The missing indicator method for response variables in binary transition models: A simulation study

Barragan Ibañez, Camila Natalia

The missing indicator method for response variables in binary transition models: A simulation study

Camila Natalia Barragán Ibáñez

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: 21 March 2022

Supervisor: Dr. Mark de Rooij

Acknowledgments

I would like to express my sincere gratitude to Prof. Dr Mark de Rooij for his patience and support. Without his advice and feedback, I would not have been able to complete this thesis. In addition, I would like to thank the professors of the specialisation Methodology and Statistics since their teachings also helped me.

I would also like to thank my family. The constant support of my father and the care of my mother helped me during this time. In the same way, I thank my brother for listening to me and helping me decide when I was in doubt.

Finally, I want to thank Andres, my partner, for his love and trust in me, which helped me overcome moments of uncertainty. Additionally, I want to thank his parents, who were a great support to get here.

Abstract

Longitudinal data are often collected in different research areas such as medicine, biology, education, and psychology. We can build a transitional model using longitudinal binary data, which aims to model the probability of transition between the response categories. In this type of data is common to find missing values due to dropouts, costs, and organizational problems. The missing-indicator model is often used as a method to handle missing values in this type of data. This method consists in creating a new category for the missing values. Therefore, the binary logistic model changes to a baseline-category logit model. This study aims to evaluate the bias of the estimated coefficients when the missing-indicator method is used in the response of a binary transitional model. Based on an empirical example, a Monte Carlo simulation with three factors is carried out: (1) type of missingness, (2) sample size, and (3) proportion of missing data. The coefficients bias from the baseline-category logit model is evaluated using boxplots and a three-way MANOVA analysis. The results suggest that sample size, the proportion of missing data, type of missingness and the interaction between sample size and proportion affect the bias of the estimated coefficients; nonetheless, the effect size is small. When each dependent variable is analysed separately using ANOVA, the effects of the proportion of missing and the interaction between sample size and proportion were statistically significant for only one coefficient. However, the effect size is still small. Therefore, the conclusion is that the estimated coefficients' bias for all the missingness types is low.

Contents

1 Introduction

Longitudinal data are often collected in different fields like medicine, biology, education, and psychology. This type of data makes it possible to study the tendencies and transitions of different phenomena over time. Furthermore, the advance of technology is starting to facilitate the collection of longitudinal data (Ginexi et al., 2014).

For binary longitudinal data, there are two approaches for the analysis. The first one has the objective of modelling the trends over time. For instance, the Generalized Estimation Equations make estimations for correlated data based on predictors (Ghahroodi et al., 2010). The second approach is transitional modelling, which aims to model the probability of a transition from one response category to another. These models use responses observed previously as predictors and may include other covariates (Azzalini, 1994; Bonney, 1987).

The first-order Markov transitional model for binary data can be written as a logistic regression. This Markov chain is characterized by the response Y_t being conditional on the previous response (Y_p) but conditionally independent of the responses further back in time. In the model in Equation 1, the response at the previous time point (y_p) and covariates (x_t, z_t) are used as predictors of the response at time point (t), this is,

$$log\frac{P(y_t = 1|y_p, x_t, z_t)}{P(y_t = 0|y_p, x_t, z_t)} = \alpha + \beta_1 y_p + \beta_2 x_t + \beta_3 z_t. \quad (1)$$

Missing values are especially common in longitudinal research affecting estimations and predictions. This missing is generated in different ways, that is, they have different mechanisms. Rubin (1976) proposed a classification of three categories according to the missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR) (Buuren, 2018). First, in the MCAR case, all cases have the same probability of being missing. Second, for MAR the probability of being missing depends on the observed data. Lastly, the probability of missing in MNAR depends on partially observed and unobserved information.

Longitudinal data often presents missing values due to dropout, costs, or organizational reasons (Madley-Dowd et al., 2019; Wang & Hsu, 2020). Various methods have been developed to handle missing values in research with binary outcomes. For instance, multiple imputations, last observation carried forward, complete case analysis, weighted estimating equation, and the missing-indicator method (Greenland & Finkle, 1995; Yang

et al., 2007). Every method mentioned has its advantages and disadvantages, but the least recommended is the missing-indicator method due to the bias generated (Greenland & Finkle, 1995; van der Heijden et al., 2006).

The missing-indicator method creates a new variable to indicate that there is a missing value in the original numeric variable, whilst, in the case of categorical variables, a new category is created. Thereby, starting with a dichotomous variable $y \in 0, 1$ would result in a new variable with three categories $1, 0, m$, where m indicates missing.

Using the missing-indicator method for missing values in the predictors results in biased parameter estimates and is recommended against its use (Donders et al., 2006; Greenland & Finkle, 1995; Knol et al., 2010; Pereira Barata et al., 2019; Song et al., 2021). Other methods like multiple imputations showed better results in handling missing data maintaining low the bias (Donders et al., 2006; Pereira Barata et al., 2019).

Nevertheless, it is also possible to find studies that supported, under certain conditions, the use of the missing-indicator method. For instance, Blake et al. (2020) concluded that the missing-indicator in the case where the covariance is missing with MNAR mechanism gives unbiased results. Some papers suggested that the missing-indicator method is appropriate for specific designs like matched case-control studies and randomized studies (Groenwold et al., 2012; Huberman & Langholz, 1999). Finally, a small proportion of studies conclude that although the results may be biased, it is an efficient method. Thus, the selection of this method depends on the balance of efficiency-accuracy that the researcher is willing to accept (Henry et al., 2013; Li et al., 2004).

In the case of transitional models, de Rooij (2018) used the missing-indicator method for missing values in the response variable and obtained almost unbiased parameters estimates for all types of missing data (i.e., MCAR, MAR and MNAR). de Rooij (2018) analyzed longitudinal multinomial data with a transitional approach in the context of an experiment comparing the transition probabilities of treatment and control conditions. de Rooij (2018), used the Point Classification model (IPC), within a distance framework, to fit a multinomial model with four categories. Three were response categories, and one was created for the missing values. The results showed that the missing-indicator method leads to unbiased parameter estimates. Despite the promising results, the reason the results are almost unbiased is still unknown.

