ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 728−736

Journal of Clinical Epidemiology

Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example

Mirjam J. Knol[a,b,*], Kristel J.M. Janssen[a], A. Rogier T. Donders[c], Antoine C.G. Egberts[b,d], E. Rob Heerdink[b], Diederick E. Grobbee[a], Karel G.M. Moons[a], Mirjam I. Geerlings[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str. 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands*
[b]*Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands*
[c]*Department of Epidemiology, Biostatistics and HTA, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands*
[d]*Department of Clinical Pharmacy, University Medical Center Utrecht, Utrecht, The Netherlands*

Accepted 11 August 2009

Abstract

Objective: Missing indicator method (MIM) and complete case analysis (CC) are frequently used to handle missing confounder data. Using empirical data, we demonstrated the degree and direction of bias in the effect estimate when using these methods compared with multiple imputation (MI).

Study Design and Setting: From a cohort study, we selected an exposure (marital status), outcome (depression), and confounders (age, sex, and income). Missing values in ''income'' were created according to different patterns of missingness: missing values were created completely at random and depending on exposure and outcome values. Percentages of missing values ranged from 2.5% to 30%.

Results: When missing values were completely random, MIM gave an overestimation of the odds ratio, whereas CC and MI gave unbiased results. MIM and CC gave under- or overestimations when missing values depended on observed values. Magnitude and direction of bias depended on how the missing values were related to exposure and outcome. Bias increased with increasing percentage of missing values.

Conclusion: MIM should not be used in handling missing confounder data because it gives unpredictable bias of the odds ratio even with small percentages of missing values. CC can be used when missing values are completely random, but it gives loss of statistical power.

Keywords: Missing; Confounder; Etiologic; Indicator method; Complete case; Multiple imputation

1. Introduction

In observational etiologic research, missing data on one or more confounders can affect the possibility to adequately adjust for confounding variables. There are many methods to handle missing data [1,2]. Commonly, researchers exclude subjects with missing data, the so-called complete case analysis (CC), because multivariable modeling in standard software packages usually excludes persons with a missing value on any of the variables in the model. Obviously, this affects the number of subjects and, thereby, the statistical power, but more importantly, it may lead to seriously biased estimates [1,3−5]. This bias occurs because missing values are typically related to other observed subject characteristics, including the outcome. Even in a follow-up study where the outcome is not yet known at baseline, missing values can be (indirectly) related to the outcome if predisposing factors are associated with the missing covariate and the outcome. For example, a question on medication use might be skipped by someone with low education level, whereas education level is also associated with mortality.

Another approach to handle missing data is the ''MIM'', which was specifically proposed for missing confounder data in etiologic research [6,7]. This method does not exclude subjects from the analysis but adds an extra variable to the statistical model to indicate that the value of a certain variable is missing. Although it has been argued that the MIM gives biased results [1,3,5,8], it is still used often. A survey of 100 articles showed that 32 of 81 articles mentioned a method for handling covariates of which four (13%) used the MIM [9]. Reasons to use the MIM might be that it is an intuitively appealing method because it seems to adjust for missing values and is easy to use. In addition, it is thought that the MIM only gives some residual confounding [10], and researchers might think that this is acceptable. Furthermore, the MIM is advocated for use of missing baseline measurements in randomized trials [11].

* Corresponding author. Tel.: +31-88-7551168; fax: +31-88-7568099.
E-mail address: m.j.knol@umcutrecht.nl (M.J. Knol).

0895-4356/$ − see front matter
doi: 10.1016/j.jclinepi.2009.08.028

What is new?

- We showed that using the missing indicator method (MIM) gives unpredictable direction of bias.
- We showed the direction and magnitude of bias resulting from using the MIM.
- We gave a clear explanation of why the MIM fails.
- We combined empirical data with simulations to convey our message.
- Implication is that the MIM should never be used to handle missing confounder data.

A more sophisticated approach to handle missing data is to impute (ie, fill in) missing values. Imputing the overall or subgroup mean commonly yields biased effect estimates as well [2]. Imputing a missing value by a value predicted by a regression model using all other observed variables in the data set including the outcome seems a better approach [1,2,4,8,12,13]. This imputation can be done once (single imputation) or multiple times (multiple imputation [MI]). If the missing values are completely random or if they depend on observed variables, single imputation gives unbiased effect estimates but overestimates the precision, that is, underestimates the standard error, of the estimate because it assumes that all data are present [1,8,13]. In MI, the missing value is imputed multiple times (usually 5–10), and the uncertainty of the imputed values is taken into account. MI has been described in numerous articles, and it has been shown to give valid estimates and valid standard errors if missing values are completely random or if they depend on observed variables [1,4,8,12–18].

