MULTIVARIATE BEHAVIORAL RESEARCH
2022, VOL. 57, NOS. 2–3, 513–523
https://doi.org/10.1080/00273171.2021.1912582

Multiple imputation in data that grow over time: a comparison of three strategies

X. M. Kavelaars[a,b], J. R. van Ginkel[c], and S. van Buuren[a,d]

[a]Department of Methodology and Statistics, Utrecht University, The Netherlands; [b]Department of Methodology and Statistics, Tilburg University, The Netherlands; [c]Department of Psychology, Methodology and statistics, Leiden University, The Netherlands; [d]The Netherlands Organization for Applied Scientific Research, The Netherlands

ABSTRACT

Multiple imputation is a recommended technique to deal with missing data. We study the problem where the investigator has already created imputations before the arrival of the next wave of data. The newly arriving data contain missing values that need to be imputed. The standard method (RE-IMPUTE) is to combine the new and old data before imputation, and re-impute all missing values in the combined data. We study the properties of two methods that impute the missing data in the new part only, thus preserving the historic imputations. Method NEST multiply imputes the new data conditional on each filled-in old data $m_2 > 1$ times. Method APPEND is the special case of NEST with $m_2 = 1$, thus appending each filled-in data by single imputation. We found that NEST and APPEND have the same validity as RE-IMPUTE for monotone missing data-patterns. NEST and APPEND also work well when relations within waves are stronger than between waves and for moderate percentages of missing data. We do not recommend the use of NEST or APPEND when relations within time points are weak and when associations between time points are strong.

KEYWORDS
Missing data; multiple imputation; nested imputation; congeniality

1. Introduction

Missing data are inevitable in empirical studies and require careful attention. When not appropriately handled, missing data can seriously affect the validity of statistical inference. Multiple imputation (Rubin, 1987) is a widely accepted technique to obtain valid inferences from incomplete data. The procedure replaces each missing value by several plausible values, thereby creating $m > 1$ completed datasets. The analyst estimates the parameters of scientific interest from each imputed dataset by conventional complete-data methods, and pools these estimates to their final point and interval values.

Longitudinal studies collect data on the same persons in multiple waves. While we consider the basic case with just two waves, wave 1 and wave 2, the problem and its solutions generalize to multiple waves. Suppose that the investigator had already imputed the missing data up to wave 1, and that the dataset for the new wave 2 becomes available. There are then three basic possibilities for imputing the latest data:

1. We re-impute the entire dataset m times thus overwriting any imputations we had in the data up to wave 1 (RE-IMPUTE);
2. We treat each of the m_1 imputed datasets up to wave 1 as complete and multiply-imputed the latest part m_2 times, resulting in $m_1 \times m_2$ nested imputed datasets (NEST);
3. The same as option 2, but then setting $m_2 = 1$, resulting in $m = m_1$ imputed datasets (APPEND).

Method RE-IMPUTE is relatively straightforward and has known statistical properties. The main downside of RE-IMPUTE is that it will create new imputations each time a new wave arrives. For reasons of reproducibility, the database manager may need to store different versions of the imputed data, which may be challenging. Additionally, two identical statistical analysis models result in non-identical point estimates (and potentially also in non-identical conclusions) when their imputation models rely on

CONTACT X. M. Kavelaars x.m.kavelaars@tilburguniversity.edu

different sets of variables. By definition, replicability of analyses from previous waves under RE-IMPUTE depend on the inclusion of future variables. Note however, that it may be defensible to update estimates when new data arrive.

Methods NEST and APPEND both preserve the existing imputations of earlier waves. Rubin (2003, p. 6) suggested the NEST method for applications "where some part of the imputation process is extremely expensive to implement, and the other part relatively inexpensive." It may, for example, be expensive to RE-IMPUTE all historical data of yearly (large scale) surveys when a new wave of data becomes available and to store all historical versions of the imputed data for the sake of reproducibility. For longitudinal data, both NEST and APPEND bypass the need for multiple versions of multiply imputed datasets. However, NEST leads to an expansion in the number of imputations. Moreover, NEST requires special pooling rules because imputations within a nest are correlated. Method APPEND solves both problems and has been used in practice (e.g. Aardoom et al., 2016a, 2016b). However, it is unknown how useful the imputations are, and what the effect is on the validity of the inferences. The objective of the current research is to find out which methods are safe to use when the data grow over time.

2. Problem illustration

The Project on Preterm and Small for Gestational Age Infants (POPS) (Veen et al., 1991; Verloove-Vanhorick et al., 1986) study included about 94% of all children born in 1983 in the Netherlands with a birth weight below 1500 g or a gestational age below 32 weeks ($n = 1338$). The study followed these children at various ages (e.g. 1, 5, 10, 14, 19, 28 and 35 years) to measure physical, cognitive and psychosocial outcomes.

Of the $n = 959$ surviving participants, $n = 596$ completed participation at age 19 and those who dropped out differed systematically from full responders in health (Hille, 2005). Van Buuren (2018, chap. 10) multiply imputed the block of 363 missing children using data from all previous waves. These imputations confirmed the suspected existence of selective dropout. Of course, there are also missing data in later waves at 28 and 35 years. If we impute these later waves, should we re-impute the block of 363 children? What happens if we preserve these imputations?

In general, repeating previous analyses with newly imputed data may result in different parameter estimates or conclusions. We demonstrate this with an adapted example from Van Dommelen et al. (2014). The researchers used the POPS data to predict the effect of early catch-up growth (developing toward the median of the growth charts in the first year of life) on health and well-being in young adulthood from the POPS data. Multiple imputation of the data available at age 14 showed that catch-up growth in weight did not predict length and weight at age 14.

However, when we ran the same analysis, but now with data from wave at age 19 included in the imputation model, we found an opposite effect. Catch-up growth did predict length and weight at age 14[1]. This example demonstrates that RE-IMPUTE potentially lacks replicability and may even lead to different conclusions. Thus, even when using the same data, the attentive reader may raise the question which of these analyses should be trusted.

We will return to the POPS dataset in the five section and evaluate the three strategies.