This study aims to establish whether using the missing-indicator method in longitu-

dinal binary data results in unbiased parameter estimations. Adding a category for the missing data changes the binary logistic model to a multinomial baseline-category logit model. The baseline-category logit model uses a binary logit model for pairs of each response category with a baseline category (Agresti, 2002). Therefore, for a model with three categories (i.e., 0, 1, m), we have two different equations. The model is

$$log\frac{P(y_t = m|y_p, x_t, z_t)}{P(y_t = 0|y_p, x_t, z_t)} = \alpha_{11} + \beta_{11}I(y_p = m) + \beta_{21}I(y_p = 1) + \beta_{31}x_t + \beta_{41}z_t \quad (2)$$

$$log\frac{P(y_t = 1|y_p, x_t, z_t)}{P(y_t = 0|y_p, x_t, z_t)} = \alpha_{12} + \beta_{12}I(y_p = m) + \beta_{22}I(y_p = 1) + \beta_{32}x_t + \beta_{42}z_t, \quad (3)$$

where 0 is the baseline category, and I is an indicator function of the previous response. Please note that the second subscript of each parameter differentiates both equations. The logarithms of the odds of being in category 1 or m instead of being in category 0 are the result of the sum of the intercept, the response at the previous time point and the two covariates at the time point t. There are two coefficients in each equation related to the previous response. The first one is β_{11} in Equation 2 or β_{12} in Equation 3 that refers to the situation where the response at the previous time point is missing. The second one is β_{21} in Equation 2 or β_{22} in Equation 3, which refers to the case when the response at the previous time point is 1.

The main objective of this thesis is to evaluate the bias of the estimated coefficients when the missing-indicator method is used in the response of a binary transitional model comparing the estimations of equation 3 with the parameters in Equation 1, That is, we will compare the corresponding parameters of each row in Table 1.

Table 1: Estimated coefficients with its corresponding population parameter.

Estimations	Population parameters	Parameter name
α_{12}	α	Intercept
β_{22}	β_1	Previous response
β_{32}	β_2	Time point
β_{42}	β_3	Maternal smoking

A Monte Carlo simulation is performed to carry out the comparison. The results are divided into two sections, one for the multivariate analysis and the second for the univariate analyses.

2 Monte Carlo Simulation

2.1 Design

The simulation compares the population model parameters with the estimated coefficients from the binary and baseline-category logit model for the different types of missing data. The factors of this simulation are as follows:

1. Type of missingness: Non-missing data, Missing Completely at Random, Missing at Random, and Missing not at Random.

2. The proportion of missing data: 25% and 40%

3. Sample size: 100 and 1000 participants.

The combination of the proportion of missing data and sample size make four design cells. For each cell, we use 100 replications creating 400 non-missing datasets in total. Complete data is generated first, and then the different types of missing values are introduced.

2.2 Population model

The population model is based on the example presented in Agresti (2002, p.480), where the model predicted the presence of respiratory illness (yes, no) in children. The children were examined annually at ages 7 through 10 and classified according to the presence or absence of respiratory illness. Predictor variables are maternal smoking (a=1 indicates regularly and a=0 otherwise), the child's age (t= 7, 8, 9, 10), and the response at the previous year (Y_p).

For the data generation, we have the following settings: (1) the data generated have two response categories, the presence and absence of the respiratory illness. (2) Illness presence at Y_7 is the result of drawing random values from a Bernoulli distribution with a probability of 0.2, equal to the proportion in the original example. (3) The mother's probability of being a smoker is 0.5; that is, a is drawn from a Bernoulli distribution with $P(a = 1) = 0.5$. Moreover, this is a time constant predictor.

Equation 4 shows the population model in which the response y_t is drawn from a Bernoulli distribution with the probability $P(y_t)$ derived from

$$log\frac{P(y_t = 1|y_p, t, a)}{P(y_t = 0|y_p, t, a)} = -0.293 + 2.21y_p - 0.243t + 0.296a. \tag{4}$$

As mentioned before, the datasets with missing data are generated from the non-missing datasets. In the non-missing datasets, a binomial logistic regression model is fitted, while in the others, the baseline-category logit model is used.

2.3 Missing data

As in de Rooij (2018), the probability of an indicator of missing data was determined by

$$Pr(s = 1) = \frac{exp(\mu)}{1 + exp(\mu)}, \tag{5}$$

where s is the missing indicator, and μ is the linear predictor of the missingness model, which the missingness mechanism will define. There are some considerations in the simulation of missing data. The first one is that it is possible that after a missing value, the subject could respond again in further time points; that is, the missing cases can be intermittent. Secondly, there are no missing data at age 7.

2.3.1 Missing Completely at Random

The probability of missing data in this mechanism is the same for every observation. Therefore, μ equals a constant that depends on the proportion of missing data, that is,

$$\mu_1 = -0.4 \tag{6}$$

$$\mu_2 = -1.1, \tag{7}$$

where μ_1 produces approximately 40% of missing data and μ_2 produces around 25% of missing data.

2.3.2 Missing at Random

In this case, the probability of missing was higher for the healthy subjects in the previous year than the ill subjects, that is,

$$\mu_1 = -0.3 - 0.2Y_p \tag{8}$$

$$\mu_2 = -1 - 0.2Y_p, \tag{9}$$

where μ_1 produces around 40% of missing data and μ_2 generates approximately 25% of missing data.

2.3.3 Missing Not at Random

In this condition, the probability of missing depends on the response at the previous time point as well as the response on the current time point. This means that there was a higher probability of missing for healthy subjects,

$$\mu_1 = -0.3 - 0.2Y_p - 0.3Y_t \tag{10}$$

$$\mu_2 = -1 - 0.2Y_p - 0.3Y_t, \tag{11}$$

where μ_1 produces around 40% of missing data and μ_2 generates around 25% of missing data.

2.4 Statistical Analysis

The estimated coefficients are compared to the actual model parameters to assess the bias of the estimations according to the type of missing data. In the case of the baseline-category logit model fitting, only the estimated coefficients resulting from the odds of having a respiratory illness (category 1) instead of the absence of illness (category 0) are considered. If the estimations are unbiased then

$$E(\hat{\beta}_{22} - 2.21) = 0$$

$$E(\hat{\beta}_{32} - (-0.243)) = 0$$

$$E(\hat{\beta}_{42} - 0.296) = 0$$

$$E(\hat{\alpha}_{12} - (-0.293)) = 0$$

After the data and the missing are created, the non-missing datasets are fitted using the binary logistic regression; meanwhile, the datasets with missing values are fitted

using the multinomial baseline-category logistic model. Then, the estimated coefficients are subtracted from their corresponding population values (see 1).