Although MI is increasingly being used (eg, [16,19,20]), CC and the MIM are still common in the epidemiologic literature [9]. Many researchers, and editors alike, appear not to be aware of the degree of bias that can result from both methods. The objective of this study was to show the direction and degree of bias when using the MIM or CC to handle missing confounder data in an etiological context. In contrast to earlier studies, we used an empirical data set and simulated different patterns of missing data in a confounder. We studied the bias in the odds ratio of the association between exposure and outcome and varied the percentage of missing values from 2.5% to 30%. We compared using the MIM and CC with the use of MI.

2. Methods

2.1. Data set

We used data from the PREDICT study, which is described in detail elsewhere [21]. In short, the PREDICT study is a European prospective cohort study aimed to develop a multifactor risk algorithm for onset of major depression over 12 months. A total of 1,338 subjects were included in the Dutch part of the study (PREDICT-NL). The outcome of interest was the occurrence of major depressive disorder according to DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, version IV) criteria using the depression section of the composite international diagnostic interview [22,23]. Many potential risk factors for depression were measured. For the present analysis, we used the baseline data of the PREDICT-NL study. We selected one particular exposure variable, one outcome variable, and three confounder variables and created missing values in one of the confounders.

As exposure, we used "marital status" where married, including living together and enduring relationships, was coded as 0 and single, including divorced and widowed, was coded as 1. The outcome major depressive disorder was coded 0 when absent and 1 when present. Age, sex, and income were selected as potential confounders. Age was used as a continuous variable; females were coded as 0 and males as 1. Income was defined as the net yearly household income and was categorized into four levels: less than 12,000 euro was coded as 0; 12,000–22,800 as 1; 22,800–50,000 as 2; and more than 50,000 as 3.

From the original PREDICT-NL data set ($n = 1,338$), we selected all subjects without missing values on the five variables ($n = 1,075$) to start the present study with a complete data set. This complete data set, further referred to as "study data set," served as our reference situation. Descriptive statistics and frequencies of the five selected variables are presented in Table 1.The overall frequency of depressive disorders was 14%. The crude odds ratio between marital status and depressive disorder was 2.0 (95% confidence interval [CI]: 1.4, 2.9), indicating that being single doubled the risk of major depressive disorder in comparison with being married. Adjustment for age and sex did not change the odds ratio much (1.8; 95% CI: 1.3, 2.6). Additional adjustment for income decreased the odds ratio to 1.4 (95% CI: 0.9, 2.2), showing that income confounded the association. The odds ratio adjusted for age, sex, and income was defined as the "true" odds ratio to which all later estimates were compared. We emphasize that the data used here are for

Table 1
Distribution of age, sex, and income (confounder variables) and major depressive disorder (outcome variable) in the study data set

	Marital status	
Characteristics	Married $N = 801$	Single $N = 274$
Age, mean (SD)	50.4 (15)	48.1 (20)
Male sex, %	43.8	29.9
Income, %		
<12,000	5.4	33.2
12,000–22,800	23.0	41.2
22,800–50,000	53.8	22.6
>50,000	17.9	2.9
Depressive disorder, %	11.6	20.8

Abbreviation: SD, standard deviation.

illustration purposes only and not to estimate the true causal association between marital status and depressive disorder.

2.2. Missing values

From the study data set, 1,000 new data sets with equal size as the study data set ($n = 1{,}075$) were created by simulating the outcome variable and keeping the other four variables unchanged. In each of the 1,000 data sets, simulation of the outcome was done by first calculating the probability of depressive disorder (P) for each patient using the "true" regression coefficients and the patient values. Subsequently, for each patient a random number from a uniform distribution in the interval [0,1] was sampled. An outcome status of 1 (depressive disorder present) was assigned when the random number was smaller than P and an outcome status of 0 otherwise. By means of this method, a subject with a high probability of disease in the original data set also had a higher chance of having the disease in the simulated data sets. As a result, for example, of all patients with a probability of 0.20 to have depression, 20% should have depression across all simulations, which will be the case when these were assigned an outcome value of 1 if the random number was 0.20 or smaller. Consequently, the resulting coefficients of the covariates were on average the same as in the original data set. Accordingly, on average 14% of the subjects (per data set) had a depressive disorder. By simulating 1,000 new data sets rather than copying the study data set 1,000 times, we mimicked taking samples from a population reflecting the process of sampling 1,000 times a (different) study population from a larger unknown source population. We explicitly mimicked taking samples from a source population because we aimed to draw inferences about a population (value) and not about a single sample. In addition, the terms bias and coverage are only defined if multiple samples are taken from a (in this case) known source population.