3. Methods

Let $Y^{(1)}$ the $n \times p_1$ matrix of partially observed data collected at wave 1. We use multiple imputation to obtain m imputed data sets $Y_\ell^{(1)}|Y_{\text{obs}}^{(1)}$ for the wave-1 data. Here $Y_{\text{obs}}^{(1)}$ are the observed data in $Y^{(1)}$ and $\ell = 1, \ldots, m$. Matrix $Y^{(2)}$ holds the $n \times p_2$ incomplete data collected at wave 2 on the same n subjects. Then, how should we impute the missing data in $Y^{(2)}$? The simplest method would be to replicate the previous method, thus yielding m imputed datasets $Y_\ell^{(2)}|Y_{\text{obs}}^{(2)}$. This approach treats $Y^{(2)}$ in complete isolation of $Y^{(1)}$. We may defend this approach if we are only interested in the imputed $Y^{(2)}$ data. The rationale for longitudinal studies is, however, making inferences across different waves. Simply concatenating the imputed data $(Y_\ell^{(1)}|Y_{\text{obs}}^{(1)}, Y_\ell^{(2)}|Y_{\text{obs}}^{(2)})$ will not work. The analysis will attenuate any estimates involving both $Y^{(1)}$ and $Y^{(2)}$ because we imputed $Y^{(1)}$ and $Y^{(2)}$ as if they were unrelated. The imputation model is more restrictive than the analysis model. This condition, known as uncongeniality (Meng, 1994), leads to biased estimates.

From a theoretical point of view, the preferred alternative is to concatenate $Y = (Y^{(1)}, Y^{(2)})$ and obtain imputed data sets $Y_\ell|Y_{\text{obs}}$ based on the combined wave-1 and wave-2 data Y. This RE-IMPUTE method can adequately take account of the relations across waves, and - when properly done - leads to unbiased and efficient estimates and correct confidence intervals. RE-IMPUTE creates new imputations

[1] $b_{\text{length 14}} = 0.25$, 95% CI $(-0.10 - 0.60)$, $b_{\text{weight 14}} = 0.21$, 95% CI $(-0.04 - 0.47)$, $b_{\text{length 19}} = 0.32$, 95% CI $(0.01 - 0.62)$, $b_{\text{weight19}} = 0.33$, 95% CI $(0.09 - 0.56)$.

Table 1. Pooling rules for multiple imputation with independent datasets (re-imputation and appended imputation) (Rubin, 1987) and two-level nested datasets (nested imputation) (Rubin, 2003; Shen, 2000).

Parameter	Independent datasets	Nested datasets
Mean estimate	$\bar{Q} = \frac{1}{m_i}\sum_{k=1}^{m_i} \hat{Q}^{(k)}$	$\bar{Q} = \frac{1}{m_1 m_2}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} \hat{Q}^{(k,l)}$
Nest mean	–	$\bar{Q}_k = \frac{1}{m_2}\sum_{l=1}^{m_2} \hat{Q}^{(k,l)}$
Sampling variance	$\bar{U} = \frac{1}{m_i}\sum_{k=1}^{m_i} \bar{U}_k.$	$\bar{U} = \frac{1}{m_1 m_2}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} \bar{U}^{(k,l)}$
Within variance	–	$W = \frac{1}{m_1(m_2-1)}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} (\hat{Q}^{(k,l)} - \bar{Q}_k)^2$
Between variance	$B = \frac{1}{m_i - 1}\sum_{k=1}^{m_i} (\hat{Q}_k - \bar{Q})^2.$	$B = \frac{m_2}{m_1 - 1}\sum_{k=1}^{m_1} (\bar{Q}_k - \bar{Q})^2$
Total variance	$T = \bar{U} + \left(1 + \frac{1}{m_i}\right)B.$	$T = \bar{U} + \frac{1}{m_2}\left(1 + \frac{1}{m_1}\right)B + \left(1 - \frac{1}{m_2}\right)W$
Degrees of freedom	$\nu^{-1} = \frac{1}{m_i - 1}\left[\frac{(1+\frac{1}{m_i})B}{T}\right]^2$	$\nu^{-1} = \frac{1}{m_1 - 1}\left[\frac{\frac{1}{m_2}(1+\frac{1}{m_1})B}{T}\right]^2 + \frac{1}{m_1(m_2-1)}\left[\frac{(1-\frac{1}{m_2})W}{T}\right]^2$

Note. $\hat{Q}^{(k,l)}$ and $\bar{U}^{(k,l)}$ represent the complete data estimate and sampling variance for dataset *l* in nest *k*, respectively.

$Y_\ell^{(1)}|Y_{\text{obs}}$, which now also incorporates future data $Y^{(2)}$ from wave 2. These future data might provide important information for imputation of previous waves, particularly when the previous data are MAR (Missing at Random) on future information. Models fitted to $Y_\ell^{(1)}|Y_{\text{obs}}^{(1)}$ differ from models fitted to $Y_\ell^{(1)}|Y_{\text{obs}}$, potentially leading to different inferences, as the example in the Section 2 illustrates.

Re-imputation may not always be desirable. Reasons for fixing the imputes of $Y^{(1)}$ include the reduction of work, the improvement of reproducibility, and the evasion to store multiple versions of the imputed values. In some cases, we may want our models for $Y^{(1)}$ to be blind to any future data. For example, imputing missing data in an individual risk prediction model during wave one may use other wave-1 data, but must ignore the not-yet-available wave-2 outcomes. In such cases, it makes sense to constrain the imputation model to the wave-1 information only. The NEST and APPEND method fix the wave-1 imputations.

The NEST imputation method conditions on both $Y_\ell^{(1)}|Y_{\text{obs}}^{(1)}$ and $Y_{\text{obs}}^{(2)}$. For every $\ell = 1, ..., m_1$, NEST produces m_2 imputed data sets $Y_\jmath^{(2)}$ with $\jmath = 1, ..., m_2$, thus resulting in $m_1 \times m_2$ imputed data sets. The variance of parameter estimates must correctly reflect the extra uncertainty of adopting imputations as data (Rubin, 1987). The pooling procedure therefore accommodates differences between imputed datasets, such that parameter estimates are confidence-valid. While re-imputation uses regular pooling rules for multiple imputation, nested multiple imputation requires specific pooling rules that respect the nested data structure (Rubin, 2003; Shen, 2000). These pooling rules for two-level nested datasets are more complex than the standard pooling rules for non-nested data, as shown in Table 1 (Shen, 2000).