The deviations, estimated coefficients minus the population parameter, are used to assess the bias with two methods. The first method is a three-way MANOVA analysis with the criteria of $\alpha = 0.05$, which determine if the combinations of the factor levels affected the variation of the bias of the coefficient estimations. Moreover, the effect size partial-eta squared allows us to inspect the proportion of the total generalized variance of the estimated coefficients for each factor while controlling for other factors and interactions (Huberty & Olejnik, 2006; Richardson, 2011). After this, an individual ANOVA is carried out for each coefficient in order to determine the effect of each independent variable and its interactions on the respective coefficient. The second method is a visual comparison of the coefficient distributions for each type of missingness in different levels of the proportion of missing data and sample size.

3 Results

The code with all the steps can be found in https://github.com/cnbi/missing-indicator-simulation-study

3.1 Multivariate analysis

First, an analysis to test the assumption of normality from MANOVA was carried out. This test showed that the assumption of normality is not possible to maintain. See Table 2. However, it is important to highlight that the condition of normality is not mandatory to perform MANOVA. Moreover, as seen in the boxplots of the intercept, previous response, time and maternal smoking, the distributions are symmetrical. Another aspect that stood out was the presence of outliers.

Table 2: Multivariate Normality Test results.

Test	Statistic	p value
Mardia Skewness	1303.097	$p < 0.01$
Mardia Kurtosis	82.853	0

The Box's M value, which tests the homoscedasticity assumption, was 4071.9 ($p < 0.05$), meaning that the covariances matrices are not equal. Nevertheless, this test is sensitive to the sample size. Therefore, it was decided to perform the MANOVA analysis despite these preliminary results.

The multivariate analysis of variance was performed on four dependent variables: intercept, the coefficient for the previous response, coefficient of the time point, and the coefficient of maternal smoking. The independent variables were the type of missingness, sample size and proportion of missing values. For this thesis, the type of missingness is the variable of paramount interest.

The results of MANOVA in Table 3 showed that with the use of Pillai's trace criterion, the composite of deviations was different across the levels of the main effects of the three independent variables and one interaction. This means that the sample size, proportion of missing values, the type of missing, and the interaction of sample size and proportion of missing values, produced bias.

Table 3: Test statistics for MANOVA including Pillai's trace, approximate F-test, its degree of freedom, its p-values, and effect size.

Effect	Pillai	approx F	Df1	Df2	p-value	$\eta^2_{partial}$
Sample	0.035	14.161	4	1581	**0.000**	0.030
Proportion	0.016	6.249	4	1581	**0.000**	0.020
Type	0.043	5.720	12	4749	**0.000**	0.010
Samp:Prop	0.017	6.969	4	1581	**0.000**	0.020
Samp:Type	0.005	0.631	12	4749	0.818	0.002
Prop:Type	0.011	1.422	12	4749	0.148	0.004
Samp:Prop:Type	0.003	0.392	12	4749	0.967	0.001
Residuals	1584					

Note. Samp stands for sample size, Prop means proportion of missing data, and Type refers to the type of missingness. In the header, df1 stands for degrees of freedom for numerator, while df2 refers to the degree of freedoms denominator. The effect size is partial η^2.

Nevertheless, the partial η^2 shows that the effect sizes of the sample size, proportion, type, and the interaction sample and proportion are no larger than 3%. That indicates that the variables do not explain more than 3% of the variation of the deviations. Specif-

ically, the type of missingness explains 1% of the variance.

3.2 Univariate analyses

3.2.1 Intercept

In this subsection, all the results are related to The results in Table 4 indicate that the bias does not change considerably according to the sample, proportion, or type of missingness. Likewise, the effect size indicates that these variables do not explain more than 0.09% of the variation.

Table 4: ANOVA results with the intercept as dependent variable.

Effect	Df	Sum Sq	Mean Sq	F value	p-value	η^2
Samp	1	0.3	0.2652	0.0795	0.7780	0.0001
Proportion	1	0.6	0.5582	0.1673	0.6826	0.0001
Type	3	2.6	0.8564	0.2567	0.8566	0.0005
Samp:Prop	1	1.1	1.0809	0.3240	0.5693	0.0002
Samp:Type	3	4.5	1.5113	0.4530	0.7152	0.0009
prop:Type	3	4.2	1.3985	0.4192	0.7393	0.0009
Samp:Prop:Type	3	3.3	1.1086	0.3323	0.8020	0.0006
Residuals	1584	5284.9	3.3364			

Note. Samp stands for sample size, Prop means proportion of missing data, and Type refers to the type of missingness. In the header, Df stands for degrees of freedom and Sq refers to squares, ergo Sum Sq is the Sum of squares and Mean Sq is the Mean squares.

In Figure 1, the bias was close to zero for every type of missingness. Furthermore, the proportion of missingness seemed to have little effect on the means. As expected, the sample size affected the height of the boxes, indicating less variability for the large sample size and, thus, more accurate estimation.

3.2.2 Previous response

In Table 5, the proportion of missing values and the interaction between sample size and proportion of missing values affect the deviations of the estimated coefficients. Moreover, the effect size shows that the proportion of variation that the mentioned variables can explain is between 0.4% and 1.2%.

