Received: 24 January 2023 | Revised: 31 October 2023 | Accepted: 7 November 2023

DOI: 10.1002/sim.9965

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Multiple imputation of incomplete multilevel data using Heckman selection models

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

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

[2]Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

[3]Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

[4]Conservatoire national des arts et métiers (CNAM), Laboratoire CEDRIC-MSDMA, Paris, France

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

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

**Correspondence**
Johanna Muñoz, Julius Center for Health Sciences and Primary Care, UMC Utrecht, Str. 6.131, P.O. Box 85500, 3508GA Utrecht, The Netherlands.
Email: j.munozavila@umcutrecht.nl

**Funding information**
Horizon 2020 Framework Programme, Grant/Award Number: ReCoDID 825746

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 variables, that is, 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 multilevel structure 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 regions 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 (ie, in IPD meta-analysis, IPD-MA). For example, the clinical practice research datalink (CRPD)[1] is an EHR 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 centers, countries or studies,

*Valentijn M. T. de Jong and Thomas P. A. Debray contributed equally to this study.

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, 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.[3] 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 missing mechanism for each incomplete variable. Rubin[4] identified three missing mechanisms where the probability of missingness: (1) is independent from observed or missing values (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.

Registries are notoriously prone to incomplete variables that are MNAR, due complex recording processes.[5] 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.

Modelling data under MNAR mechanism implies to specify information about the missingness process in addition to assumptions about the observed data. Two major approaches have been used to address MNAR mechanism: pattern mixture models[6] and selection models.[7] One of the most popular techniques within selection models is the one proposed by Heckman.[8] Briefly, the Heckman selection model corrects for selection bias by estimating two linked equations: an outcome equation, where the missing variable is associated with predictors, and a selection equation, which 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[9] implemented a chained equations imputation method for continuous variables, which was extended to binary and categorical variables by employing copula estimates.[10] Also, Ogundimu and Collins[11] 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.[12] 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[13] 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, which 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 our 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 our results, outline limitations and propose future extensions of our method.

## 2 | THE HECKMAN MODEL

The Heckman selection model was initially proposed as a method to correct for selection bias, that is, which individuals are not randomly selected from the population, leading to inconsistent estimates and erroneous conclusions.[8]

Selection bias occurs when the inclusion of an observation into the sample is influenced by unobserved variables (eg, 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 (eg, the result of a blood test), or are related to other unobserved variables that influence the outcome of interest.[7]

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 $\boldsymbol{u_{ij}}$ an unobserved or unrecorded variable, with $i \in [1, 2, \dots, N]$ and $j \in [1, 2, \dots, n_i]$. Here, both latent variables are related to the sets of predictor covariates $\boldsymbol{x^O_{ij}}$ and $\boldsymbol{x^S_{ij}}$. 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 $\boldsymbol{y^*_i} = (y^*_{i1}, y^*_{i2}, \dots, y^*_{in_i})^T$ and $\boldsymbol{r^*_i} = (r^*_{i1}, r^*_{i2}, \dots, r^*_{in_i})^T$ 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 ($\boldsymbol{X^O_i} = (\boldsymbol{x^O_{i1}}, \boldsymbol{x^O_{i2}}, \dots, \boldsymbol{x^O_{in_i}})^T$), 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 ($\boldsymbol{X^S_i} = (\boldsymbol{x^S_{i1}}, \boldsymbol{x^S_{i2}}, \dots, \boldsymbol{x^S_{in_i}})^T$).

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

Here $\boldsymbol{\beta^O_i}$ and $\boldsymbol{\beta^S_i}$ are $p \times 1$ and $q \times 1$ coefficient vectors and $\boldsymbol{\epsilon^O_i} = \left(\epsilon^O_{i1}, \epsilon^O_{i2}, \dots, \epsilon^O_{in_i}\right)^T$ and $\boldsymbol{\epsilon^S_i} = \left(\epsilon^S_{i1}, \epsilon^S_{i2}, \dots, \epsilon^S_{in_i}\right)^T$ 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 $\boldsymbol{X^O_i}$ and $\boldsymbol{X^S_i}$. However, to avoid multicollinearity problems,[14] it is recommended to include in $\boldsymbol{X^S_i}$ 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 $\boldsymbol{r^*_i}$ (relevance condition) but not with the actual observation $\boldsymbol{y^*_i}$ (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 $\boldsymbol{\beta^O_i}$ if no efforts are made to adjust for the non-representativeness of the observed $\boldsymbol{X^O_i}$ and $\boldsymbol{y_i}$ 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[8] assumes that residual terms have a bivariate normal distribution (BVN),

$$\begin{pmatrix} \boldsymbol{\epsilon^O_i} \\ \boldsymbol{\epsilon^S_i} \end{pmatrix} \sim N_{n_i}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_i & \rho_i\sigma_i \\ \rho_i\sigma_i & 1 \end{pmatrix}\right),$$

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

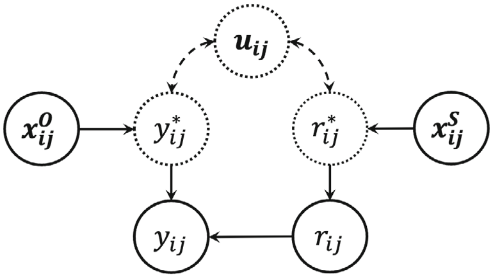

FIGURE 1 Graphical representation of the Heckman selection model: here the nodes (dotted lines = latent variables, continuous lines = observed variables) describe the relationship between $y^*_{ij}$ the latent response and $r^*_{ij}$ the latent selection variables of a $j$th unit in a $i$th cluster,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

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 $\boldsymbol{x^O_{ij}}$). Conversely, when $\rho_i \neq 0$, this suggests that data are MNAR.

## 2.1 | Heckman model estimation

Under the BVN distribution, the parameters of the Heckman model can be estimated using the two-step Heckman method[8] or the full information maximum likelihood method.[17] However, both methods may lead to inconsistent estimators when the true underlying distribution is not a 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 cumulative joint distribution $F\left(\epsilon^S_{ij}, \epsilon^O_{ij}\right)$ of the error terms is given by:

$$F\left(\epsilon^S_{ij}, \epsilon^O_{ij}\right) = C\left(F_S\left(\epsilon^S_{ij}\right), F_O\left(\epsilon^O_{ij}\right); \rho_i\right),$$

where $F_S\left(\epsilon^S_{ij}\right)$ and $F_O\left(\epsilon^S_{ij}\right)$ are the marginal distributions of the error terms for both equations and C is a copula function with with correlation parameter $\rho_i$. Thus, to estimate the parameters of the Heckman method, it is sufficient to specify the marginal distributions of the error terms and 'connect' them with a suitable copula function. In our imputation method, we estimate the Heckman model using a Gaussian copula and employing the method for non-random selection,[19] which is a flexible approach that allows the specification of different parametric distributions of the selection and outcome variables and different types of dependence structure between the two equations.

## 2.2 | Hierarchical model

The Heckman model can be extended to hierarchical settings, that is, in cases when individuals or sampling units are nested within clusters, as is the case in EHR or IPD meta-analysis. 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, that is, that make the same assumptions about the data.