In each simulated data set, missing values were created in the confounder variable "income" using four different scenarios (Table 2). In scenario 1, we created missing values according to the missing completely at random (MCAR) mechanism [8]. In randomly selected subjects, the confounder "income" was set to missing irrespective of the values of the other variables. In scenarios 2, 3, and 4 (Table 2) we created missing values according to the missing at random (MAR) mechanism [8]. Missing data are denoted as MAR when missing data occur in relation to other but observed patient characteristics. We created missing values in the confounder "income" in relation to the exposure value and the outcome value. It may seem odd to create missing values in relation to the outcome but, as said above, missing values in a confounder are often indirectly related to the outcome. We distinguished four subgroups of exposure and outcome in which we created missing values in the confounder "income": (1) married and no depressive disorder; (2) married and depressive disorder; (3) single and no depressive disorder; (4) single and depressive disorder. To define the distribution of missing

Table 2
Characteristics of the four different scenarios of creating missing values in the confounder "income"

Characteristics	Scenario 1	Scenario 2			Scenario 3			Scenario 4		
Mechanism	MCAR	MAR			MAR			MAR		
Missing values	Created in randomly selected subjects	Associated with marital status (exposure) and major depressive disorder (outcome)								
Distribution of missing values	Not applicable		Depressive disorder			Depressive disorder			Depressive disorder	
		Marital status	No (0)	Yes (1)	Marital status	No (0)	Yes (1)	Marital status	No (0)	Yes (1)
		Married (0)	3	2	Married (0)	5	1	Married (0)	1	5
		Single (1)	3	1	Single (1)	1	5	Single (1)	5	1
Explanation for distribution of missing values	Not applicable	Realistic distribution of missing values in "income" as encountered in the original PREDICT-NL data set			Missing values were mainly created in married nondepressed subjects and single depressed subjects			Missing values were mainly created in married depressed subjects and single nondepressed subjects		

Abbreviations: MCAR, missing completely at random; MAR, missing at random.

values of ''income'' over the four categories, we used odds ratios of missingness (Table 2). We defined three different sets of odds ratios. In scenario 2 (Table 2), the set of odds ratios was 3-2-3-1 representing the distribution of missing values of the variable ''income'' in the original PREDICT-NL data set. In scenario 3 (Table 2), we used a set of odds ratios of 5-1-1-5 where missing values were particularly created in the married nondepressed subjects (exposure = 0, outcome = 0) and the single depressed subjects (exposure = 1, outcome = 1). In scenario 4, (Table 2), a set of odds ratios of 1-5-5-1 was used to create missing values mainly in the married depressed subjects (exposure = 0, outcome = 1) and the single nondepressed subjects (exposure = 1, outcome = 0). In scenarios 3 and 4, we created a large contrast in the percentage of missing income values between the categories that determine that being *single* is a risk factor for depression and those that determine that being *married* is a risk factor for depression. By creating this contrast, the possible bias of the methods to handle missing data becomes most apparent.

Five different percentages of missing values in the confounder ''income'' were simulated namely 2.5%, 5%, 10%, 20%, and 30%. These percentages are not uncommon in etiologic research.

We did not create missing values according to the missing not at random (MNAR) mechanism. MNAR implies that the probability that an observation is missing depends on unobserved characteristics, such as the value of the observation itself.

2.3. Methods to handle missing data

We compared three methods to handle missing data: the MIM, CC, and MI. With MIM, the missing values in ''income'' were recoded as 0, and an extra indicator variable was created and coded as 1 if the value on ''income'' was missing and 0 otherwise. Both the recoded ''income'' variable and indicator variable were included in the logistic regression model. Note that this is the same as adding an extra category for missing values to the categorical ''income'' variable.

In CC, all subjects with missing values were excluded from the logistic regression analysis.

In MI, all observed data were used to estimate the missing values with regression analysis using multivariate imputation by chained equations [24]. This estimation was performed five times to get variation in the imputed values. Then conventional logistic regression analysis was performed in each of the five data sets, and the results were pooled in a way that reflects the extra variability because of uncertainty of the imputed values [25].

