Journal of Clinical Epidemiology 102 (2018) 107–114

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study

Frans E.S. Tan[a,*,1], Shahab Jolani[a,1], Hilde Verbeek[b]

[a]Department of Methodology and Statistics, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands
[b]Department of Health Service Research, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

Accepted 14 June 2018; Published online 28 June 2018

Abstract

Objectives: We provide guidelines for handling the most common missing data problems in repeated measurements in observational studies and deal with practicalities in producing imputations when there are many partly missing time-varying variables and repeated measurements.

Study Design and Setting: The Maastricht Study on long-term dementia care environments was used as a case study. The data contain 84 momentary assessments for each of 115 participants. A continuous outcome and several time-varying covariates were involved containing missing observations varying from 4% to 25% per time point. A multiple imputation procedure is advocated with restrictions imposed on the relation within and between partially missing variables over time.

Results: Multiple imputation is a better approach to deal with missing observations in both outcome and independent variables. Furthermore, using the statistical package R-MICE, it is possible to deal with the limitations of current statistical software in imputation of missing observations in more complex data.

Conclusion: In observational studies, the direct likelihood approach (i.e., the standard longitudinal data methods) is sufficient to obtain valid inferences in the presence of missing data only in the outcome. In contrast, multiple imputation is required when dealing with partly missing time-varying covariates and repeated measurements.

Keywords: Longitudinal design; Multiple imputation; Observational study; Partly missing time-varying covariates; Overparametrization; R-MICE

1. Introduction

A major advantage of analyzing longitudinal data over cross-sectional data is the possibility to describe individual profiles over time. Because characteristics of subjects may vary over time, measuring the outcome and time-varying characteristics of the subjects repeatedly enables us to better evaluate the effect of them on the outcome for an arbitrary subject [1]. There are many examples, for instance in health care practice, that demonstrate the merits of longitudinal data [2–4]. However, analyzing longitudinal data typically needs advanced approaches when compared to standard cross-sectional data.

Missing data are one of the central problems that one encounters during the analysis of longitudinal data. Subjects may drop out due to, for example, sudden severe illness, death, or inability to locate by the researchers, or a measurement may be missing due to reasons that are unknown to or known but not measured by the researcher. Missing data are a unique challenge all researchers face from time to time, especially those in health care practice [5]. As research designs have become more complex and often multicentered, the problem of missing data has become much more common and complicated. Therefore, statisticians have been addressing this problem over decades and developed solutions that can stand the scrutiny of statistical theory [6–8].

Popular solutions include excluding from the analysis those subjects who have missing observations (i.e., complete cases analysis), simple substitution methods, and advanced approaches like the direct maximum likelihood and multiple imputation (MI) [9]. Although applied researchers may know the existence of these methods, they

Conflict of interest: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[1] The first two authors contributed equally.

* Corresponding author: Frans E.S. Tan PhD, Associate Professor Methodology and Statistics, Maastricht University, Faculty of Health, Medicine and Life Sciences, CAPHRI Care and Public Health Research Institute, Department of Methodology and Statistics, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. Tel.: +31433882278; Fax: +31433618388.

E-mail address: frans.tan@maastrichtuniversity.nl (F.E.S. Tan).

https://doi.org/10.1016/j.jclinepi.2018.06.006
0895-4356/

What is new?

Key findings

- When analyzing longitudinal data with missing observations, two situations require different approaches. If the missing data are in the outcome only (and the independent variables are fully observed), the direct likelihood method will produce unbiased estimates under the missing at random assumption, and thus multiple imputation is not necessary. If, on the other hand, some of the independent variables contain missing observations too, imputation of missing data is then advantageous.
- A problem arises if there are more columns (variables per time point) than rows (subjects) when the data are constructed for the imputation purpose (i.e. the data are converted to the wide format). With no restrictions, imputing missing data cannot be performed and any software packages will simply crash or stop imputing. Therefore extra restrictions should be imposed while preserving as much as possible the correlation structure of the data, given the imputation model.
- The R- MICE package is useful to successfully deal with such complex longitudinal data.

What this adds to what is known?

