Multilevel Multivariate Imputation by Chained Equations through Bayesian Additive Regression Trees[⋆,⋆⋆]

Heleen Brüggen[a,1,2]

[a]*Methodology and Statistics for the Behavioural, Biomedical and Social Sciences, Utrecht Universty, The Netherlands*

Abstract

Defining a congenial imputation model for a hierarchically structured dataset is a tough process due to complicated non-linear relationships. Using a non-parametric, tree-based Bayesian Additive Regression Trees (BART) model as the imputation model might solve this. This study investigates if, in a multilevel context, the use of BART in Multiple Imputation (MI) improves the bias, coverage, and variance estimates of model parameter estimates compared to current practices. The performance of multilevel-BART models were evaluated with a statistical simulation in a multilevel imputation context. The population data-generating mechanism was a multilevel linear model with random intercepts, slopes, and cross-level interactions. The multilevel-BART imputation models were compared to single-level predictive mean matching (PMM), multilevel PMM, single-level BART, and complete case analysis. The results show that the multilevel PMM model performed best in terms of bias, coverage, and variance of the parameter estimates. However, the multilevel-BART model showed promising results for the random intercepts and slopes. Nevertheless, in its current form, multilevel-BART models do not offer an improvement over the multilevel PMM.

Keywords: bayesian additive regression trees, BART, multilevel-BART, multilevel multiple imputation, multilevel missing data, chained equations

[⋆]**Word count** = 5949

[⋆⋆]**FETC Case Number** = 23-1778

Candidate Journal = Computational Statistics & Data Analysis

Research Archive = `github.com/heleenbrueggen/masterthesis`

Email address: `h.bruggen@uu.nl` (Heleen Brüggen)

[1]**Student number** = 6474292

[2]**Supervisors** = T. Volker MSc., Dr. G. Vink, H. Oberman MSc.

Contents

1. Introduction

Incomplete data is a common challenge in many fields of research. Frequently used ad hoc strategies to deal with missing data, such as complete case analysis or mean imputation, often lead to erroneous inferences in realistic situations. Missingness can follow a multivariate mechanism that may depend on observed data — or even unobserved data. Ad hoc strategies don't account for those mechanisms, leading to biased estimates and inaccurate variance estimates (Austin et al., 2021; Enders, 2017; Kang, 2013; Little and Rubin, 2002; van Buuren, 2018). Multiple imputation (MI; Rubin, 1987) is proven to be an effective method for dealing with multivariate incomplete data supported by a considerable amount of methodological research (Audigier et al., 2018; Austin et al., 2021; Burgette and Reiter, 2010; Enders, 2017; Grund et al., 2021; Hughes et al., 2014; Little and Rubin, 2002; Mistler and Enders, 2017a; Van Buuren, 2007; van Buuren, 2018).

MI separates the missing data problem from the analysis problem (Audigier et al., 2018; Austin et al., 2021; Bartlett et al., 2015; Burgette and Reiter, 2010; Carpenter and Kenward, 2013; Enders, 2017; Grund et al., 2021; Hughes et al., 2014; Little and Rubin, 2002; Mistler and Enders, 2017a; Van Buuren, 2007; van Buuren, 2018). A statistical model specifying the variables and their relationships used for imputation — i.e. the imputation model — is defined for every variable with missing values. Each missing value in the dataset is imputed m times by drawing values from their posterior predictive distribution conditional on the observed data and parameters from the imputation model. By repeatedly drawing values from the posterior predictive distributions — in other words, the distribution of plausible replacement values — the necessary variation associated with the missingness problem is considered. After imputation, each of the imputed datasets are analyzed according to the model of interest, i.e. the substantive analysis model. Then, their corresponding model parameters are pooled together according to Rubin's rules (Rubin, 1987). One central requirement for MI is the concept of congeniality: the imputation model should be at least as general as the analysis model and preferably all-encompassing (Bartlett et al., 2015; Enders et al., 2018a; Grund et al., 2016, 2018b; Little and Rubin, 2002; Meng, 1994). If not, the imputation model will not be compatible with the analysis model and the pooled estimates of the latter may be biased.

When MI is applied in a multilevel data context, concerns regarding congeniality become more pronounced (Audigier et al., 2018; Dong and Mitani, 2023; Enders et al., 2020, 2018a,b, 2016; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017a; Quartagno and Carpenter, 2022; Resche-Rigon and White, 2018; Taljaard et al., 2008; van Buuren, 2018). Multilevel data is hierarchically structured, where, for example, students are nested within classes within schools (Hox and Roberts, 2011; Hox et al., 2017). When analyzing multilevel data, this hierarchical structure should be taken into consideration. Ignoring it will underestimate the intra-class correlation (ICC) — the proportion of the total variance at the grouping level (Gulliford et al., 2005; Hox and Roberts, 2011; Shieh, 2012) — and standard errors, as conventional statistical analyses assume independence of observations (Hox and Roberts, 2011; Lüdtke et al., 2017; Taljaard et al., 2008; van Buuren, 2018). Accounting for this structure can be done using multilevel models (MLMs; Hox and

Roberts, 2011; Hox et al., 2017; Lüdtke et al., 2017). MLMs can contain variables relating to the individual level — level-1 variables — or to the grouping structure — level-2 variables or potentially higher order structures. For example, imagine a case where students are nested within classes. Here, the academic performance of a student is a level-1 variable, whereas the teacher's experience is a level-2 variable. Additionally, MLMs allow you to specify random intercepts, indicating that some classes have students that significantly perform better or worse academically on average; random slopes, indicating that the relationship between the performance of students and the outcome variable differs between classes; and cross-level interactions, indicating that the effect of performance of students can differ with the teacher's experience (Hox and Roberts, 2011; Hox et al., 2017). Typically, the complexity of the multilevel analysis model is built stepwise with non-linearities: predictors, random intercepts, random slopes, and cross-level interactions are added incrementally to the model. Hence, the analysis model is not determined beforehand (Hox and Roberts, 2011; Hox et al., 2017). Thus, ensuring congeniality for the imputation model can be complex, since the final analysis model is not pre-determined. Furthermore, including the hierarchical structure along with cross-level interactions or other complicated non-linearities in imputation models is quite challenging (Burgette and Reiter, 2010; Hox and Roberts, 2011; van Buuren, 2018), also because very complex models might not converge (van Buuren, 2018).

A popular and flexible implementation of MI in a multilevel context, is fully conditional specification (FCS), otherwise known as chained equations (Audigier et al., 2018; Burgette and Reiter, 2010; Grund et al., 2018a; Van Buuren, 2007). FCS employs univariate linear mixed models to account for the hierarchical structure of multilevel models (Enders et al., 2018a; Mistler and Enders, 2017a; Resche-Rigon and White, 2018) and iteratively imputes each incomplete variable conditional on observed and previously imputed variables (Enders et al., 2018a,b, 2016; Grund et al., 2018a; Hughes et al., 2014; Mistler and Enders, 2017a; van Buuren, 2018). Furthermore, it can impute non-linearities, such as cross-level interactions, by using "passive imputation" or defining a separate imputation model for the non-linearities (Grund et al., 2018b; van Buuren, 2018). However, including these non-linearities in FCS is still complicated (Grund et al., 2018b, 2021; van Buuren, 2018). FCS can also handle random intercepts and slopes, yet, once again, correctly specifying an imputation model accounting for these random effects can be challenging (Grund et al., 2018b, 2021; van Buuren, 2018).

Non-parametric, tree-based models might alleviate these complexities when defining imputation models. They do not assume a specific data distribution. So, they implicitly model non-linear relationships and can simultaneously handle continuous and categorical variables (Breiman et al., 1984; Burgette and Reiter, 2010; Chipman et al., 2010; Hill et al., 2020; James et al., 2021; Lin and Luo, 2019; Salditt et al., 2023). The use of tree-based, non-parametric models like regression trees, random forests, or Bayesian Additive Regression Trees (BART) in imputation of single-level data simplified the imputation process and improved model parameter estimates (Burgette and Reiter, 2010; Silva and Gutman, 2022; Waljee et al., 2013; Xu et al., 2016). Specifically, the imputations showed better confidence interval coverage of the parameters, lower variance and lower bias, especially in non-linear or interactive contexts (Burgette and Reiter, 2010; Silva and Gutman, 2022; Xu et al., 2016). Waljee et al. (2013) also found lower misclassification error rate for the predicted class as

well as lower imputation error when imputing with a random forest algorithm compared to multivariate imputation by chained equations (`mice`) using linear, logistic, and polytomous logistic regression imputation models, K-nearest neighbors (KNN) and mean imputation.

Despite these promising findings, BART models have yet to be implemented in a multilevel imputation context. In prediction, multilevel-BART (M-BART) models have predominantly been implemented with random intercepts only (Chen, 2020; Tan et al., 2016; Wagner et al., 2020; Wundervald et al., 2022). Wagner et al. (2020) have found that this random intercept R-BART model provided better predictions with a lower mean squared error (MSE) compared to a parametric MLM; Tan et al. (2016) found higher area under the curve (AUC) values compared to a single-level BART model and linear logistic random intercept model; and Chen (2020) found better predictions and better coverage of the parameter estimates compared to parametric models and a single-level BART model. Other researchers modeled the random intercept as an extra split on each terminal node and found a lower MSE compared to a standard BART model and parametric MLMs (Wundervald et al., 2022). Dorie et al. (2022) developed an M-BART model that included random intercepts, random slopes and cross-level interactions by modeling these random parts with a Stan (Lee et al., 2017) model and the fixed parts with a BART model. Their results showed that their algorithm `stan4bart` showed better coverage of the sample average treatment effect (SATT) and lower root mean squared error (RMSE) compared to BART models with varying intercept, BART models ignoring the multilevel structure, bayesian causal forests, and parametric MLMs.

Considering all these findings, my research question will be: *How can multivariate imputation by chained equations through a multilevel bayesian additive regression trees model improve the bias, variance, and coverage of the multilevel model parameter estimates compared to current practices?* Given the success of non-parametric models in single-level MI, I anticipate that employing M-BART models in a multilevel missing data context will reduce bias, accurately model variance, and improve estimate coverage compared to conventional implementations of multilevel MI, single-level MI, and complete case analysis with the R package `mice` (Buuren and Groothuis-Oudshoorn, 2011).