The APPEND imputation method is similar to NEST, but with $m_2 = 1$. The latter setting has several effects. The number of imputed datasets will remain constant as we add waves. We may use the conventional pooling rules during analysis. And finally, APPEND may take less work than RE-IMPUTE or NEST. On the other hand, APPEND may be unable to propagate uncertainty correctly, potentially leading to confidence intervals that are too short.

The relative pros and cons of the NEST and APPEND methods are not yet well understood. An essential theoretical result (Rubin, 1987) is that we can impute variables that have a monotone missing-data pattern sequentially without the need to iterate. More in particular, if all subjects with observed wave-2 data have complete wave-1 data, then the missing-data pattern is monotone. Monotone missing-data patterns are not uncommon in longitudinal data, as they result from panel attrition and drop-out. If the missing data are monotone, then NEST and APPEND will be as good as RE-IMPUTE. In that case, we may generate the imputations very fast with one pass through the data.

In contrast, under non-monotone missingness, there are subjects with missing wave-1 data and observed wave-2 data. RE-IMPUTE will impute the missing wave-1 data given the wave-2 information, possibly leading to sharper inferences for analyses involving only wave-1 data (Xie & Meng, 2016). On the other hand, NEST and APPEND will preserve any previously made imputations, which may attenuate the later parameter estimates involving wave-1 and wave-2 data. Attenuation is likely to be more severe if relations among wave-1 variables are weaker, and associations across wave-1 and wave-2 are stronger.

In summary, methods NEST and APPEND make more restrictive assumptions about the missingness pattern of incomplete data than RE-IMPUTE. Also, APPEND may underestimate the variance. The simulation study in the following section will highlight the circumstances in which NEST and APPEND may become problematic.

4. Simulation study

4.1. Objectives

The simulation studies address the following questions:

1. In which situations will NEST and APPEND affect the validity of statistical inferences?
2. Do the three methods perform similarly under a monotone missing data mechanism?
3. Does APPEND result in confidence–valid parameter estimates?

We evaluated the performance of the imputation methods with two different models:

1. A relatively simple situation with a linear regression model based on two timepoints which clarifies the influence of the missingness pattern and the correlation structure within the data on the validity of each method;
2. A more complex growth model with time-varying covariate to explore how performance generalizes to the multilevel context and to more timepoints.

These two models shed light on the potential of RE-IMPUTE, NEST, and APPEND in a variety of longitudinal applications.

4.2. Linear regression model

4.2.1. Simulation setup

We evaluated the performance of RE-IMPUTE, NEST and APPEND in a two-stage setup, resembling a longitudinal design with two waves. The wave-1 data consisted of two incomplete covariates: x_1 and y_1. The wave-2 data had two incomplete variables x_2 and y_2.

Table 2. Different correlation structures used in the simulation study.

	ρ_{within}		ρ_{between}			
Scenario	$\rho_{x_1y_1}$	$\rho_{x_2y_2}$	$\rho_{x_1x_2}$	$\rho_{y_1y_2}$	$\rho_{y_1x_2}$	$\rho_{x_1y_2}$
1.	0.10	0.10	0.10	0.10	0.10	0.10
2.	0.10	0.10	0.30	0.30	0.30	0.30
3.	0.10	0.10	0.50	0.50	0.50	0.50
4.	0.10	0.10	0.70	0.70	0.70	0.70
5.	0.30	0.30	0.10	0.10	0.10	0.10
6.	0.30	0.30	0.30	0.30	0.30	0.30
7.	0.30	0.30	0.50	0.50	0.50	0.50
8.	0.30	0.30	0.70	0.70	0.70	0.70
9.	0.50	0.50	0.10	0.10	0.10	0.10
10.	0.50	0.50	0.30	0.30	0.30	0.30
11.	0.50	0.50	0.50	0.50	0.50	0.50
12.	0.50	0.50	0.70	0.70	0.70	0.70
13.	0.70	0.70	0.10	0.10	0.10	0.10
14.	0.70	0.70	0.30	0.30	0.30	0.30
15.	0.70	0.70	0.50	0.50	0.50	0.50
16.	0.70	0.70	0.70	0.70	0.66	0.66

We manipulated the correlation structure and the missingness pattern of the data. Since the strength of the relation between wave-1 and wave-2 data is crucial for valid inference, we set correlations between variables within waves ($\rho_{\text{within}} = \rho_{x_1y_1}, \rho_{x_2y_2}$) and correlations between variables between waves ($\rho_{\text{between}} = \rho_{x_1x_2}, \rho_{x_1y_2}, \rho_{y_1x_2}, \rho_{y_1y_2}$). We specified ρ_{within} and ρ_{between} at four different values: $0.10, 0.30, 0.50,$ and 0.70, which resulted in a 4×4 factorial design of 16 correlation structures presented in Table 2. Note that we slightly lowered $\rho_{y_1x_2}$ and $\rho_{x_1y_2}$ in condition 16 to ensure a positive definite covariance matrix.

4.2.1.1. *Data generation.* For each correlation structure, we drew 2000 samples from the multivariate standard normal distribution with a sample size of $n = 425$. This sample size is the minimum sample size required to detect the smallest specified population effect with 80% power at the $\alpha = 0.05$ level.

4.2.1.2. *Missingness and imputation procedure.* We made the data incomplete under monotone and non-monotone missingness (see Table 3). Since the missingness mechanism was not the focus of this study, we created missing data under one missingness mechanism only. For simplicity, we implemented a Missing Completely at Random (MCAR) mechanism and make use of the theoretical result that validity of multiple imputation under MCAR and MAR are known to be similar (e.g. Rubin, 1976). Each combination of missing values had the same probability. Every sample had a total of either 20% or 50% incomplete cases.

Table 3. Implemented combinations of complete (1) and incomplete (0) variables for monotone and non-monotone missing data patterns.

