Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Multiple imputation of incomplete multilevel data using Heckman selection models

J. Munoz* | V. de Jong | T. Debray

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

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

Present Address
Julius Center for Health Sciences and Primary Care, Utrecht Medical Center, Universiteitsweg 100, 3584 CG Utrech, 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 not justified when observations are clustered (e.g., within studies) or when the presence of missingness depends on unobserved information. Unfortunately, this situation is increasingly common, and typically arises when individual participant data (IPD) are combined from multiple studies. While several imputation methods have been proposed for addressing individual studies where data are missing not at random (MNAR), their application and validity in large datasets with clustering remains unclear. We therefore explored the consequence of MNAR data in IPDMA in-depth, and proposed novel multilevel imputation methods for common missing patterns in clustered datasets. These methods build upon the principles of Heckman selection models, and adopt a two-stage meta-analysis approach for imputing binary and continuous variables. After evaluating the proposed imputation models in simulated scenarios, we illustrated their use in a real IPDMA based on a malaria study in five subregions in Uganda, where we focused on estimating the prevalence of parasitemia for children aged 2-10 years.

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 large combined datasets derived from electronic health records or from individual participant data (IPD). For example, the clinical practice research datalink (CRPD)[1] is an electronic health record (EHR) dataset in the UK, and 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 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, and are therefore more alike than individuals from another cluster. Sometimes, clusters may also differ in participant eligibility criteria, follow-up length, predictor and outcome definitions, or in the quality of applied measurement methods. Hence, correlation is likely to be present between observations from the same cluster, which can lead to differences or 'heterogeneity' between clusters regarding baseline patient characteristics and subsequent outcomes.

Clustered datasets often contain many incomplete variables. For example, in registry data it is common that test results are not available in registry data 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 meta-analysis, it is common that studies 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 across clusters, 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 non-random; MNAR). Traditional imputation methods are designed to address incomplete data sets where variables are MCAR or MAR. Their implementation is reasonable when there is no obvious mechanism of missingness, or when the observed data strongly relate to unobserved information.

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 that are based on 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 (partially) depend on unobserved information.

One could see, a missing variable is a mixture of MAR and MNAR, depending on the available information, making the MAR mechanism more likely as more information is recorded. For instance, healthcare professionals may decide what type of measurement on indication, and depending on their personal experience or behavior decide whether or not to record the reasons for the measurement.

Registries are notoriously prone to incomplete variables that are MNAR, due to the complex recording process[6], e.g. laboratory tests are taken only in certain patients based on sign and symptom information that is often incomplete or not recorded. Also IPD may suffer from MNAR, for example when study participants who experience unfavorable results drop out of the study or also due to heterogeneity of the primary objective or resources of the studies involved, which may result in variables relevant to explain the missing process not being recorded.

The MNAR mechanism is considered as non-ignorable because to deal with this type of missingness it is necessary to specify information about the missingness process in addition to assumptions about the observed data. A common approach to address MNAR is to adopt 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 main 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 its implementation does not require data to be MNAR, and can also be used when data are MCAR or MAR. It therefore offers an appealing solution to incomplete data sets where the missingness mechanism is unknown.

Over the past few years, several Heckman selection models 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 needed 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 assume that the effect of explanatory variables on the missingness mechanisms is common across clusters.

Therefore, the aim of this paper is to develop a multilevel imputation method for incomplete continuous and binary variables that are MNAR. In section 2 we provide an introduction to the Heckman model and its estimation, and we extended to a hierarchical setting. Then in the section 3 we define the main steps of the proposed imputation method. Posteriorly, in the section 4 we provide the settings and results of a Montecarlo simulation study used to evaluate the performance of the proposed impuatation 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 we provide a discussion in section 6 about the results, 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 selection of a subject or his observation into the sample is influenced by unobserved variables (e.g., the respondent's level of trust toward healthcare entities may cause him or her to self-select out of the study or refuse to sign consent for a test), which in turn are correlated with unobserved variables related to a variable of interest (e.g., the result of a blood test)[13].