2.4. Bias and coverage

For all situations (ie, four different scenarios of creating missing values, five different missing value percentages, and three different methods to handle these missing values), we calculated the mean odds ratio over the 1,000 simulations for the association between marital status and depressive disorder adjusted for age, sex, and income. These mean odds ratios were compared with the true odds ratio estimated from the study data set, which was 1.4 (95% CI: 0.9, 2.2). We assessed the coverage of the 95% CI of the odds ratio in each situation. To this aim, we calculated the 95% CI around the estimated odds ratio in each simulated data set and assessed whether the true odds ratio was included. The coverage was then calculated as the percentage of 95% CIs over the 1,000 simulations that indeed included the true odds ratio. Coverage of 0.95 represents correct coverage. Note that coverage is not the same as the CI around the estimated effect. It is possible that the CIs around the effect estimates in the 1,000 data sets in some of the scenarios are wider with than without missing values (eg, CC), but yet the coverage is 95%. Furthermore, note that the coverage we calculated is the overall coverage. It could be asymmetrical in that there is not 2.5% on each side, but there is, for example, 1.0% on one side and 4.0% on the other side.

All analyses were performed with R2.4.1 (R Development Core Team, Vienna, Austria) [26].

3. Results

3.1. Scenario 1—missing completely at random

Using the MIM gave an overestimation of the odds ratio of marital status (Fig. 1A). This bias increased with an increasing percentage of missing values. Both CC and MI gave an unbiased odds ratio of exposure. The coverage of the 95% CI was around 0.95 for all methods of handling missing data (Fig. 1B).

3.2. Scenario 2—MAR, odds ratios of 3-2-3-1

Using MIM resulted in a small overestimation of the odds ratio of exposure with 30% of missing values (Fig. 2A). CC gave an overestimation of the odds ratio with more bias in higher percentages of missing values up to an odds ratio of 1.6 (95% CI: 0.98, 2.6) in case of 30% missing values. The estimates of the odds ratio after MI were close to the true odds ratio for all percentages of missing values. MIM, CC, and MI all showed coverage of about 0.95 (Fig. 2B).

3.3. Scenario 3—MAR, odds ratios of 5-1-1-5

There was an overestimation of the odds ratio using MIM and an underestimation of the odds ratio using CC (Fig. 3A). With both methods, the bias increased with an increasing percentage of missing data up to an odds ratio of 1.7 (95% CI: 1.1, 2.6) for MIM and 0.68 (95% CI: 0.41, 1.1) for CC with 30% missing values. MI resulted in an unbiased odds ratio for all percentages of missing

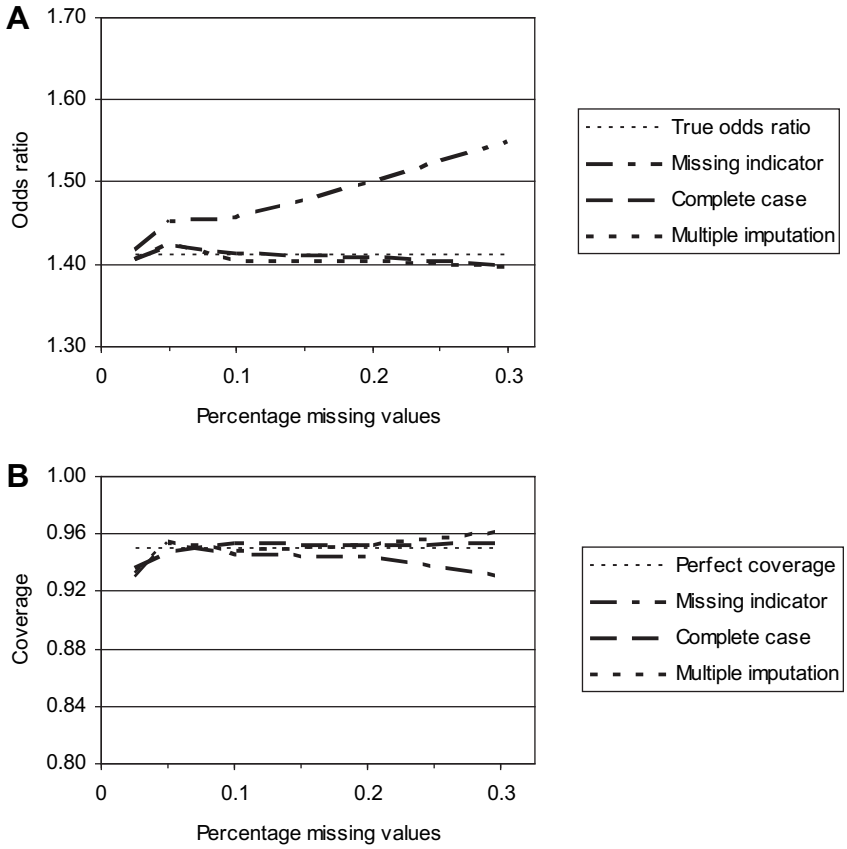

Fig. 1. Scenario 1—odds ratio (A) of marital status and coverage of 95% confidence interval (B) for different methods of handling missing data when missing values in income were created according to the missing completely at random mechanism compared with the true odds ratio (1.4) and correct coverage (0.95).