2. Method

2.1. Theoretical background

2.1.1. Bayesian Additive Regression Trees (BART)

BART is a sum-of-trees model proposed by Chipman et al. (2010) with regression trees as its building blocks (Chipman et al., 2010; Hill et al., 2020; James et al., 2021). Regression trees recursively split the data into binary subgroups based on the predictors included in the model. At each step down the tree, the splits are based on the predictor that minimizes the variability within the subgroups. Observations are then assigned to a certain subgroup according to these splits. This is continued until a certain stopping criterion is reached; for example, we desire a minimal number of observations within a subgroup (Breiman et al., 1984; Hastie, 2017; James et al., 2021; Salditt et al., 2023). Recursive binary partitioning of the predictor space doesn't assume a specific data form. This makes regression trees, and as a consequence, BART, non-parametric models (Breiman et al., 1984; Hastie, 2017;

James et al., 2021; Salditt et al., 2023). It allows regression trees to model non-linearities and other complicated relationships well and automatically (Burgette and Reiter, 2010; Hill et al., 2020). Chipman et al. (2010) define the BART model as:

$$f(\mathbf{x}) = \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k), \tag{1}$$

where $f(\mathbf{x})$ is the overall fit of the model: the sum of K regression trees; $\mathbf{x}$ are the predictor variables; T_k is the k$^{\text{th}}$ tree; and M_k is the collection of leaf parameters within the k$^{\text{th}}$ tree, i.e. the collection of predictions for its terminal nodes (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021). The data are assumed to arise from a model with additive normally distributed errors: $Y = \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$. Next to the sum-of-trees model, BART also includes a regularization prior that constrains the size and fit of each tree so that each contributes only a small part of the variation in the outcome variable to prevent overfitting. The prior is imposed over all parameters of the sum-of-trees model, specifically, $(T_1, M_1), \ldots, (T_K, M_K)$ and σ. However, the specification of the regularization prior is simplified by a series of independence assumptions:

$$\begin{aligned} p((T_1, M_1), \ldots, (T_K, M_K), \sigma) &= \Big[\prod_k p(T_k, M_k)\Big] p(\sigma), \\ &= \Big[\prod_k p(M_k|T_k) p(T_k)\Big] p(\sigma), \\ p(M_k|T_k) &= \prod_h p(\mu_{hk}|T_k), \end{aligned} \tag{2}$$

where $\mu_{hk} \in M_k$. These assumptions state that the tree components (T_k, M_k) and the standard deviation (σ) are independent of each other, and the leaf parameters within every tree $(\mu_{hk}|T_k)$ are independent of each other. Thus, priors only need to be specified for those parameters (Chipman et al., 2006, 1998, 2010; Hill et al., 2020). Chipman et al. (1998) define an independent prior for each tree. The probability that a node splits at depth d is defined as:

$$\alpha(1+d)^{-\beta}, \alpha \in (0,1), \beta \in [0, \infty), \tag{3}$$

where the default specification put forth by Chipman et al. (2006, 2010) is $\alpha = .95$ and $\beta = 2$. This specification sets the probability of a tree with 1, 2, 3, 4, and 5 nodes at .05, .55, .28, .09, and .03 respectively. Thus, smaller trees are favored. Chipman et al. (2006, 2010) also provide a default specification for the prior for the leaf parameters. They propose to rescale the response value to the interval $[-.5, .5]$ and then, the leaf parameter prior is defined as:

$$\mu_{hk} \sim \mathcal{N}(0, \sigma_\mu^2), \text{with } \sigma_\mu^2 = \frac{.5}{t\sqrt{K}}, \tag{4}$$

where t is a preselected number and K is the number of trees. This prior shrinks the leaf parameters μ_{hk} towards 0, decreasing their individual effect. If t or K increases, more shrinkage is applied. Chipman et al. (2006, 2010) recommend using $t = 2$ — or values between 1 and 3 — as a default. Furthermore, Chipman et al. (2006, 2010) propose the conjugate inverse chi-square distribution as the prior for the residual standard deviation — $\sigma^2 \sim \nu\lambda/\chi^2_\nu$. They represent λ, the degrees of freedom, as the probability that σ — from BART —, is less than the estimated residual standard deviation from a linear regression model, $\hat{\sigma}_{\text{OLS}}$. Their default specification is $\nu = 3$ and $\Pr(\sigma < \hat{\sigma}_{\text{OLS}}) = \lambda = .9$ (Chipman et al., 2006, 1998, 2010; Hill et al., 2020).

BARTs are estimated using the Bayesian back-fitting Markov Chain Monte Carlo (MCMC) algorithm with a Metropolis-within-Gibbs sampler (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021). Each tree is initialized with the mean response value divided by the number of trees as its root node — $\hat{f}^1_k(x) = \frac{1}{nK}\sum_{i=1}^{n} y_i$, with sample size n. Then, each pair (T_k, M_k) is updated considering the remaining trees, their associated parameters, and σ by sampling from the following conditional distribution:

$$(T_k, M_k)|T_{k'}, M_{k'}, \sigma, y, \tag{5}$$

where, $T_{k'}$ is every tree expect the k$^{\text{th}}$ tree. This conditional distribution only depends on $(T_{k'}, M_{k'}, y)$ through the partial residuals:

$$r_i = y_i - \sum_{k'<k} \hat{f}^b_{k'}(x_i) - \sum_{k'>k} \hat{f}^{b-1}_{k'}(x_i), \text{with } i = 1, \ldots, n, \tag{6}$$

where $\hat{f}^b_k(x_i)$ is the prediction of the k^{th} tree in the b^{th} iteration for person i and sample size n. Thus, updating each pair (T_k, M_k) simplifies to proposing a new tree fit to the partial residuals, r_i, treating them as the data, by perturbing the tree from the previous iteration. Perturbations entail either *growing*, *pruning*, or *changing* a tree. *Growing* means adding additional splits, *pruning* removes splits, and *changing* changes decision rules. The algorithm stops after the specified number of iterations (Chipman et al., 2006, 1998, 2010; Hill et al., 2020; James et al., 2021).

2.1.2. Random intercept BART (R-BART)

Tan et al. (2016); Wagner et al. (2020) and Dorie et al. (2024) define an R-BART model including a random intercept. The BART model (1) is extended to include a random intercept by:

$$f(\mathbf{x}) = \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k) + \alpha_j, \tag{7}$$

where, now, $f(\mathbf{x})$ is the overall fit of the model incorporating random intercept α_j for cluster

j. So, the data are now assumed to arise from the following model:

$$Y_{ij} = \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k) + \alpha_j + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0,\, \sigma^2),\;\; \alpha_j \sim \mathcal{N}(0,\, \tau^2), \tag{8}$$

where $\alpha_j \perp \epsilon_{ij}$. Now the joint prior distribution (2) becomes:

$$\begin{aligned} p((T_1, M_1), \ldots, (T_K, M_K), \sigma) &= \Big[\prod_k p(T_k, M_k)\Big] p(\sigma) p(\tau), \\ &= \Big[\prod_k p(M_k|T_k) p(T_k)\Big] p(\sigma) p(\tau), \\ p(M_k|T_k) &= \prod_h p(\mu_{hk}|T_k). \end{aligned} \tag{9}$$

A Metropolis-within-Gibbs algorithm is used to draw values from the posterior. First, the Gibbs sample for σ, τ, and α_j are obtained from their respective posterior distributions. Then, we obtain $\tilde{Y}_{ij} = Y_{ij} - \alpha_j$ and view $\tilde{Y}_{ij}|\boldsymbol{x}_j$ as a BART model. So, $\tilde{Y}$ is now used as the outcome variable in the BART algorithm described in the previous section, 2.1.1. (Tan et al., 2016; Wagner et al., 2020). Dorie et al. (2024) implemented this algorithm within the R package `dbarts` with the function `rbart_vi()`. Where, the default prior for the random intercept is $\tau \sim \text{Cauchy}(0, 2.5)$: a Cauchy distribution with a scale parameter 2.5 times the original scale.

2.1.3. stan4bart

Dorie et al. (2022) developed a M-BART model that included random intercepts, random slopes, and cross-level interactions. They extend a Bayesian linear mixed model with a BART model (1):

$$f(\mathbf{x}) = \boldsymbol{\beta}\mathbf{x}^{\beta} + \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k) + \boldsymbol{\lambda}\mathbf{w}, \tag{10}$$

where $\boldsymbol{\beta}$ is a vector of linear, parametric coefficients; $\mathbf{x}^{\beta}$ the design matrix for the linear predictors; $\boldsymbol{\lambda}$ a vector of the parametric random effects; $\mathbf{w}$ the design matrix for the random effects; and $\sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k)$ is a non-parametric, sum-of-trees BART model (Dorie et al., 2022). So, the data are assumed to arise from the following model:

$$Y_{ij} = \boldsymbol{\beta}\mathbf{x}^{\beta} + \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k) + \boldsymbol{\lambda}\mathbf{w} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0,\, \sigma^2),\;\; \boldsymbol{\lambda} \sim \mathcal{N}(0,\, \boldsymbol{\Sigma}_{\lambda}), \tag{11}$$

where $\boldsymbol{\Sigma}_{\lambda}$ is the variance-covariance matrix for the random intercept and slopes. The model is implemented as a Gibbs sampler: a Hamiltonian Monte Carlo, no-U-turn sampler with a diagonal Euclidean adaptation matrix is used to jointly sample the linear, parametric

components given the non-parametric components. The non-parametric components are sampled using the BART algorithm described in section 2.1.1. To accomplish this, a parametric Stan model (Lee et al., 2017) fits equation 10 with $\sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k)$ as a generic linear offset. Dorie et al. (2022) combine a custom mutable Stan sampler object with a BART sampler with a fixed variance and offset term. First, the Stan sampler collects the current draws of the BART model into $vec_i \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k)$ and uses this to draw $\beta, \lambda, \sigma, \Sigma_\lambda | \mathbf{Y}, vec_i \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k)$. Then, σ and $vec_i \left[\boldsymbol{\beta}\mathbf{x}_i^{\beta} + \boldsymbol{\lambda}\mathbf{w}_i\right]$ are passed to BART, which produces $M_k, T_k | \mathbf{Y}, vec_i \left[\boldsymbol{\beta}\mathbf{x}_i^{\beta} + \boldsymbol{\lambda}\mathbf{w}_i\right], \sigma, M_{k'}, T_{k'}$. The cycle is completed by passing $vec_i \sum_{k=1}^{K} g(\mathbf{x}; T_k, M_k)$ back to Stan. The process is continued for the set number of posterior samples which are intended for inference. This algorithm is implemented in the R package `stan4bart` (Dorie, 2023).