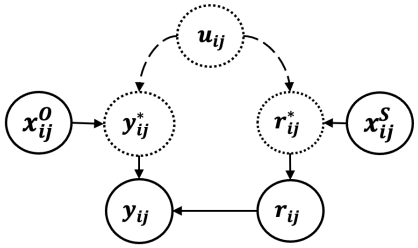


FIGURE 1 DAG Heckman selection model: here the nodes (dotted = latent, continuous = observable) describe the relationship between y^*_{ij} the latent response and r^*_{ij} the latent selection variables

This can be better visualized in Figure 1, where for the jth individual or unit within the ith 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 x^O_{ij} and x^S_{ij} sets of predictor covariates. 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 $y^*_i = (y^*_{i1}, y^*_{i2}, ..., y^*_{in_i})^T$ and $r^*_i = (r^*_{i1}, r^*_{i2}, ..., r^*_{in_i})^T$ the vectors of latent outcomes and latent selections in the i cluster, then Heckman's model is defined by two main equations: the outcome equation (1), which describes the relation latent outcome-exposure, and the selection equation (2) which details the likelihood that the outcome is observed in the sample.

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

$$r^*_i = X^S_{ij} \beta^S_i + \epsilon^S_i \tag{2}$$

Here β^O_i and β^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, and should only be associated with the selection in the sample r^*_{ij} but not with the actual observation y^*_{ij}.

In the presence of selection bias, aforementioned outcome equation will yield biased estimates of β^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),

$$\begin{pmatrix} \epsilon^O_{ij} \\ \epsilon^S_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_i & \rho_i \sigma_i \\ \rho_i \sigma_i & 1 \end{pmatrix} \right)$$

where σ_i corresponds to the variance of the error in the outcome equation and ρ_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 ρ_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)[16]. However, both methods can lead to inconsistent estimators when this underlying distribution cannot be assumed[17].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 copula, that joins the marginal distribution of the error terms of the selection and outcome equation which are specified separately.

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

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 many data could not follow a BVN distribution. This facilitates the Heckman model estimation and makes it more robust to deviations from the assumptions regarding the distribution of the data.

2.2 | Hierarchical model

The Heckman model can be extended to hierarchical settings, i.e., in data where individuals or sampling units are nested within groups, as is the case in EHR or IPD. In this case, sample units from the same group are expected to share some characteristics (e.g., distribution of variables, relationships between exposure and outcomes, missing processes, missing mechanisms) given only the fact that they belong to the same group.

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 observation process, but also in the missing data process, thus requiring imputation models that are congenial to the analysis model, i.e., that have the same assumptions about the data.

Different procedures can be adopted to combine information between groups; however, in our imputation method we opted for the two-stage approach that is often used in meta-analyses. This is because such an approach is less computationally 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 θ_i the cluster specific parameters of the Heckman model,i.e., the parameters of the two equations in each study: the outcome model and the selection model, for each of the N clusters are estimated separately and then, in a second stage, all θ_i are combined using a random effects model.

In a random effects model, θ_i parameters are assumed to be drawn independently and identically from an imaginary distribution of parameters with a θ_m population mean and a ψ population variance[19]. Thus, the $\theta_i = \theta_m + b_i$ can be specified 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 | METHOD

We follow a similar approach proposed by Resche-Rigon and White (2018)[20] for multilevel imputation of data. Briefly, his method was developed to impute variables from a hierarchical structure (i.e., when there are unit samples grouped within a cluster or group). 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 group, the imputation values are drawn from a (generalized) linear model conditional on θ_m the marginal population parameters,

i.e., those estimated after pooling the cluster-specific parameters θ_i.On the other hand, when the variable is sporadically missing within the group, 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 θ_i towards θ_m.

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 θ_m parameter set at the marginal or population level. Our method is basically a univariate imputation method, but since it is implemented in a Gibbs sampling procedure, it can also be used to impute multiple incomplete variables in a data set.

3.1 | Imputation of univariate incomplete dataset

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 θ. 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 θ parameter vector, θ^*, 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 θ^* vector.

Below we describe each step in depth:

3.1.1 | Draw the θ^* parameter vector

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

Initially, we estimate by the copula method 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 y_{ij}^{obs} observable measurements within each i group . The Heckman model is estimated with the **gjrm** function of the GJRM R package 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 groups with observable information, i.e., those that are not systematically missing and have sufficient information to estimate the heckman model. In particular, we pooled the p coefficients of the β^O outcome equation and estimated a multivariate random effects meta-analysis model with them, similarly we combined all q coefficient parameters of the β^S selection equation. We also performed a univariate random effects meta_analysis on σ' the log-transformed parameter of σ and another on ρ' the fisher-transformed parameter of ρ.