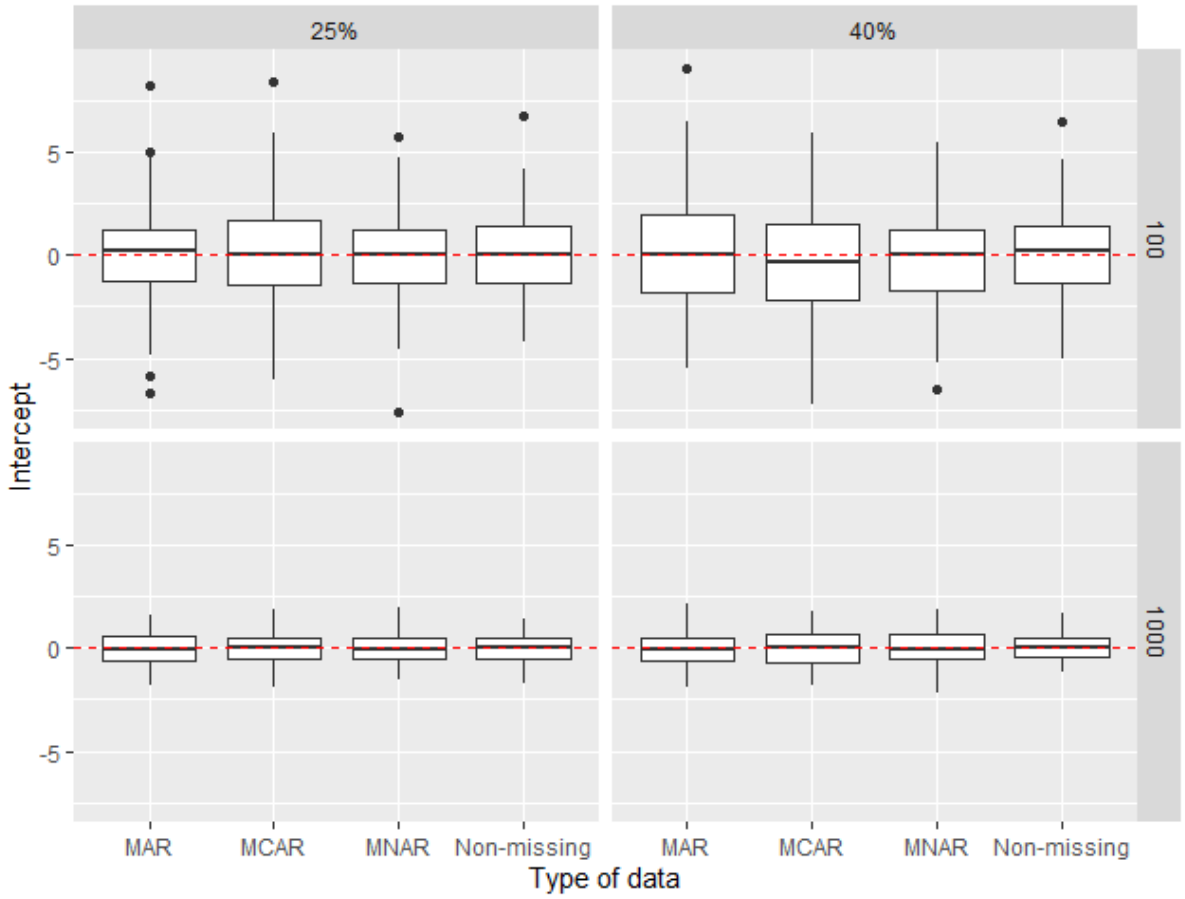

Figure 1: Comparison of the distributions of the deviations of the intercept coefficients for different types of missingness. The rows represent each level of the sample size, and the columns are the proportion of missing values. The dashed line marks zero, which indicates that there is no bias.

Table 5: ANOVA results with the previous response as dependent variable.

Effect	Df	Sum Sq	Mean Sq	F value	p-value	η^2
Sample	1	0.364	0.364	2.8042	0.0942	0.002
Proportion	1	0.769	0.769	5.9275	**0.0150**	0.004
Type	3	0.243	0.081	0.6256	0.5985	0.001
Samp:Prop	1	2.544	2.544	19.6228	**0.0000**	0.012
Samp:Type	3	0.231	0.077	0.5950	0.6183	0.001
prop:Type	3	0.447	0.149	1.1500	0.3276	0.002
Samp:Prop:Type	3	0.337	0.112	0.8657	0.4583	0.002
Residuals	1584	205.397	0.13			

Note. Samp stands for sample size, Prop means proportion of missing data, and Type refers to the type of missingness. In the header, Df stands for degrees of freedom and Sq refers to squares, ergo Sum Sq is the Sum of squares and Mean Sq is the Mean squares.

In Figure 2, the bias of the estimated coefficients for the previous response was close to 0 for all the types of missingness. Thus we can conclude that for our variable of interest, the type of missingness did not have a main effect or an interaction effect that could explain the variance in the deviations. When we explore the interaction between the sample size and the proportion of missing data with boxplots, we can see that the bias in the case of 100 participants with 40% of missing data is slightly larger than zero (see Figure 3). However, we have to remember that this effect only explains 1.2%. Moreover, When the proportion of missing is the only variable in the boxplot, it can be noted that the bias in the two situations is similar (see Figure 4).

The sample size affected the variation of the coefficients, so the variation was smaller when there were 1000 participants. The proportion of missing values seems to have a negligible effect on the means. In the cases where the proportion is 40%, the variation of the estimated is slightly bigger.