	Wave 1		Wave 2		Monotone	Non-monotone
	x_1	y_1	x_2	y_2		
1.	1	1	1	1	✓	✓
2.	1	1	1	0	✓	✓
3.	1	1	0	1	✓	✓
4.	1	1	0	0	✓	✓
5.	1	0	1	1		✓
6.	1	0	1	0		✓
7.	1	0	0	1		✓
8.	1	0	0	0	✓	✓
9.	0	1	1	1		✓
10.	0	1	1	0		✓
11.	0	1	0	1		✓
12.	0	1	0	0	✓	✓

We imputed the 2000 samples with Bayesian normal linear regression using the RE-IMPUTE, NEST and APPEND methods. Under RE-IMPUTE, we imputed datasets with incomplete wave-1 and wave-2 data $m = 5$ times. For NEST and APPEND, we imputed each wave-1 dataset $m_1 = 5$ times and then concatenated the incomplete wave-2 data to each completed wave-1 dataset. We imputed these partially completed datasets $m_2 = 5$ (NEST) and $m_2 = 1$ (APPEND) times. Thus, we obtained five imputed datasets after RE-IMPUTE and APPEND, and 25 datasets after NEST.

After imputation, we fitted a linear regression model on each completed dataset:

$$\hat{y}_2 = b_0 + b_{x_1}x_1 + b_{y_1}y_1 + b_{x_2}x_2$$

The results were pooled using the method-specific pooling rules presented in Table 1.

4.2.1.3. Outcomes of interest.

Quantities of interest were variable means and regression coefficients of the fitted model. True values were population means $\boldsymbol{\mu}$ and population regression coefficients derived from the correlation structure of the data.

We considered imputations as valid if pooled parameter estimates were unbiased, and if the coverage of the confidence intervals was at a nominal 95% level. The relative efficiency referred to the width of the 95% confidence interval compared to the other imputation strategies.

We performed the simulation study in `R` (R Core Team, 2016). Data amputations and imputations were performed with `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) using method *norm()* with 25 iterations (Oberman et al., 2020; Van Buuren, 2018).

4.2.2. Results

4.2.2.1. Validity.

For 50% missing data, all analyses under RE-IMPUTE had estimates close to the desired bias and coverage (bias: ≤ 0.01, coverage: $93.3 - 96.0\%$). As expected, the validity of the NEST and APPEND methods depended on the missingness pattern and the correlation structure of the data. The performance was satisfactory under monotone missingness in all cases (bias: ≤ 0.01, coverage: $93.8 - 96.1\%$). We identified some issues for non-monotone missing data. When relations were weaker within than between time points (i.e. Scenarios 2–4, 7, 8, 12), regression coefficient estimates were biased (bias: $0.03 - 0.52$). Coverage was low (coverage: $2.3 - 95.2\%$). On the other hand, the bias was small (bias: $0.01 - 0.03$) and the coverage was good (coverage: $94.6 - 98.3\%$) when relations within waves were equivalent to or stronger than relations between waves (i.e. scenarios 1, 5, 6, 9–11, 13–16). The numerical results for one of the regression coefficients are shown in Table 4 (monotone missingness) and Table 5 (non-monotone missingness) respectively, and are graphically presented in Figure 1.

Table 6 shows that results are similar for 20% missing data, but bias and undercoverage were much less pronounced for the scenarios with stronger relations between timepoints (bias: $0.01 - 0.19$, coverage: $43.6\% - 96.0\%$). Results for other regression coefficients are similar and available upon request.

4.2.2.2. Efficiency.

The 95% confidence intervals of variable means had similar relative widths (NEST vs. RE-IMPUTE: $0.95 - 1.05$; APPEND vs. RE-IMPUTE: $1.00 - 1.10$). Due to the larger number of imputed data sets, NEST was slightly more efficient compared to RE-IMPUTE ($0.90 - 0.96$) and APPEND ($0.89 - 0.94$).

4.2.2.3. Summary.

When taken together, we found that:

1. All techniques perform similarly under monotone missingness;
2. NEST and APPEND are on par on all scenarios;
3. NEST and APPEND perform well with respect to bias and coverage when relations within time points are strong;
4. NEST and APPEND produce bias estimates and low coverages when relations between time points are strong and relations within time points are weak.
5. Biases and low coverages of NEST and APPEND quickly taper off for smaller amounts of missing data.

4.3. Growth model

4.3.1. Simulation setup

We also evaluated the performances of RE-IMPUTE and APPEND in an unconditional growth model with

Table 4. Monotone missingness, 50% missing data. Population regression coefficients (Pop) and regression coefficients of the 95% confidence interval after RE-IMPUTE (*R*), NEST (*N*) and APPEND (*A*) methods. Regression coefficient of the incomplete wave-1 variable (b_{y_1}).

Scenario		Regression coefficients				Coverage 95% CI			Width 95% CI		
	$\rho_{within} = 0.10$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
1.	$\rho_{between} = 0.10$	0.08	0.08	0.08	0.08	95.5	95.1	94.8	0.27	0.25	0.27
2.	$\rho_{between} = 0.30$	0.23	0.23	0.23	0.23	94.5	94.9	94.2	0.26	0.24	0.26
3.	$\rho_{between} = 0.50$	0.42	0.42	0.42	0.42	94.8	94.8	94.6	0.25	0.23	0.25
4.	$\rho_{between} = 0.70$	1.17	1.17	1.17	1.17	93.8	93.8	93.6	0.10	0.09	0.10
	$\rho_{within} = 0.30$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
5.	$\rho_{between} = 0.10$	0.07	0.07	0.07	0.07	95.4	95.5	95.2	0.29	0.26	0.29
6.	$\rho_{between} = 0.30$	0.19	0.19	0.19	0.19	94.8	95.3	94.9	0.27	0.25	0.27
7.	$\rho_{between} = 0.50$	0.31	0.31	0.31	0.31	95.2	94.9	95.0	0.25	0.23	0.25
8.	$\rho_{between} = 0.70$	0.57	0.57	0.57	0.57	95.0	94.7	94.7	0.22	0.20	0.22
	$\rho_{within} = 0.50$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
9.	$\rho_{between} = 0.10$	0.06	0.06	0.06	0.06	94.9	94.4	94.8	0.31	0.29	0.32
10.	$\rho_{between} = 0.30$	0.16	0.16	0.16	0.16	94.5	95.2	94.7	0.30	0.28	0.30
11.	$\rho_{between} = 0.50$	0.25	0.25	0.25	0.25	94.5	95.0	94.8	0.27	0.25	0.27
12.	$\rho_{between} = 0.70$	0.40	0.40	0.40	0.40	95.5	96.1	94.4	0.23	0.21	0.23
	$\rho_{within} = 0.70$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
13.	$\rho_{between} = 0.10$	0.05	0.05	0.05	0.05	95.2	95.8	95.7	0.38	0.35	0.38
14.	$\rho_{between} = 0.30$	0.14	0.14	0.14	0.14	94.6	94.5	94.3	0.36	0.33	0.36
15.	$\rho_{between} = 0.50$	0.21	0.21	0.21	0.21	95.4	95.1	95.0	0.32	0.30	0.33
16.	$\rho_{between} = 0.70$	0.35	0.35	0.35	0.35	94.8	95.2	94.5	0.27	0.25	0.27