Different procedures can be adopted to combine information across clusters; however, in our imputation method we opted for the two-stage approach that is often used in meta-analyses.[20] 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 one-stage methods. Our method is based on the following conditional imputation model, whose parameters are $\boldsymbol{\theta} = \{\boldsymbol{\beta^O}, \boldsymbol{\beta^S}, \boldsymbol{\Psi^S}, \boldsymbol{\Psi^O}, (\sigma_i, \rho_i)_{1<=i<=N}\}$

$$\begin{aligned}
\boldsymbol{y^*_i} &= \boldsymbol{X^O_i}(\boldsymbol{\beta^O} + b^O_i) + \boldsymbol{\epsilon^O_i},\\
\boldsymbol{r^*_i} &= \boldsymbol{X^S_i}(\boldsymbol{\beta^S} + b^S_i) + \boldsymbol{\epsilon^S_i},\\
b^O_i &\sim N_p(0, \boldsymbol{\Psi^O}),\\
b^S_i &\sim N_q(0, \boldsymbol{\Psi^S}),
\end{aligned}$$

$$\begin{pmatrix} \boldsymbol{\epsilon^O_i} \\ \boldsymbol{\epsilon^S_i} \end{pmatrix} \sim N_{n_i}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_i & \rho_i\sigma_i \\ \rho_i\sigma_i & 1 \end{pmatrix}\right).$$

Briefly, at the first stage we estimate the cluster-specific parameters of the Heckman model $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} = \{\widehat{\boldsymbol{\beta^O_i}}, \widehat{\boldsymbol{\beta^S_i}}, \widehat{\sigma}_i, \widehat{\rho}_i\}$, for clusters, that is, for all $i$ in 1, 2, …,$N$. That is, for each cluster, we estimate separately the parameters of the two equations, the outcome model (1) and the selection model (2) via a copula model.

At the second stage, we fit a random effects multivariate meta-analysis model for the coefficients. Here $\boldsymbol{\beta^O_i}$ and $\boldsymbol{\beta^S_i}$ parameters are assumed to be drawn independently and identically from a latent multivariate normal distribution of parameters with across-cluster means $\boldsymbol{\beta}^{\boldsymbol{O}}$ and $\boldsymbol{\beta}^{\boldsymbol{S}}$ and across-cluster variance covariance $\boldsymbol{\psi}^{\boldsymbol{O}}$ and $\boldsymbol{\psi}^{\boldsymbol{S}}$ respectively.[21] Here we assume that $\sigma_i$ and $\rho_i$ are random variables that follow a distribution across clusters, and we estimate random effects univariate meta-analysis models for these parameters.

## 3 | USING THE HECKMAN MODEL TO IMPUTE MISSING DATA

When dealing with MNAR data, it is necessary to make assumptions about the distribution of the missing data. Unfortunately, these assumptions cannot directly be verified from the observed data. Generally, MNAR imputation methods can be categorized into two distinct approaches: pattern mixture models and selection models, which rely on specific assumptions (please refer to the Appendix for details). Our imputation method is based in the Heckman selection model, which is estimated through a Gaussian copula model, and inherently embodies certain assumptions (as outlined in Table 1).

We follow a similar approach proposed by Resche-Rigon and White[22] for multilevel data imputation, which allows us to impute values in very common scenarios in IPD meta-analysis, for example, sporadic and systematic missingness patterns. Our imputation method was created to impute datasets with a single missing outcome variable or covariate, but it can also be used to impute multiple incomplete variables in a dataset, as it can be implemented under a multivariate imputation by chained equations (MICE) approach.

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 $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{O}}$ as the set of covariates associated with it. On the other hand, in the selection equation (2), $\boldsymbol{r}_{\boldsymbol{i}}^{*}$ corresponds to the latent variable of selection of the incomplete predictor variable together with its $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{S}}$ associated set of predictors.

### 3.1 | Univariate imputation

Given an outcome variable $\boldsymbol{y} = (y_1, y_2, \dots, y_N)^T$, which 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 $y_{ij}^{obs}$.

$$p\left(y_{ij}^{miss}\middle|y_{ij}^{obs}\right) = \int_{\boldsymbol{\theta}} p\left(y_{ij}^{miss}\middle|\boldsymbol{\theta}, y_{ij}^{obs}\right) p\left(\boldsymbol{\theta}\middle|y_{ij}^{obs}\right) \mathrm{d}\boldsymbol{\theta}.$$

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

TABLE 1 Assumptions of the 2l.2stage.heckman imputation method.

| Imputation method assumptions |
|---|
| The MNAR mechanism can be characterized as indirectly non-ignorable, that is, the likelihood of a missing value in the incomplete variable is not determined by the variable's own value (ie, self-masking), but instead depends on other unobserved variables that are also correlated with the incomplete variable |
| The error terms of both selection and outcome equations are connected by a Gaussian distribution and they are linearly related through a correlation parameter $\rho$ |
| An exclusion restriction variable, that is, a predictor variable that appears in the selection equation but is not included in the outcome equation |
| Hierarchical Level-2 structure: cluster-specific parameters also follow a Gaussian underlying distribution and the coefficients in the outcome equation are uncorrelated with those in the selection equation |

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

Below we describe each step in depth.

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

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

Initially, we use the copula method to estimate the set of cluster-specific parameters, $\widehat{\boldsymbol{\theta_i}} = \{\widehat{\boldsymbol{\beta_i^O}}, \widehat{\boldsymbol{\beta_i^S}}, \widehat{\sigma}_i, \widehat{\rho}_i\}$, using all $j$ units with observable measurements $y_{ij}^{obs}$ within cluster $i$. Here we obtain the estimates not only the parameters' point estimates $\widehat{\boldsymbol{\theta_i}}$, but also their corresponding $\widehat{\boldsymbol{S}(\boldsymbol{\theta_i})}$ within-cluster variance-covariance matrix.

*Fit a meta-analysis model*

In this step, we pool the parameters $\widehat{\boldsymbol{\theta_i}}$ with a random effects meta-analysis model using only the clusters with observable information, that is, those with no systematically missing outcome. In particular, we pool the $p$ coefficients of $\boldsymbol{\beta^O}$ in the outcome equation and fit a multivariate random effects meta-analysis model with them. Similarly we combine all $q$ coefficients of $\boldsymbol{\beta^S}$ in the selection equation. Thus, we can denote the study specific coefficients

$$\widehat{\boldsymbol{\beta_i^O}} = \boldsymbol{\beta^O} + \boldsymbol{b_i^O} + \boldsymbol{\epsilon_i'^O},$$
$$\widehat{\boldsymbol{\beta_i^S}} = \boldsymbol{\beta^S} + b_i^S + \boldsymbol{\epsilon_i'^S}$$

using the random effects $b_i^O \sim N(0, \boldsymbol{\Psi^O})$ and $b_i^S \sim N(0, \boldsymbol{\Psi^S})$ with sampling errors $\epsilon_i'^O$ and $\epsilon_i'^S$.

We assume that $\sigma_i$ and $\rho_i$ are random variables coming from an across-cluster distribution, so we can express them as:

$$log(\sigma_i) \sim N(log(\sigma), \psi^\sigma),$$

$$tanh^{-1}(\rho_i) \sim N(tanh^{-1}(\rho), \psi^\rho).$$

For each of them, we perform a univariate random effects meta-analysis. The model for $\sigma_i$ is given by the model:

$$log(\widehat{\sigma}_i) = \sigma + b_i^\sigma + \epsilon_i^\sigma$$

with $b_i^\sigma \sim N(0, \psi^\sigma)$, and $\epsilon_i^\sigma \sim N(0, var(log(\widehat{\sigma}_i)))$. In the case of an incomplete binary variable, the $\sigma_i$ parameter is not specified, so it is not necessary to perform a random effects meta-analysis model of $\sigma_i$ or to include it in any of the following steps in the imputation model.