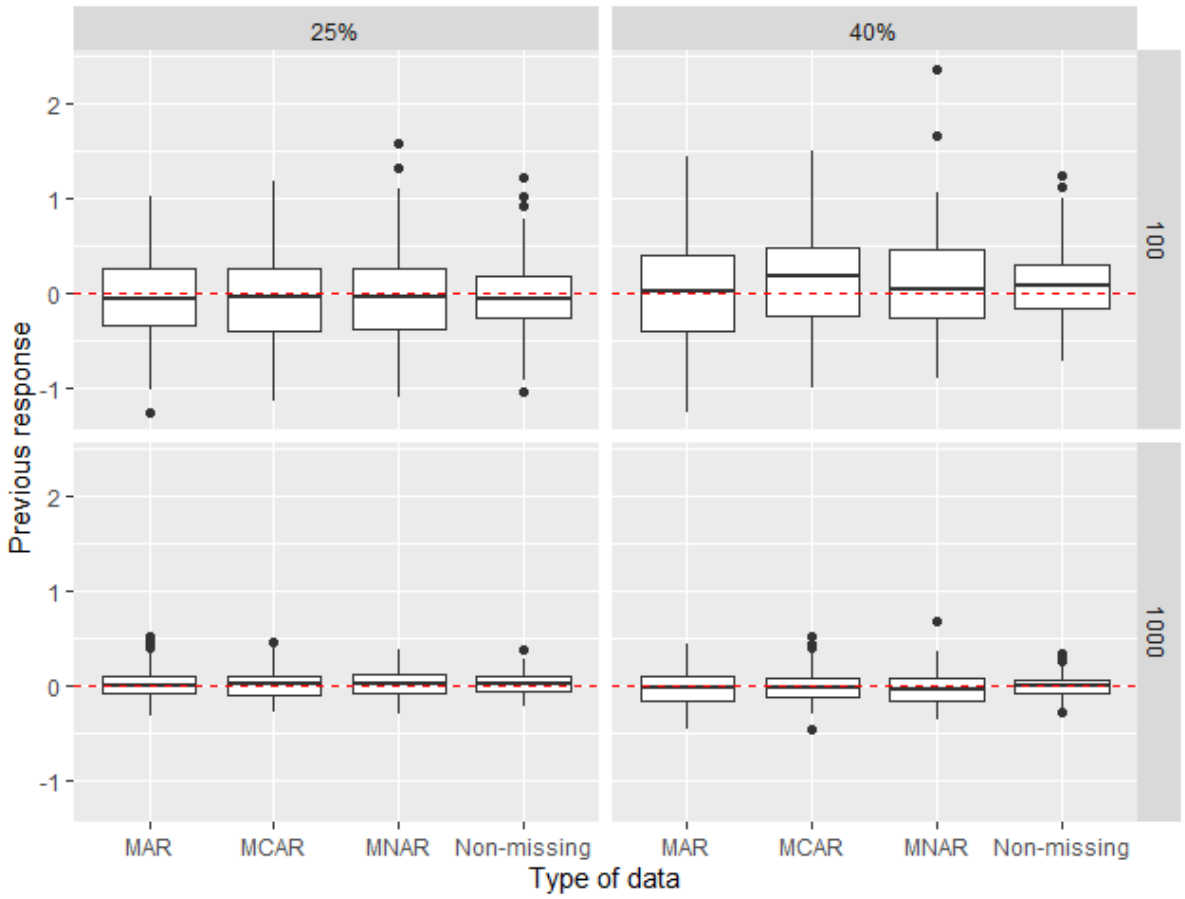

Figure 2: Comparison of the distributions of the bias for the previous response for each type of missingness according to the sample size and the proportion of missing values. The dashed line marks the zero, which indicates that there is no bias.

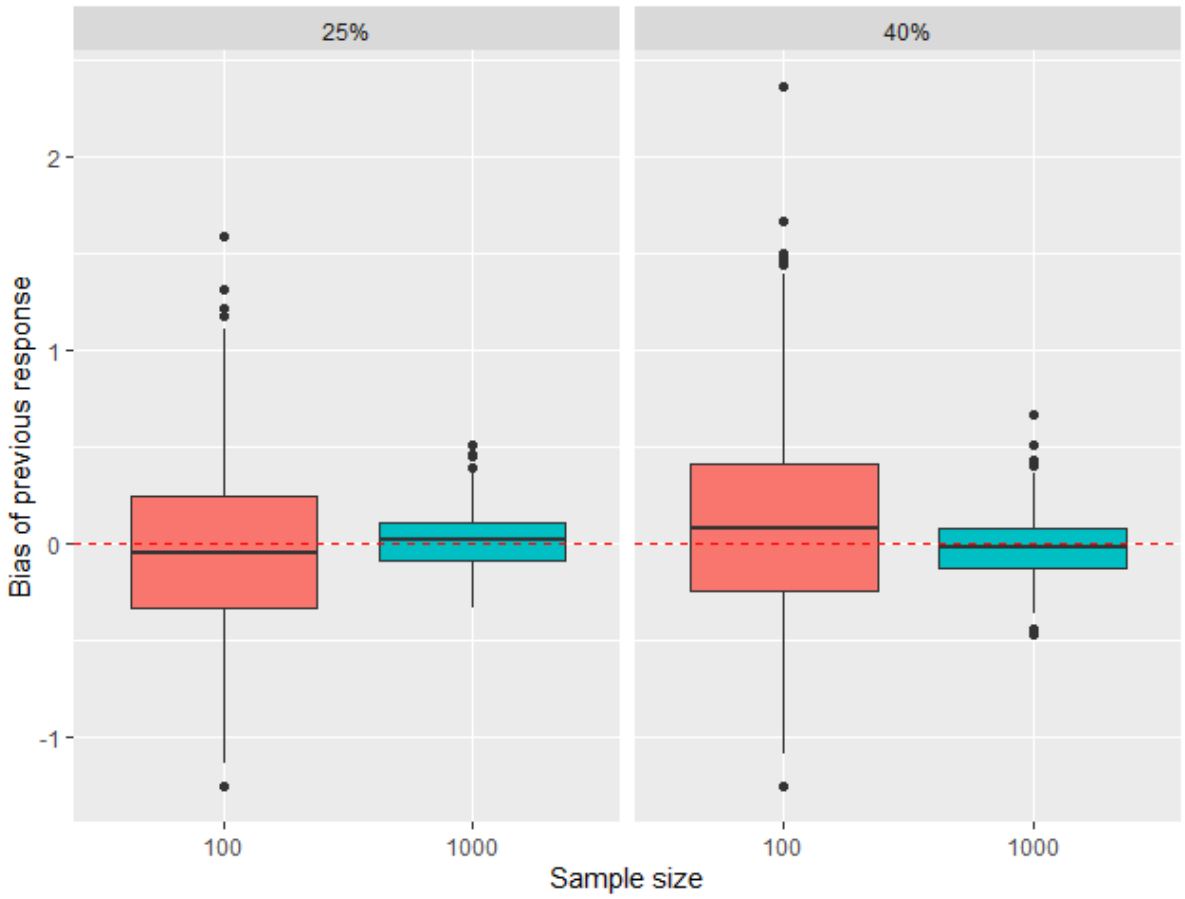

Figure 3: Comparison of the distributions of the bias for the previous response according to the sample size and the proportion of missing values. The dashed line marks the zero, which indicates that there is no bias.