values. Coverage was lower than 0.95 in MIM and CC with coverage of even 0.85 and 0.17, respectively, for 30% missing values (Fig. 3B), indicating that the true odds ratio was not included in the 95% CI in 15% and 83% of the simulations, whereas this should be only 5%. MI gave good coverage for all percentages of missing values.

3.4. Scenario 4—MAR, odds ratios of 1-5-5-1

The odds ratio was underestimated when using MIM and largely overestimated when using CC and showed increasing bias with an increasing percentage of missing values (Fig. 4A). MI gave unbiased odds ratios. Coverage was lower than 0.95 using MIM and very low using CC, whereas MI gave good coverage (Fig. 4B).

4. Discussion

This study demonstrates the degree of bias in the effect estimate of exposure when using the MIM and CC for missing confounder data in comparison with MI. MIM and CC gave a biased odds ratio in almost all situations of missing confounder data. The direction and degree of that bias depended on how the missing values were related to exposure and outcome. The bias was already present with a small percentage of missing values and increased when the percentage of missing values increased. Coverage of the 95% CI when MIM and CC were used was often too low meaning that the CI was too liberal. This implies that using MIM and CC to handle missing confounder data would lead to a statistical significant odds ratio of the exposure at interest, whereas in reality there is no association. MI gave unbiased odds ratios and good coverage in all scenarios up to 30% of missing confounder values.

Why do MIM and CC result in an over- or underestimation of the odds ratio of marital status? If missing values are MCAR, MIM results in incomplete adjustment for the confounder because all missing values are set to 0, and, therefore, in an overestimation of the exposure effect. Although uncommon in practice, when missing values are MCAR the direction of bias when using MIM is indeed predictable. Obviously, the degree of bias when using MIM for missing confounder data depends on the strength of that confounder.

If missing values depend only on other variables in the data set (ie, MAR), the bias when using MIM and CC analysis is not predictable. The degree and direction of the bias depends on the distribution of missing values over the exposure–outcome categories. In scenario 2, the missing values were quite similarly distributed over the four categories, which resulted in minimal bias when using MIM and overestimation when using CC. When the missing