The model for $\rho_i$ is given by:

$$tanh^{-1}(\widehat{\rho}_i) = \rho + b_i^\rho + \epsilon_i^\rho$$

with $b_i^\rho \sim N(0, \psi^\rho)$, and $\epsilon_i^\rho \sim N(0, var(tanh^{-1}(\widehat{\rho}_i)))$.

*Draw the marginal parameters* $\boldsymbol{\Theta}$

From the meta-analysis model, we obtain the marginal estimates $\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\beta^O}}, \widehat{\boldsymbol{\beta^S}}, \widehat{\sigma}, \widehat{\rho}\}$ and the between-cluster variance matrix $\widehat{\boldsymbol{\psi}} = \{\widehat{\boldsymbol{\Psi^O}}, \widehat{\boldsymbol{\Psi^S}}, \widehat{\psi^\sigma}, \widehat{\psi^\rho}\}$, that is, variance-covariance matrix of the random effects, with their corresponding variance-covariance matrices $\widehat{\boldsymbol{S_\Theta}}$ and $\widehat{\boldsymbol{S_\psi}}$, which are used to draw the $\boldsymbol{\Theta}^*$ and $\boldsymbol{\psi}^*$ parameters, from their posterior distribution as follows:[3]

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

*Draw the cluster parameters $\boldsymbol{\theta}_{\boldsymbol{i}}^{*}$*

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

$$\theta_i^{*} \sim N\left(\left(\boldsymbol{\psi}^{*-1} + \widehat{\boldsymbol{S}_{\theta_i}}^{-1}\right)^{-1}\left(\boldsymbol{\psi}^{*-1}\boldsymbol{\Theta}^{*} + \widehat{\boldsymbol{S}_{\theta_i}}^{-1}\widehat{\boldsymbol{\theta}_{\boldsymbol{i}}}\right), \left(\boldsymbol{\psi}^{*-1} + \widehat{\boldsymbol{S}_{\theta_i}}^{-1}\right)^{-1}\right).$$

As can be seen, the mean and variance of the posterior distribution is a combination of the estimated marginal and cluster-specific parameters. Here the weights on the cluster-specific parameters $\widehat{\boldsymbol{\theta}_{\boldsymbol{i}}}$ and the marginal parameters $\boldsymbol{\Theta}^{*}$ are inversely proportional to the within cluster variance $\widehat{\boldsymbol{S}_{\theta_i}}$ and between clusters variance $\boldsymbol{\psi}^{*}$. For example, when $\widehat{\boldsymbol{S}_{\theta_i}} < \boldsymbol{\psi}^{*}$ the mean of the conditional distribution gives more weight to the estimated cluster-specific parameter. Conversely, when $\widehat{\boldsymbol{S}_{\theta_i}} > \boldsymbol{\psi}^{*}$, more weight is given to the estimated marginal parameters. Therefore, in case of a cluster with systematic missingness, the within-cluster variance is effectively infinite ($\widehat{\boldsymbol{S}_{\theta_i}} \rightarrow \infty$), so that 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.

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

Having estimated $\boldsymbol{\theta}_{\boldsymbol{i}}^{*}$, 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 $p(y_{ij}^{miss}|\boldsymbol{\theta}_{\boldsymbol{i}}^{*})$, their predictive distribution given $\boldsymbol{\theta}_{\boldsymbol{i}}^{*}$, as follows:

### 3.2.1 | Continuous missing variable

The imputed value of $y_{ij}^{miss}$ can be drawn from the conditional expectation of $y_{ij}$ on unobserved measurements:[23]

$$\mu = E[y_{ij}|r_{ij} = 0, \boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{O}*}, \boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{S}*}, \rho_i^{*}, \sigma_i^{*}],$$

$$\mu = \boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{O}}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{O}*} + \rho_i^{*}\sigma_i^{*}\frac{-\phi\left(\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{S}}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{S}*}\right)}{\Phi\left(-\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{S}}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{S}*}\right)},$$

$$y_{ij}^{miss} \sim N(\mu, \sigma_i^{*2}),$$

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

### 3.2.2 | 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:[23]

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

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

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

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

### 3.3 | 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,[24] 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 other 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 or predictor of the main analysis model, of for auxiliary variables. 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,[25] to impute datasets with multiple incomplete variables that differ in type and missing mechanism.

### 3.4 | Technical details of the implementation

The Heckman model is estimated with the **gjrm** function of the GJRM R package,[26] under the bivariate sample selection model (BSS) specification with a bivariate normal error distribution (BivD=“N”). The meta-analysis model is estimated with the **mixmeta** function of the R package mixmeta,[27] 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.[28]

## 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.

### 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 a “basic scenario”, that is, where the database collected information from $N = 10$ clusters of $n_i = 1000$ individuals. Subsequently, we altered both $N$ and $n_i$ in additional analyses (Table 2).

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 $\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.2 & 0.015 \\ 0.015 & 0.2 \end{pmatrix}\right)$.

We then simulated for each cluster a baseline covariate $X_{2i} \sim N(\mu_2, 1)$ and an 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, that is, $\boldsymbol{X_i^O} = [1, \boldsymbol{X_{1i}}, \boldsymbol{X_{2i}}]$. 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:

$$\boldsymbol{y_i^*} = \boldsymbol{X_i^O} \boldsymbol{\beta_i^O} + \boldsymbol{\epsilon_i^O},$$
$$\boldsymbol{r_i^*} = \boldsymbol{X_i^O} \boldsymbol{\beta_i^S} + \boldsymbol{\epsilon_i^S},$$

TABLE 2 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 | 0 (MAR), 0.3,0.6,0.9 (MNAR) | 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),<br>Heckman (t-skew),<br>Self-masking |

where $\boldsymbol{\beta_i^O} = (\beta_{i0}^O, \beta_{i1}^O, \beta_{i2}^O)^T$ and $\boldsymbol{\beta_i^S} = (\beta_{i0}^S, \beta_{i1}^S, \beta_{i2}^S, \beta_{i3}^S)^T$. Here we assumed that the coefficient k varied across studies, by including cluster-specific random effects as:

$$\beta_{ik}^O = \beta_k^O + b_{ik}^O, k = 0, 1, 2,$$
$$\beta_{ik}^S = \beta_k^S + b_{ik}^S, k = 0, 1, 2, 3.$$

Denoting $\boldsymbol{\beta^O}$=$(\beta_0^O, \beta_1^O, \beta_2^O)^T$ and $\boldsymbol{\beta^S}$=$(\beta_0^S, \beta_1^S, \beta_2^S, \beta_3^S)^T$, we fixed coefficients $\boldsymbol{\beta^O}$=$(0.3, 1, 1)$ and $\boldsymbol{\beta^S}$=$(-0.8, 1.3, -0.7, 1.2)$ 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 ($b_{i0}^O \perp\!\!\!\perp b_{i1}^O \perp\!\!\!\perp b_{i2}^O$ and $b_{i0}^S \perp\!\!\!\perp b_{i1}^S \perp\!\!\!\perp b_{i2}^S \perp\!\!\!\perp b_{i3}^S$), but were linked between both selection and outcome equations through a bivariate normal distributed as:

$$\begin{pmatrix} b_{ik}^O \\ b_{ik}^S \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{kk}^O & \rho_b\sqrt{\psi_{kk}^O}\sqrt{\psi_{kk}^S} \\ \rho_b\sqrt{\psi_{kk}^O}\sqrt{\psi_{kk}^S} & \psi_{kk}^S \end{pmatrix}\right)$$

with $k = 0, 1, 2$, $\psi_{00}^S = \psi_{00}^O = \psi_{11}^S = \psi_{11}^O = \psi_{22}^S = \psi_{22}^O = 0.4$ and $\rho_b = \rho * 0.4$. 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 $b_{i3}^S \sim N(0, 0.2)$ 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 2).