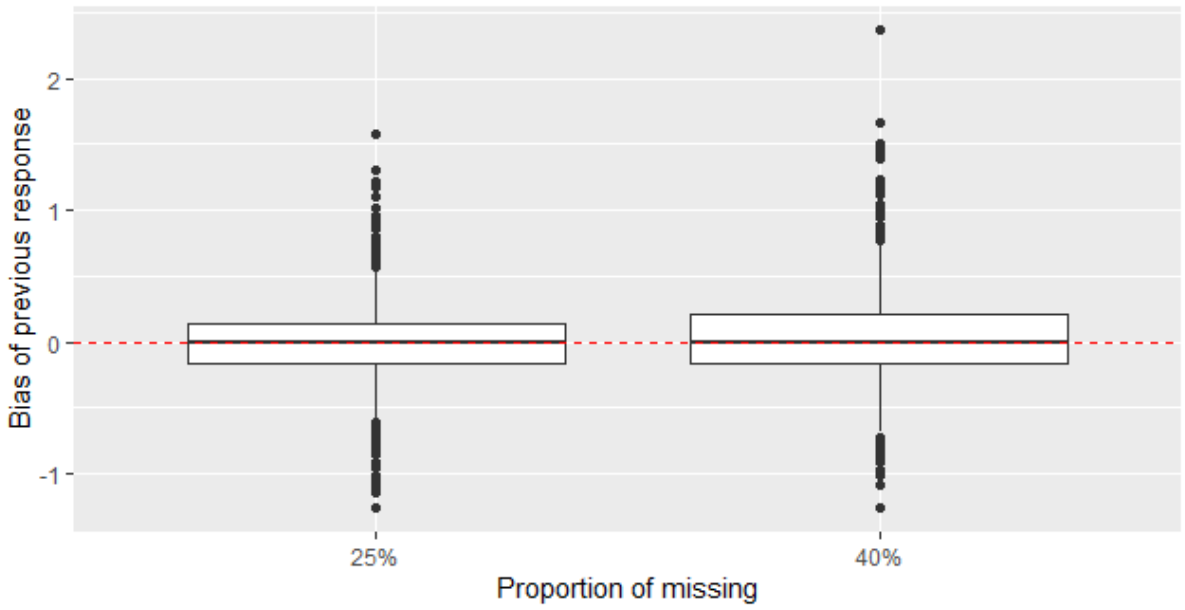

Figure 4: Distributions of the bias for the previous response according to the proportion of missing values. The dashed line marks the zero, which indicates that there is no bias.

3.2.3 Time point

In Table 6 it can be seen that there is no difference in the bias of the coefficients estimated. Thereby, the change in sample size, the proportion of missing values or the type of missing do not generate a variation in the bias.

Table 6: ANOVA results with the time point as dependent variable.

Effect	Df	Sum Sq	Mean Sq	F value	p-value	η^2
Samp	1	0.062	0.0623	1.4173	0.2340	0.0009
Proportion	1	0.009	0.0093	0.2104	0.6465	0.0001
Type	3	0.060	0.0199	0.4539	0.7145	0.0009
Samp:Prop	1	0.007	0.0071	0.1613	0.6881	0.0001
Samp:Type	3	0.046	0.0154	0.3491	0.7898	0.0007
Prop:Type	3	0.066	0.0218	0.4968	0.6846	0.0009
Samp:Prop:Type	3	0.036	0.0119	0.2707	0.8466	0.0005
Residuals	1584	69.644	0.0440			

Note. Samp stands for sample size, Prop means proportion of missing data, and Type refers to the type of missingness. In the header, Df stands for degrees of freedom and Sq refers to squares, ergo Sum Sq is the Sum of squares and Mean Sq is the Mean squares.

Figure 5 also shows that independent of the sample size, type of missingness and proportion of missing values; all biases are around 0. The difference that stood out in the boxes is its height, indicating that the estimations with 1000 participants were more accurate than 100 participants. Also, having 40% of missing data generates more variation than 25%.

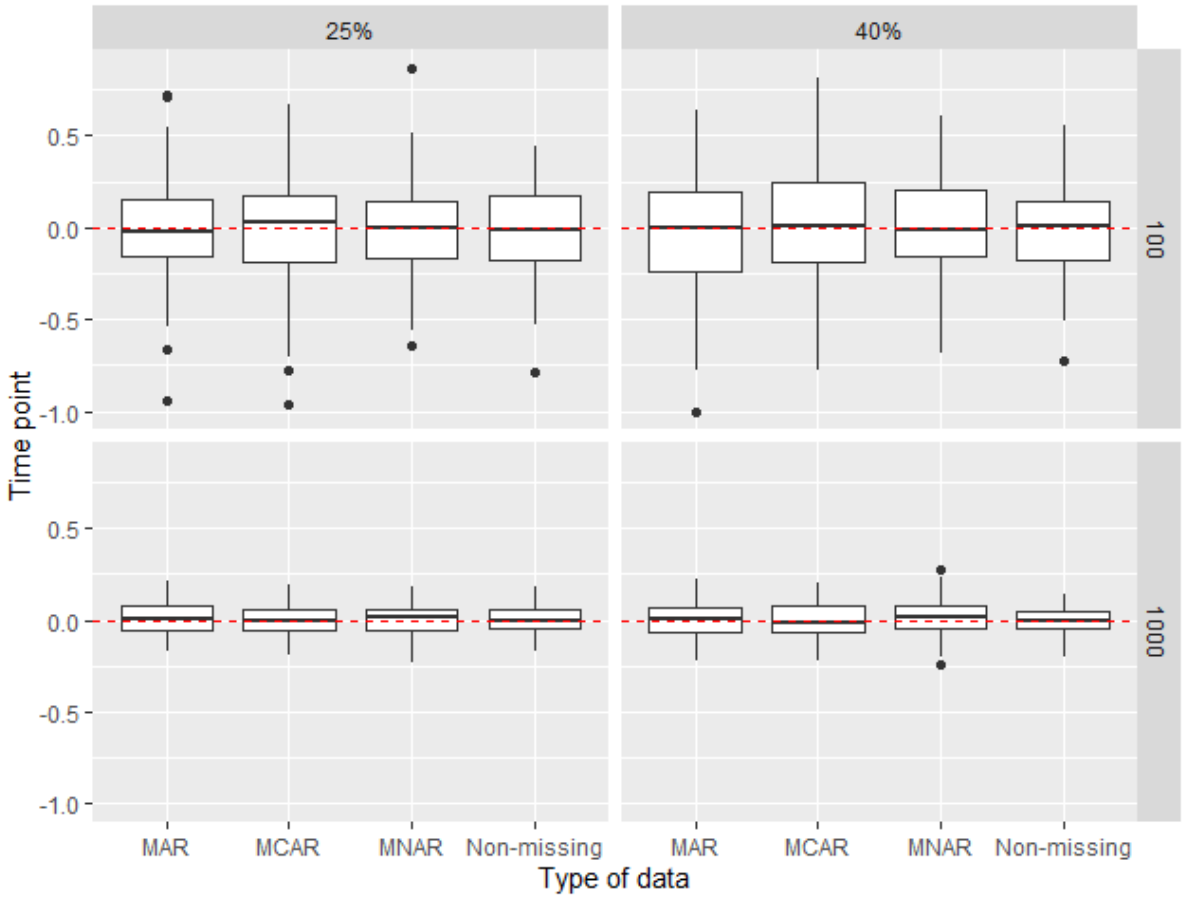

Figure 5: Comparison of the distributions of the deviations of the estimated for the time point for each type of missingness according to the sample size and the proportion of missing values. The dashed line marks the zero, which indicates that there is no bias.