2.2. Simulation study

2.2.1. Data generating mechanism

A simulation study is assembled to evaluate the performance of M-BART models in a multilevel imputation context. The population data-generating mechanism is based on the following MLM:

$$y_{ij} = \beta_{0j} + \sum_{f=1}^{7} \beta_{fj} X_{fij} + \epsilon_{ij}, \qquad X_{fij} \sim \mathcal{MVN}(0, \boldsymbol{\Sigma}_x), \tag{12a}$$

$$\beta_{0j} = \gamma_{00} + \sum_{q=1}^{2} \gamma_{0p} Z_{pj} + \upsilon_{0j}, \tag{12b}$$

$$\beta_{fj} = \gamma_{f0} + \sum_{q=1}^{2} \gamma_{fp} Z_{pj} + \upsilon_{fj}, \qquad Z_{pj} \sim \mathcal{MVN}(0, \boldsymbol{\Sigma}_z), \tag{12c}$$

where y_{ij} is a continuous level-1 outcome variable for person i in group j and X_{fij} are 7 continuous level-1 variables and Z_{pj} are 2 continuous level-2 variables. The predictors are multivariate normally distributed with means of 0 and variance-covariance matrix $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_z$, respectively:

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 6.25 & & & & & & \\ 2.25 & 9 & & & & & \\ 1.5 & 1.8 & 4 & & & & \\ 2.25 & 3.06 & 2.04 & 11.56 & & & \\ 1.5 & 1.8 & 1.2 & 2.04 & 4 & & \\ 1.125 & 1.35 & 0.9 & 1.53 & .9 & 2.25 & \\ 3.3 & 3.96 & 2.64 & 4.488 & 2.64 & 1.98 & 19.36 \end{pmatrix}, \tag{13a}$$

$$\boldsymbol{\Sigma}_z = \begin{pmatrix} 1 & \\ .48 & 2.56 \end{pmatrix}. \tag{13b}$$

The covariances between the variables are calculated such that the correlation between the variables is .3, aligned with Cohen's (1990) medium effect size benchmark. The residuals are normally distributed as,

$$\epsilon_{ij} \sim \mathcal{N}(0, 25). \tag{14}$$

The random intercept β_{0j} is determined by the overall intercept γ_{00}, the 2 group-level effects $\gamma_{0p}Z_{pj}$ and the group-level random residuals υ_{0j}. The overall intercept γ_{00} is set to 10 and the group-level effects γ_{01} and γ_{02} to .5. The 7 regression coefficients β_{fj} for the continuous variables X_{fij} depend on the intercepts γ_{f0}, the cross-level interactions $\gamma_{fp}Z_{pj}$, and the random slopes υ_{fj}. The 7 intercepts, or within-group effect sizes, γ_{f0} are set to .5, the cross-level interactions γ_{11}, γ_{21}, and γ_{32} are set to .35.

$$\gamma_{00} = 10, \;\; \boldsymbol{\gamma}_{0p} = \begin{pmatrix} .5 \\ .5 \end{pmatrix}, \;\; \boldsymbol{\gamma}_{f0} = \begin{pmatrix} .5 \\ .5 \\ .5 \\ .5 \\ .5 \\ .5 \\ .5 \end{pmatrix}, \;\; \boldsymbol{\gamma}_{fp} = \begin{pmatrix} .35 & 0 \\ .35 & 0 \\ 0 & .35 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{15}$$

The random slopes are multivariate normally distributed with a mean of 0 and a variance-covariance matrix $\mathbf{T}$ shown in equation 16. Again, the covariances are calculated to yield a correlation of .3.

$$\boldsymbol{\upsilon}_j \sim \mathcal{MVN}(0, \mathbf{T}), \;\; \mathbf{T} = \begin{pmatrix} t_{00} & & & & & & & \\ .3 & 1 & & & & & & \\ .3 & .3 & 1 & & & & & \\ .3 & .3 & .3 & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{16}$$

The variance of υ_{0j}, the group-level random residuals t_{00}, are scaled such that the ICC was .5. The following formula is used to calculate υ_{0j} following the variance decomposition from Rights and Sterba (2019):

$$\text{ICC} = \frac{\boldsymbol{\gamma}^{b\prime}\boldsymbol{\phi}^{b}\boldsymbol{\gamma}^{b} + \tau_{00}}{\boldsymbol{\gamma}^{w\prime}\boldsymbol{\phi}^{w}\boldsymbol{\gamma}^{w} + \boldsymbol{\gamma}^{b\prime}\boldsymbol{\phi}^{b}\boldsymbol{\gamma}^{b} + tr(\mathbf{T}\boldsymbol{\Sigma}) + \tau_{00} + \sigma^2}, \tag{17}$$

where $\boldsymbol{\gamma}^{b}$ and $\boldsymbol{\gamma}^{w}$ are the level-1 and level-2 fixed effects; $\boldsymbol{\phi}^{b}$ is the variance-covariance matrix of a vector with 1 — for the intercept — and all level-2 predictors; $\boldsymbol{\phi}^{w}$ is the variance-

covariance matrices of all cluster-mean-centered level-1 predictors; τ_{00} is the variance of the random intercept; $\mathbf{T}$ is the variance-covariance matrix of the random intercept and slopes; $\mathbf{\Sigma}$ is the variance-covariance matrix of a vector containing 1 — for the intercept — and the level-1 variables; and σ^2 is the residual variance. The value for τ_{00} is calculated using the function `uniroot()` in R (R version 4.3.2 (2023-10-31); R Core Team, 2023).

2.2.2. Simulation design

Table 1 shows the design factors considered in the simulation study. These factors are either grounded in prior research or deemed realistic in real-world applications (Enders et al., 2020, 2018b; Grund et al., 2018b; Gulliford et al., 1999; Hox et al., 2017; Murray and Blitstein, 2003). According to Kreft and de Leeuw (2007), 30 groups is the smallest acceptable number in multilevel research and 50 groups is frequent in organizational research (Maas and Hox, 2005). Group sizes of 15 are typical in educational research (Lüdtke et al., 2017) and group sizes of 50 and an ICC of .5 are often used in simulation studies (Akkaya Hocagil and Yucel, 2023; Enders et al., 2020, 2018a,b; Grund et al., 2018b; Maas and Hox, 2005; Mistler and Enders, 2017b; Salditt et al., 2023). Oberman and Vink (2023) recommend including both Missing Completely At Random (MCAR) and Missing At Random (MAR) missingness mechanisms in simulation studies. The statistical properties of the imputation method are not deemed sound if it cannot yield valid inferences under MCAR and including MAR is important in evaluating the imputation method's performance. The amount of missingness in datasets is varied between 0% — as an additional benchmark — and 50%, which is often used in simulation studies as a high amount of missingness (Grund et al., 2016; Lüdtke et al., 2017; Schouten and Vink, 2021). 5 different imputation methods are compared:

Table 1: Simulation design

Design factors	**Values**
Number of clusters (J))	30, 50
Within-cluster sample size (n_j)	15, 50
Intraclass Correlation (ICC)	.5
Missing data mechanism	MCAR, MAR
Amount of missingness	0%, 50%

1. conventional single-level imputation with PMM (predictive mean matching),
2. conventional multilevel imputation with PMM,
3. single-level BART imputation,
4. M-BART imputation accounting for random intercepts (Tan et al., 2016; Wagner et al., 2020),
5. M-BART imputation accounting for random effects and cross-level interactions (Dorie et al., 2022).

For each combination of design factors, 100 datasets are simulated for the first 4 methods. A differing number of datasets are simulated for the 5$^{\text{th}}$ method due to time restrictions, which can be seen in table 2.

The first and second methods are implemented with the method `pmm` from `mice` and `2l.pmm` in combination with `2lonly.mean` — for the level-1 and level-2 variables — from `miceadds` (Robitzsch et al., 2024) respectively.

The third, fourth, and fifth methods are implemented by writing new method-functions in R for the package `mice`. The functions `bart()` and `rbart_vi()` from the `dbarts` package were used for the third and fourth methods (Dorie et al., 2024). The function `stan4bart()` from the package `stan4bart` was used for the fifth method (Dorie, 2023). The functions were written such that they can be used as imputation methods in the `mice` package as follows: for every variable to be imputed, a respective BART model is fitted — with the default specifications from the `bart` function — based on the predictor matrix. Then, the fitted values — the posterior means — are extracted for the observed and missing values. Imputations for the missing values are then obtained using predictive mean matching: a set of candidate donors are obtained by matching the predicted values for the observed cases that are closest to the predicted values for the missing cases — i.e. type 0 matching (van Buuren, 2018). Then, the observed value of one randomly selected donor is used as the imputed value for the missing case. The code for these functions can be found in Appendix A.

For all imputation methods, the incomplete datasets are imputed 5 times with 10 iterations each. Then, each of the 5 imputed datasets are analyzed using the R package `lme4` (Bates et al., 2015) with an MLM reflecting the population generating mechanism. The estimates from the 5 imputed datasets are pooled together using the R package `mice` (Buuren and Groothuis-Oudshoorn, 2011). These pooled estimates are compared on the bias, coverage, and the width of the 95% confidence intervals.

Table 2: Number of simulated datasets used in evaluation of the stan4bart imputation method for every design factor.

	Number of groups: 50 Group size: 15	**Number of groups: 30 Group size: 50**	**Number of groups: 50 Group size: 15**	**Number of groups: 50 Group size: 50**
MAR	20	20	20	20
MCAR	100	40	100	20

As an additional benchmark, the imputation methods are compared to analyses using listwise deletion — i.e. complete case analysis — and using the true data without missing values — i.e. the comparative truth.

2.2.3. Missing data generation

Missing values in the variables are introduced by multivariate amputation using the function `ampute()` (Schouten et al., 2018) from package `mice`. As can be seen in table 1, the missing data mechanism is either Missing Completely At Random (MCAR) or Missing At Random (MAR). The missing data mechanism is said to be MCAR when the cause of the missing data is unrelated to the data and MAR when the missing data is related to the observed data (Rubin, 1976). The amount of missingness is either 0% or 50%, which is defined as the percentage of cases that have at least one missing value.

For both MCAR and MAR, all possible patterns with 1 to 5 missing values out of the 10 variables (*x1, x2, x3, x4, x5, x6, x7, z1, z2, and y*) per case are generated. They have the same relative frequency of occurrence in the datasets.

For the MAR mechanism, the weighted sum of scores on the observed variables is used to predict the probability of missingness for a case. When they remain observed in a specific pattern, the weights of the variables *x4* and *z1* are set to 2 and 1.5 and the weights of the other variables are set to 1. The type of missingness is set to `'RIGHT'` meaning that cases with a higher weighted sum of scores have a higher probability of becoming incomplete. So, this means that cases with higher values on *x4* and *z1* are more likely to become incomplete.

In summary, either no missing values are introduced (0%), or up to 5 missing values are introduced in 50% of the cases. When data is MAR, the probability of a value being missing depends on the observed values of all other variables, with variables *x4* and *z1* having a greater influence on this probability.

2.2.4. Evaluation

The estimates from the analysis models are evaluated in terms of absolute bias, coverage of 95% confidence intervals, with their respective Monte Carlo SE (MCSE), and the width of the 95% confidence intervals (Morris et al., 2019; Oberman and Vink, 2023):

$$\text{Bias} = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \theta), \qquad \text{MCSE}_{\text{Bias}} = \sqrt{\frac{\sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_t - \bar{\theta})^2}{n_{\text{sim}}(n_{\text{sim}} - 1)}}, \tag{18}$$

$$\text{Coverage} = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} 1(\hat{\theta}_{\text{low,i}} \leq \theta \leq \hat{\theta}_{\text{upp,i}}), \quad \text{MCSE}_{\text{Cov}} = \sqrt{\frac{\hat{\text{Cov}}\text{erage}(1 - \hat{\text{Cov}}\text{erage})}{n_{\text{sim}}}}, \tag{19}$$

$$\text{CIW} = \frac{1}{n_{\text{sim}}} \sum_{t=1}^{n_{\text{sim}}} (\hat{\theta}_{\text{upp,i}} - \hat{\theta}_{\text{low,i}}), \tag{20}$$

where $\hat{\theta}_t$ is the estimated parameter in simulation t, θ is the true value, $\bar{\theta}$ is the mean of $\hat{\theta}_t$, and n_{sim} is the number of simulated datasets. The lower and upper bounds of the 95% confidence intervals are denoted as $\hat{\theta}_{\text{low,i}}$ and $\hat{\theta}_{\text{upp,i}}$ respectively. The coverage is the proportion of the 95% confidence intervals that contain the true value.