- Analysis of the aforementioned complex observational longitudinal data, with many repeated measurements and partly missing time-varying covariates, can be analyzed using the R-MICE package by imposing extra restrictions on the relation within partly missing variables over time.
- When missing data are in the independent variables, the direct likelihood removes subjects with missing observations, which results in biased estimates.

What is the implication and what should change now?

- Care should be taken when analyzing longitudinal data with partly missing observations in the covariates. Moreover, standard software like SPSS and SAS may fail to deliver estimates if there are many time points and time- varying covariates. The guidelines as proposed in the article may be useful for a successful analysis.

may be less aware of the advantages and disadvantages of them depending on the design and underlying missing data mechanisms. Moreover, longitudinal data may have many time points and often contain time-varying independent variables with missing observations [10] so that imputation of missing data using standard software like SPSS and SAS may fail in such complex designs.

The purpose of this article is to provide researchers with practical guidelines to handle the most common missing repeated measurements data problems in observational studies. Many researchers, for example, in health care research and health services, use standard techniques as offered in software like SPSS without realizing the problems that may occur in their particular data. We specifically aim to address

- The important problem of how to analyze longitudinal data if there are missing observations in the outcome only and/or if missing observations are extended to independent variables too. These two situations require different approaches.
- Practicalities in producing imputations when there are many time-varying variables and repeated measurements, such that the imputation task will be impossible without making extra restrictions.
- The difficulties with common and ready-to-use imputation routines in statistical packages SPSS, SAS, and R.

In Section 2, we introduce the Maastricht Study on long-term dementia care environments (MLTD) as a case study and elaborate on its missing data structure. Using this structure as a reference, several potential problems have been considered. In Section 3, a brief review of possible solutions to handle missing data is given. Moreover, a limited simulation study is conducted to further elaborate on performance of different methods based on bias and coverage aspects of the estimates. In Section 4, an outline is given about the statistical analysis of the MLTD study. In Subsection 4.1, tips and tricks are given of how to implement the state-of-the-art method to handle missing observations. In Subsection 4.2, we describe the software limitations by comparing SPSS, SAS, and R-MICE. In Section 4.3 the suggested approach to deal with missing observations is applied to the MLTD study and the results are presented. In section 5, the article ends with a discussion.

2. Missing data structure of the MLTD study

As a motivation example, the MLTD study has a longitudinal design aiming at investigating the effect of innovative dementia care environments (i.e., small scale, homelike) in comparison with traditional nursing homes (large scale) on residents' daily life [3]. In this case study, we are interested to compare the mood between the elderly living in traditional large-scale wards (LSW = 1; 29 wards) and innovative small-scale wards (LSW = 0; 86 wards). A randomized observation schedule was performed, such that every participant was observed for 1 minute during every 20-minute period within a 4.5-hour observation block (there was a half hour break in each block). Each participant was then observed on 7 days: two weekday mornings (07:00–11:30), two weekday afternoons (11:30–16:00), two weekday evenings (16:00–20:30), and one Saturday afternoon (11:30–16:00). In total, 12 (observation minutes per block in a day) × 7(observation days) = 84 momentary assessments were recorded per participant. We focus on mood and engagement in activity (activity), which were assessed by the Maastricht Electronic Daily Life Observation tool [11]. Mood was observed using a 7-point rating scale, ranging from 1 = great signs of negative mood to 7 = very high positive mood, and the variable activity measures, for example, household activity, or musical activity. The MLTD study is an example where there are many missing observations in the outcome (mood) and in the time-varying independent variables (like activity engagement). Supplementary Table 1 shows the frequency of missing data across dayparts and time.

The main reason for missing data was the residents' inability to locate for observations (e.g., they were on a trip with family, had health care appointments, or were in their private rooms and thus unable to be observed). The way that researchers should deal with missing data to obtain unbiased estimates depends on the statistical model that they want to analyze and the type of missing data mechanism, which will be elaborated in the next section.

3. Methods

Following Rubin [6], three types of missing data mechanism are missing completely at random (MCAR), missing at random (MAR), and missing not at random (see Supplementary Material for details).