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

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

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

## 4.3 | Additional 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 was binary. Therefore we simulated $y_i$ as a binary incomplete variable, by keeping similar parameters to the ones used in the simulation of a missing continuous variable, but here after defining the observed 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, performing 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. **Self-masking:** In addition we simulated a self-masking 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 $r_i^* = 0.3y_i^* + \epsilon_i^S$. These settings led to around 60% missingness of the outcome variable across 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 $\boldsymbol{\beta^O} = (\beta_0^O, \beta_1^O, \beta_2^O)^T$, with special emphasis on the treatment effect parameter $\beta_1^0$. We also report the estimated standard deviation from the covariance matrix of the random effects, that is, $\sqrt{\psi_{00}^O}, \sqrt{\psi_{11}^O}, \sqrt{\psi_{22}^O}$.

### 4.4.1 | Estimating procedures

After the imputation procedure with incomplete continuous outcome, we estimated the following mixed linear effect model using the **lmer()** function from the lme4 R package.[29]

$$\boldsymbol{y_i = X_i^O \beta_i^O + \epsilon_i^O}$$

with $\beta_i^O = \beta^O + b_i^O$, $b_i^O \sim N(0, \Psi^O)$ and $\epsilon_i^O \sim N(0, \Sigma_i)$.

In case of an incomplete binary outcome, we used the same matrix of predictors but we fit on a generalized linear mixed model 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.

## 4.5 | Imputation 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 for 1-level incomplete variable, that is, with no study specification, following the imputation method proposed by Galimard et al.[9]

- **2l.MAR**: Multiple imputation assuming MAR with 2-level hierarchical datasets. We used the multilevel imputation model (2l.2stage.norm and 2l.2stage.bin) from the micemd R package,[25] which are described by Audigier et al.[12]
- **2l.2stage.heckman**: The proposed imputation method based on the Heckman model for hierarchical datasets.

## 4.6 | Performance measures

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

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

In addition, on the GitHub repository (https://github.com/johamunoz/Statsmed_Heckman), we reported the empirical standard errors (EmpSE), Monte Carlo standard errors (ModSE) on the coefficients, average processing time (time in seconds) and the percentage of datasets where the imputation method converged (run), that is, the imputation method generated an output.

## 4.7 | Software

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

The 2l.2stage.heckman imputation method is available in the micemd R package[25] (as **mice.2l.2stage.heckman()**) and also on the github repository, specified above, where we also provide all codes accompanying this paper as well as a toy example that explains how to implement the method in mice.

## 4.8 | Results from the simulation study

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

*Continuous incomplete variable*

Figure 2 shows the results of simulations where the incomplete outcome was continuous. In the MAR scenario, that is, when $\rho = 0$, all imputation methods provided similar unbiased estimates of the coefficients $\beta^O$, but as $\rho$ increased, that is, the mechanism became MNAR, the estimates for the complete-case analysis and the 2l.MAR imputation method became further away from the true value. As expected, both Heckman-based imputation methods (1l.heckman and 2l.2stage.heckman) gave less biased estimates of the coefficients $\beta^O$ in MNAR scenarios, but the estimates of the random effects parameters $\sqrt{\psi^O_{kk}}$ with $k = \{0, 1, 2\}$ in the 1l.heckman method were the most unbiased. This could be explained by the fact that 1l.heckman does not take into account any grouping information.

In terms of coverage, the 2l.2stage.heckman imputation method gave the best coverage values in all $\rho$ scenarios with a coverage level close to the nominal 95% value. Although 2l.2stage.heckman was not properly a randomization-valid method,[31] as it was not unbiased and had a coverage above 95% across all $\rho$ values, this method gave better results in terms of bias and coverage when compared to the other methods evaluated. The 2l.2stage.heckman method, also resulted in better $\beta^O$ estimates in terms of RMSE than the other methods evaluated, through a better compromise between bias and variance.

In particular, the 2l.2stage.heckman method provides an advantage over the 1l.heckman method on data sets with systematically missing variables, by including cluster information in the imputation model. This can be seen in the random effects parameter estimates $\sqrt{\psi^O_{kk}}$ of the 1l.heckman method, which were more biased than those of the other methods where cluster information was included, that is, 2l.MAR and 2l.2stage.heckman.