The meta-analysis model is performed with the **mixmeta** function of the R package mixmeta, which allows the use of maximum likelihood (ML), restricted maximum likelihood (REML) and moments estimation methods. For the simulation and illustrative study, we used the restricted REML estimation method, which is recommended as it has a good balance between insensitivity and efficiency[21].

Draw θ_m^* the marginal parameters

From the meta-analysis model, we obtain the marginal estimates $\widehat{\theta_m}$ and the between-cluster variance $\widehat{\psi}$ with their corresponding variance-covariance matrices $\widehat{S_{\theta_m}}$ and $\widehat{S_\psi}$, which are used to draw the θ_m^* and ψ^* parameters as follows:

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

Draw θ_i^* the cluster parameters

We draw θ_i^* shrunk-cluster parameters for each group i from the following posterior distribution conditional on θ_m^* and ψ^*.

$$\theta_i^* \sim N\left(\frac{\theta_m^*/\psi^* + \widehat{\theta}_i/\widehat{S_{\theta_i}}}{1/\psi^* + 1/\widehat{S_{\theta_i}}}, \frac{1}{1/\psi^* + 1/\widehat{S_{\theta_i}}}\right)$$

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 $\widehat{\theta}_i$ cluster-specific and the θ_m^* marginal parameters are inversely proportional to the $\widehat{S_{\theta_i}}$ within and ψ^* between clusters variances. 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. In the case of systematic missingness, it is like considering the within-cluster variance to be infinite ($\widehat{S_{\theta_i}} \to \infty$), then letting all parameters rely only on the marginal estimates.

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

Having θ_i^* the shrunk-cluster parameters vector for each group, we back transform σ^* and ρ^* 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 θ_i^* as follows:

Continuous missing variable

The imputed value of the y_{ij}^{miss} missing observation can be drawn from the conditional expectation of y_{ij} on unobserved measurements:

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

Binary missing variable

The missing y_{ij}^{miss} is drawn from a Bernoulli distribution with p_{ij}^* proportion parameter given by $P[y_{ij} = 1|r_{ij} = 0]$, the conditional probability that $y_{ij} = 1$ given that the measure is unobservable ($r_{ij} = 0$), in a bivariate probit model[22]:

$$\begin{aligned}
p_{ij}^* &= P[y_{ij} = 1|r_{ij} = 0, \beta_i^{O^*}, \beta_i^{S^*}, {\rho_i}^*] \\
p_{ij}^* &= \frac{\Phi_2(x_{ij}^O \beta_i^{O^*}, -x_{ij}^S \beta_i^{S^*}, -\rho_i)}{\Phi(-x_{ij}^S \beta_i^{S^*})} \\
y_{ij}^{miss} &\sim Ber(p_{ij}^*)
\end{aligned}$$

3.2 | Imputation of multivariate incomplete dataset

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 , that allows imputing multiple incomplete predictors and covariates in a given dataset.

The MICE package allows to specify imputation methods to each of the missing variables by setting the method and the predictive matrix for each of the missing variables. To use our method, it is necessary to specify for the MNAR missing variable the method **2l.heckman** in the mice methods vector. Furthermore, in the prediction matrix, the group or cluster variable should be specified as **'-2'**, all predictor variables belonging to the selection and outcome as **'1'**, the exclusion restrictions or predictor variables that are only included in the selection equation as **'-3'** and those that are only included in the outcome equation as **'-4'**. Please refer to the toy example in the attached github repository to better understand how to implement the imputation model.

4 | SIMULATION STUDY

Aim

We design a simulation study aimed to compare the performance of imputation methods when they impute a missing variable that comes from a hierarchical dataset and follows a MNAR mechanism.

Data-generation mechanism

We generated the data from the Heckman selection model with bivariate normal distribution error terms. For simplicity we assume that the database collects information from $N = 10$ clusters of equal number of individuals $n_i = 1000$. 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 $\mu_h \sim N(\binom{0}{0}, \left(\begin{smallmatrix} 0.2 & 0.015 \\ 0.015 & 0.2 \end{smallmatrix}\right))$, with $h = \{2, 3\}$. We then simulated for each cluster a baseline covariate $X_{2i} \sim N(\mu_2, 1)$ and a exclusion restriction $X_{3i} \sim N(\mu_3, 0.5^2)$.