3.1. Solutions to handle missing data

Throughout the years, many solutions have been advocated to deal with missing data. In this article, we have summarized the most popular methods including complete case analysis (CCA), available cases (ACs), mean substitution (MS), missing indicator method (MIM) and last observation carried forward together with the state-of-the-art MI [7,12,13]. Which method is preferred depends on the design of the study and the underlying missing data mechanism. Here, we concentrate on observational studies and will argue that some popular methods are not recommendable for observational longitudinal designs.

The usual way to compare the performance of the different strategies is via a Monte Carlo simulation study because the true correlations and effects are known. We have therefore performed a limited simulated longitudinal study to show the consequence of choosing a strategy on parameter estimates. In short, four scenarios were investigated:

- Missing observations on both outcome and independent variables under MCAR
- Missing observations only on the outcome under MAR
- Missing observations only on the independent variables under MAR
- Missing observations on both outcome and independent variables under MAR

Details of the simulation study are provided in the Supplementary Material. It should also be noted that when performing MI, we put some extra restrictions comparable to that of the MLTD data to evaluate our proposed restricted MI procedure (see Section 4.1 for more details). Moreover, we emphasize that the results of this simulation study may not be applied to experimental designs.

3.2. Naïve methods

3.2.1. Complete case analysis

When using CCA, all cases with missing values are deleted. This approach is default in statistical software packages, which makes its use very easy and unnoticed. It can be safely used under MCAR because restricting the analysis to only those participants with no missing observations can be considered as taking another random sample from the original population [14]. However, as opposed to cross-sectional designs, the CCA can produce biased estimates under MAR. This is confirmed in the simulation study because the CCA only produced unbiased estimates under MCAR (see Table 1 scenario 1). Another point is that throwing away information leads to less efficient results when adopting the CCA. For instance, removing 50% of participants in the MLTD study may have a large impact on variance estimates as well as statistical power.

3.2.2. AC analysis

In this approach, each statistic will be calculated using the observed values of the relevant variable(s). If only the outcome has missing observations, the AC implies the direct likelihood approach. Consequently, AC gives unbiased estimates (as opposed to CCA) [15]. In contrast, when missing data are in the independent variables, the AC removes subjects with missing observations, which results in biased estimates (as the remaining subjects are not necessarily a representative of the target population). In the simulation study, the estimates of the regression coefficients were unbiased when the outcome had missing observations, while it leaded to biased estimates and lower coverage rates with missing data in the independent variables (see Table 1 scenarios 2, 3 and 4, particularly for β_0).

Table 1. Simulation results from five replications with a sample size of n = 115 and three repeated measurements

Scenario 1: MCAR—x and y missing within total 50% incomplete.

	Statistics					
Method	$\widehat{\beta}_0$	$se(\widehat{\beta}_0)$	**95% CI coverage rate of** β_0	$\widehat{\beta}_1$	$se(\widehat{\beta}_1)$	**95% CI coverage rate of** β_1
REF	2.01	0.095	0.95	0.50	0.101	0.95
CCA	2.01	0.136	0.95	0.50	0.145	0.95
AC	2.01	0.101	0.95	0.50	0.114	0.95
MS	2.05	0.091	0.90	0.42	0.104	0.88
MIM	-	-	-	-	-	-
LOCF	2.05	0.097	0.92	0.42	0.102	0.87
MI	2.03	0.101	0.95	0.45	0.117	0.93

Scenario 2: MAR—y missing with approximately 50% of the outcome variable incomplete

	Statistics					
Method	$\widehat{\beta}_0$	$se(\widehat{\beta}_0)$	**95% CI coverage rate of** β_0	$\widehat{\beta}_1$	$se(\widehat{\beta}_1)$	**95% CI coverage rate of** β_1
REF	2.00	0.095	0.95	0.51	0.101	0.94
CCA	2.66	0.116	0.00	0.41	0.135	0.88
AC	2.00	0.100	0.96	0.50	0.112	0.94
MS	2.16	0.084	0.55	0.41	0.102	0.86
MIM	-	-	-	-	-	-
LOCF	2.00	0.096	0.95	0.40	0.096	0.82
MI	2.03	0.100	0.94	0.46	0.115	0.95

