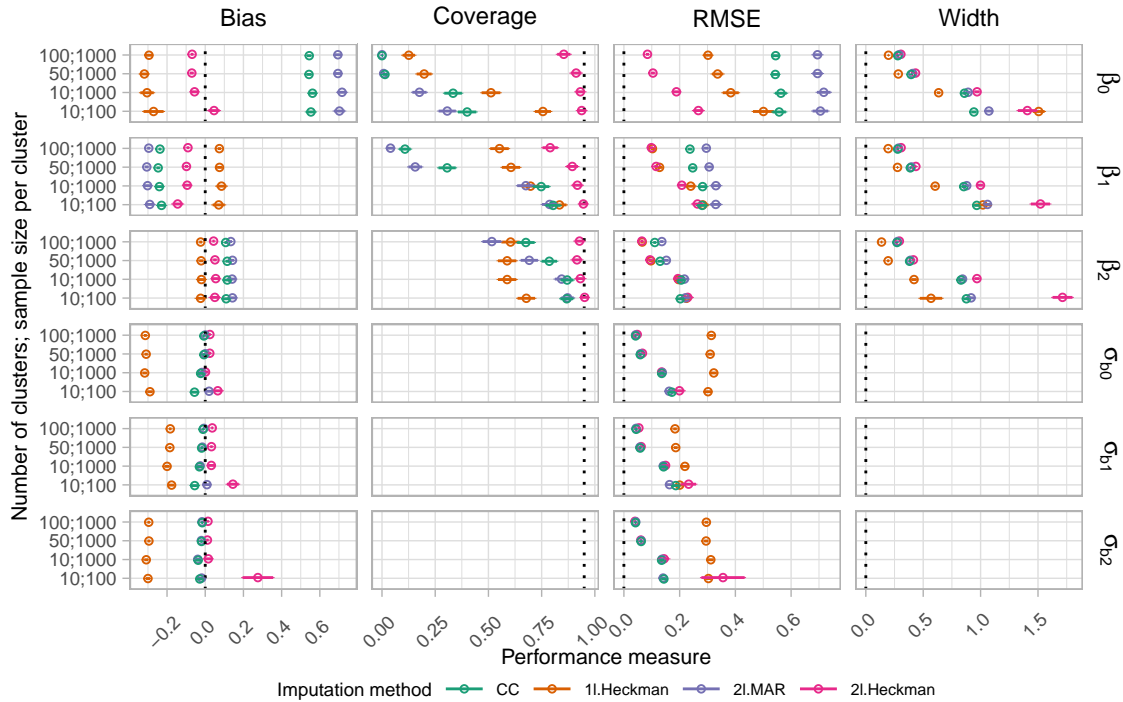


**FIGURE 4** Continuous incomplete variable under systematical missingness by varing: number of clusters (N); sample size per cluster ($n_i$), where dashed lines depict the target performance criteria value

### 4.9.4 | Sensitivity analysis: distributional assumptions

When using the Heckman model in an explicit MNAR process, i.e. when the probability of loss is associated with the value of the missing variable, we observe that under all imputation methods we obtain biased estimators, with poor coverage. Particularly when estimating the intercept we observed more biased estimates compared to those obtained for the other estimated parameters B2 with coverage below 60%.

Our model may not be fully suitable for this type of scenario, especially if the main analysis is focused on estimating absolute prevalence estimates, as it is greatly affected by the bias of the intercept parameter ($\beta_0$) and the width of the 95th percentile CI of all coefficient parameters. Also the imputation method was affected in terms of the bias of the random-effects parameters $\sigma_{b2}$.