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

$$y_i^* = \beta_i^O X_i^O + \epsilon_i^O$$
$$r_i^* = \beta_i^S X_i^O + \epsilon_i^S$$

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

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

We used these fixed coefficients $\beta_h^O = \{2, 1, 1\}$ and $\beta_h^S = \{0.1, 1.5, -0.7, 1.8\}$ in order to get around 40% of sporadically missing values on the response y_{ij} in the entire data set. Additionally, we made that the y_{ij}^* observations were systematically missing in 20% of the clusters included in data set. 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 $\sigma_{b0}^2 = \sigma_{b1}^2 = \sigma_{b2}^2 = 0.5$. We considered that the correlation parameter of the random effects between equations is a 40% of the value of the assumed correlation parameter between error terms ρ. In addition, we included a random effect on the exclusion restriction variable given by $b_3 \sim N(0, 0.3)$ assuming that the intracluster variation in the exclusion restriction effect is lower than the variation on other coefficient parameters effects. The ρ parameter adopts different values depending on the simulated missing mechanism (See below additional scenarios) .

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 σ_i^2 is variable across clusters and distributed as $log(\sigma_i) \sim N(log(0.7), 0.05)$.

Adittional scenarios To investigate the performance of the imputation methods under the following scenarios:

- **M(N)AR scenarios:** We assessed whether the model performed well in terms of bias and coverage when the data followed a missing MAR mechanism ($\rho = 0$), and when it followed a MNAR mechanism with a low ($\rho = 0.3$), intermediate ($\rho = 0.6$) and strong correlation ($\rho = 0.9$) between y^* and r^*.
- **Influence of sample size**: Model sensitivity was analysed with respect to $n_i = \{50, 100, 1000\}$ the number of patients per cluster and $N = \{10, 50, 100\}$ the number of studies. We consider the
- **Violation of distributional assumptions:** To assess how imputation models behave in the face of deviations from normality assumptions, we simulated data in which the errors followed a skewed t-distribution and also in which the missing process follows a MNAR mechanism with an explicit truncated model, i.e., the participation is directly related to the value of the outcome variable.

- **Binary response**: We evaluated the imputation method when missing variables is binary. Therefore we simulated y_i binary incomplete variables, the marginal set of parameters were fixed to $B_h^O = \{0.5, 1, 1\}$ and $B_h^S = \{0.1, 1.5, -0.7, 1.2\}$, keeping the other parameters similar to the ones used in the simulation of missing continuous variables. The observable binary variables was defined as:

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

Estimand

The estimads 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 β_1^0. We also report in the estimated variance of the random effects and residual errors σ_{b0}^2,σ_{b1}^2,σ_{b2}^2,σ_e^2.

After the imputation procedure, we estimated the following (generalized) mixed linear effect model using the lmer() function. $y_i = \beta_i^O X_i^O + \epsilon_i^O$ In case of missing binary variable, we used the same matrix of predictors but on a binary model estimated with glmer() functions. Then, we pooled the estimates of the β_i^O and the variance of random effect and residual errors of the multiple imputed datasets according to Rubin's rule[23], over which we calculated the performance measures on estimands.

For coverage estimation, we obtain the confidence intervals from the Wald method. For the random effect variance, even when it is possible to estimate the confidence intervals through the profile or bootstrap method, we did not estimated them for computational reasons.

Method

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 proposed by Resche-Rignon and (2018)[20].
- **2l.Heckman**: The proposed imputation method based on the Heckman model for hierarchical datasets.

Performance measures

We calculated the following evaluation criteria according to the formulas provided in Morris et al.(2019)[24]:

- **Bias:** Bias on the coefficient and random effect parameters.
- **EmpSE:** Empirical standard errors of the estimates on the coefficient and random effect parameters.
- **ModSE:** Monte Carlo standard errors on the coefficient parameters.
- **MSE:** Mean squared error of the coefficient and random effect parameters.
- **Coverage(%):** Coverage of the 95% confidence intervals for the coefficient parameters. In addition we reported the average time processing and the percentage of datasets where the imputation method converged, i.e., the imputation method generated an output.

Software

For simulation study and illustrative examples we used R version 4.0.4 in a linux environment. The Heckman 2L imputation method is available in mice R package (**mice.2l.heckman**) and also on the github repository https://github.com/johamunoz/Heckman2l where you can also find all the codes accompanying this paper and a toy example that explains how to implement the method in mice.