Scenario 3: MAR—approximately 40% of the independent variables was incomplete

	Statistics					
Method	$\widehat{\beta}_0$	$se(\widehat{\beta}_0)$	**95% CI coverage rate of** β_0	$\widehat{\beta}_1$	$se(\widehat{\beta}_1)$	**95% CI coverage rate of** β_1
REF	2.00	0.095	0.95	0.50	0.102	0.96
CCA	2.53	0.097	0.00	0.38	0.112	0.80
AC	2.19	0.092	0.45	0.43	0.105	0.90
MS	2.07	0.101	0.92	0.35	0.109	0.72
MIM	1.16	0.123	0.00	0.42	0.097	0.85
LOCF	2.03	0.099	0.95	0.42	0.110	0.89
MI	2.04	0.102	0.94	0.43	0.128	0.93

Scenario 4: MAR—approximately 50% of the dependent and independent variables was incomplete

	Statistics					
Method	$\widehat{\beta}_0$	$se(\widehat{\beta}_0)$	**95% CI coverage rate of** β_0	$\widehat{\beta}_1$	$se(\widehat{\beta}_1)$	**95% CI coverage rate of** β_1
REF	2.01	0.095	0.95	0.50	0.102	0.95
CCA	2.61	0.106	0.00	0.39	0.123	0.83
AC	2.10	0.097	0.81	0.46	0.110	0.94
MS	2.10	0.093	0.79	0.39	0.106	0.84
MIM	-	-	-	-	-	-
LOCF	2.02	0.097	0.93	0.41	0.104	0.84
MI	2.04	0.101	0.93	0.44	0.122	0.94

Abbreviations: AC, available cases; CCA, complete case analysis; CI, confidence interval; LOCF, last observation carried forward; MI, multiple imputation; MIM, missing indicator method; MS, mean substitution; REF, reference.

True regression model: $Y_{it} = 2 + 0.5X_{it} + u_i + \varepsilon_{it}, \;\; i = 1, \ldots, \; 115, \;\; t = 1, \; 2, \; 3.$

Specification of regression model: $Y_{it} = \beta_{0i} + \beta_1 X_{it} + \varepsilon_{it}, \ldots\ldots \;\; i = 1, \ldots, \; 115, \;\; t = 1, \; 2, \; 3..$

3.2.3. Mean substitution

The MS implies each missing value of a variable is replaced by the arithmetic mean of that variable. In general, it can produce biased estimates with lower coverage rates even under MCAR (see Table 1 scenario 1, where the coverage rate was equal to 0.88).

3.2.4. Missing indicator method

This method fills missing observations with a fixed number and then adds a dummy variable to the analysis model to indicate whether the value of that variable was missing [16]. The MIM is attractive because of adjusting for an incomplete independent variable, but it is only valid if the missing data mechanism is independent of the outcome conditional on the other independent variables (see Table 1 scenario 3, where it was badly biased with zero coverage). Also, it is not designed to handle missing observations in the outcome.

3.2.5. Last observation carried forward

The last observed value of a variable is used as an imputed value for the follow-up missed observations in this method. Despite simplicity, it is severely criticized because of leading to biased estimates in nearly all situations [9,13] (see Table 1, all scenarios for $\widehat{\beta}_1$).

Because almost all naïve methods fail when there are missing observations in the independent variables, we briefly discuss MI as an advanced method in the next section.

3.3. An advanced method: MI

When using MI, each missing entry is replaced with more than one imputed value, randomly drawn from a distribution of possible values that is determined using information from the data. This leads to multiply imputed data sets each of which can be then analyzed separately with standard statistical procedures. Finally, the results of separate fits are combined to form a single inference using a combination rule known as Rubin's rule [12]. MI also provides a solution to the problem of imputation uncertainty existing in the single imputation methods. The standard implementations of MI assume the missing data mechanism is MAR so that it provides valid inferences under MAR (and also MCAR) (see Table 1 all scenarios, where MI led to estimates with negligible bias and acceptable coverage rates [~95%]).