Published simulation studies frequently consider a relative bias — which is the absolute bias divided by the true parameter value multiplied by 100 — of 10% as acceptable bias (Enders et al., 2020, 2018a,b; Finch et al., 1997). Enders et al. (2018a); Morris et al. (2019); Oberman and Vink (2023); van Buuren (2018) suggest that a coverage of 95% is acceptable. Poor coverage, i.e. below 95%, indicates biased estimates or too narrow intervals. While, coverage above 95% indicates that efficiency could still be gained. Furthermore, Bradley (1978) suggests a liberal criterion for coverage between 92.5% and 97.5% being acceptable (Enders et al., 2020, 2018a,b). In this study coverage is only considered for the fixed effects, since literature suggests that symmetric confidence intervals for the random parts is unsuitable (Enders et al., 2020, 2018a,b; Maas and Hox, 2005). The width of the confidence intervals is a measure of the statistical precision of the estimates: a smaller width indicates a more precise estimate (Oberman and Vink, 2023; van Buuren, 2018).

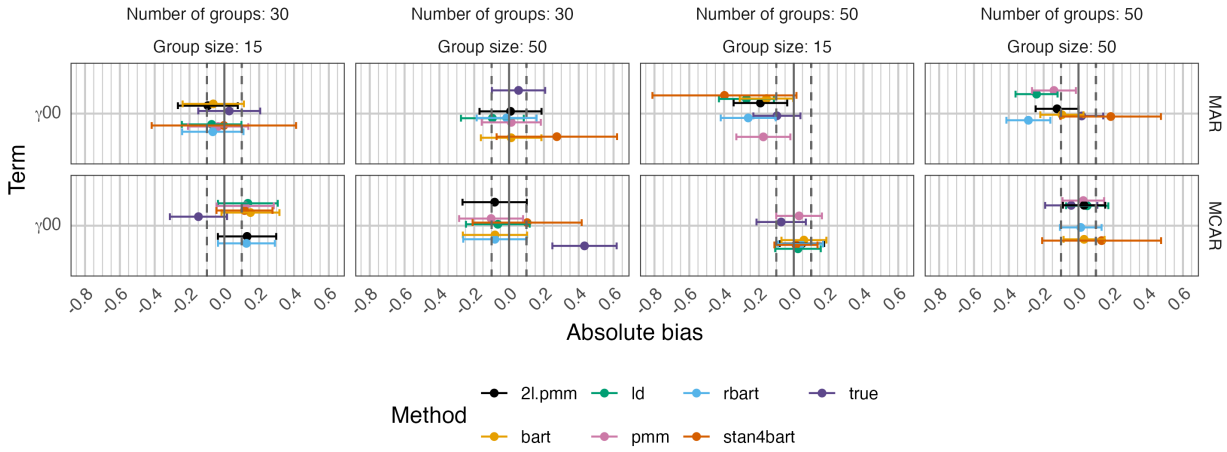
3. Results

3.1. Bias

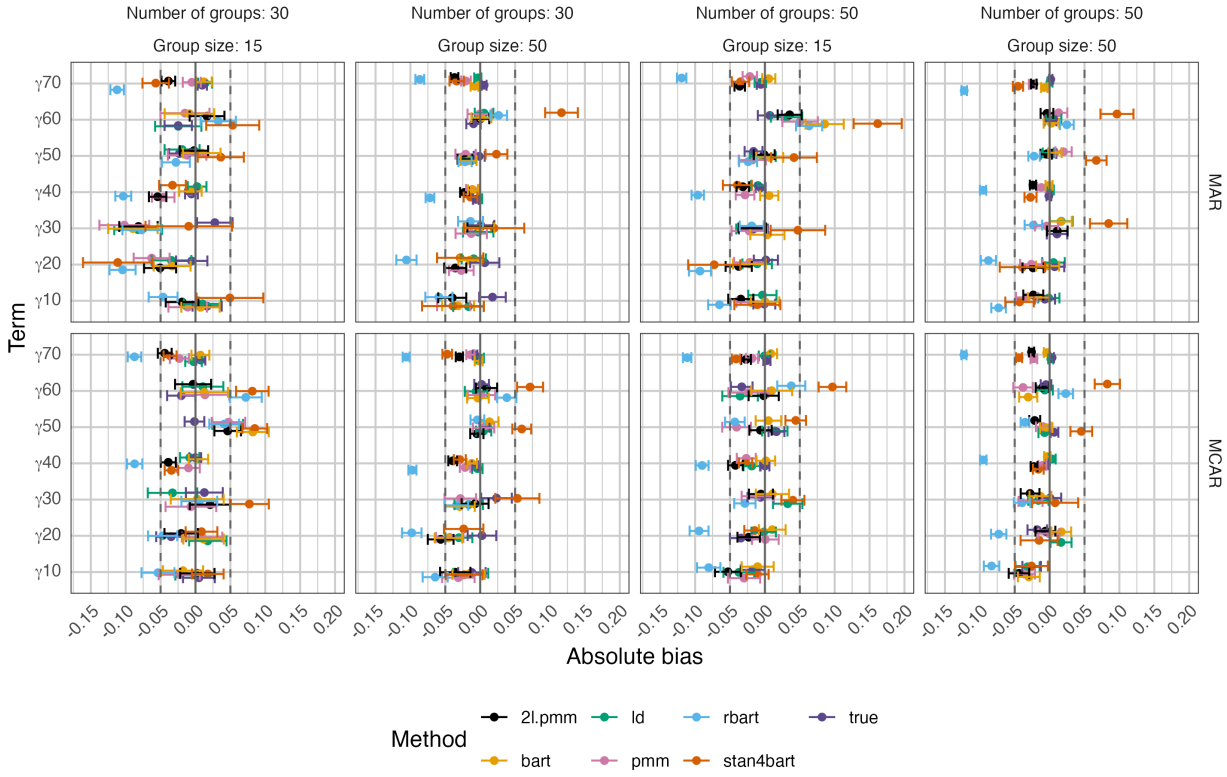
Figures 1 through 3 show the absolute bias and corresponding Monte Carlo SE for the estimates of the linear mixed model for all imputation methods in consideration: the overall intercept — γ_{00}; the level-1 effects — $\gamma_{10}, \gamma_{20}, \ldots, \gamma_{70}$; the level-2 effects — γ_{01} and γ_{02}; cross-level interactions — γ_{11}, γ_{21}, and γ_{32}; the random slopes — υ_1, υ_2, and υ_3; and the residual and intercept variance — $\upsilon 0$ and ϵ_{ij}.

Figure 1a shows that for most imputation methods when the data is MAR, the overall intercept is approximately unbiased when there are 30 groups — the absolute biases fall between the 10% relative bias lines — but is underestimated when there are 50 groups. stan4bart is unbiased with the smallest sample size and overestimates the intercept when there are larger groups. Nonetheless, the simulation uncertainty for stan4bart still encompasses the zero-bias line. Listwise deletion underestimates the intercept compared to the 0% missing, "true" data when the data is MAR. When the data is MCAR, there seem to be fewer differences in performance between the imputation methods: all approximately overestimate the intercept with 30 groups of size 15 and are approximately unbiased with the other group conditions, except for stan4bart, which overestimates the intercept with 50 groups of size 50.

Figure 1b shows the absolute bias of the level-1 effects — $\gamma_{10}, \gamma_{20}, \ldots, \gamma_{70}$. When the sample size is smallest — i.e. 30 groups of size 15 — fluctuations in bias are frequent for all methods when the data is MAR and MCAR. 2l.pmm, pmm, and bart seem to perform best out of all imputation methods for both missingness mechanisms, especially with a larger total sample size. stan4bart seems to overestimate some fixed effects, which increases with the total sample size and when the data is MAR. Overall, rbart has the worst performance in terms of absolute bias, consistently underestimating some level-1 effects.



(a) Overall intercept



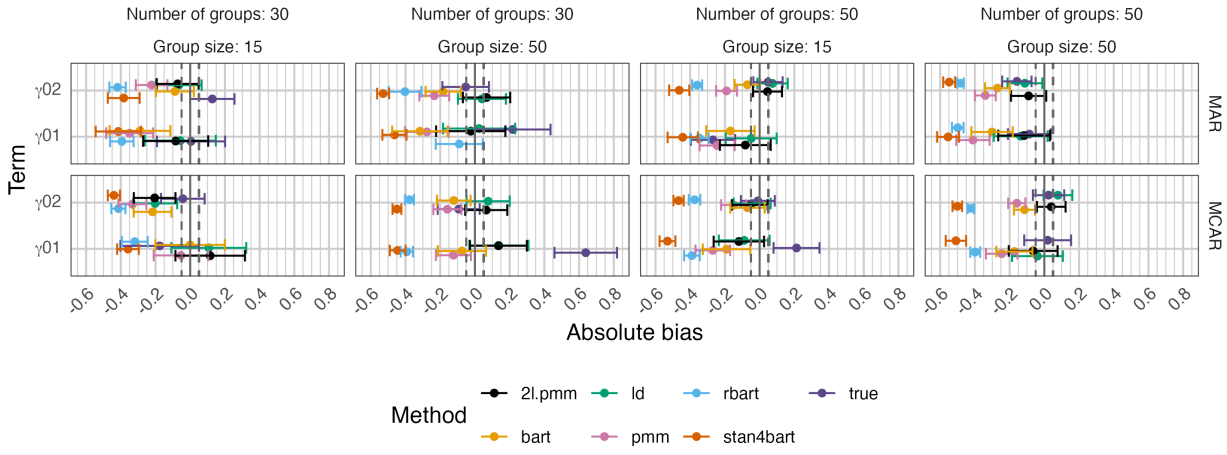
(b) Level-1 effects

Figure 1: Absolute bias of the overall intercept and level-1 fixed effects of the linear mixed model with respective Monte Carlo SE for all simulated datasets over 100 simulations with ICC = .5. The dashed lines represent ±10% relative bias. Method stan4bart is based on a differing number of dataset simulations described in table 2.

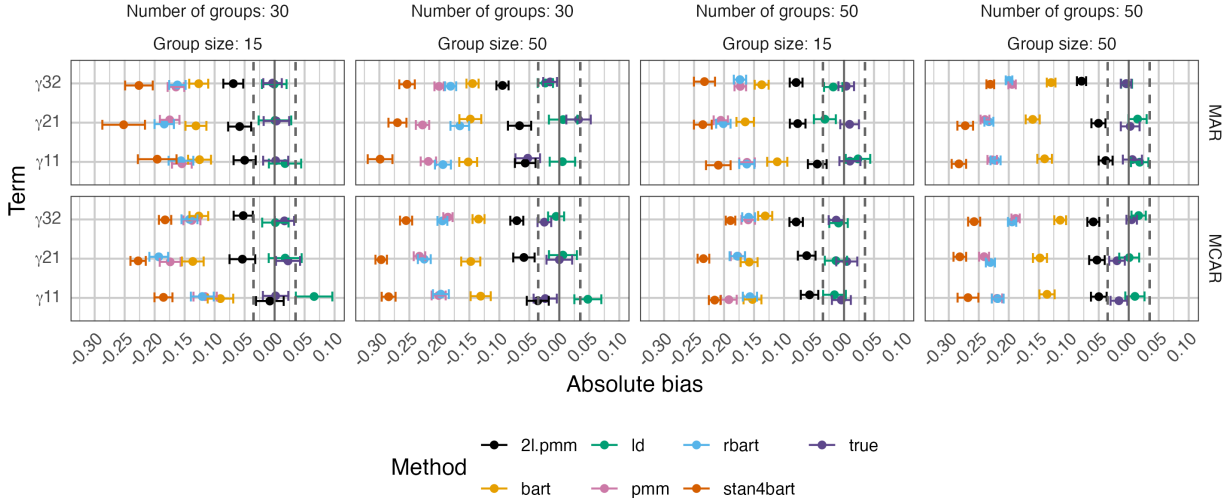
Considering the level-2 effects — γ_{01} and γ_{02} — from figure 2a, stan4bart performs the worst out of all imputation methods: underestimating the level-2 effects. 2l.pmm performs

best out of all imputation methods, even though still over- or underestimating the level-2 effects at times and seemingly performing slightly worse under MCAR. pmm and rbart consistently underestimate the level-2 effects for all conditions. bart performs slightly better, at times even mimicking the performance of 2l.pmm.