Results

Descriptive results We generated data sets of 10 groups of 1000 patients each in each scenario. For example, for the scenario in which the error terms followed a normal distribution, out of the 500 datasets generated, we obtained that on average 48.2% 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 Y.

–> –> –>

5 | AN ILUSTRATIVE STUDY

Malaria is a mosquito-borne disease and, especially in children and pregnant women, is the leading cause of illness and death in Africa. 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, and 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 were 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 parasitaemia and an etymological study was conducted to estimate mosquito prevalence, details of the project and survey are provided elsewhere[25].

For this example, we used data accessed directly from CliniEpiDB[26], 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 the table 1, which were used as predictors of the imputation model.

Subregion	Districts (N)	Children (N)	Age (years)	Log10 Female Anopheline	Wealth index	Bednet (%)	Females (%)	Holidays (%)	Missing test(%)
North_East	5	794	5.50	2.67[1.5,4.3]	-0.45[-1.2,2.2]	10.70	49.00	31.90	17.50
Mid_Eastern	8	1354	5.61	0.84[0.1,2.5]	-0.14[-1,2.5]	9.30	48.10	32.90	25.60
South_Western	14	3596	5.69	0.27[0.1,1.3]	0.18[-1,2.9]	23.80	49.40	66.50	21.10
Mid_Western	12	3172	5.66	1.27[0.1,3.2]	-0.03[-1,2.8]	13.30	48.90	62.90	20.50
East_Central	9	2152	5.61	2.74[0.4,6.3]	0.01[-1.1,3.1]	13.20	51.60	51.60	16.00

TABLE 1 Descriptive analysis, predictor variables

To illustrate our method, following the article by **(author?)**[27], 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: MCAR, MAR and MNAR.

In the MCAR approach, prevalence was calculated on the basis of the actual test, i.e., we only included patients with a test result. In the case of MAR, 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, two-person mosquito net. In addition, age was included as a power 3 spline function, the cluster-level Log10 mean of the number of female anopheline mosquitoes per household estimated from the etymological 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 used under the MAR approach. In addition, we included a holiday indicator variable as ERV, which 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), conditionated on the remaining imputation predictors. The model results in Table 2 indicate that the holiday indicator could be a plausible ERV variable, as it was significantly associated with ry, but not with y.

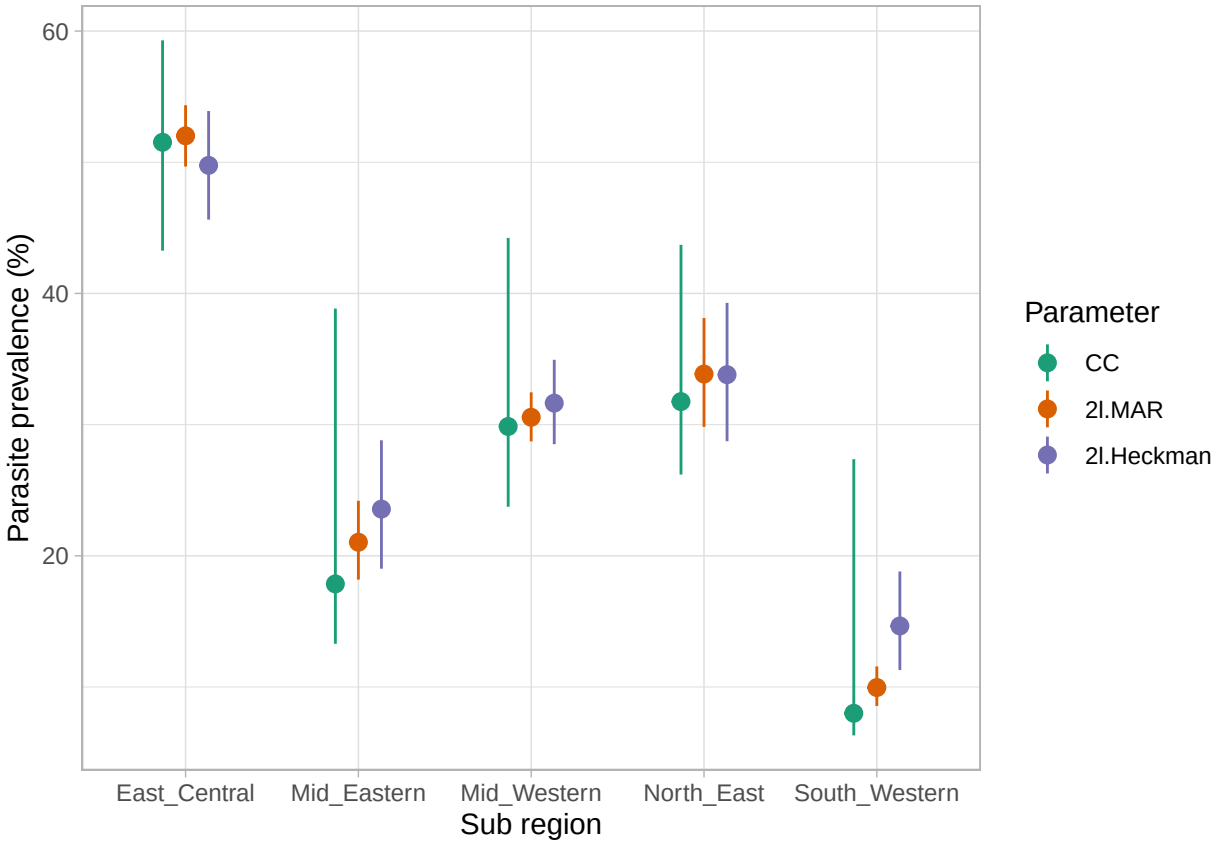
According to our imputation approach, non-participants were estimated to have a higher prevalence of malaria than participants in more than half of the districts analyzed. As can be seen in Figure **??** in terms of subregion level prevalence the estimates of the approaches do not differ significantly between methods, however except for the East-Central