It is worth mentioning again that the AC method through direct maximization of the likelihood function (e.g., the mixed function in SPSS for longitudinal data) provides valid results only if the missing data are in the outcome (under MAR assumption) [15]. In fact, the results of MI in such cases are similar, if a little less efficient, than the direct likelihood method when the number of imputations is large ([17], pp. 525). Therefore, there is little gain from MI in such circumstances, and the direct likelihood analysis is preferred (as it is relatively easy to do). Nevertheless, when the missing data appear in the independent variables (or simultaneously in the outcome and independent variables), standard procedures like the mixed function in SPSS, by default, remove cases with missing data in the independent variables, and therefore, the estimates are biased. MI, on the other hand, becomes useful because all missing values are automatically imputed, and multiple (completed) imputed data sets can then be easily analyzed using the standard procedures (e.g., the mixed function in SPSS).

A popular approach, known as Fully Conditional Specification (FCS) [18] or chain equation [19], imputes missing data on a variable-by-variable basis. For each incomplete variable, an appropriate regression model is specified conditional on the other variables, that is, an ordinary linear regression for a continuous variable, a logistic regression for a binary variable, and a multinomial model for a categorical variable ([13] chapter. 4). Starting from an initial imputation, missing values of each incomplete variable are imputed in turn using its corresponding model and the most updated imputations of the other variables. The whole cycle is iterated a number of times to stabilize the results, assuming the algorithm has converged to a stationary distribution.

The FCS approach is nowadays popular in practice because of ease and availability in major statistical software packages. In the next two sections, the MLTD data will be analyzed using the MI approach and compared to the CCA method.

4. Statistical analysis of the MLTD study

To compare the average mood of participants in the large- and small-scale wards, two models were considered: a substantive model (i.e., the analysis model) and an imputation model. For analysis, we used a random intercept model with "Mood" as the outcome and the independent variables were large-scale ward indicator ("LSW" = 1), Activity indicator (activity = 1), part of the day ("Daypart" with seven categories), and the repeated measurement of the participant within the Daypart ("Time" treated as continuous). All (first order) interactions were also specified. Note that the model had missing observations in the covariate "Activity" and interactions with "LSW", "Daypart," and "Time".

The model that is used for imputing missing data is called the imputation model. In general, generating imputations for missing observations is difficult. Most notably, the imputation model should contain all the relations between variables specified in the substantive model [20] to prevent the congeniality issue (i.e., inconsistencies between the substantive model and the imputation model). If the analyst wishes to include interactions and other nonlinear terms in the substantive model, these should be reflected in the imputation model too. Failure to include such terms causes overestimation or underestimation of the parameters.

Using the FCS approach to impute the missing observations may raise problems when there are many time-dependent variables and repeated measurements like in the MLTD study. These will be elaborated in the next subsection and we suggest tips and tricks on how to overcome such issues.

4.1. Tips and tricks of how to implement MI

An attractive feature of MI is that it can incorporate additional information (auxiliary variables) available in the data but usually unused in the analysis. This is very relevant because the standard implementations of MI assume the missing data mechanism is MAR, and inclusion of additional variables increases the plausibility of the MAR assumption. Collins et al. [21] and Schafer [20] recommended an inclusion strategy to have a richer imputation model than the substantive model.

The analysis of data with repeated measures using random-effect models requires the data being in the so-called "long" format, which implies a separate record is allocated to each time point per person (i.e., each person may have multiple records in the data set). However, MI is conveniently done when the data are in the "wide" format [13]. The latter implies that each person occupies only one record in the data set, and observations made at different time points are coded as different columns. This feature is attractive within the FCS framework because the imputation model for each incomplete variable is directly specified from the other variables.

When converting the MLTD data to the wide format, we encounter a specific computational issue for generating imputations. More specifically, the number of repeated measurements in the MLTD study is 84 (i.e., 84 measurements for Mood, 84 measurements for Activity, 84 measurements for social interaction, and 84 measurements for the interaction where Activity and social interaction are involved in the imputation model). As a result, there will be in total more than 300 time-varying variables in the wide format, while only 115 subjects are available in the study. Therefore, an imputation model that includes all other variables as predictors for a particular variable cannot be fitted due to overparameterization. For example, suppose Mood at time 1 needs to be imputed. Using the FCS principle, the predictors in the imputation model for Mood at time 1 are as follows: Mood, Activity, all interactions between Activity and time at time 2, at time 3, …, at time 84 (there are 84 repeated measurements if we combine Dayparts and time). The MI procedure in SPSS (and any other software) will simply crash when such imputation model is defined. We therefore need to customize the imputation model such that it is not overfitted and general enough to produce good imputations. We propose the following procedure. First, it is plausible to assume that all variables at the same time point are related with each other while unrelated to the other variables in the other time points. We make an exception that each variable at a particular time point is a predictor of the same variable at the other time points (e.g., Mood at time 2, Mood at time 3, and so on are predictors of Mood at time 1). Hence, the imputation model for Mood at time 1 includes Activity, social interaction, and the interaction with LSW at time 1, Mood at time 2, 3, …, 84 as predictors. Background variables such as sex and location, and the time-independent variable LSW are included in the imputation model too. It should be noted that we used the same restriction rule when specifying the imputation model in the simulation study.