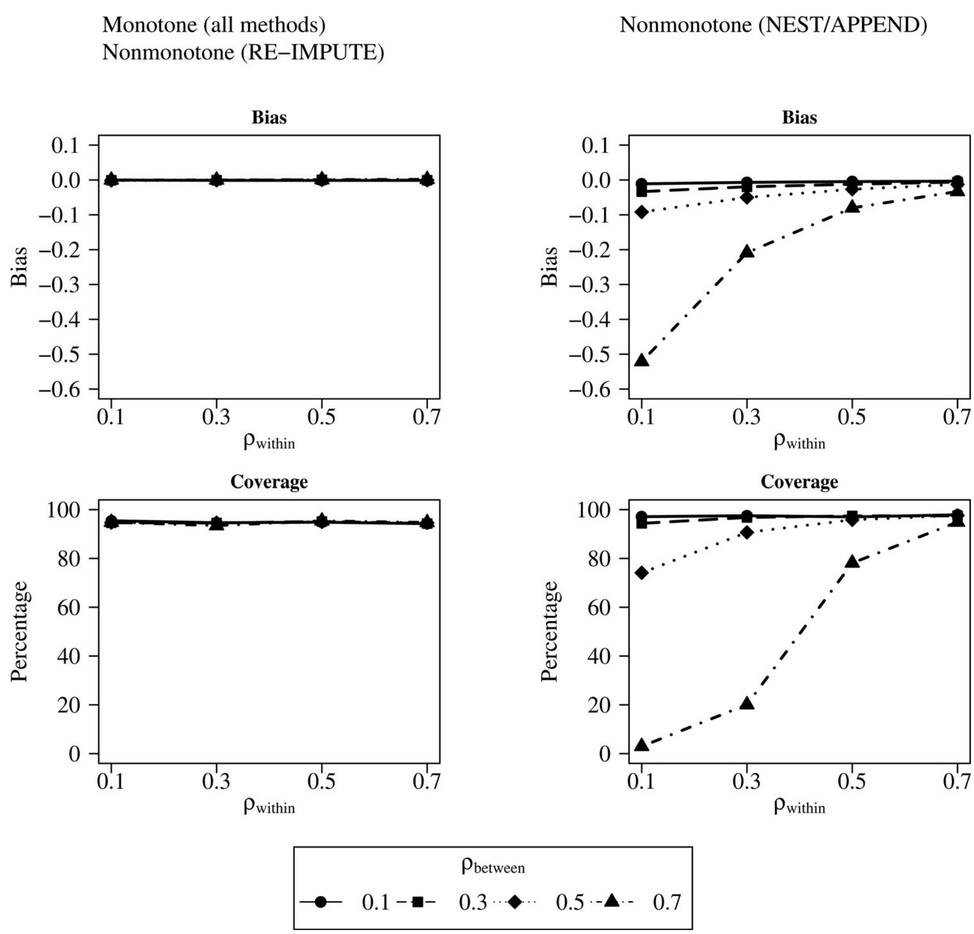

Figure 1. Observed patterns of bias and coverage of regression coefficients by correlation structure (ρ_{within} = correlation within time points; $\rho_{between}$ = correlation between time points), aggregated over conditions with similar results. Left: monotone missingness with RE-IMPUTE/NEST/APPEND; and nonmonotone missingness with RE-IMPUTE; right: nonmonotone missingness with NEST/APPEND.

Table 5. Non-monotone missingness, 50% missing data. Population regression coefficients (Pop) and regression coefficients of the 95% confidence interval after RE-IMPUTE (*R*), NEST (*N*) and APPEND (*A*) methods. Regression coefficient of the incomplete wave-1 variable (b_{y_1}).

Scenario		Regression coefficients				Coverage 95% CI			Width 95% CI		
	$\rho_{within} = 0.10$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
1.	$\rho_{between} = 0.10$	0.08	0.08	0.07	0.07	94.6	96.8	96.8	0.26	0.25	0.26
2.	$\rho_{between} = 0.30$	0.23	0.23	0.20	0.20	93.7	95.2	94.0	0.25	0.24	0.26
3.	$\rho_{between} = 0.50$	0.42	0.42	0.33	0.33	94.9	72.3	74.0	0.23	0.25	0.26
4.	$\rho_{between} = 0.70$	1.17	1.17	0.65	0.65	94.7	2.3	4.8	0.10	0.47	0.51
	$\rho_{within} = 0.30$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
5.	$\rho_{between} = 0.10$	0.07	0.07	0.06	0.06	94.7	98.0	97.2	0.28	0.26	0.27
6.	$\rho_{between} = 0.30$	0.19	0.19	0.17	0.17	95.9	97.3	96.2	0.26	0.25	0.26
7.	$\rho_{between} = 0.50$	0.31	0.31	0.26	0.26	95.2	90.8	91.9	0.24	0.24	0.25
8.	$\rho_{between} = 0.70$	0.57	0.57	0.36	0.36	94.2	15.4	24.1	0.20	0.29	0.31
	$\rho_{within} = 0.50$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
9.	$\rho_{between} = 0.10$	0.06	0.06	0.06	0.06	94.2	97.4	97.1	0.31	0.29	0.31
10.	$\rho_{between} = 0.30$	0.16	0.16	0.15	0.15	94.3	97.0	97.2	0.29	0.28	0.29
11.	$\rho_{between} = 0.50$	0.25	0.25	0.22	0.22	94.7	95.2	95.3	0.26	0.26	0.27
12.	$\rho_{between} = 0.70$	0.40	0.40	0.32	0.32	95.4	74.5	76.7	0.22	0.24	0.25
	$\rho_{within} = 0.70$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
13.	$\rho_{between} = 0.10$	0.05	0.05	0.05	0.05	94.8	98.3	98.1	0.39	0.36	0.38
14.	$\rho_{between} = 0.30$	0.14	0.14	0.13	0.13	93.8	97.2	97.4	0.36	0.34	0.36
15.	$\rho_{between} = 0.50$	0.21	0.21	0.19	0.19	94.7	97.7	97.6	0.32	0.31	0.32
16.	$\rho_{between} = 0.70$	0.35	0.35	0.32	0.32	95.7	95.8	94.6	0.26	0.26	0.28