	Test result (y)	Missing indicator (ry)
(Intercept)	-1.074(0.051)***	1.353(0.046)***
Female Anopheline/house (Log10 mean)	0.73(0.026)***	0.151(0.021)***
Household wealth index	-0.551(0.039)***	0.04(0.026)
Bednet for 2 person-Yes	-0.298(0.082)***	0.704(0.08)***
Sample taken in holidays-Yes	-0.044(0.054)	0.19(0.05)***
Sex-Male	0.103(0.054)	0.054(0.05)
s(Age)	1.839(1.974)***	1.019(1.037)***

***p<0.01

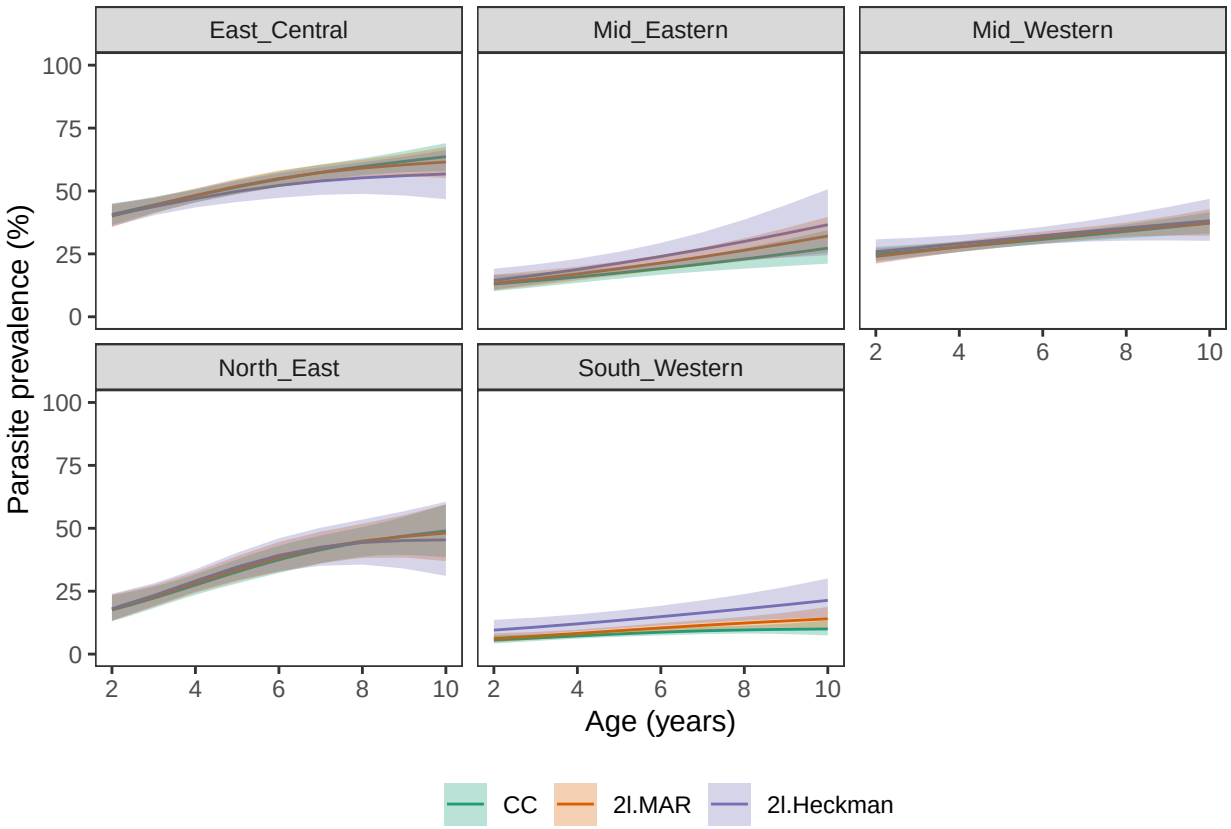
TABLE 2 Evaluation of Holidays as exclusion restriction variable