Another important issue is the imputation of interaction terms with missing values. In the MLTD study, for example, Activity has missing observations, and hence, its interaction with LSW has missing observations too. One solution to this is to "passively" impute the interaction term after imputing the main variable. In the MLTD data, for instance, the variable Activity can be first imputed and the imputed value is multiplied by the value of LSW to from the interaction term. The latter can then be used as a predictor in the imputation model of another variable (e.g., Mood). A second approach is to consider the interaction term as "Just Another Variable" (JAV) and impute it separately. A recent simulation study by White et al. [22] showed that the JAV method performed better than the passive imputation approach in many settings. In this study, we applied JAV to impute the interaction between Activity (social interaction) and LSW.

It should be noted that the missing values could also be imputed when the data are in the long format. Nevertheless, it requires a form of multilevel imputation, which is in turn more sophisticated and requires experts' knowledge. For a gentle introduction of multilevel imputation see Hox, van Buuren, and Jolani [23].

4.2. Software limitations

In this section, we briefly compare the MI procedure in SPSS, SAS, and R-MICE. In general, no statistical software and packages are able to perform MI for the MLTD study without making extra restrictions as the number of parameters in the imputation model exceeds the number of subjects (the so-called overparameterization). SPSS uses by default the FCS approach to generate imputations. It also allows including two-way interactions (only for categorical variables). However, it is unclear how the interactions of categorical variables with missing values are handled (i.e., a passive imputation or JAV approach). It is also possible to add constraints to limit the role of variables during imputation, but the procedure is not flexible enough to customize the variable's role in the imputation model.

In SAS, the FCS approach can optionally be used, but the user should control the role of each variable separately to pose restrictions on imputation models. This is very time consuming in the MLTD data, for example, as more than

Table 2. Relevant parameter estimates for Ward-effect with and without multiple imputation

Est LSW × effect	Substantive model	
Time, ward, activity, all first-order interactions as independent		
	Complete case analysis (No multiple imputation)	MI, pooled estimates
LSW	**−0.29 (0.46)**	**−0.11 (0.47)**
LSW × Daypart 1: Morning 1	0.67 (0.41)	0.54 (0.43)
LSW × Daypart 2: Morning 2	0.89 (0.41)	0.69 (0.45)
LSW × Daypart 3: Afternoon 1	0.19 (0.42)	0.10 (0.46)
LSW × Daypart 4: Afternoon 2	1.12 (0.41)	1.27 (0.44)
LSW × Daypart 5: Afternoon 3	**−0.11 (0.41)**	**−0.02 (0.44)**
LSW × Daypart 6: Evening 1	−0.60 (0.42)	−0.50 (0.44)
LSW × Daypart 7: Evening 2	-	
LSW × time	−0.10 (0.03)	−0.09 (0.04)
LSW × Activity	**−0.09 (0.26)**	**−0.27 (0.32)**

MI, multiple imputation.
Mood × 10 is the outcome.
Standard errors between brackets.

250 imputation models must be specified separately. Furthermore, interaction terms are passively imputed.

The R-package MICE imputes missing data using the FCS approach. To create imputations, users can customize the role of variables in the so-called predictor matrix. Here, the predictor matrix regulates each variable's role in the imputation model. Interaction terms can also be imputed either using a passive or the JAV method.