Table 6. Non-monotone missingness, 20% missing data. Population regression coefficients (Pop) and regression coefficients of the 95% confidence interval after RE-IMPUTE (R), NEST (N) and APPEND (A) methods. Regression coefficient of the incomplete wave-1 variable (b_{y_1}).

Scenario		Regression coefficients				Coverage 95% CI			Width 95% CI		
	$\rho_{within} = 0.10$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
1.	$\rho_{between} = 0.10$	0.08	0.08	0.08	0.08	94.5	95.4	95.5	0.21	0.21	0.21
2.	$\rho_{between} = 0.30$	0.23	0.23	0.22	0.22	95.0	96.0	95.3	0.20	0.20	0.20
3.	$\rho_{between} = 0.50$	0.42	0.42	0.39	0.39	94.6	92.4	92.0	0.19	0.19	0.20
4.	$\rho_{between} = 0.70$	1.17	1.17	0.98	0.98	94.8	43.6	45.7	0.07	0.34	0.35
	$\rho_{within} = 0.30$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
5.	$\rho_{between} = 0.10$	0.07	0.07	0.07	0.07	95.5	96.4	96.5	0.22	0.22	0.22
6.	$\rho_{between} = 0.30$	0.19	0.19	0.18	0.18	95.5	96.0	95.9	0.21	0.21	0.21
7.	$\rho_{between} = 0.50$	0.31	0.31	0.30	0.30	94.8	94.5	94.5	0.19	0.19	0.20
8.	$\rho_{between} = 0.70$	0.57	0.57	0.50	0.50	94.7	78.3	78.5	0.16	0.20	0.21
	$\rho_{within} = 0.50$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
9.	$\rho_{between} = 0.10$	0.06	0.06	0.06	0.06	95.2	96.7	96.5	0.25	0.24	0.25
10.	$\rho_{between} = 0.30$	0.16	0.16	0.16	0.16	95.5	96.5	96.5	0.23	0.23	0.23
11.	$\rho_{between} = 0.50$	0.25	0.25	0.24	0.24	94.2	95.7	95.3	0.21	0.21	0.21
12.	$\rho_{between} = 0.70$	0.40	0.40	0.37	0.37	95.5	94.2	93.8	0.18	0.19	0.19
	$\rho_{within} = 0.70$	Pop	*R*	*N*	*A*	*R*	*N*	*A*	*R*	*N*	*A*
13.	$\rho_{between} = 0.10$	0.05	0.05	0.05	0.05	95.0	96.7	96.7	0.30	0.29	0.30
14.	$\rho_{between} = 0.30$	0.14	0.14	0.14	0.14	94.7	96.1	96.3	0.28	0.28	0.28
15.	$\rho_{between} = 0.50$	0.21	0.21	0.21	0.21	95.0	96.5	96.5	0.25	0.25	0.25
16.	$\rho_{between} = 0.70$	0.35	0.35	0.34	0.34	95.2	95.8	96.4	0.21	0.21	0.21

one time-varying covariate x_t and outcome variable y_t at timepoint $t = (1, ..., T)$:

$$y_{it} = \beta_{0i} + \beta_{1i} x_{it} + e_{it} \quad (1)$$

where
$$\beta_{0i} = \gamma_0 + u_{0i}$$
$$\beta_{1i} = \gamma_1 + u_{1i}$$

We applied a specialized design of the simulation study in Section 4 and varied three conditions:

1. Cross-correlation between x_t and y_t (ρ_{cross}): Low (0.10) vs. high (0.50)
2. Lag-1 autocorrelations of x_t and y_t (ρ_{auto}): Low (0.10) vs. high (0.50)
3. Missingness pattern: monotone vs. nonmonotone.

4.3.1.1. Data generation. We generated data of five timepoints ($T = 5$) with residual variance $e_t \sim N(0, 1)$, fixed intercept $\gamma_0 = 1$ and random intercept variance $u_0 \sim N(0, 0.20)$. Fixed slope γ_1 followed from cross-correlation ρ_{cross}, which had a truncated normal distribution with untruncated normal mean $\mu_\rho = \rho_{cross}$, untruncated normal variance V_ρ and truncation range $(-1, 1)$. Variance V_ρ was chosen such that $\text{Var}(u_1) = 0.20$ and fixed slope effect $\gamma_1 = E[\rho_{cross}]$. Further, we assumed independence between random intercept variance u_0 and random slope variance u_1.

Table 7. Regression coefficients and (the width of) their 95% confidence intervals of early catch-up growth predicting length at age 19, after RE-IMPUTE, NEST and APPEND regression coefficients were unadjusted and adjusted for potential confounders.

Unadjusted	RE-IMPUTE					NEST					APPEND				
	B	95% CI			Width	*B*	95% CI			Width	*b*	95% CI			Width
Weight	0.47	0.30	–	0.63	0.33	0.46	0.29	–	0.62	0.33	0.47	0.28	–	0.66	0.38
Length	0.58	0.46	–	0.71	0.25	0.57	0.44	–	0.69	0.25	0.56	0.42	–	0.70	0.27
HC	0.29	0.12	–	0.45	0.33	0.26	0.10	–	0.42	0.32	0.26	0.05	–	0.46	0.41
WL	−0.45	−0.81	–	−0.10	0.71	−0.41	−0.77	–	−0.06	0.72	−0.41	−0.78	–	−0.04	0.74
Adjusted	RE–IMPUTE					NEST					APPEND				
	B	95% CI			Width	*B*	95% CI			Width	*b*	95% CI			Width
Weight	0.27	0.13	–	0.42	0.29	0.26	0.10	–	0.41	0.31	0.27	0.08	–	0.46	0.38
Length	0.42	0.31	–	0.53	0.23	0.40	0.28	–	0.52	0.24	0.39	0.24	–	0.54	0.30
HC	0.12	−0.05	–	0.29	0.34	0.09	−0.06	–	0.23	0.29	0.09	−0.10	–	0.29	0.39
WL	−0.40	−0.76	–	−0.04	0.73	−0.36	−0.73	–	0.00	0.73	−0.33	−0.72	–	0.05	0.78

Note. HC = head circumference, WL = weight adjusted for length.