With respect to the Skewed-t scenario, only the bias of $\beta_0$ was affected for all imputation methods applied. But $\beta_0$ estimates for both the 1l.Heckman and 2l.Heckman were drastically affected in terms of bias, with very poor coverage (below 25%).Therefore, the applicability of our method could be questionable in scenarios that do not conform to the BNV assumption.
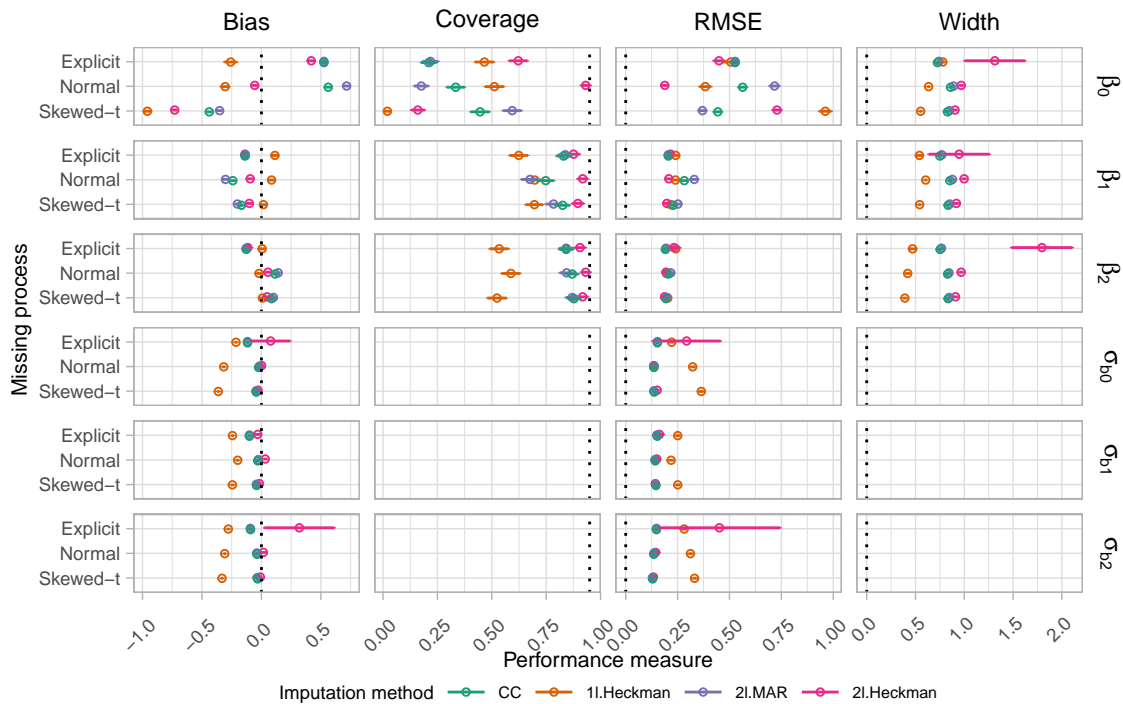
**FIGURE 5** Continuous incomplete variable with deviations in distribution assumptions, where dashed lines depict the target performance criteria value

## 5 | AN ILLUSTRATIVE STUDY

Malaria is a mosquito-borne disease and is the leading cause of illness and death in Africa, especially in children and pregnant women. To prevent the spread of the disease, long-lasting nets (LLINs) and indoor residual spraying (IRS) in at-risk households are used as control measures.

Specifically, in Uganda, under the Uganda LLIN evaluation project, a LLINS distribution campaign was conducted between 2013 and 2014. In 2017, the effect of LLIN control together with insecticides was assessed through a cross-sectional community survey in 104 health sub-districts in 48 districts located within 5 sub-regions of Uganda.

In each sub-district, a sample of households with at least one child aged 2-10 years was surveyed, where information was collected on household conditions and use of preventive measures. In addition, finger prick blood samples were taken from each child to determine the prevalence of parasitemia and an entomological study was conducted to estimate mosquito prevalence. Details of the project and survey are provided elsewhere.[34]

For this example, we used data accessed directly from CliniEpiDB,[35] where data were collected from 5195 households with verified consent, inhabited by 11137 residents aged 2-10 years. Blood samples were only taken from 8846 children, as 69 were excluded from the study due to lack of consent and 2222 were not present at the time of the survey. Although the original data set consists of 164 variables, here we only consider the variables described in Table 2, which were used as predictors in the imputation model and are fully observed in the dataset. In this dataset, the parasitaemia test is an incomplete binary result (1=positive test, 0=negative test), which is missing 21% across the whole dataset.

To illustrate our proposed method, following the article by Rugnao et al. (2019),[36] we estimated the prevalence of parasitemia by subregion and by age after approximately 3 years of LLIN campaigns started. We estimated parasitemia prevalence using 3 approaches that made different assumptions on the missingness mechanism: MCAR, MAR and MNAR.