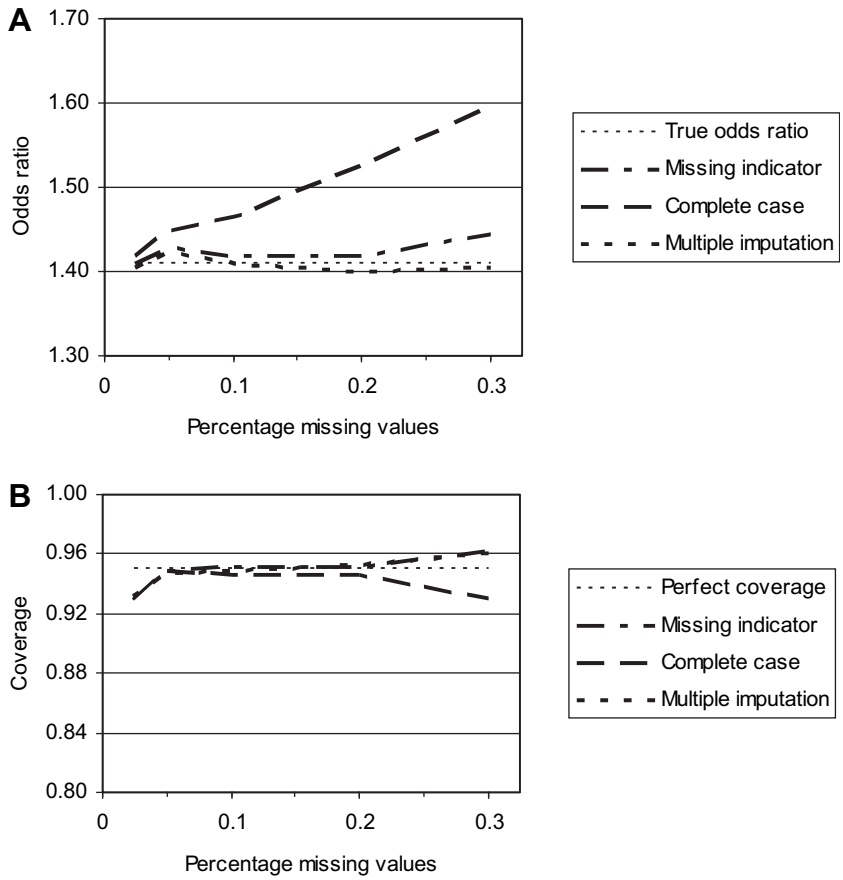

Fig. 2. Scenario 2—odds ratio (A) of marital status and coverage of 95% confidence interval (B) for different methods of handling missing data when missing values in income were realistically created according to the missing at random mechanism with odds ratios for missing values of 3-2-3-1 compared with the true odds ratio (1.4) and correct coverage (0.95).

values were mainly created in the categories that determine that being *single* (marital status = 1) is a risk factor for depressive disorder (scenario 3), MIM gave an overestimation and CC an underestimation of the odds ratio of marital status. Having relatively more missing values in these categories resulted in incomplete adjustment for income and, thus, overestimation of the odds ratio for marital status when using MIM. CC resulted in an underestimation in this scenario because the number of analyzable subjects that determine the association between being *single* and depressive disorder decreased. Creating the missing values mainly in the categories that determine that being *married* is a risk factor for depressive disorder, MIM resulted in an underestimation and CC in an overestimation (scenario 4). Creating missing values particularly in these categories, MIM again resulted in incomplete adjustment for income but now yielded an underestimation of the effect as being married was inversely associated (ie, odds ratio < 1) with depressive disorder. CC resulted in this scenario in an overestimation because there were fewer subjects contributing to the association between being married and depressive disorder.

Hence, in scenarios with a large contrast in number of missing values between the categories that determine that being *single* is a risk factor for depression and those that determine that being *married* is a risk factor for depression, such as how we created in scenarios 3 and 4, it seemed possible to predict the degree and direction of bias. However, if the contrast is not so clear, such as in scenario 2, the direction of bias of using CC and MIM is less predictable. In addition, commonly the missing value mechanism is not explicitly known.

Previous studies also evaluated the potential bias of CC or the MIM [1,3,5,17,27]. These articles also showed that CC is unbiased when missing values are MCAR and biased when missing values are MAR [1,3,5,27]. They also showed that MIM is biased when missing values are MCAR or MAR [1,3,5]. It is hard to compare the amount of bias found in our article with that found in other articles because the amount (and direction) of bias heavily depends on the mechanism of missingness, which we showed in our article.

What should researchers do with missing confounder data? If missing values depend on observed data (ie, MAR) but the missing value mechanism is not exactly known, MIM and CC give an unpredictable degree and direction of the bias and should not be used. MIM might be used safely if missing values are MCAR, the percentage of missing values is low ($<5\%$), and the confounder is weak. However, the situation that all the three criteria are met is unlikely to occur. In