Similar to the simulation in Section 4, we generated 2,000 datasets of $n = 425$ subjects, thereby exceeding the minimum sample advised size of $n = 100$ for growth models (Curran et al., 2010).

4.3.1.2. Missingness and imputation procedure.

We created 50% casewise missingness under MCAR (missing completely at random) according to either a monotone or a nonmonotone missingness pattern. Since multilevel imputation with incomplete covariates is not straightforward, we limited incomplete data to outcome variable y_t (Grund et al., 2018). To obtain an identifiable imputation model with random effects, we started imputing at $t = 3$ and ensured that at least two timepoints were observed for every subject. We imputed missing data via RE-IMPUTE ($m = 5$ for all imputations) and APPEND ($m_1 = 5$ and $m = 1$ for all later imputations). Since NEST and APPEND resulted in similar performance in the previous simulation, we omitted NEST in the current simulation. We imputed missing data via the `2l.pmm()` method in the `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) and `miceadds` (Robitzsch et al., 2020) packages. After imputation, we estimated fixed and random effects using the `nlme` (Pinheiro, 2020) package and pooled estimates with the `mitml` (Grund et al., 2019) package.

4.3.1.3. Outcomes of interest.

We evaluated three outcomes of interest:

1. Bias between pooled parameter estimates after imputation and true values of both fixed effects (γ_0 and γ_1) and random variances (U_0, U_1, e_t).
2. Coverage of the true parameter value in the 95% confidence interval of the fixed effects.
3. Relative width of the pooled 95% confidence interval.

4.3.2. Results

4.3.2.1. Fixed effects.

Both imputation methods estimated fixed effects γ_0 and γ_1 without bias under all conditions at all timepoints ($< |.002|$). Coverage was good and exceeded the nominal 95% ($97.0 - 99.0\%$). RE-IMPUTE and APPEND were equally efficient with a relative efficiency of $0.99 - 1.01$.

4.3.2.2. Random effects.

Random slope variance u_1 could be estimated unbiasedly (bias: <.010). However, RE-IMPUTE and APPEND systematically overestimated random intercept variance U_0 (bias: $0.003 - 0.213$) in autocorrelated data. This bias does not indicate a difference between RE-IMPUTE and APPEND, but rather originates from the underlying imputation procedure that could not deal with autocorrelations properly.

4.3.2.3. Summary.

When taken together, we can conclude that

1. Both RE-IMPUTE and APPEND could estimate almost all parameters unbiasedly and with accurate coverage.
2. Autocorrelations are not handled accurately by the underlying imputation procedure of both RE-IMPUTE and APPEND, resulting in bias in the random intercept variance.

5. Data application

In practice, NEST and APPEND may perform better in real life than in the simulation. Real datasets often contain auxiliary variables that provide extra information that reduces the impact of ignoring future data (Daniels et al., 2014; Xie & Meng, 2016). Let us revisit the POPS dataset (Verloove-Vanhorick et al., 1986).

5.1. Method

The analysis models specified by Van Dommelen et al. (2014) predict five outcomes at age 19 (wave-2 data: length, cognition, health-related quality of life, internalizing problems, and externalizing problems) from four types of catch-up growth (wave-1 data: weight, length, head circumference, and weight–length). The model was adjusted for potential confounders (also wave-1 data), which includes gestational age, sex, maternal age at birth, maternal smoking during pregnancy, maternal diabetes, socioeconomic status, parity, ethnicity, and target length. The researchers selected $n = 334$ cases born small for gestational age without severe complications ($n = 228$ for weight, $n = 203$ for length, $n = 178$ for head circumference, and $n = 64$ for weight adjusted for length) from the incomplete POPS cohort.

We completed the dataset using the three imputation strategies. Similar to the original article, we re-imputed the incomplete wave-1 and wave-2 data $m = 10$ times (Van Dommelen et al., 2014). In addition, we imputed the wave-1 data $m_1 = 10$ times, and imputed the wave-2 data for each completed wave-1 dataset with $m_2 = 10$ (NEST) and $m_2 = 1$ (APPEND).

After multiple imputation we fitted eight linear regression models per outcome variable. The models predicted the outcome from catch-up growth in weight, length, head circumference or weight adjusted for length, either unadjusted or adjusted for potential confounders. Quantities of scientific interest were regression coefficients of catch-up growth predictors and their 95% confidence intervals. We imputed data with mice (Van Buuren & Groothuis-Oudshoorn, 2011) using predictive mean matching similar to the original study.

5.2. Results

5.2.1. Missing data and correlation patterns

None of the data selections followed a strictly monotone missingness pattern, but at least 60% of missing values in the wave-1 data corresponded to cases without observations in wave-2 data, except for the weight–length predictor ($45.1 - 64.7\%$ monotone).

Potential problems arising from non-monotone missingness may be mitigated by strong correlations within waves. Each catch-up growth predictor had at least one correlation with another catch-up growth predictor (i.e. within wave 1) that exceeded the correlation with the outcomes (i.e. between wave 1 and wave 2). Hence, we considered NEST and APPEND methods appropriate for these data.

5.2.2. Parameter estimates

Table 7 presents the regression coefficients of catch-up growth predicting length at age 19. The three imputation strategies produce similar point estimates and confidence intervals. We found agreement between conclusions in seven out of eight models (weight, length, head circumference; either adjusted or unadjusted, and weight–length unadjusted). The adjusted model for weight–length produced divergent estimates: Weight–length predicted length after RE-IMPUTE (CI: $-0.76 - -0.04$), but not after NEST or APPEND (CI nested: $-0.73 - 0.00$; CI appended: $-0.72 - 0.05$). Across all models, NEST and RE-IMPUTE were approximately equally efficient (relative width 95% CI NEST vs. RE-IMPUTE: $0.85 - 1.07$) and more efficient than APPEND. Results for other outcome variables were qualitatively similar and are available upon request.