region, prevalence estimates under the Heckman approach are higher than those estimated using MAR or CC aproaches.



In terms of prevalences by age, there are no significant differences between methods overall. The prevalence estimates for children aged 2 to 6 years according to the approaches evaluated in all regions are very similar (Figure

??, which could be partly explained by the mobility of children at this age compared to that of school-age 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. Possible reasons for selection bias in surveys of this type have been suggested[28], 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. Nevertheless, we were unable to find information that suggests or confirms 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

This work was done in order to extend multiple imputation for clustered datasets, in situations where some incomplete variables follow a MNAR mechanism. For clustered datasets, only imputation methods under the MAR mechanism have been previously proposed and although imputation methods exist to handle MNAR they have only been designed for individual studies, which makes them limited in common IPD situations such as systematic missingness or when the proportion of missingness of a variable is very high in one of the included studies. In this context, a new multiple imputation method was proposed to handle continuous and binary MNAR covariates specifically for clustered dataset, which also allows appropriate borrowing of information between the clusters to obtain more reliable imputation results at the individual cluster level.

From the results of the simulations we can observe that the imputation method we propose can be valid for the imputation of continuous and binary type missing variables that follow a MNAR mechanism according to the Heckman model and that come from multilevel data such as those used in the IPDMA studies.

Overall the method produced unbiased estimates with convergence close to 95% for the fixed effects parameters with variation at the cluster level and also unbiased estimators for the random effects parameters.

Empirically, the method was shown to be robust to systemic and sporadic missingness in individual studies. This method, in particular, could provide more robust imputation values compared to individual-level imputation methods, as it not only allows for imputation of missing values in clusters with systematic losses, but can also shrink the values of individual clusters towards

the global mean of studies, particularly advantageous in studies with extreme values or with values far away from those found on average at the global level.

The advantage of the proposed method over the previous ones 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 to say that under the specification of a validity exclusion variable the method determines by itself which is the most adjustable correlation parameter between equations (ρ), or in general terms the missingness mechanism (MAR or MNAR), in each of the clusters evaluated. The imputation method is built on the mice package, which allows, first of all, to be used both on the outcome and on the covariates and, in addition, offers the option of being used simultaneously with other imputation methods of the package, 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 variables with heterogeneous error variance and for binary variables.

One of the major limitations 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. Also the method is sensitive to the value of the correlation between the selection equation and outcome (rho), and in general it is observed that it can lead to biassed results on fixed global parameters i.e. without variation across clusters.Similarly, 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 it requires a minimum sample size to estimate all the parameters of the Heckman model which can be at least twice the number of parameters required in a MAR prediction model. 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.

The data were simulated by attributing a constant correlation across all clusters in order to evaluate the performance against M(N)AR assumptions, but in practice this parameter is variable across clusters which can considerably affect the performance of the method. So in future research the effect of this parameter could be more deeply evaluated.

On the other hand, 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.

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

17. 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

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

19. 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

20. 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

21. 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

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

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

24. 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

25. 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

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

27. 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

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