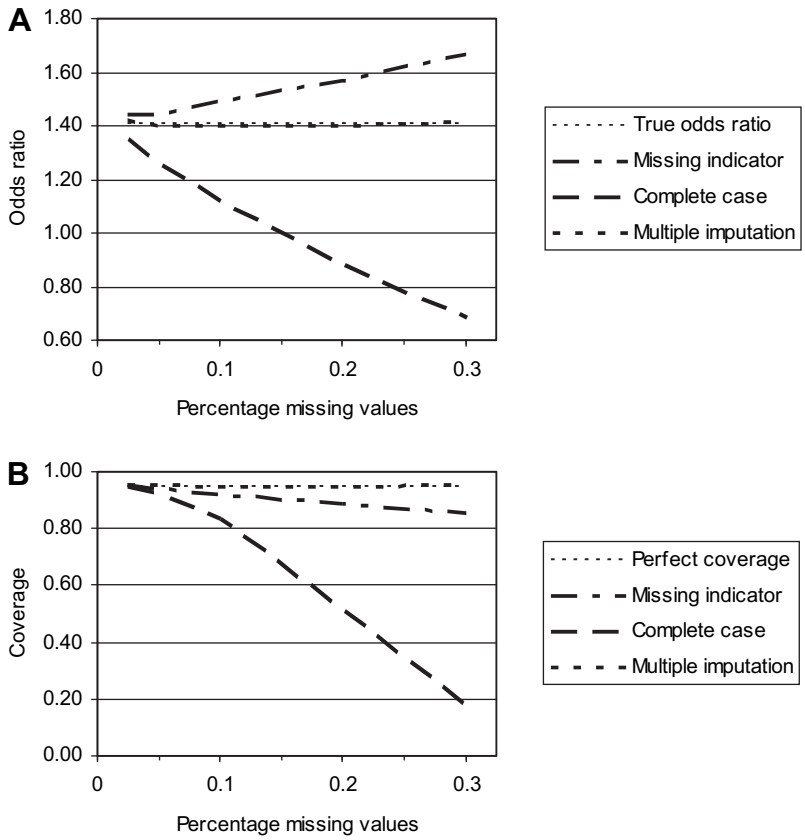

Fig. 3. Scenario 3—odds ratio (A) of marital status and coverage of 95% confidence interval (B) for different methods of handling missing data when missing values in income were created according to the missing at random mechanism with odds ratios for missing values of 5-1-1-5 compared with the true odds ratio (1.4) and correct coverage (0.95).

addition, CC performs equally well as MIM in this situation and is easier to perform. MI performs well when missing values are MCAR or MAR even up to 30% of missing values and is, therefore, the method of choice. Researchers might be reluctant to apply MI because it is believed to be time consuming, difficult to perform, and a self-fulfilling prophecy (because information on exposure and outcome is used to impute missing values and subsequently the exposure–outcome relationship is estimated using the imputed values). However, an increasing number of studies appearing on the application of MI and software has much improved [28] making the method better accessible. An alternative that is still better than MIM and CC analysis [8,12] might be to perform single imputation. An advantage of single imputation is that it is easier to understand and that user-friendly statistical packages can be used. A disadvantage is that it still produces too precise estimates (too small standard errors), which increases the chance of a type I error [1].

In this study, we made certain choices in creating missing values. Missing data patterns might be more complex in practice and, therefore, the degree and direction of bias might be even more unpredictable. First, we let the missing values depend on the exposure and outcome only, whereas missing values can also be related to more variables. Second, we created missing values in one confounder variable, whereas data can be missing in more than one confounder or in the exposure or outcome variable. In addition, a low percentage of missing data in several variables will easily result in a large percentage of subjects with at least one missing value. Third, we created missing values according to the MCAR and MAR mechanism and not according to the MNAR mechanism. MNAR implies that the probability that an observation is missing depends on unobserved characteristics, such as the value of the observation itself. Missing values in medical research are typically neither MCAR nor MNAR but often MAR [8]. In medical research, usually many variables are measured in the participants and, therefore, the chance that the missing values depend on observed data (ie, MAR) is quite high, allowing for sophisticated imputation to reduce the potential for biased results. However, it is a serious problem if missing values are MNAR because there is no general method to handle these missing values [4,8,14,29].

In conclusion, we showed that one should not use the MIM to handle missing confounder data because it gives a biased estimation of the odds ratio of exposure even with small percentages of missing values. More importantly, the direction of the bias is unpredictable. CC can be used when missing values are MCAR. This, however, is hardly ever the case. Moreover, CC always leads to loss of statistical power. MI gives unbiased effect estimates when missing