3.2.4 Maternal smoking

Like other estimations, Table 7 shows that all biases are similarly independent of the sample size, the proportion of missing values and type of missingness. In Figure 6, the biases for maternal smoking are close to 0. As before, the large sample size led to more accurate estimated coefficients. Moreover, the datasets with 40% of missing values had a larger variation.

Table 7: ANOVA results for maternal smoking as dependent variable.

Effect	Df	Sum Sq	Mean Sq	F value	p-value	η^2
Samp	1	0.033	0.0334	0.3286	0.5665	0.0002
Proportion	1	0.194	0.1943	1.9104	0.1671	0.0010
Type	3	0.113	0.0378	0.3715	0.7736	0.0007
Samp:Prop	1	0.022	0.0222	0.2178	0.6408	0.0001
Samp:Type	3	0.115	0.0384	0.3772	0.7695	0.0007
Prop:Type	3	0.029	0.0098	0.0966	0.9619	0.0002
Samp:Prop:Type	3	0.033	0.0112	0.1097	0.9544	0.0002
Residuals	1584	161.110	0.1017			

Note. Samp stands for sample size, Prop means proportion of missing data, and Type refers to the type of missingness. In the header, Df stands for degrees of freedom and Sq refers to squares, ergo Sum Sq is the Sum of squares and Mean Sq is the Mean squares.

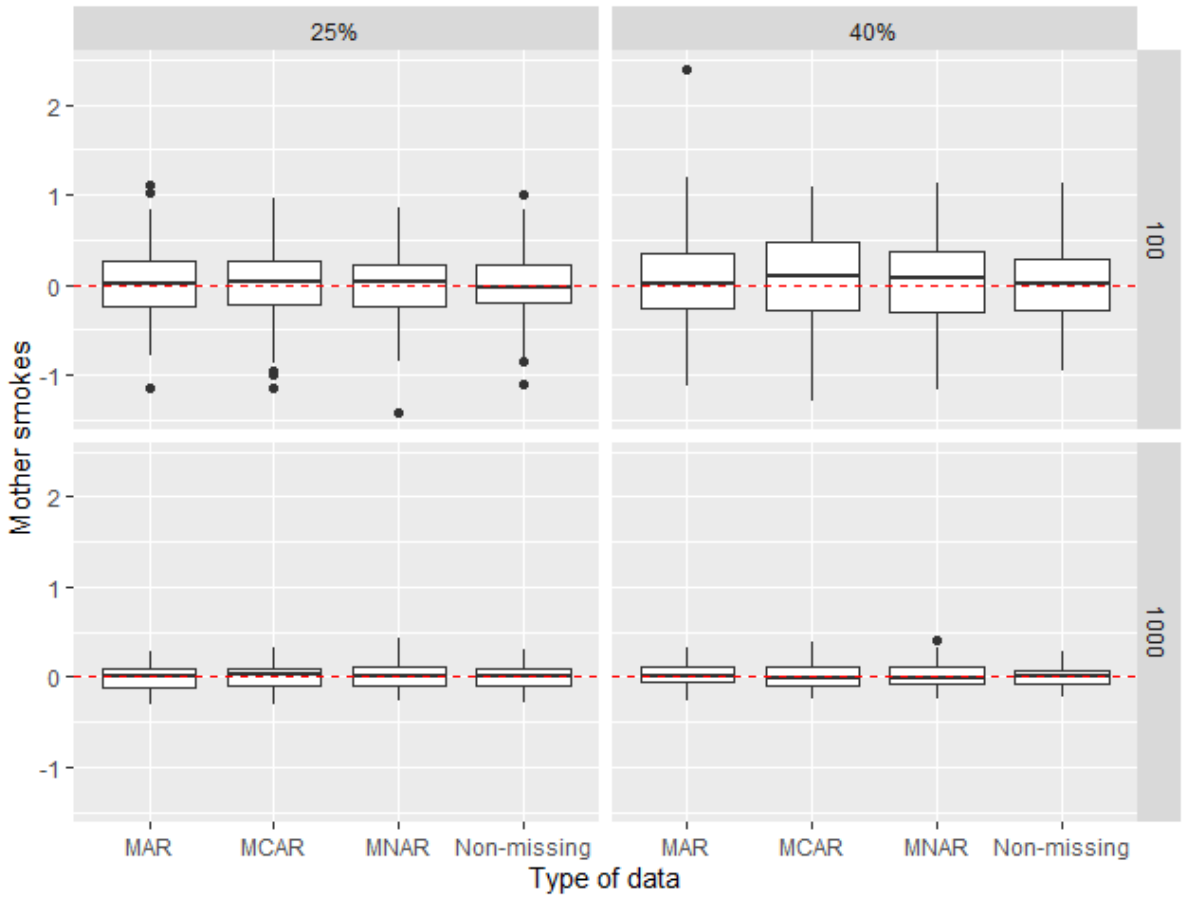

Figure 6: Comparison of the distributions of the deviations of estimated of maternal smoking for each type of missingness according to the sample size and the proportion of missing values. The dashed line marks the zero, which indicates that there is no bias.

4 Discussion and conclusions

Using the missing indicator method in longitudinal binary transitional models leads to estimated coefficients that, on average, are close to the population values regardless of the type of missingness. The simulation of an empirical example with different proportions of missing data, sample size and type of missingness indicated that the deviations of the estimated coefficients on average are close to zero.

These results indicated a difference statistically significant in the bias due to the main effects of sample size, the proportion of missing values, type of missingness and the interaction of sample and proportion. Nevertheless, the effect sizes of these variables were small according to Cohen's rule of thumb (Field, 2009). Concerning the type of

missingness, the differences between MCAR, MAR, MNAR and non-missingness only explained 1% of the deviations variation.

When the coefficients were examined individually, only the bias of the estimated of the previous response was significantly different for the different levels of proportion and the interaction between sample size and proportion of missing values. This result is not surprising, considering that this variable was the predictor with missing values. However, the effect size is small, and the variable of interest is the type of missingness. Therefore, the effect of sample size and the interaction are not in the scope of this study.