For the cross-level interactions — γ_{11}, γ_{21}, and γ_{32} —, stan4bart performs worst out of all methods. Figure 2b shows that stan4bart consistently underestimates the cross-level interactions. For both MAR and MCAR, listwise deletion performs largely acceptable in terms of bias. 2l.pmm performs best out overall, nonetheless still underestimating the cross-level interactions regularly. Furthermore, 2l.pmm performs slightly better under MCAR than MAR. bart outperforms pmm and rbart, but still underestimates the cross-level interactions. Additionally, bart, rbart and pmm perform worse with larger groups.



(a) Level-2 effects



(b) Cross-level interactions

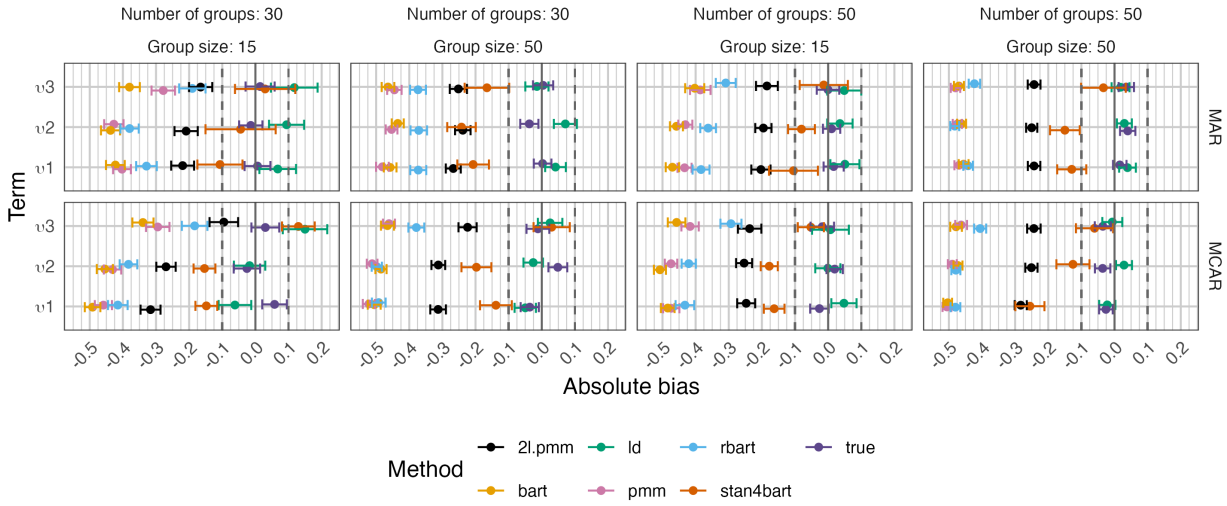
Figure 2: Absolute bias of the fixed level-2 effects and cross-level interactions of the linear mixed model with respective Monte Carlo SE for all simulated datasets over 100 simulations with ICC = .5. The dashed lines represent ±10% relative bias. Method stan4bart is based on a differing number of dataset simulations described in table 2.

The absolute bias for the random slopes — υ_1 to υ_3 — in figure 3a show that stan4bart performs best overall. When the data is MAR, stan4bart provides acceptable biases when group sizes are 15 performing better with more but smaller groups. pmm, bart and rbart perform worst: consistently underestimating the random slopes for all factor conditions. 2l.pmm performs better than pmm, bart, and rbart, but still underestimates the random slopes for most conditions. Listwise deletion performs largely acceptable in terms of bias, most of the time staying within the 10% relative bias lines for both MAR and MCAR.

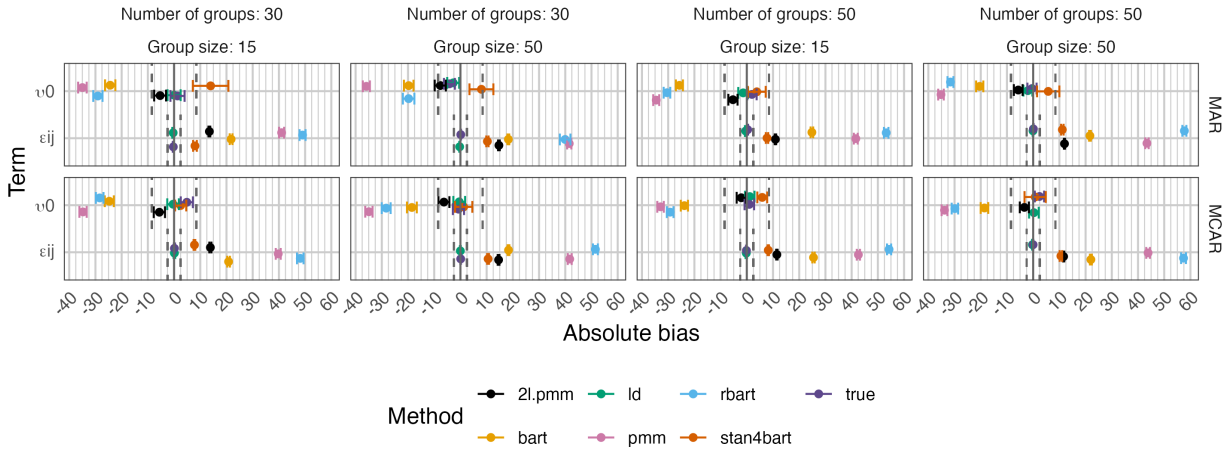
From figure 3b it can be seen that stan4bart and 2l.pmm have an acceptable bias for the intercept variance for most conditions. stan4bart seems to slightly overestimate the

intercept variance when the data is MAR compared to MCAR. Additionally, stan4bart improves in bias when there are more groups in the dataset. 2l.pmm slightly underestimates the intercept variance to an acceptable extent for all conditions. pmm, rbart and bart routinely underestimate the intercept variance. Listwise deletion shows a very minor bias, underestimating the intercept variance when the data is MAR.

Looking at the residual variance, listwise deletion is the only method that is approximately unbiased for all — or any — condition. All other imputation methods routinely overestimate the residual variance. stan4bart has the best performance followed by 2l.pmm, bart, pmm, and rbart, in that order. Overall, the bias seems consistent across all conditions. Aside from stan4bart that increases in bias when the total sample size increases and 2l.pmm which decreases in bias with more groups.



(a) Random slopes



(b) Residual and intercept variance

Figure 3: Absolute bias of the random effects of the linear mixed model with respective Monte Carlo SE for all simulated datasets over 100 simulations with ICC = .5. The dashed lines represent ±10% relative bias. Method stan4bart is based on a differing number of dataset simulations described in table 2.

3.2. Coverage

Figures 4 and 5 show the coverage of the 95% confidence intervals of the fixed effects — the overall intercept, level-1 effects, level-2 effects, and cross-level interactions — of the linear mixed model for all imputation methods in consideration with corresponding Monte Carlo SE.

Figure 4 shows that the coverage of the overall intercept, γ_{00}, is best for imputation method 2l.pmm when the data is MAR. However, when the data is MCAR, stan4bart performs best. 2l.pmm shows undercoverage in the smallest sample size when the data is MCAR and stan4bart shows fluctuations in coverage when the data is MAR. pmm and rbart

show undercoverage for most conditions. bart undercovers the intercept for all conditions unless the data is MCAR with 50 groups.

Next, the coverage of the level-1 effects in figure 4 shows that 2l.pmm performs best. stan4bart and pmm also show a good performance in terms of coverage, however, they regularly overcover and sometimes undercover the level-1 effects. bart shows undercoverage more consistently. rbart has most fluctuations in coverage: when group sizes are 50, it shows undercoverage as low as around 47,5% and overcoverage as high as around 98%.

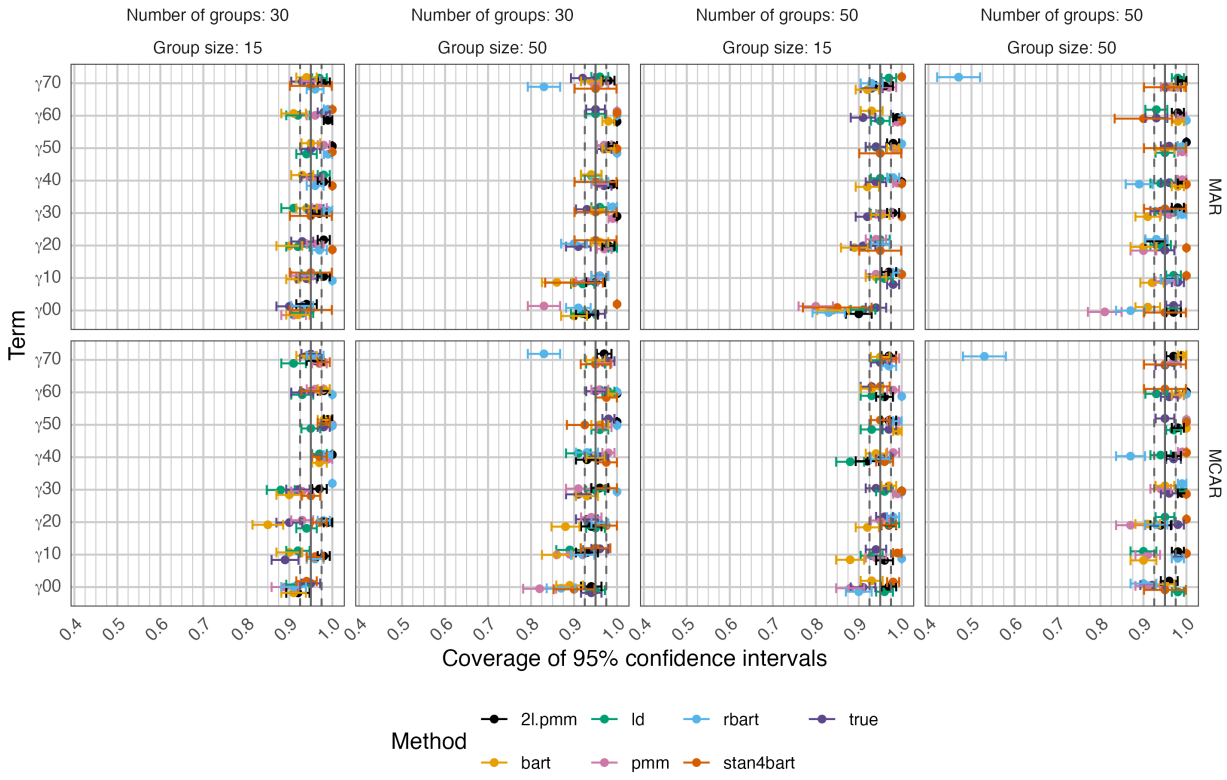


Figure 4: Coverage of the 95% confidence intervals of the intercept and level-1 effects of the linear mixed model with respective Monte Carlo SE for all simulated datasets over 100 simulations with ICC = .5. Method stan4bart is based on a differing number of dataset simulations described in table 2. The solid line represents the nominal 95% coverage, and the dashed lines at .925 and .975 represent the liberal criterion from Bradley (1978).