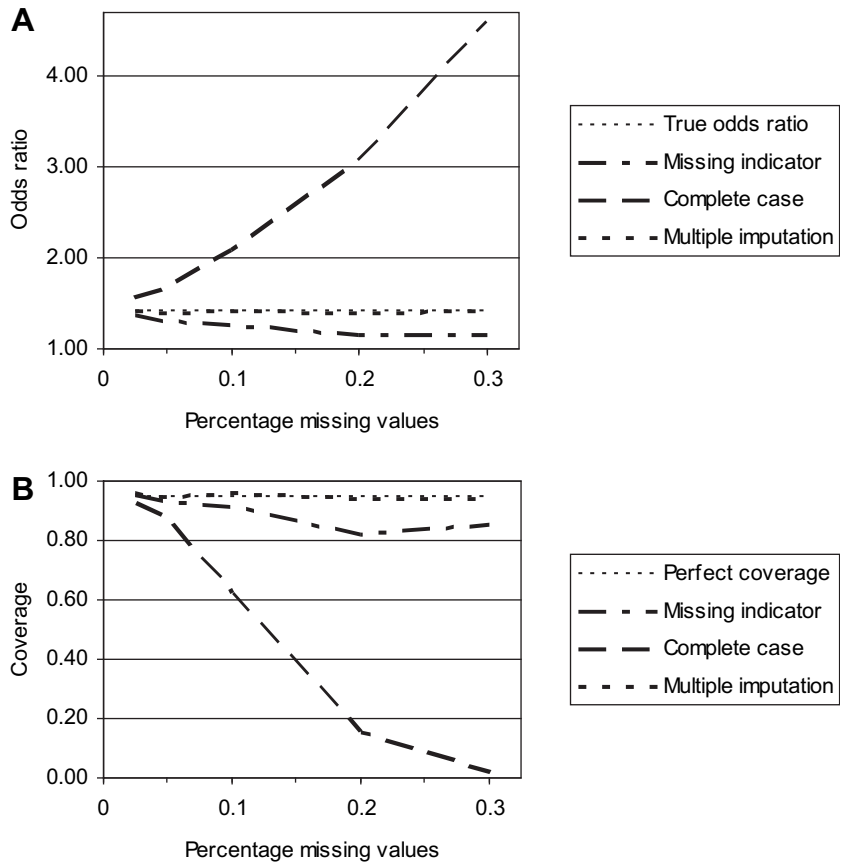

Fig. 4. Scenario 4—odds ratio (A) of marital status and coverage of 95% confidence interval (B) for different methods of handling missing data when missing values in income were created according to the missing at random mechanism with odds ratios for missing values of 1-5-5-1 compared with the true odds ratio (1.4) and correct coverage (0.95).

values are MCAR and when they depend on observed data (MAR). The latter is often the case in medical research, and, hence, MI can be used in most situations with missing confounder data.

Acknowledgments

This research was funded by an unrestricted grant from Novo Nordisk and the Scientific Institute of the Dutch Pharmacists (WINAp) and by the Netherlands Organization for Scientific Research, a VIDI grant (NWO: project no. 917-66-311). The PREDICT study was funded by The European Commission, reference: PREDICT-QL4-CT2002-00683.

References

[1] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255−64.

[2] Little RJ. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227−37.

[3] Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. J Am Stat Assoc 1996;91:222−30.

[4] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147−77.

[5] Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895−907.

[6] Anderson AB, Basilevsky A, Hum DPJ. Missing data: a review of the literature. In: Rossi PH, Wright JD, Anderson A, editors. Handbook of survey research. New York, NY: Academic Press; 1983. p. 415−92.

[7] Regression analysis. In: Miettinen OS Theoretical epidemiology: principles of occurrence research. New York, NY: John Wiley & Sons; 1985. p. 216−43.

[8] Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087−91.

[9] Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer 2004;91:4−8.

[10] Vandenbroucke JP, von Elm EE, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Epidemiology 2007;18:805−35.

[11] White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. Stat Med 2005;24:993−1007.

[12] Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092−101.

[13] Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473−89.

[14] Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10:585−98.

[15] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18:681−94.
[16] Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. Am J Epidemiol 2003;157:74−84.
[17] van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59:1102−9.
[18] Crawford SL, Tennstedt SL, McKinlay JB. A comparison of anlaytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995;48:209−19.
[19] Goldstein AM, Martinez M, Tucker MA, Demenais F. Gene-covariate interaction between dysplastic nevi and the CDKN2A gene in American melanoma-prone families. Cancer Epidemiol Biomarkers Prev 2000;9:889−94.
[20] Knol MJ, Heerdink ER, Egberts AC, Geerlings MI, Gorter KJ, Numans ME, et al. Depressive symptoms in subjects with diagnosed and undiagnosed type 2 diabetes. Psychosom Med 2007;69:300−5.
[21] King M, Weich S, Torres F, Svab I, Maaroos H, Neeleman J, et al. Prediction of depression in European general practice attendees: the PREDICT study. BMC Public Health 2006;6:6.
[22] World Health Organization. Composite international diagnostic interview (CIDI). Version 2.1. Geneva, Switzerland: WHO; 1997.
[23] Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, et al. The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry 1988;45:1069−77.
[24] Buuren van S, Oudshoorn K. Flexible multivariate imputation by MICE. Leiden, The Netherlands: TNO Prevention and Health; 1999.
[25] Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, NJ: John Wiley & Sons; 1987.
[26] R Development Core Team. R: a language and environment for statistical computing. ISBN 3-900051-07-0. Available at. Vienna, Austria: R Foundation for Statistical Computing; 2006. http://www.R-project.org.
[27] Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006;59:1115−23.
[28] Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61:79−90.
[29] Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 1987.