The results presented are in line with previous research. The bias that can be seen in the estimated coefficient of the previous response depended mainly on the sample size and the proportion of missing values; the same was found in Rambhadjan (2016) where the missing-indicator was tested in longitudinal data with the three missing mechanisms (MCAR, MAR, and MNAR). Moreover, in the context of transitional modelling de Rooij (2018) found that the use of the missing-indicator also led to unbiased estimations.

The main strength of this study is that the population model is based on an empirical example. Using an actual situation that the researcher can find gives ecological validity to the conclusions. Furthermore, the simulation allowed to test the effect of the types of missing values in scenarios with different proportions of missing data and sample sizes.

It is also important to note that this study has some limitations. The main limitation of this study is that the assumptions for MANOVA and ANOVA were not met. The assumptions of multivariate normality, homogeneity of the covariance matrices, and the absence of outliers were not met. The final decision of performing MANOVA despite the assumptions being not met was based on the literature. Huberty and Olejnik (2006) indicated that the normality is not necessary in the case that the data are symmetrical, which can be seen in the boxplots. Secondly, Box's test used for testing the equality of the covariances matrices is unstable. As a countermeasure, we used Pillai's statistics which is considered as robust (Field et al., 2012; Tabachnick & Fidell, 2013).

Other limitations are related to the simulation. The study included only one set of population parameters. Also, the generalization is limited to other cases that have the same characteristics as the example used. For this reason, the test of the missing indicator in other data and under diverse conditions is essential.

In conclusion, this study showed that it is possible to use the missing-indicator method

in longitudinal binary data in the context of a transitional model with a first-degree Markov chain to obtain unbiased estimations. When the response is missing, the missing-indicator method can handle different types of missing data resulting in unbiased estimated coefficients on average. The generalization of these results is limited to the characteristics of the model and the simulation. For this reason, it is necessary to test the missing-indicator method in other types of data and situations. Also, it is essential to test the missing-indicator method with other methods different to ANOVA, like analyzing the standards errors, the type I error, and the classification performance.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed). Wiley-Interscience.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, *81*(4), 767–775. https://doi.org/10.1093/biomet/81.4.767

Blake, H. A., Leyrat, C., Mansfield, K. E., Tomlinson, L. A., Carpenter, J., & Williamson, E. J. (2020). Estimating treatment effects with partially observed covariates using outcome regression with missing indicators. *Biometrical Journal*, *62*(2), 428–443. https://doi.org/10.1002/bimj.201900041

Bonney, G. E. (1987). Logistic Regression for Dependent Binary Observations. *Biometrics*, *43*(4), 951–973. https://doi.org/10.2307/2531548

Buuren, S. v. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor; Francis Group.

de Rooij, M. (2018). Transitional modeling of experimental longitudinal data with missing values. *Advances in Data Analysis and Classification*, *12*(1), 107–130. https://doi.org/10.1007/s11634-015-0226-6

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Field, A. P. (2009). *Discovering statistics using SPSS: And sex, drugs and rock 'n' roll* (3rd ed). SAGE Publications.

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R* [OCLC: ocn760970657]. Sage.

Ghahroodi, Z. R., Ganjali, M., & Kazemi, I. (2010). Models for longitudinal analysis of binary response data for identifying the effects of different treatments on insomnia. *Applied Mathematical Sciences*, *4*(62), 3067–3082. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.426.9420&rep=rep1&type=pdf

Ginexi, E. M., Riley, W., Atienza, A. A., & Mabry, P. L. (2014). The Promise of Intensive Longitudinal Data Capture for Behavioral Health Research. *Nicotine & Tobacco Research*, *16*(Suppl 2), S73–S75. https://doi.org/10.1093/ntr/ntt273

Greenland, S., & Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264. https://doi.org/10.1093/oxfordjournals.aje.a117592

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., & Moons, K. G. (2012). Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, *184*(11), 1265–1269. https://doi.org/10.1503/cmaj.110977

Henry, A. J., Hevelone, N. D., Lipsitz, S., & Nguyen, L. L. (2013). Comparative methods for handling missing data in large databases. *Journal of Vascular Surgery*, *58*(5), 1353–1359.e6. https://doi.org/10.1016/j.jvs.2013.05.008

Huberman, M., & Langholz, B. (1999). Application of the Missing-Indicator Method in Matched Case-Control Studies with Incomplete Data. *American Journal of Epidemiology*, *150*(12), 1340–1345. https://doi.org/10.1093/oxfordjournals.aje.a009966

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed) [OCLC: ocm61463874]. Wiley-Interscience.

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G., & Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal of Clinical Epidemiology*, *63*(7), 728–736. https://doi.org/10.1016/j.jclinepi.2009.08.028

Li, X., Song, X., & Gray, R. H. (2004). Comparison of the Missing-Indicator Method and Conditional Logistic Regression in 1:m Matched Case-Control Studies with

Missing Exposure Values. *American Journal of Epidemiology*, *159*(6), 603–610. https://doi.org/10.1093/aje/kwh075

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Pereira Barata, A., Takes, F. W., van den Herik, H. J., & Veenman, C. J. (2019). Imputation methods outperform missing-indicator for data missing completely at random. *2019 International Conference on Data Mining Workshops (ICDMW)*, 407–414.

Rambhadjan, J. (2016). *Using the missing indicator method on a dependent binary variable, a simulation study on bias and equivalence of model parameter estimates.* (PhD Thesis). Universiteit Leiden. Netherlands. https://studenttheses.universiteitleiden.nl/access/item%3A2631459/view

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

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

Song, M., Zhou, X., Pazaris, M., & Spiegelman, D. (2021). The missing covariate indicator method is nearly valid almost always. *arXiv preprint arXiv:2111.00138*.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6. ed., international ed). Pearson.

van der Heijden, G. J., T. Donders, A. R., Stijnen, T., & Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, *59*(10), 1102–1109. https://doi.org/10.1016/j.jclinepi.2006.01.015

Wang, C.-Y., & Hsu, L. (2020). Multinomial logistic regression with missing outcome data: An application to cancer subtypes. *Statistics in Medicine*, *39*(24), 3299–3312. https://doi.org/10.1002/sim.8666

Yang, X., Shoptaw, S., Kun Nie, Juanmei Liu, & Belin, T. R. (2007). Markov transition models for binary repeated measures with ignorable and nonignorable missing

values. *Statistical Methods in Medical Research*, *16*(4), 347–364. https://doi.org/10.1177/0962280206071843