Under the MCAR assumption, prevalence was calculated on the basis of the recorded tests, i.e., we only included patients with a test result. Under the MAR assumption, the test values of children who were not present during the survey were imputed with the 2l.2stage.bin method of the MICEMD package, where the community was taken as the cluster and the following factors previously associated with parasitemia were used as predictors in the imputation model: sex, bednet ( indicator of whether only

**TABLE 2** Descriptive analysis, predictor variables

| Sub-region | District (N) | Children (N) | Age mean (years) | Log10 Female Anopheline | Wealth index | Bednet (%) | Girls (%) | Holiday (%) | No test (%) |
|---|---|---|---|---|---|---|---|---|---|
| North East | 5 | 794 | 5.50 | 2.67[1.5,4.3] | -0.45[-1.2,2.2] | 10.7 | 49.0 | 31.9 | 17.5 |
| Mid Eastern | 8 | 1354 | 5.61 | 0.84[0.1,2.5] | -0.14[-1.0,2.5] | 9.3 | 48.1 | 32.9 | 25.6 |
| South Western | 14 | 3596 | 5.69 | 0.27[0.1,1.3] | 0.18[-1.0,2.9] | 23.8 | 49.4 | 66.5 | 21.1 |
| Mid Western | 12 | 3172 | 5.66 | 1.27[0.1,3.2] | -0.03[-1.0,2.8] | 13.3 | 48.9 | 62.9 | 20.5 |
| East Central | 9 | 2152 | 5.61 | 2.74[0.4,6.3] | 0.01[-1.1,3.1] | 13.2 | 51.6 | 51.6 | 16.0 |

**TABLE 3** Evaluation of Holidays as exclusion restriction variable

| Predictors/response | Test result (y) | Test taken (ry) |
|---|---|---|
| (Intercept) | -1.07(0.05)*** | 1.35(0.05)*** |
| Log10 Female Anopheline | 0.73(0.03)*** | 0.15(0.02)*** |
| Wealth index | -0.55(0.04)*** | 0.04(0.03) |
| Bednet-Yes | -0.30(0.08)*** | 0.70(0.08)*** |
| Holidays-Yes | -0.04(0.05) | 0.19(0.05)*** |
| Girls-No | 0.10(0.05) | 0.05(0.05) |
| s(Age) | 1.84(1.97)*** | 1.02(1.04)*** |

*Note:* ***p<0.01

two or fewer persons share a mosquito bed net). In addition, we included age as a power 3 spline function, the cluster-level Log10 mean of the number of female anopheline mosquitoes per household estimated from the entomological survey, and the household wealth index from principal components analysis calculated specifically for the surveyed households.

Under the MNAR assumption, we used the proposed 2l.Heckman method to impute missing test values. The selection and outcome equation included the same predictor variables as used under the MAR approach. In addition, we included a holiday indicator variable as ERV. This was calculated according to school vacation calendars and public holidays in Uganda in 2017. We examined the association of this ERV with the outcome variable (y) and with the selection indicator (ry), conditioned on the remaining imputation predictors. The model results in Table 3 indicate that the holiday indicator could be a plausible ERV variable, as there was strong evidence of an association with ry, but no evidence of an association with y.

According to our imputation approach, non-tested children were estimated to have a higher prevalence of malaria than participants in more than half of the districts analyzed. As can be seen in Figure 6, for each subregion the prevalence estimates of the approaches did not differ significantly between methods. However, prevalence estimates under the MNAR assumption (i.e. 2 level Heckman) were higher than those estimated under the MAR or MCAR aproaches, except for the East-Central region.

In terms of prevalence by age, there were no significant differences between methods (Figure 7). The prevalence estimates for children aged 2 to 6 years were very similar in all regions under the different assumptions. Assuming that children start going to school after the age of 6, the results could be partly explained by the mobility of 2-6 year olds compared to school-age children, i.e. school-age children spend more time outdoors and travel more than younger children.