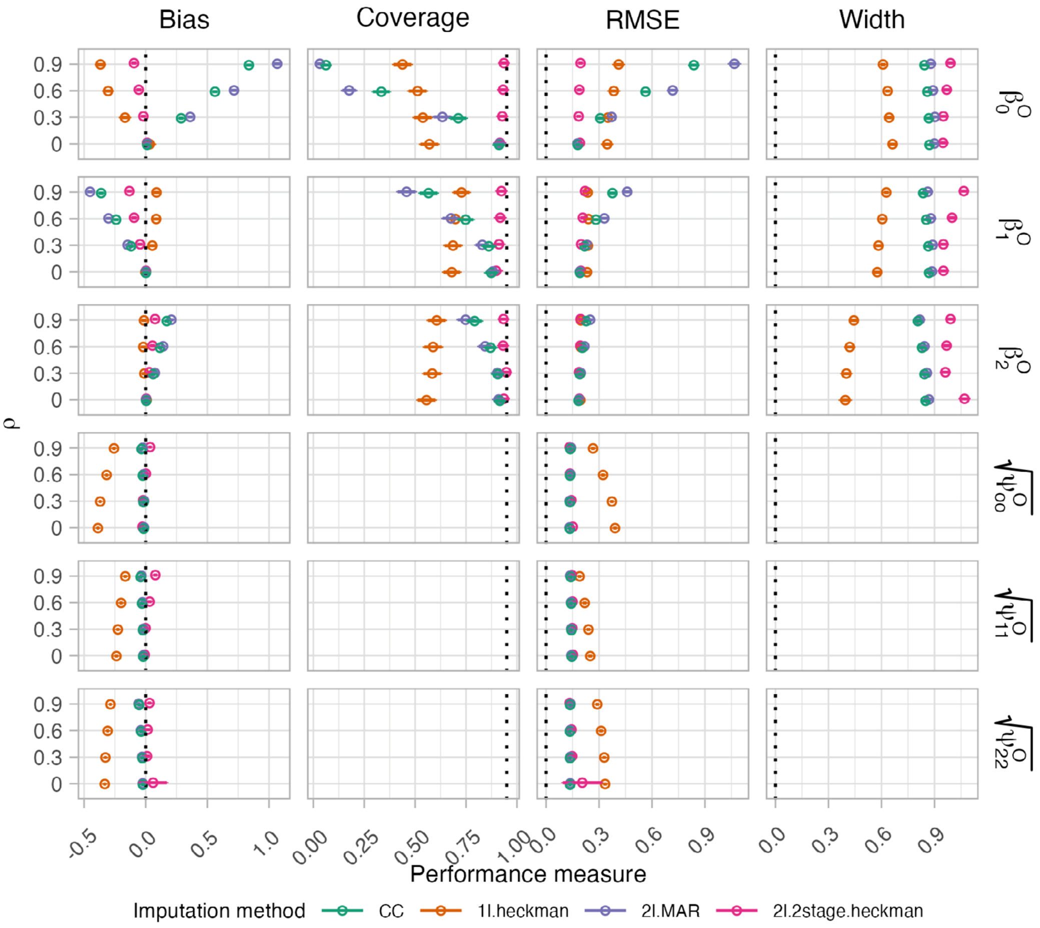

**FIGURE 2** Continuous incomplete variable by varying $\rho$, where dashed lines depict the target performance criteria value.

Regarding the width of the 95% confidence interval (CI) of the $\beta^O$ parameters, we observed that using the 2l.2stage.heckman method we obtained the widest CIs of $\beta^O$ among the methods evaluated. The width of these CI generally increased as $\rho$ moved away from zero.

*Bivariate incomplete variable*

In the case of an incomplete bivariate outcome, we also found (Figure 3) that the 2l.2stage.heckman method provided the least unbiased results on the coefficients $\beta^O$ with the best coverage values. As for the random effect parameters, we observed unbiased estimates especially $\sqrt{\psi_{22}^O}$ at large values of $\rho$.

### 4.8.2 | Sensitivity analysis: Number of clusters and sample size of clusters

We assessed the robustness of our method to variations in the number of clusters and also in the cluster sample size (Figure 4).

By increasing N, the number of clusters, from 10 to 100, we observed that the bias was not affected, but the 95%CI width decreased (higher precision) and the RMSE decreased as well. On the other hand, as we reduced the number of units per cluster (from $n_i$=1000 to $n_i$=100) precision decreased for all coefficients, the bias of the $\beta^O$ coefficient estimates was not drastically affected, but the bias of the $\sqrt{\psi_{kk}^O}$ random effects parameters was.

When we reduced the sample size to 50 patients per study, bias and RMSE of the $sqrt\psi_{22}^O$ were drastically affected (not shown here, but in the Appendix). This could be partly explained by the sparse information on certain clusters, which directly affects the estimation of the Heckman model on these clusters.

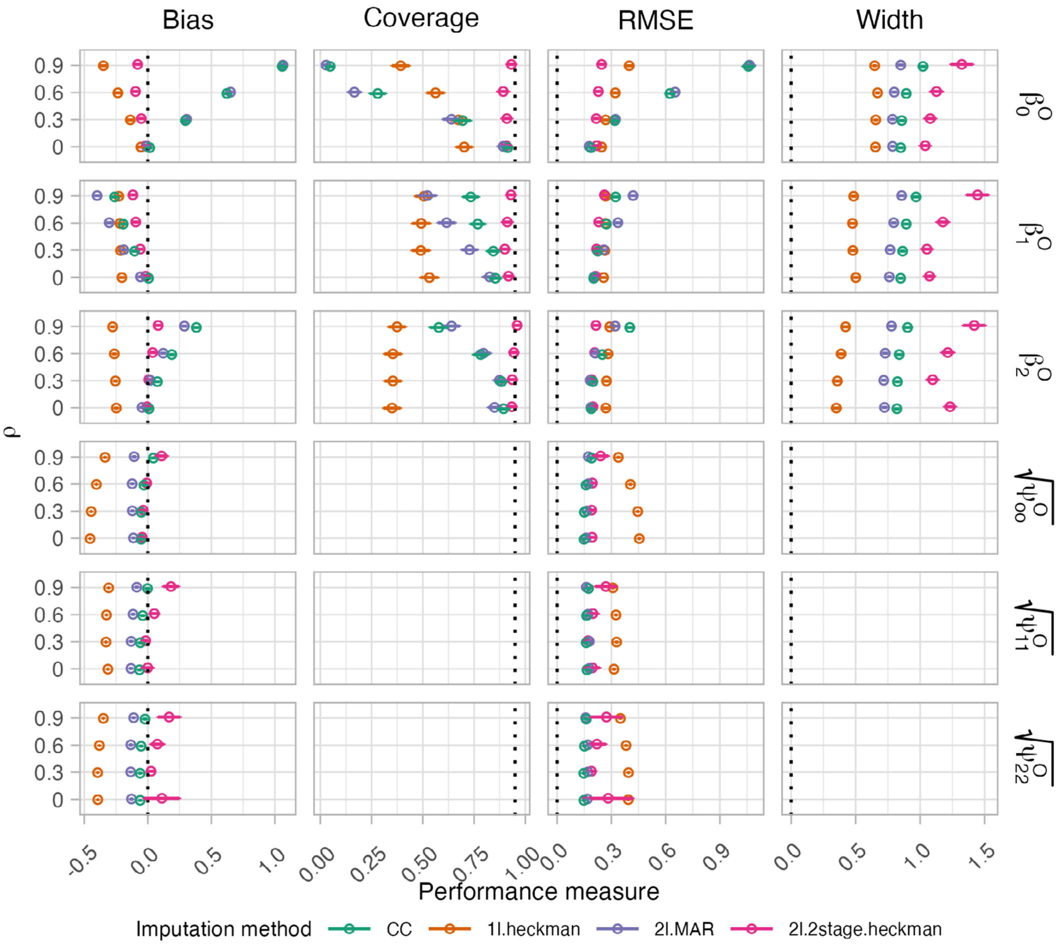

**FIGURE 3** Binary incomplete variable by varying $\rho$, where dashed lines depict the target performance criteria value.

### 4.8.3 | Sensitivity analysis: Distributional assumptions

We also assessed the performance of our method when faced with deviations in distribution assumptions (Figure 5). When using the Heckman model in an self-masking MNAR process, that is, when the probability that an observation is missing is associated with the value of the observation, we observe that under all imputation methods we obtain biased estimators, with poor coverage. In particular, when estimating the parameter intercept $\beta_0^O$ we observe more biased estimates compared to the estimates of the other coefficients, with a coverage of less than 60

It is possible that our model is not entirely suitable for this type of scenario, especially if the main analysis is focused on estimating absolute prevalence estimates, as it could be greatly affected by the bias of the ($\beta_0^O$) parameter. Also the imputation method was strongly affected in terms of the bias in the $\sqrt{\psi_{22}^O}$ parameter.

With respect to the Skewed-t scenario, only the bias of $\beta_0^O$ was affected for all imputation methods applied. But the estimates of $\beta_0^O$ for both 1l.heckman and 2l.2stage.heckman were drastically affected in terms of bias, with a very poor coverage (below 25%). Therefore, the applicability of our method might be questionable in scenarios that do not conform to the BNV assumption.

Please note that our approach relies on specific assumptions, and any deviation from these assumptions may have an impact on its performance. Therefore, it is advisable for users to carefully assess these assumptions and determine whether they align with a plausible missing mechanism in their data.