Lastly, the coverage of the level-2 effects and cross-level interactions in figure 5 shows that bart has the overall worst coverage, with coverages ranging from 65% to 82.5%. pmm also routinely undercovers the level-2 effects, which worsens with group sizes of 50. rbart tends to overcover the level-2 effects, but also undercovers them at times. Overall, 2l.pmm demonstrates the best coverage, despite exhibiting slight undercoverage when the data is MAR and there are 50 groups. stan4bart shows under- and overcoverage for the level-2 effects, but performs better with smaller groups and when the data is MCAR.

2l.pmm also has the best coverage of the cross-level interactions. Albeit showing under-

or overcoverage at times. Listwise deletion also performs considerably good, showing better coverage when the data is MAR compared to MCAR. When the group size is 15, pmm; bart; and rbart, show an acceptable coverage — sometimes slightly under- or overcovering the cross-level interactions —, but considerable undercoverage when the group size is increased to 50. stan4bart has the worst coverage of the cross-level interactions: showing considerable undercoverage, especially when the group size is increased to 50.

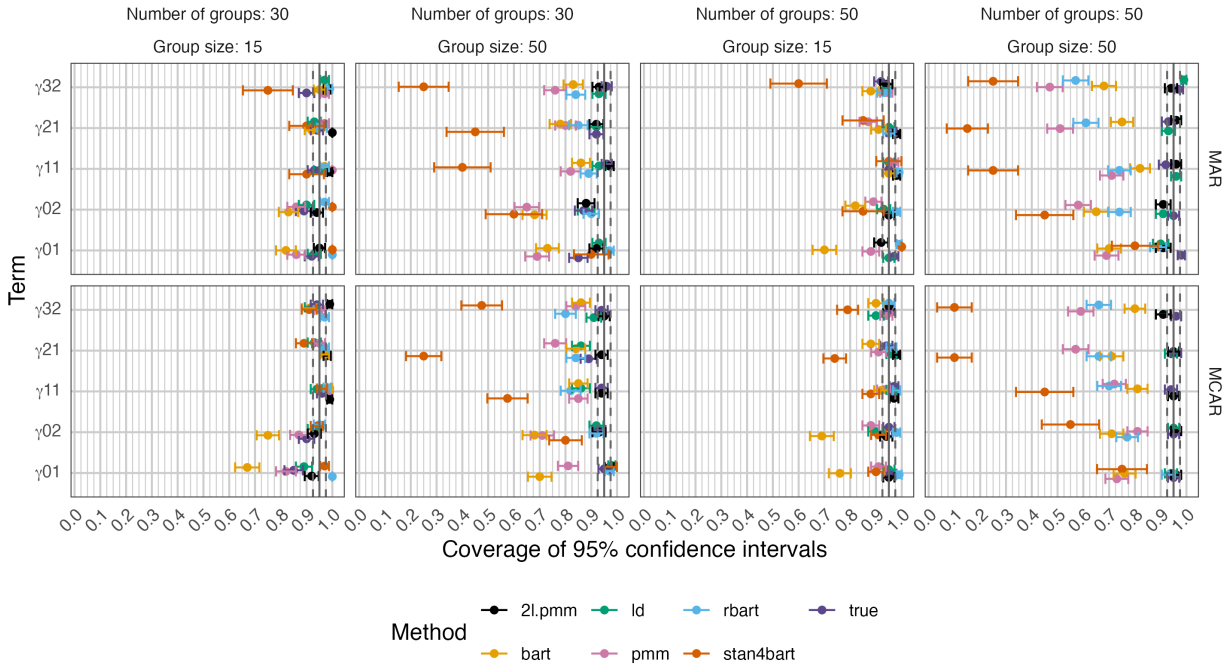


Figure 5: Coverage of the 95% confidence intervals of the level-2 and cross-level effects of the linear mixed model with respective Monte Carlo SE for all simulated datasets over 100 simulations with ICC = .5. Method stan4bart is based on a differing number of dataset simulations described in table 2. The solid line represents the nominal 95% coverage, and the dashed lines at .925 and .975 represent the liberal criterion from Bradley (1978).

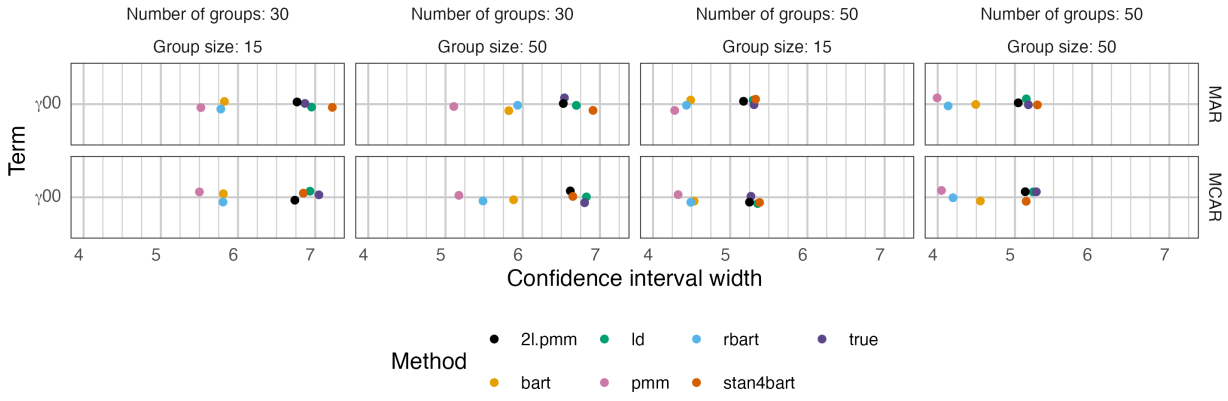
3.3. Confidence interval width

Figures 6 and 7 show the 95% confidence interval width (CIW) of the estimates of the fixed effects for all imputation methods in consideration.

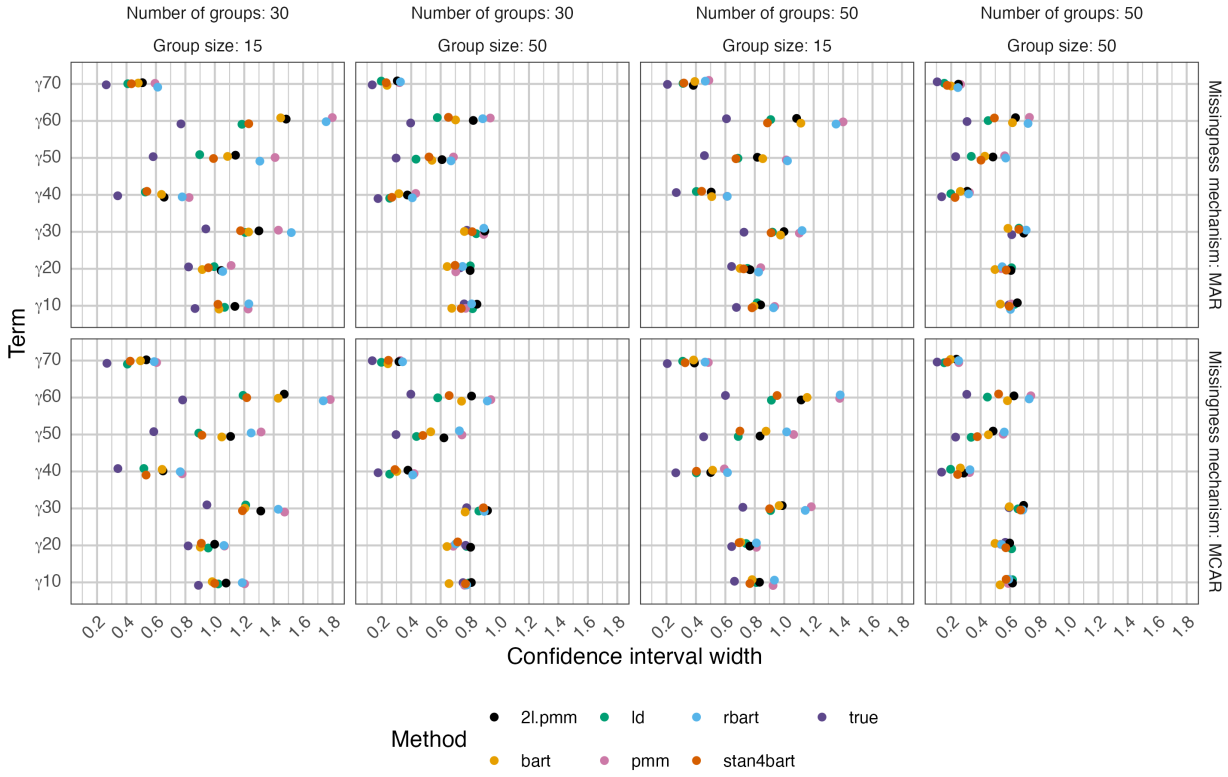
Figure 6a shows that the CIW of the intercept is smallest for pmm. However, paired with the coverage estimates from figure 4, pmm seems to be efficient but routinely undercovers the intercept. The same pattern can be seen for bart and rbart: showing smaller confidence intervals widths, they routinely undercover the intercept. 2l.pmm, stan4bart and listwise deletion show larger CIWs — somewhat mimicking the true data — oft paired with acceptable coverage.

For the level-1 effects, $\gamma_{10}, \gamma_{20}, \ldots, \gamma_{70}$, the CIW of the true data is oft smallest. Listwise deletion and stan4bart are closest in mimicking the width of the true data paired with acceptable coverages in figure 4. 2l.pmm and bart have overall slightly larger confidence

intervals, and, as mentioned before, 2l.pmm has the best coverage of the level-1 effects while bart shows more undercoverage. pmm and rbart show the largest confidence intervals. pmm shows both under- and overcoverage in figure 4. While rbart shows considerable undercoverage when its confidence intervals are smallest and overcoverage when its confidence intervals are largest — for example for effect γ_{70} and γ_{60} when the data is MAR with 50 groups of 50. Lastly, the CIW for all methods decreases with an increase in total sample size.