However for school children, prevalences estimated with the Heckman method were found to be higher in the Mid-East and Southwest regions than those obtained with the other methods, whereas in the East-Central region the estimates with the Heckman method are lower. A possible reason for selection bias in surveys of this type is, for example, that daytime visits might favor measurement in sick school children who stay home, leading to overestimated prevalence results as found in the East-Central region.[37] Nevertheless, we were unable to find information confirming the direction in which malaria prevalence is driven by selection bias in this Uganda study or in other studies similar to this one.
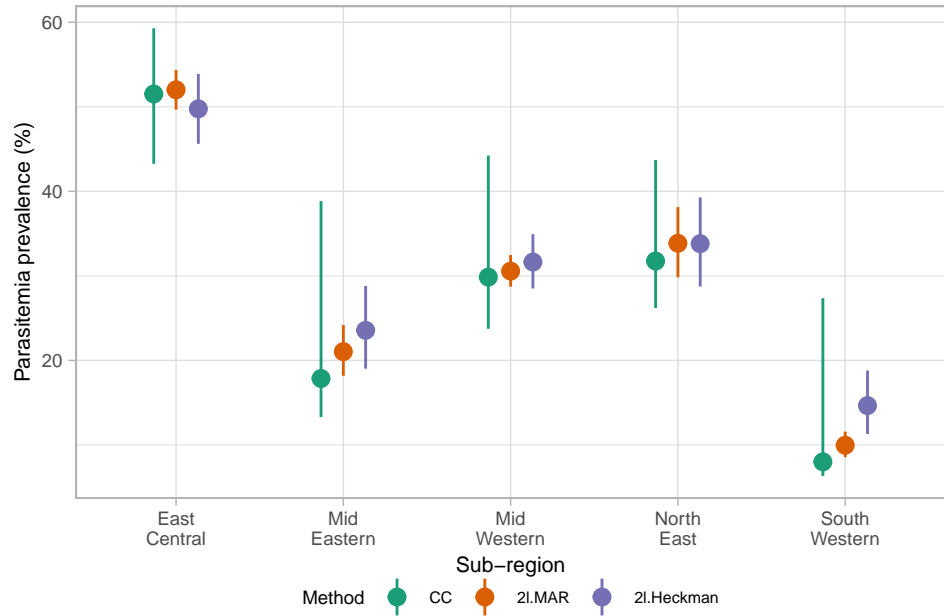
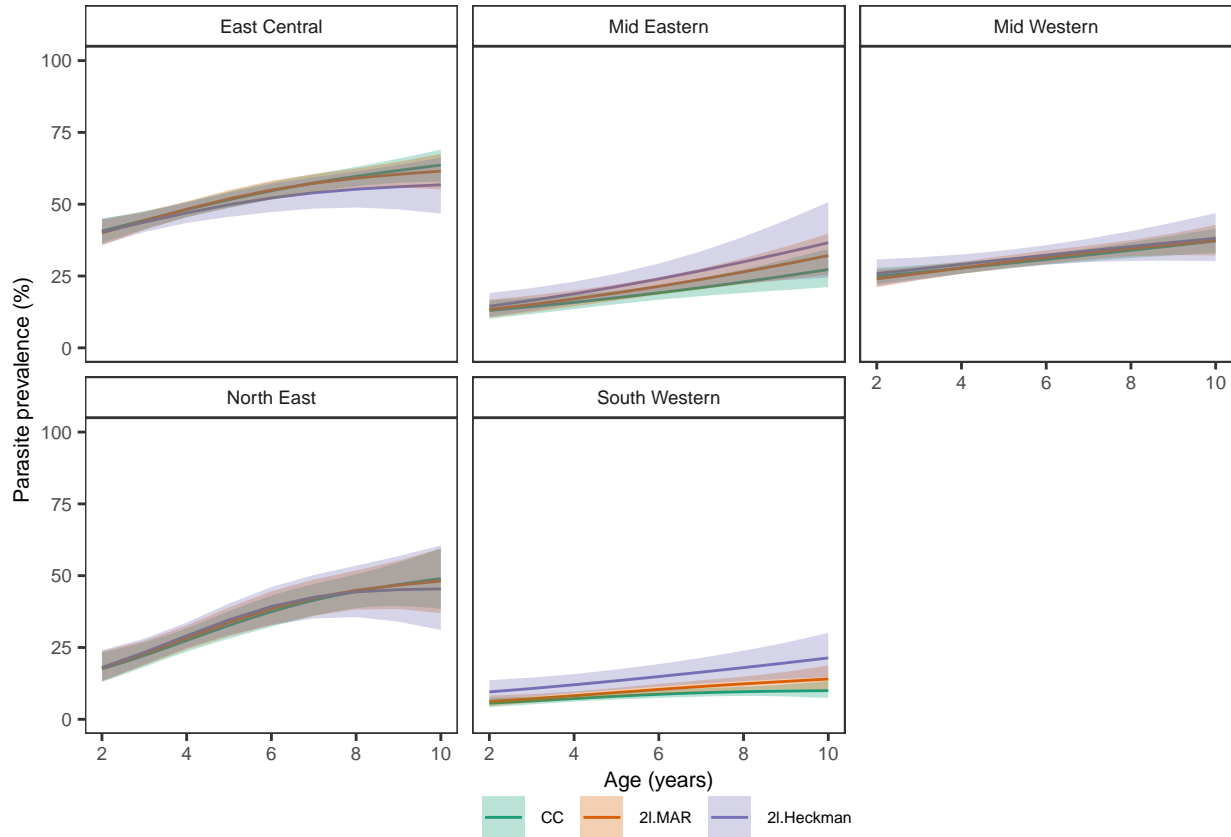**FIGURE 6** Estimates of malaria prevalence by sub-region