As we employ a Gaussian copula function to estimate the Heckman model, it is important to recognize that deviations from normality assumptions could lead to biased and less accurate estimates.[34] In certain situations, it is recommended to compare the results with those obtained using alternative imputation methods, such as predicted mean matching (PMM), which can be especially beneficial to prevent the generation of imputation values that fall outside the range of observable

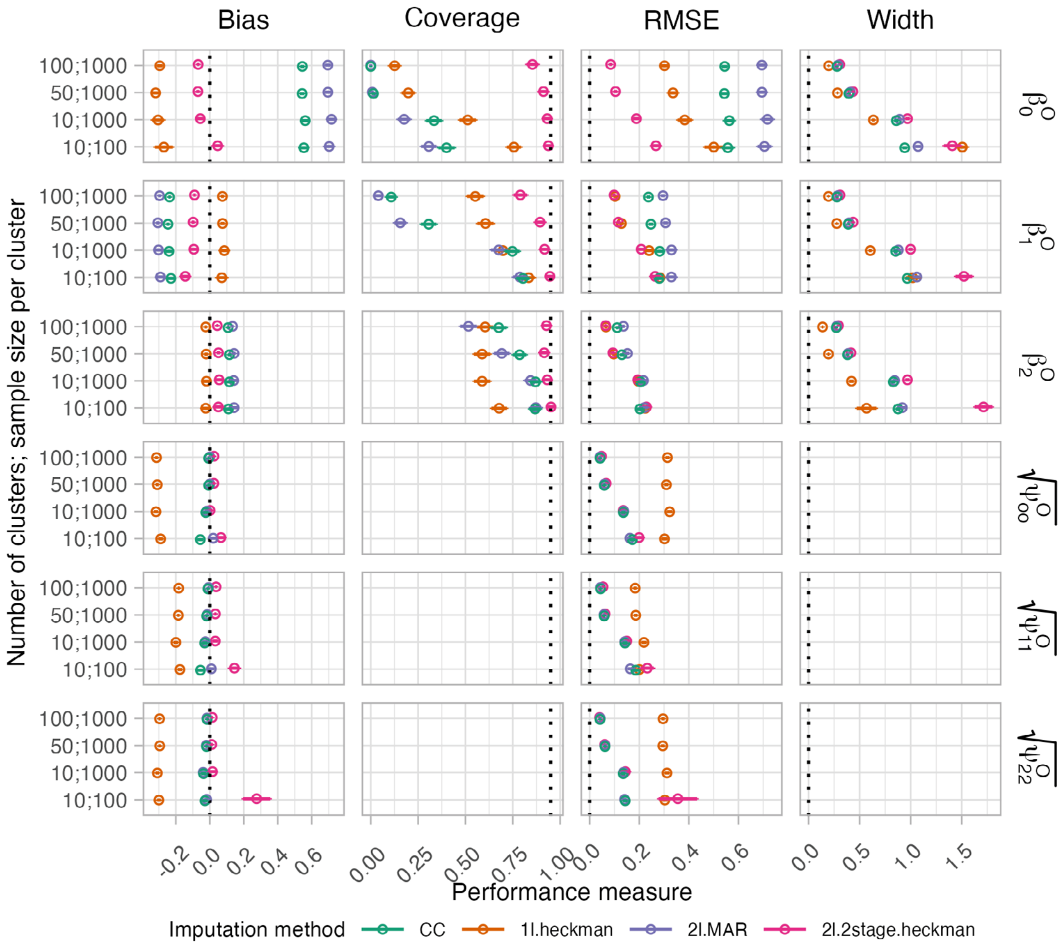

FIGURE 4 Continuous incomplete variable under systematic missingness by varying: number of clusters (N); sample size per cluster ($n_i$), where dashed lines depict the target performance criteria value.

values. Alternatively, one may explore the possibility of refining our method by utilizing a copula function that better fits the error dependence distribution of the data.

## 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 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.[35]

For this example, we used data accessed directly from ClinEpiDB,[36] where data were collected from 5195 households with verified consent, inhabited by 11 137 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 3, which were used

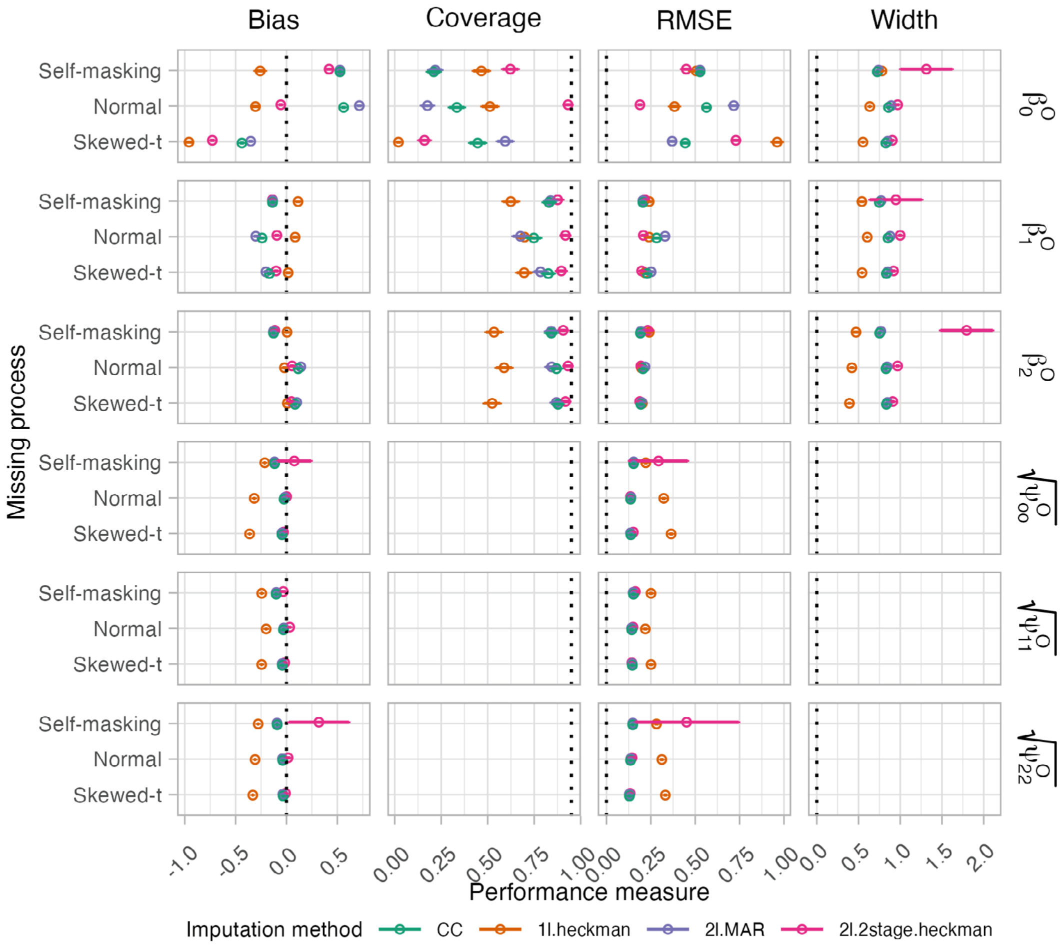

**FIGURE 5** Continuous incomplete variable with deviations in distribution assumptions, where dashed lines depict the target performance criteria value.