4.3. Results of the MLTD study

Table 2 shows the results of the analysis with and without MIs. The number of imputations is set to m = 20 (The R-MICE algorithm for the MLTD data is available upon request). The outcome is Mood10 (Mood10: mood multiplied by a factor 10). Because the main interest lies in a comparison between the large- and small-scale wards, Table 2 only presents the estimates of the relevant regression coefficients. Coefficients that are associated with other variables than LSW or interactions between variables other than LSW do not contribute to the difference between the two types of wards and thus can be left out.

As can be seen from the table, the estimates do differ indicating that the missing data mechanism was not likely to be MCAR. The most noticeable difference was regarding LSW ($\widehat{\beta}_{LSW-NOMI} = -0.29(0.46)$ versus $\widehat{\beta}_{LSW-MI} = -0.12(0.47)$), the interaction LSW × Daypart 5 ($\widehat{\beta}_{LSW-Acitivity-NOMI} = -0.11(0.41)$ versus $\widehat{\beta}_{LSW-Acitivity-MI} = -0.02(0.44)$), and the interaction LSW × Activity ($\widehat{\beta}_{LSW-Activity-NOMI} = -0.09(0.26)$ versus $\widehat{\beta}_{LSW-Activity-MI} = -0.27(0.32)$). The estimated coefficients differed from each other almost a factor 3 and for LSW × Daypart 5, a factor 5. Note that the standard errors (between brackets) from MI were a bit larger than those without MI. This is because, the MI procedure accounts for the uncertainty of the imputed values. Figure 1 also shows the estimated profiles per day part. The most noticeable difference was for evening 1, where the estimated profile was decreased after MI while it was increased without MI. The difference between the two types of wards, however, did not change that much.

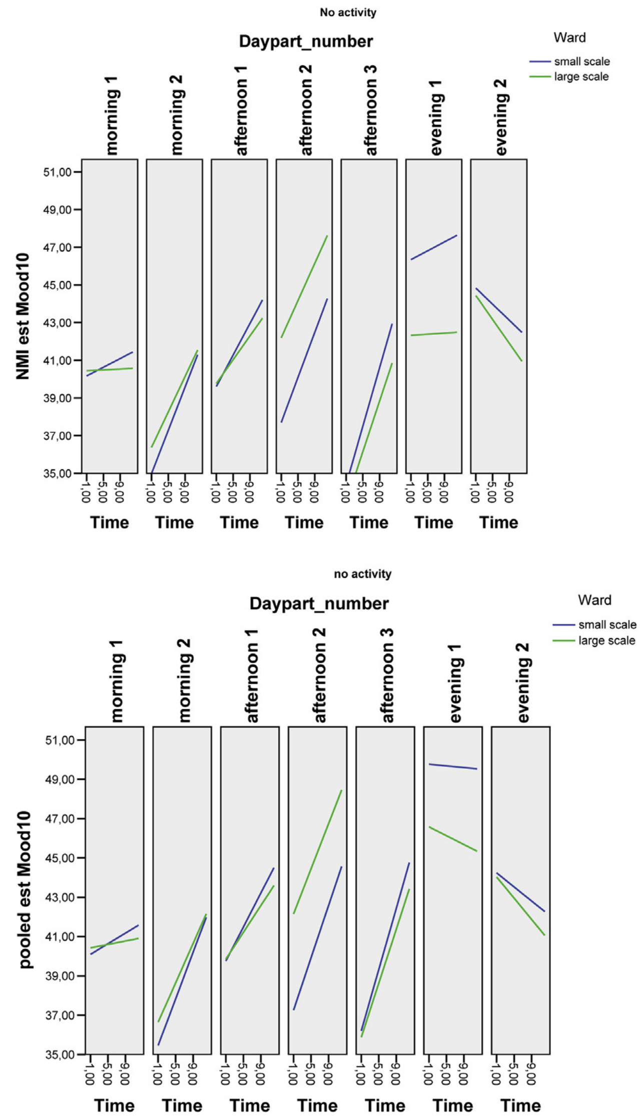

Fig. 1. Estimated difference between large- and small-scale wards and activity when not imputed and when imputed. NMI, no missing imputation.

5. Discussion

This article presented a proper method to handle complex observational longitudinal data with many time points in the presence of missing data in the outcome as well as in the independent variables. We demonstrated the performance of different methods using a simulation study and the data from the Maastricht long-term dementia care environments study.