**FIGURE 7** Estimates of malaria prevalence by sub-region and age

## 6 | DISCUSSION

We have extended and evaluated methods for multiple imputation of clustered datasets, in situations where some incomplete variables follow a MNAR mechanism. For clustered datasets, only imputation methods under the MAR mechanism had previously been proposed. Although there are imputation methods that can handle MNAR they can only handle the case of individual studies. This puts limit to their use in situations common in IPD-MA such as when there is systematic missingness or when the proportion of missingness of a variable is very high in one of the included studies. To address this gap, we proposed a new multiple imputation method for continuous and binary MNAR covariates, that is specifically designed for clustered datasets. Our method, allows borrowing of information between the clusters to obtain more reliable imputation results at the individual cluster level.

In our simulations we observed that the imputation method we proposed id optimal for the imputation of continuous and binary missing variables that follow a MNAR mechanism according to the Heckman model and that come from multilevel data, such data commonly used in the IPDMA.

Overall, our new method produced unbiased estimates with convergence close to 95% for the coefficient parameters. It also resulted in less biased estimates for the random effects parameters compared to the other methods evaluated.

Empirically, we showed that the proposed method could be applicable when there is systematic missingness at the cluster level. This method, in particular, could provide less biased imputation values compared with individual-level imputation methods, as it not only allows imputation of missing values in clusters with systematic missingness, but can also shrunk the values of individual clusters toward the overall study mean. This is especially advantageous in studies with small sample sizes, where an analysis approach that ignores data from other studies may lead to extreme estimates of prevalence. By using our proposed method, these would be reduced to the overall mean.

The advantage of the proposed method over methods that assume MAR is that it allows the imputation of variables from cluster level data following a MAR or MNAR mechanism according to the Heckman model. That is to say that under the specification of a valid exclusion variable the method determines which is the most adjustable correlation parameter between equations ($\rho$), or in general terms the missingness mechanism (MAR or MNAR), in each of the clusters evaluated.

Our implementation of the imputation method was built according to the specifications of the mice R package and is available in the MICEMD package, which allows, first of all, to be used both on the outcome and on the covariates. In addition, it offers the option of being used simultaneously with other imputation methods implemented in the MICE package, which is advantageous in databases containing missing variables with different prediction methods and models. Finally, the method can be used on systematically and sporadically missing clusters, both for continuous and binary missing variables with heterogeneous effects and error variances.

### 6.1 | Limitations and future directions

A major limitation of our method is that it needs a valid restriction variable, which in some contexts is difficult to establish at the individual study level and can be even more challenging if one tries to find a valid exclusion variable across clusters.

In addition, to estimate the marginal estimates, the method only uses clusters with observable information, i.e. that are not systematically missing or have sufficient information to estimate the Heckman model. The latter might restrict the evaluation of the Heckman model at the cluster level to a certain number of predictors depending on the sample size of the cluster. This method also requires at least some of the studies to have no systematic missigness in the incomplete variable in question. That is, if all studies have the incomplete variable systematically missing, the method is unusable, which can be an important limitation in practice.