**TABLE 3** Descriptive analysis, predictor variables.

| 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 |

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,[37] we estimated the prevalence of parasitemia by region 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, that is, 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 two or fewer persons share a mosquito bed net). In addition, we included age as a cubic 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.2stage.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 ($r$), conditioned on the remaining imputation predictors. The model results in Table 4 indicate that the holiday indicator could be a plausible ERV variable, as there was strong evidence of an association with $r$, but no evidence of an association with $y$.

TABLE 4 Evaluation of holidays as exclusion restriction variable.

| Predictors/response | Test result ($y$) | Test taken ($r$) |
|---|---|---|
| (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*: ***$p < 0.01$. s(Age) denotes the cubic spline of Age.

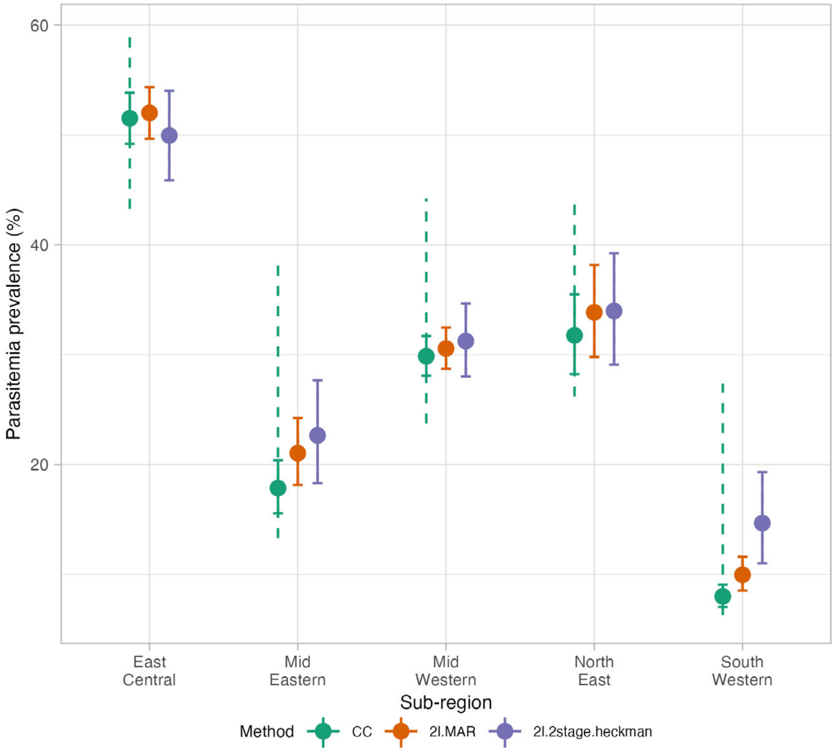

FIGURE 6 Estimates of prevalence of malaria parasitemia by region. The ranges of plausible prevalence values are represented by dashed lines behind the CC method.

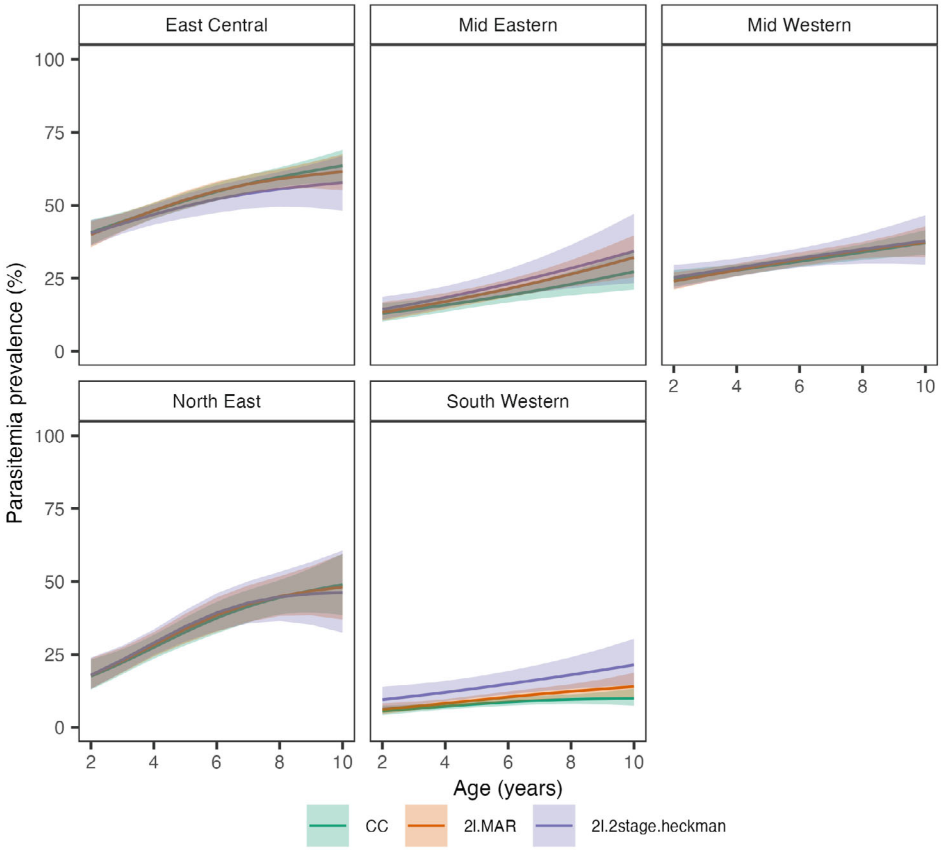

FIGURE 7 Estimates of prevalence of malaria parasitemia by region and age.

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 analysed. As can be seen in Figure 6, for each region the prevalence estimates of the approaches did not differ significantly between methods. However, prevalence estimates under the MNAR assumption (ie, 2 level Heckman) were higher than those estimated under the MAR or MCAR approaches, 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 old compared to school-age children, that is, 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 South-Western 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.[38] 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.

## 6 DISCUSSION

We have extended and evaluated methods for multiple imputation of clustered datasets, in situations where some incomplete variables follow a MNAR mechanism. Although there are imputation methods that address incomplete MNAR variables, these can only deal with the case of individual studies. This limits their use in common situations 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 incomplete continuous and binary MNAR variables with sporadic as well as systematic missingness, applicable for the case of clustered datasets with heterogeneous effects and error variances.

In a simulation study, we found that our proposed imputation method is well suited for the imputation of both continuous and binary incomplete variables following a MNAR mechanism indirectly non-ignorable. The simulation study showed that the application of the proposed method on clustered datasets with heterogeneous effects and error variances and with sporadically and systemically missing variables resulted in less biased estimates of effects with a convergence close to 95% compared to those obtained using other imputation methods evaluated.

Our method may yield better results than Heckman-based imputation methods for individual studies, as it not only enables the imputation of missing values within clusters exhibiting systematic missingness but also has the capacity to adjust values within individual clusters, aligning them more closely with the overall study mean. This can be especially advantageous in studies with small sample sizes, where an analysis approach that ignores data from other studies may lead to extreme effect estimates.