6. Discussion

We studied a multiple imputation problem when the data grow over time. The newly arriving data contain missing values that need to be imputed. The standard method (RE-IMPUTE) is to combine the new and old data before imputation, and re-impute all missing values in the combined data. We investigated the properties of two methods that impute the missing data in the new part only, thus preserving the historic imputations. Method NEST multiply imputes the new data conditional on each filled-in old data $m_2 > 1$ times. Method APPEND is the special case of NEST with $m_2 = 1$, thus appending each filled-in data by single imputation.

An attractive feature of methods NEST and APPEND is that they keep the old imputations in place, thereby preserving the results of statistical analyses performed on the earlier waves. Method APPEND is more convenient than NEST, but less efficient. We found that NEST and APPEND have the same validity as RE-IMPUTE for monotone missing data patterns. NEST and APPEND also work well when the relations within waves are stronger than between waves and for moderate percentages of missing data, say up to 25%. We do not recommend the use of NEST or APPEND when relations within time points are weak and when associations between time points are strong. We found a slight drop in efficiency of APPEND relative to NEST, so there is a (small) price to pay for its operational advantages.

The current paper was limited to exploring validity of RE-IMPUTE, NEST, and APPEND under relatively simple scenarios, aiming to clarify whether - and if so when - caution is warranted in the first place. In practice, more complex situations with larger numbers of variables and timepoints might be of interest. We did observe that the underlying imputation procedures do not handle autocorrelated data well, which remains be resolved for longitudinal data in general.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by funding.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgements: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

Data availability

The data that support the findings of this study are available in Figshare at https://doi.org/10.6084/m9.figshare.9902276. For the subset of the POPS data used in this study, the administrator of the POPS cohort, Dr. S.M. van der Pal at TNO Child Health, Leiden, The Netherlands (see http://www.tno.nl/pops) can be contacted. The data are not publicly available due to privacy and ethical restrictions.

ORCID

X. M. Kavelaars http://orcid.org/0000-0003-1600-3153
J. R. van Ginkel http://orcid.org/0000-0002-4137-0943
S. van Buuren http://orcid.org/0000-0003-1098-2119

References

Aardoom, J. J., Dingemans, A. E., Spinhoven, P., van Ginkel, J. R., de Rooij, M., & van Furth, E. F. (2016a). Web-based fully automated self-help with different levels of therapist support for individuals with eating disorder symptoms: A randomized controlled trial. *Journal of Medical Internet Research*, *18*(6), e159. https://doi.org/10.2196/jmir.5709.

Aardoom, J. J., Dingemans, A. E., van Ginkel, J. R., Spinhoven, P., Van Furth, E. F., & Van den Akker-van Marle, M. E. (2016b). Cost-utility of an internet-based intervention with or without therapist support in comparison with a waiting list for individuals with eating disorder symptoms: A randomized controlled trial. *International Journal of Eating Disorders*, *49*(12), 1068–1076. https://doi.org/10.1002/eat.22587.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition & Development: Official Journal of the Cognitive Development Society*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969.

Daniels, M. J., Wang, C., & Marcus, B. H. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics*, *70*(1), 62–72. https://doi.org/10.1111/biom.12121.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Grund, S., Robitzsch, A., Ludtke, O. (2019). *mitml: Tools for multiple imputation in multilevel modeling.* R package version 0.3-7. https://CRAN.R-project.org/package=mitml.

Hille, E. T. M. (2005). Nonresponse bias in a follow-up study of 19-year-old adolescents born as preterm infants. *Pediatrics*, *116*(5), e662–e666. https://doi.org/10.1542/peds.2005-0682.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558. https://doi.org/10.1214/ss/1177010269

Oberman, H. I., Van Buuren, S., Vink, G. (2020). *Missing the point: Non-convergence in iterative imputation algorithms*. https://openreview.net/pdf?id=fHSVg6mVqpw.

Pinheiro, J. (2020). *nlme: Linear and nonlinear mixed effects models.* R package version 3.1-150. https://CRAN.R-project.org/package=nlme.

R Core Team (2016). *R: A language and environment for statistical compu*ting. R Foundation for Statistical Computing. http://www.R-project.org/.

Robitzsch, A., Grund, S., Henke, T. (July, 2020). *miceadds: Some additional multiple imputation functions, especially for mice.* R package version 3.10-28. https://CRAN.R-project.org/package=miceadds.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Ed. By Donald B. Rubin. (Vol. 81). John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, *57*(1), 3–18. https://doi.org/10.1111/1467-9574.00217

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581.

Shen, Z. (2000). *Nested multiple imputation* [PhD thesis]. Harvard University.

Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03.

van Dommelen, P., Van der Pal, S. M., Bennebroek Gravenhorst, J., Walther, F. J., Wit, J. M., & van der Pal de Bruin, K. M. (2014). The effect of early catch-up growth on health and well-being in young adults. *Annals of Nutrition & Metabolism*, *65*(2–3), 220–226. https://doi.org/10.1159/000364884.

Veen, S., Ens-Dokkum, M. H., Schreuder, A. M., Verloove-Vanhorick, S. P., Ruys, J. H., Verloove-Vanhorick, S. P., & Brand, R. (1991). Impairments, disabilities, and handicaps of very preterm and very-low-birthweight infants at five years of age. *The Lancet*, *338*(8758), 33–36. https://doi.org/10.1016/0140-6736(91)90015-H.

Verloove-Vanhorick, S. P., Verwey, R. A., Brand, R., Bennebroek Gravenhorst, J., Keirse, M. J. N. C., & Ruys, J. H. (1986). Neonatal mortality risk in relation to gestational age and birthweight: results of a national survey of preterm and very-low-birthweight infants in the Netherlands. *The Lancet*, *327*(8472), 55–57. https://doi.org/10.1016/S0140-6736(86)90713-0.

Xie, X., Meng, X.-L. (2016). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, *27*, 1485–1594. https://doi.org/10.5705/ss.2014.067