Also, the method can be sensitive to both the sample size and the number of studies included in the database. On the one hand, a small sample size at the individual study level can affect not only the precision of estimates but also the convergence of the method since sample size needed to estimate the parameters of the Heckman model which can be at least twice the number of parameters required to estimate in an imputation model that assumes MAR. On the other hand, a high number of studies that may improve the precision of the estimators may also make the estimation of the marginal parameters more difficult and also considerably increase the processing time of our method.

In our simulation study, data were generated by assuming a constant correlation across all clusters in order to evaluate the performance against M(N)AR assumptions, in practice, however this parameter can be variable across clusters and this can considerably affect the performance of our method. Therefore, in future research the effect of this parameter could be further

evaluated. One might also consider relaxing this assumption of constant correlation to allow for a random effects distribution for the correlation parameter.

Further, the method can also be extended to other types of variables such as count or ordinal variables. Similarly, less restrictive Heckman based models can be considered in terms of normality distribution of errors and no specification of exclusion variables such as those proposed by Ogundimu & Collins (2019). [10].

## 6.2 | Conclusion

We have proposed an extension to the Heckman model that can account for MNAR, MAR or MCAR of a continuous or binary variable in clustered data sets. Our simulations showed that it can have favorable statistical properties, when its assumptions were met, and provided that the sample size is sufficiently large. Regarding deviations from distributional assumptions of the error terms, the coefficient parameters were fairly robust in terms of bias, but the intercept was not.

## FOOTNOTES

### Disclaimer

The views expressed in this paper are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the authors are employed/affiliated.

### Acknowledgements

## References

1. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology* 2015; 44(3): 827–836. doi: 10.1093/ije/dyv098

2. The Emerging Risk Factors Collaboration . The Emerging Risk Factors Collaboration: Analysis of Individual Data on Lipid, Inflammatory and Other Markers in over 1.1 Million Participants in 104 Prospective Studies of Cardiovascular Diseases. *European Journal of Epidemiology* 2007; 22(12): 839–869. doi: 10.1007/s10654-007-9165-7

3. Rubin DB. Inference and Missing Data. *Biometrika* 1976; 63(3): 581–592. doi: 10.1093/biomet/63.3.581

4. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 1988; 83(404): 1198–1202. doi: 10.1080/01621459.1988.10478722

5. Enders CK. *Applied Missing Data Analysis*. Methodology in the Social SciencesNew York: The Guilford Press. second edition ed. 2022.

6. Liu D, Oberman HI, Muñoz J, Hoogland J, Debray TPA. Quality Control, Data Cleaning, Imputation. In: arXiv. 2021.

7. Heckman JJ. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. In: NBER. 1976 (pp. 475–492).

8. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A Multiple Imputation Approach for MNAR Mechanisms Compatible with Heckman's Model. *Statistics in Medicine* 2016; 35(17): 2907–2920. doi: 10.1002/sim.6902

9. Galimard JE, Chevret S, Curis E, Resche-Rigon M. Heckman Imputation Models for Binary or Continuous MNAR Outcomes and MAR Predictors. *BMC Medical Research Methodology* 2018; 18(1): 90. doi: 10.1186/s12874-018-0547-1

10. Ogundimu EO, Collins GS. A Robust Imputation Method for Missing Responses and Covariates in Sample Selection Models. *Statistical Methods in Medical Research* 2019; 28(1): 102–116. doi: 10.1177/0962280217715663

11. Audigier V, White IR, Jolani S, et al. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science* 2018; 33(2). doi: 10.1214/18-STS646

12. Hammon A, Zinn S. Multiple Imputation of Binary Multilevel Missing Not at Random Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2020; 69(3): 547–564. doi: 10.1111/rssc.12401

13. Vella F. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* 1998; 33(1): 127. doi: 10.2307/146317

14. Puhani PA. Foul or Fair? The Heckman Correction for Sample Selection and Its Critique. A Short Survey. Tech. Rep. 97-07, ZEW - Leibniz Centre for European Economic Research; Leibniz, Germany: 1997.