(a) Overall intercept



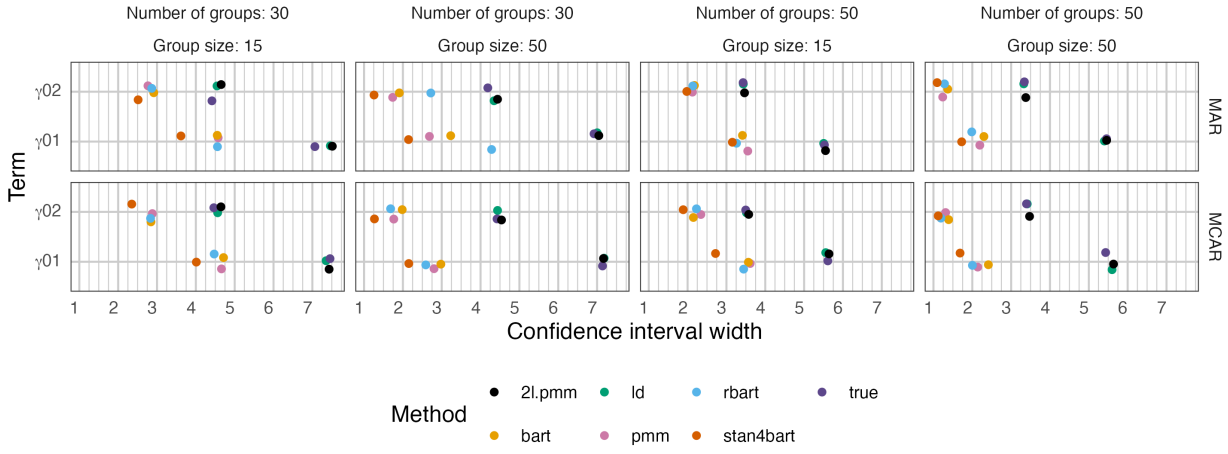
(b) Level-1 effects

Figure 6: Width of the 95% confidence intervals for the intercept and level-1 effects of the linear mixed model for all simulated datasets over 100 simulations with ICC = .5. Method stan4bart is based on a differing number of dataset simulations described in table 2.

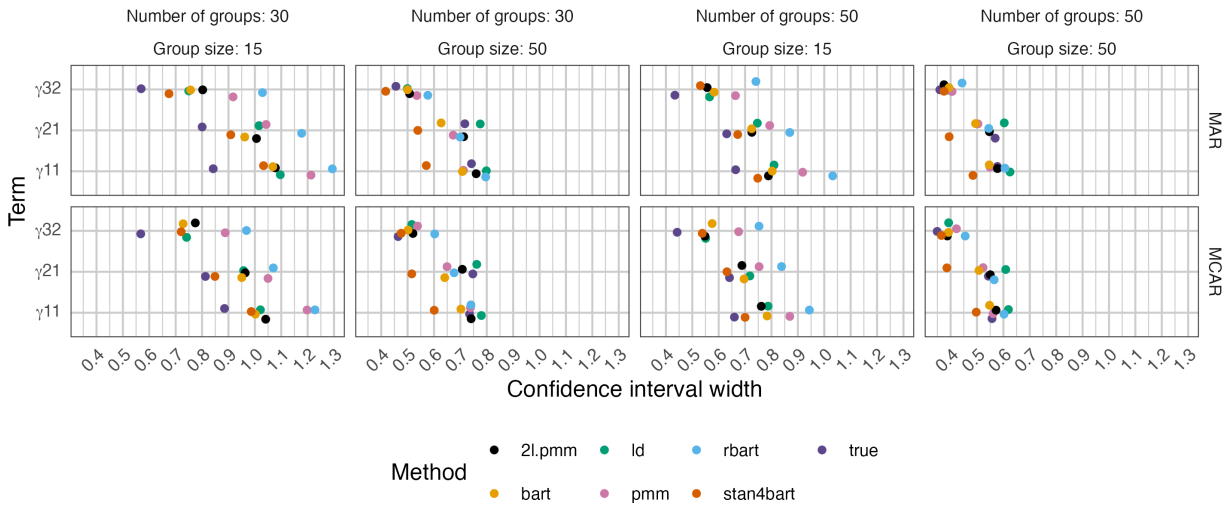
From figure 7a we can see that 2l.pmm and listwise deletion have largest CIWs for the level-2 effects, mimicking the true data. The other imputation methods, pmm; bart; rbart; and stan4bart, show smaller confidence intervals that decrease when the total sample size increases, leading to more undercoverage — as can be seen in figure 5. At the same time,

2l.pmm and listwise deletion show a decrease in width when there are more groups in the data, which is a pattern mirrored by the true data as well.

The width of the 95% confidence intervals for the cross-level interactions — shown in figure 7b show a similar pattern for all methods: when group sizes are 15, all methods have confidence intervals larger than the true data. However, when group sizes are 50, the confidence intervals decrease in size for all methods, either being smaller than or similar to the true data. This pattern is also reflected in figure 5, where these smaller intervals result in undercoverage.



(a) Level-2 effects



(b) Cross-level interactions

Figure 7: Width of the 95% confidence intervals for the level-2 and cross-level effects of the linear mixed model for all simulated datasets over 100 simulations with ICC = .5. Method stan4bart is based on a differing number of dataset simulations described in table 2.

4. Discussion

The goal of this study was to investigate whether the use of multilevel-BART (M-BART) models in MI could improve performance in the context of missing multilevel data. Even though MI has been implemented in a multilevel context (Audigier et al., 2018; Dong and Mitani, 2023; Enders et al., 2020, 2018a,b, 2016; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017a; Quartagno and Carpenter, 2022; Resche-Rigon and White, 2018; Taljaard et al., 2008; van Buuren, 2018), issues rise with the current implementation. Since MLMs are built step-wise with non-linearities, ensuring congeniality between the imputation model and analysis model is difficult. Additionally, mirroring the hierarchical structure of the data in the imputation models is also challenging (Burgette and Reiter, 2010; Hox and Roberts, 2011; van Buuren, 2018) and can result in complex models that might not converge (van Buuren, 2018). This study aimed to solve these problems by using BART models in the imputation process. These models — being non-parametric and tree-based — are able to implicitly model non-linearities and interactions (Breiman et al., 1984; Burgette and Reiter, 2010; Chipman et al., 2010; Hill et al., 2020; James et al., 2021; Lin and Luo, 2019; Salditt et al., 2023), alleviating the problems when defining a multilevel imputation model.

The simulations indicate that out of all investigated imputation methods, the straightforward multilevel imputation method — 2l.pmm — had the best overall performance. It showed the least overall bias and most consistent coverage. Furthermore, the width of its confidence intervals oftentimes mimicked the "true" data with 0% missing. On the other hand, the random effects — i.e. the random slopes, random intercept, and residual variance — and the cross-level interactions were often biased. Prior research has shown that passive imputation — where the transformation is done on-the-fly, so it can be included in the imputation model — yields biased estimates because it doesn't accurately reflect the complex joint distribution (Grund et al., 2018b; Seaman et al., 2012; Vink and Van Buuren, 2013). Furthermore, some coverage estimates for the fixed effects suggested some efficiency could still be gained. Also, 2l.pmm — which is based on `lme4::lme()`, a linear mixed model — might have an unfair advantage due to the data generating mechanism being based on a multivariate linear mixed model (Oberman and Vink, 2023).

Together with 2l.pmm, stan4bart was a model that could incorporate the most multilevel structure present in the data. While showing some promising results, stan4bart also showed some considerable biases and undercoverages. It seemed to be unable to recover the level-2 effects — fixed and cross-level. However, we need to consider that stan4bart, compared to 2l.pmm, didn't constrain the level-2 imputations to be the same within groups, thus estimating their variance inaccurately. We can see this, for stan4bart, in the too small estimated variance for the level-2 effects and the overestimated residual variance (van Buuren, 2018). Thus, possibly underestimating the level-2 effects and, as a consequence, the cross-level interactions. However, it did show the best performance in terms of the random structure of the data, outperforming 2l.pmm. Still, a major disadvantage of this method was its extensive computational time: the imputation of 20 datasets with 50 groups of size 50 took around 3

days to complete[3].

Most surprising was the performance of rbart, which is a model that should be able to account for some random structure in the data — namely, intercept variances. However, it had the worst overall performance. Remarkably, it was unable to capture the intercept variance — underestimating it. rbart only showed acceptable results for some — not all — fixed level-1 effects and was outperformed by bart at times. It would be interesting to evaluate the rbart imputation method with data only including random intercepts — since it is meant to account for that.

Unsurprisingly, single-level imputation methods — pmm and bart — were unable to accurately capture the multilevel structure of the data: underestimating level-2 effects, cross-level interactions and the random effects, while still showing acceptable results for the overall intercept and level-1 effects. The undercoverage of the overall intercept and level-1 effects γ_{10}, γ_{20} and γ_{30} as well as their inaccurate confidence interval widths are in line with prior research. Namely, ignoring the multilevel structure results in underestimating the ICC, standard errors, random intercepts, random slopes, and overestimating the residual variance (Enders et al., 2016; Hox and Roberts, 2011; Lüdtke et al., 2017; Taljaard et al., 2008; van Buuren, 2018).

Lastly, listwise deletion showed better results than expected: it outperformed 2l.pmm for most parameters. This would be expected under MCAR but not under MAR (Austin et al., 2021; Carpenter and Kenward, 2013; Enders et al., 2018b; Grund et al., 2018b, 2021; Little and Rubin, 2002; Lüdtke et al., 2017; Peeters et al., 2015; Schouten and Vink, 2021; van Buuren, 2018). Thus, indicating that a special case could have been generated — much like those mentioned in (van Buuren, 2018, §2.7) —, where listwise deletion outperforms the other imputation methods when the missing data was generated as MAR.

This study had a few limitations. Firstly, due to time restrictions paired with extensive computational time for the imputation methods, only 100 repetitions were used for evaluation. Morris et al. (2019) define a minimum of repetitions in a simulation study based on the required level of precision — MCSE — and expected coverage. For an MCSE of 0.5% and expected coverage of 95%, they pose that 1900 repetitions are needed. As a result, especially our 95% confidence interval coverage estimates are variable. Secondly, in the current implementation, the bart imputation models — single-level bart, rbart and stan4bart — are computationally expensive; taking, at the least, several hours to impute one dataset. Then, this study considered a limited number of factors resulting in a limited picture of the performance of the imputation methods. For example, the amount of missingness was fixed at 50% and the ICC at 0.5 and, according to previous methodological research, these factors can influence the performance of the imputation methods (Akkaya Hocagil and Yucel, 2023; Enders et al., 2020, 2018a,b; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017b). Lastly, the generated MAR mechanism did not mimic the expected characteristic associated with MAR. So, the performance of the imputation methods were not evaluated under a strong MAR mechanism, which is common when evaluating imputation methods (Austin et al., 2021; Carpenter and Kenward, 2013; Enders et al., 2018b; Grund

[3]2 x Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, 48 cores, 64GB Mem, Nvidia GTX 1080Ti

et al., 2018b, 2021; Little and Rubin, 2002; Lüdtke et al., 2017; Peeters et al., 2015; Schouten and Vink, 2021; van Buuren, 2018). In addition to important because, with real-life data, MCAR can rarely be assumed (Kang, 2013; Little and Rubin, 2002; Oberman and Vink, 2023; van Buuren, 2018).

So, avenues for future research include exploring the performance of the M-BART model — stan4bart — under a stricter MAR mechanism, or, a MNAR mechanism, which Oberman and Vink (2023) pose to be a likely real-life missingness mechanism. Thus, for a method to perform well in real-life situations, it should preferably perform well under both MAR and MNAR (Oberman and Vink, 2023). Furthermore, the robustness of stan4bart could be evaluated under different amounts of missingness, different ICCs, and more differing sample sizes to reflect other realistic situations. For example, group sizes of 5 are often found in family-type research (Maas and Hox, 2005). Also, Lower ICCs, lower sample sizes, and a higher amount of missingess could introduce other biases and complexities (Akkaya Hocagil and Yucel, 2023; Enders et al., 2020, 2018a,b; Grund et al., 2016, 2018a,b, 2021; Lüdtke et al., 2017; Mistler and Enders, 2017b). Research could be conducted into the reduction of computational time for stan4bart: reducing the number of posterior samples, thinning samples, burn-in samples, or other specifications for the stan4bart method. Then, lastly, the matching procedure for the predictive mean matching step could be explored. Research could examine how different matching methods to better implement the sampling variability (van Buuren, 2018) could be implemented and how they would affect the performance of the stan4bart method.