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 Heckman's model. That is, under the specification of a valid exclusion variable the method determines which is the most adjustable inter-equation correlation parameter ($\rho$), or in general terms the missingness mechanism (MAR or MNAR), in each of the clusters evaluated.

Finally, our imputation method was built according to the specifications of the R mice package and is available in the micemd package. This allows the method to be easily and simultaneously used with other imputation methods implemented for the MICE package, which is advantageous in databases containing several incomplete variables that require different imputation methods and imputation models.

## 7 | 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 marginal estimates, the method only uses clusters with observable information, that is, 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.

Also, the method can be sensitive to both the sample size of the individual studies, as well as 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 the sample size required to estimate the parameters of the Heckman model can be at least twice the number of parameters required to estimate in an imputation model that assumes MAR. On the other hand, a large number of studies, which may improve precision of the estimations may also make the estimation of the marginal parameters more difficult and may 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 which can considerably affect the performance of our method. Therefore, the effect of this parameter could be further evaluated in future research. One might also consider relaxing the assumption of constant correlation to allow for a random effects distribution of the correlation parameter.

Further, the method can also be extended to other copula models for non-random selection, with different distributions of the selection and outcome equations and dependency structure.[19] 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 and Collins.[11]

## 8 | 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 estimated parameters were fairly robust in terms of bias, but the intercept was not.

## ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under ReCoDID grant agreement No 825746. In addition, we are grateful for the support of Professor Matthias Egger in conceptualizing this article and helping us to find a suitable data set for the illustrative study. We would like to thank the Associate Editor, and the reviewer for their helpful comments that have substantially improved the manuscript.

## 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.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Clinical Epidemiology Database, ClinEpiDB. Data are from Sarah Staedke, Martin Donnelly, Janet Hemingway, et al. Dataset: LLINEUP Cluster Randomized Trial. ClinEpiDB. 05 November 2020.Release 14 (https://clinepidb.org/ce/app/workspace/analyses/DS_7c4cd6bba9/new) and can be downloaded by the registered user immediately after submission of the request.

## ORCID

*Johanna Muñoz* https://orcid.org/0000-0002-2384-5415
*Orestis Efthimiou* https://orcid.org/0000-0002-0955-7572
*Vincent Audigier* https://orcid.org/0000-0002-4169-7866
*Valentijn M. T. de Jong* https://orcid.org/0000-0001-9921-3468
*Thomas P. A. Debray* https://orcid.org/0000-0002-1790-2719

## REFERENCES

1. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol*. 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. *Eur J Epidemiol*. 2007;22(12):839-869. doi:10.1007/s10654-007-9165-7
3. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med*. 2013;32(28):4890-4905. doi:10.1002/sim.5894
4. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581
5. Liu D, Oberman HI, Muñoz J, Hoogland J, Debray TPA. Quality control, data cleaning. 2021.
6. Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*. 1996;52(1):98. doi:10.2307/2533148
7. Vella F. Estimating models with sample selection bias: a survey. *J Hum Resour*. 1998;33(1):127. doi:10.2307/146317
8. 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*, NBER; 1976;5:475-492.
9. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat Med*. 2016;35(17):2907-2920. doi:10.1002/sim.6902
10. Galimard JE, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Med Res Methodol*. 2018;18(1):90. doi:10.1186/s12874-018-0547-1
11. Ogundimu EO, Collins GS. A robust imputation method for missing responses and covariates in sample selection models. *Stat Methods Med Res*. 2019;28(1):102-116. doi:10.1177/0962280217715663
12. Audigier V, White IR, Jolani S, et al. Multiple imputation for multilevel data with continuous and binary variables. *Stat Sci*. 2018;33(2):160-183. doi:10.1214/18-STS646
13. Hammon A, Zinn S. Multiple imputation of binary multilevel missing not at random data. *J R Stat Soc Ser C Appl Stat*. 2020;69(3):547-564. doi:10.1111/rssc.12401
14. Puhani PA. *Foul or Fair? The Heckman Correction for Sample Selection and its Critique. A Short SurveyTech*. Rep. 97-07. Leibniz, Germany: ZEW—Leibniz Centre for European Economic Research; 1997.

15. Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. *J Econom Perspect*. 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. *Stat Med*. 2020;39(11):1658-1674. doi:10.1002/sim.8504
17. Amemiya T. Tobit models: a survey. *J Econ*. 1984;24(1):3-61. doi:10.1016/0304-4076(84)90074-5
18. Smith MD. Modelling sample selection using Archimedean copulas. *Econom J*. 2003;6(1):99-123. doi:10.1111/1368-423X.00101
19. Wojtyś M, Marra G, Radice R. Copula based generalized additive models for location, scale and shape with non-random sample selection. *Computat Stat Data Anal*. 2018;127:1-14. doi:10.1016/j.csda.2018.05.001
20. 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. *Clin Trials*. 2005;2(3):209-217. doi:10.1191/1740774505cn087oa
21. Higgins JPT, Thompson SG, Spiegelhalter DJ. A Re-evaluation of random-effects meta-analysis. *J R Stat Soc A Stat Soc*. 2009;172(1):137-159. doi:10.1111/j.1467-985X.2008.00552.x
22. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649. doi:10.1177/0962280216666564
23. Greene WH. *Econometric Analysis*. 8th ed. New York, NY: Pearson; 2018.
24. Buuren VS, Groothuis-Oudshoorn K, Vink G, et al. Mice: Multivariate Imputation by Chained Equations. https://CRAN.R-project.org/package=mice, 2021.
25. Audigier V, Resche-Rigon M. micemd: Multiple Imputation by Chained Equations with Multilevel Data. https://CRAN.R-project.org/package=micemd, 2022.
26. Radice GMaR. GJRM: Generalised Joint Regression Modelling. 2021, https://CRAN.R-project.org/package=GJRM.
27. Gasparrini A, Sera F. Mixmeta: An extended mixed-effects framework for meta-analysis. https://CRAN.R-project.org/package=Mixmeta 2021.
28. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293. doi:10.3102/10769986030003261
29. Bates D, Maechler M, Bolkeraut 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*. Hoboken, NJ: John Wiley & Sons, Inc.; 1987.
31. Buuren vS. *Flexible Imputation of Missing Data. Chapman and Hall/CRC Interdisciplinary Statistics SeriesBoca*. 2nd ed. Raton: CRC Press, Taylor and Francis Group; 2018.
32. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
33. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.
34. Ogundimu EO, Hutton JL. A sample selection model with skew-normal distribution: modelling sample selection and skewness. *Scand J Stat*. 2016;43(1):172-190. doi:10.1111/sjos.12171
35. 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
36. Staedke S, Donnelly M, Hemingway J, Kamya M, Dorsey G. ClinEpiDB.Study: LLINEUP cluster randomized trial. https://clinepidb.org/ce/app/workspace/analyses/DS_7c4cd6bba9/new/details 2021.
37. 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. *Malar J*. 2019;18(1):207. doi:10.1186/s12936-019-2838-3
38. Program TD. *DHS Survey Design: Malaria Parasitemia* tech. rep. Washington, D.C.: U.S. Agency for International Development (USAID); 2020.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Muñoz J, Efthimiou O, Audigier V, de Jong VMT, Debray TPA. Multiple imputation of incomplete multilevel data using Heckman selection models. *Statistics in Medicine*. 2024;43(3):514-533. doi: 10.1002/sim.9965