15. Angrist JD, Krueger AB. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 2001; 15(4): 69–85. doi: 10.1257/jep.15.4.69

16. Gomes M, Kenward MG, Grieve R, Carpenter J. Estimating Treatment Effects under Untestable Assumptions with Nonignorable Missing Data. *Statistics in Medicine* 2020; 39(11): 1658–1674. doi: 10.1002/sim.8504

17. Amemiya T. Tobit Models: A Survey. *Journal of Econometrics* 1984; 24(1): 3–61. doi: 10.1016/0304-4076(84)90074-5

18. Smith MD. Modelling Sample Selection Using Archimedean Copulas. *The Econometrics Journal* 2003; 6(1): 99–123. doi: 10.1111/1368-423X.00101

19. Simmonds MC, Higginsa JPT, Stewartb LA, Tierneyb JF, Clarke MJ, Thompson SG. Meta-Analysis of Individual Patient Data from Randomized Trials: A Review of Methods Used in Practice. *Clinical Trials* 2005; 2(3): 209–217. doi: 10.1191/1740774505cn087oa

20. Higgins JPT, Thompson SG, Spiegelhalter DJ. A Re-Evaluation of Random-Effects Meta-Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172(1): 137–159. doi: 10.1111/j.1467-985X.2008.00552.x

21. Resche-Rigon M, White IR. Multiple Imputation by Chained Equations for Systematically and Sporadically Missing Multilevel Data. *Statistical Methods in Medical Research* 2018; 27(6): 1634–1649. doi: 10.1177/0962280216666564

22. Radice GMaR. GJRM: Generalised Joint Regression Modelling. https://CRAN.R-project.org/package=GJRM; 2021.

23. Gasparrini A, Sera F. Mixmeta: An Extended Mixed-Effects Framework for Meta-Analysis. https://CRAN.R-project.org/package=Mixmeta; 2021.

24. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics* 2005; 30(3): 261–293. doi: 10.3102/10769986030003261

25. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, on behalf of the PROG-IMT Study Group . Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data. *Statistics in Medicine* 2013; 32(28): 4890–4905. doi: 10.1002/sim.5894

26. Greene WH. *Econometric Analysis*. New York, NY: Pearson. eighth edition ed. 2018.

27. Buuren vS, Groothuis-Oudshoorn K, Vink G, et al. Mice: Multivariate Imputation by Chained Equations. https://CRAN.R-project.org/package=mice; 2021.

28. Audigier VCM, Resche-Rigon MIE. Micemd: Multiple Imputation by Chained Equations with Multilevel Data. https://CRAN.R-project.org/package=micemd; 2021.

29. Bates D, Maechler M, Bolker [aut B, et al. Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4. https://CRAN.R-project.org/package=lme4; 2022.

30. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and StatisticsHoboken, NJ, USA: John Wiley & Sons, Inc. . 1987

31. Buuren vS. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC Interdisciplinary Statistics SeriesBoca Raton: CRC Press, Taylor and Francis Group. second edition ed. 2018.

32. Morris TP, White IR, Crowther MJ. Using Simulation Studies to Evaluate Statistical Methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. doi: 10.1002/sim.8086

33. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.

34. Staedke SG, Kamya MR, Dorsey G, et al. LLIN Evaluation in Uganda Project (LLINEUP) – Impact of Long-Lasting Insecticidal Nets with, and without, Piperonyl Butoxide on Malaria Indicators in Uganda: Study Protocol for a Cluster-Randomised Trial. *Trials* 2019; 20(1): 321. doi: 10.1186/s13063-019-3382-8

35. Staedke S. ClinEpiDB. https://clinepidb.org/ce/app/workspace/analyses/DS_7c4cd6bba9/new/details; .

36. Rugnao S, Gonahasa S, Maiteki-Sebuguzi C, et al. LLIN Evaluation in Uganda Project (LLINEUP): Factors Associated with Childhood Parasitaemia and Anaemia 3 Years after a National Long-Lasting Insecticidal Net Distribution Campaign: A Cross-Sectional Survey. *Malaria Journal* 2019; 18(1): 207. doi: 10.1186/s12936-019-2838-3

37. Program TD. DHS Survey Design: Malaria Parasitemia. tech. rep., U.S. Agency for International Development (USAID); USA: 2020.