5. Conclusion

To sum up, this study aimed to ease the complexities associated with multiple imputation in a multilevel context by using BART models as the imputation models. Being non-parametric and tree-based, BART models were expected to make defining a congenial and well performing imputation model for multilevel data easier. However, the current implementation of a multilevel-BART imputation method did not improve on the current implementation of multilevel predictive mean matching. Meanwhile, also showing inviting results for future research into the multilevel-BART imputation method.

References

Akkaya Hocagil, T. and Yucel, R. M. (2023). A computationally efficient sequential regression imputation algorithm for multilevel data. *Journal of Applied Statistics*, pages 1–21.

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2).

Austin, P. C., White, I. R., Lee, D. S., and Van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, 37(9):1322–1331.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and for the Alzheimer's Disease Neuroimaging Initiative* (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2):144–152.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Routledge, 1 edition.

Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9):1070–1076.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). **Mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. Wiley, 1 edition.

Chen, S. (2020). *A New Multilevel Bayesian Nonparametric Algorithm and Its Application in Causal Inference*. PhD thesis, Texas A&M University.

Chipman, H., George, E., and Mcculloch, R. (2006). Bayesian Ensemble Learning. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

Cohen, J. (1990). Statistical power analysis for the behavioral sciences. *Computers, Environment and Urban Systems*, 14(1):71.

Dong, M. and Mitani, A. (2023). Multiple imputation methods for missing multilevel ordinal outcomes. *BMC Medical Research Methodology*, 23(1):112.

Dorie, V. (2023). *Stan4bart: Bayesian Additive Regression Trees with Stan-Sampled Parametric Extensions*.

Dorie, V., Chipman, H., McCulloch, R., Dadgar, A., Team, R. C., Draheim U., G., Bosmans, M., Tournayre, C., Petch, M., Valle, R. d. L., Johnson G., S., Frigo, M., Zaitseff, J., Veldhuizen, T., Maisonobe, L., Pakin, S., and Daniel G., R. (2024). Dbarts: Discrete Bayesian Additive Regression Trees Sampler.

Dorie, V., Perrett, G., Hill, J. L., and Goodrich, B. (2022). Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning. *Entropy*, 24(12):1782.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98:4–18.

Enders, C. K., Du, H., and Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, 25(1):88–112.

Enders, C. K., Hayes, T., and Du, H. (2018a). A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. *Multivariate Behavioral Research*, 53(5):695–713.

Enders, C. K., Keller, B. T., and Levy, R. (2018b). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23(2):298–317.

Enders, C. K., Mistler, S. A., and Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21(2):222–240.

Finch, J. F., West, S. G., and MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2):87–107.

Grund, S., Lüdtke, O., and Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, 48(2):640–649.

Grund, S., Lüdtke, O., and Robitzsch, A. (2018a). Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *Journal of Educational and Behavioral Statistics*, 43(3):316–353.

Grund, S., Lüdtke, O., and Robitzsch, A. (2018b). Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organizational Research Methods*, 21(1):111–149.

Grund, S., Lüdtke, O., and Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, 53(6):2631–2649.

Gulliford, M., Adams, G., Ukoumunne, O., Latinovic, R., Chinn, S., and Campbell, M. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58(3):246–251.

Gulliford, M. C., Ukoumunne, O. C., and Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, 149(9):876–883.

Hastie, T. J., editor (2017). *Statistical Models in S*. Routledge, 1st edition.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

Hox, J. and Roberts, J. K., editors (2011). *Handbook of Advanced Multilevel Analysis*. Routledge, 0 edition.

Hox, J. J., Moerbeek, M., and Van De Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Routledge, Third edition. | New York, NY : Routledge, 2017. |, 3 edition.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1):28.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.

Kreft, I. and de Leeuw, J. (2007). *Introducing Multilevel Modeling*. Introducing Statistical Methods. SAGE, Los Angeles, Calif., reprinted edition.

Lee, D., Carpenter, B., Li, P., Morris, M., Betancourt, M., Maverickg, Brubaker, M., Trangucci, R., Inacio, M., Kucukelbir, A., Buildbot, S., Bgoodri, Seantalts, Arnold, J., Tran, D., Hoffman, M., Margossian, C., Modrák, M., Adler, A., Sakrejda, K., Stukalov, A., Lawrence, M., Goedman, R. J., Van Horn, K. S., Vehtari, A., Gabry, J., Casallas, J. S., and Bales, B. (2017). Stan-dev/stan: V2.17.1. Zenodo.

Lin, S. and Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 54(4):578–592.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Lüdtke, O., Robitzsch, A., and Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, 22(1):141–165.

Maas, C. J. M. and Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, 1(3):86–92.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, pages 538–558.

Mistler, S. A. and Enders, C. K. (2017a). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466.

Mistler, S. A. and Enders, C. K. (2017b). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics*, 42(4):432–466.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Murray, D. M. and Blitstein, J. L. (2003). Methods To Reduce The Impact Of Intraclass Correlation In Group-Randomized Trials. *Evaluation Review*, 27(1):79–103.

Oberman, H. I. and Vink, G. (2023). Toward a standardized evaluation of imputation methodology. *Biometrical Journal*, page 2200107.

Peeters, M., Zondervan-Zwijnenburg, M., Vink, G., and Van De Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, 12(4):377–394.

Quartagno, M. and Carpenter, J. R. (2022). Substantive model compatible multilevel multiple imputation: A joint modeling approach. *Statistics in Medicine*, 41(25):5000–5015.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649.

Rights, J. D. and Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24(3):309–338.

Robitzsch, A., Simon Grund, and Henke, T. (2024). Miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Salditt, M., Humberg, S., and Nestler, S. (2023). Gradient Tree Boosting for Hierarchical Data. *Multivariate Behavioral Research*, pages 1–27.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.

Schouten, R. M. and Vink, G. (2021). The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions. *Sociological Methods & Research*, 50(3):1243–1258.

Seaman, S. R., Bartlett, J. W., and White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, 12(1):46.

Shieh, G. (2012). A comparison of two indices for the intraclass correlation coefficient. *Behavior Research Methods*, 44(4):1212–1223.

Silva, G. C. and Gutman, R. (2022). Multiple imputation procedures for estimating causal effects with multiple treatments with application to the comparison of healthcare providers. *Statistics in Medicine*, 41(1):208–226.

Taljaard, M., Donner, A., and Klar, N. (2008). Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal*, 50(3):329–345.

Tan, Y. V., Flannagan, C. A. C., and Elliott, M. R. (2016). Predicting human-driving behavior to help driverless vehicles drive: Random intercept Bayesian Additive Regression Trees.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition.

Vink, G. and Van Buuren, S. (2013). Multiple Imputation of Squared Terms. *Sociological Methods & Research*, 42(4):598–607.

Wagner, J., West, B. T., Elliott, M. R., and Coffey, S. (2020). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics*, 36(4):907–931.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ*

Open, 3(8):e002847.

Wundervald, B., Parnell, A., and Domijan, K. (2022). Hierarchical Embedded Bayesian Additive Regression Trees.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602.

Appendix A. Imputation functions

Appendix A.1. BART

Listing 1: Imputation function for single-level BART

```
mice.impute.bart <- function(y, ry, x, wy = NULL, use.matcher = FALSE,
donors = 5L, ...) {
    install.on.demand("dbarts", ...)
    if (is.null(wy)) {
        wy <- !ry
    }

    # Parameter estimates
    fit <- dbarts::bart(x, y, keeptrees = TRUE, verbose = FALSE)

    yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
    yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

    # Find donors
    if (use.matcher) {
        idx <- matcher(yhatobs, yhatmis, k = donors)
    } else {
        idx <- matchindex(yhatobs, yhatmis, donors)
    }

    return(y[ry][idx])
}
```

Appendix A.2. R-BART

Listing 2: Imputation function for random intercept BART

```
mice.impute.2l.rbart <- function(y, ry, x, wy = NULL, type, use.
matcher = FALSE, donors = 5L, ...) {
    install.on.demand("dbarts", ...)
    if (is.null(wy)) {
        wy <- !ry
    }

    clust <- names(type[type == -2])
    effects <- names(type[type != -2])
    X <- x[, effects, drop = FALSE]

    model <- paste0(
        "y ~ ", paste0(colnames(X), collapse = " + ")
    )

    fit <- dbarts::rbart_vi(formula = formula(model), group.by = clust
, data = data.frame(y, x), verbose = FALSE, n.threads = 1, n.samples =
500L, n.burn = 500L, ...)

```

```
        yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
        yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

        # Find donors
        if (use.matcher) {
            idx <- matcher(yhatobs, yhatmis, k = donors)
        } else {
            idx <- matchindex(yhatobs, yhatmis, donors)
        }

        return(y[ry][idx])
    }
```

Appendix A.3. stan4bart

Listing 3: Imputation function for multilevel-BART with random effects and cross-level interactions

```
mice.impute.2l.bart <- function(y, ry, x, wy = NULL, type, intercept =
TRUE, use.matcher = FALSE, donors = 5L, ...) {
    install.on.demand("stan4bart", ...)
    if (is.null(wy)) {
        wy <- !ry
    }

    if (intercept) {
        x <- cbind(1, as.matrix(x))
        type <- c(2, type)
        names(type)[1] <- colnames(x)[1] <- "(Intercept)"
    }

    clust <- names(type[type == -2])
    rande <- names(type[type == 2])
    fixe <- names(type[type > 0])

    lev <- unique(x[, clust])

    X <- x[, fixe, drop = FALSE]
    Z <- x[, rande, drop = FALSE]
    xobs <- x[ry, , drop = FALSE]
    yobs <- y[ry]
    Xobs <- X[ry, , drop = FALSE]
    Zobs <- Z[ry, , drop = FALSE]

    # create formula
    fr <- ifelse(length(rande) > 1,
        paste0("+ (1 +", paste(rande[-1L], collapse = "+")),
        " + (1 "
    )
    randmodel <- paste0(
        "y ~ bart(", paste0(fixe[-1L], collapse = " + "), ")",
        fr, "| ", clust, ")"
    )
```

```
    fit <- eval(parse(text = paste("stan4bart::stan4bart(", randmodel,
        ", data = data.frame(y, x),
            verbose = -1,
            bart_args = list(k = 2.0, n.samples = 500L, n.burn = 500L, n.
    thin = 1L, n.threads = 1))",
        collapse = ""
    )))

        yhatobs <- fitted(fit, type = "ev", sample = "train")[ry]
        yhatmis <- fitted(fit, type = "ev", sample = "train")[wy]

        # Find donors
        if (use.matcher) {
            idx <- matcher(yhatobs, yhatmis, k = donors)
        } else {
            idx <- matchindex(yhatobs, yhatmis, donors)
        }

        return(y[ry][idx])
    }
```