If the missing data are in the outcome only (and the independent variables are fully observed), the direct likelihood method will produce unbiased estimates under the MAR, and thus MI is not necessary. If, on the other hand, some of the independent variables contain missing data too, imputation of missing data is then advantageous.

Moreover, a problem with longitudinal data like the MLTD is the existence of much more columns (variables per time point) than rows (subjects) when the data are constructed for the imputation purpose (i.e., the data are converted to the wide format). With no restrictions, imputing missing data cannot be performed and any software/packages will simply crash or stop imputing. Therefore, extra restrictions should be imposed while preserving as much as possible the correlation structure of the data, given the imputation model. The R-MICE package is useful to successfully deal with such complex longitudinal data.

In the MLTD study, persons are nested within several locations so that a three-level model would ideally be more appropriate. An analysis with a three-level model ignoring missing data, however, did not reveal a three-level factor, but it is not known if the same result could have been obtained had the missing data been imputed using a multilevel imputation model (preferably a three-level). Unfortunately, the methodology to deal with multilevel imputation procedures is still underdeveloped, and standard packages like SPSS lack such extensions. A promising future research will be to use multilevel imputations instead of standard FCS that may also deal with the problem of higher level imputation models.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.06.006.

References

[1] Tan FES. Best practices in analysis of longitudinal data: a multilevel approach. In: Osborne J, editor. Best practices in quantitative methods. Los Angeles: Sage; 2008:451−70.
[2] Meijers JMM, Tan FES, Schols JMGA, Halfens RJG. Nutritional care; do process and structure indicators influence malnutrition prevalence over time? Clin Nutr 2014;33(3):459−65.
[3] Beerens HC, Zwakhalen SMG, Verbeek H, Tan FES, Jolani S, Downs M, et al. The relation between mood, activity, and interaction in long-term dementia care. Aging Ment Health 2018;22(1):26−32.
[4] Springelkamp H, Wolfs RC, Ramdas WD, Hofman A, Vingerling JR, Klaver CC, et al. Incidence of glaucomatous visual field loss after two decades of follow-up: the Rotterdam Study. Eur J Epidemiol 2017;32(8):691−9.
[5] Penny KI, Atkinson I. Approaches for dealing with missing data in health care studies. J Clin Nurs 2012;21:2722−9.
[6] Rubin DB. Inference and missing data. Biometrika 1976;63:581−92.
[7] Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473−89.
[8] Carpenter JR, Kenward MG. Multiple imputation and its application. New York: Wiley & Sons; 2013.
[9] Molenberghs G, Kenward M. Missing data in clinical studies. West Sussex: Wiley & Sons; 2007.
[10] de Boer B, Hamers JPH, Beerens HC, Zwakhalen SMG, Tan FES, Verbeek H. Living at the farm, innovative nursing home care for people with dementia - study protocol of an observational longitudinal study. BMC Geriatr 2015;15:144.
[11] de Boer B, Beerens HC, Zwakhalen SM, Tan FE, Hamers JP, Verbeek H. Daily lives of residents with dementia in nursing homes: development of the Maastricht electronic daily life observation tool. Int Psychogeriatr 2016;28(8):1333−43.
[12] Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley& Sons; 1987.
[13] Van Buuren S. Flexible imputation of missing data. London: Chapman & Hall; 2012.
[14] Little RJ, Rubin DB. Statistical analysis with missing data. New York: Wiley & Sons; 2002.
[15] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New-York: Springer-Verlag; 2000.
[16] Donders ART, van der Heijden GJ, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087−91.
[17] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. John Wiley & Sons; 2011.
[18] Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. J Stat Comput Simul 2006;76:1048−64.
[19] Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol 2001;27:85−95.
[20] Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. Stat Neerl 2003;57:19−35.
[21] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6:330−51.
[22] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377−99.
[23] Hox J, van Buuren S, Jolani S. Incomplete multilevel data: problems and solutions. In: Harring JR, Staplecton LM, Beretvas SN, editors. Advances in multilevel modeling for educational research: addressing practical issues found in real-world applications. Charlotte, NC: Information Age Publishing, Inc; 2015:39−62.