Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study

Antoine Lewin, Ruben Brondeel, Tarik Benmarhnia, Frédérique Thomas, Basile Chaix

Type of manuscript: Original article

Attrition bias related to missing outcome data: a longitudinal simulation study

Antoine Lewin,[a,b] Ruben Brondeel,[a,b,c] Tarik Benmarhnia,[d] Frédérique Thomas,[e] Basile Chaix,[a,b]

[a]*Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, 75012, Paris, France*

[b]*Inserm, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, 75012, Paris, France*

[c]*EHESP School of Public Health, 35000, Rennes, France*

[d]*Department of Family Medicine and Public Health & Scripps Institution of Oceanography, UC San Diego, CA*

[e]*Centre d'Investigations Préventives et Cliniques, 6 rue La Pérouse, 75116 Paris, France*

Correspondence: Antoine Lewin, Inserm U1136, Faculté de Médecine Saint-Antoine, 27 rue Chaligny, 75012, Paris, France. Phone: +33(0)1 44 73 89 54. Fax: +33 (0)1 44 73 84 62. Email: antoine.lewin@iplesp.upmc.fr

Running head: Attrition bias in longitudinal studies

Financial support

This work was supported by a doctoral grant of Région Île-de-France attributed to Antoine Lewin. The RECORD study is funded by the Institute for Public Health Research (IReSP, Institut de Recherche en Santé Publique); the National Institute for Prevention and Health Education (INPES, Institut National de Prévention et d'Education pour la Santé) (Prevention Program 2007; 2010–2011 financial support; 2011–2013 financial support; 2012-2014 financial support); the National Institute of Public Health Surveillance (InVS, Institut de Veille Sanitaire) (Territory and Health Program); the French Ministries of Research and Health (Epidemiologic Cohorts Grant 2008); the National Health Insurance Office for Salaried Workers (CNAM-TS, Caisse Nationale d'Assurance Maladie des Travailleurs Salariés); the Ile-de-France Regional Health Agency (ARS, Agence Régionale de Santé); the Ile-de-France Regional Council (Conseil Régional d'Île-de-France, DIM SEnT and CODDIM); the National Research Agency (ANR, Agence Nationale de la Recherche) (Health–Environment Program 2005); the City of Paris (Ville de Paris); and the Ile-de-France Youth, Sports, and Social Cohesion Regional Direction (DRJSCS, Direction Régionale de la Jeunesse, des Sports et de la Cohésion Sociale).

Acknowledgments

We are grateful to INPES (and Pierre Arwidson) for its continued support since the beginning of the study. We are grateful to Insee, the French National Institute of Statistics and Economic Studies, which provided support for the geocoding of the RECORD participants and allowed us to access to relevant geographical data (with special thanks to Pascale Breuil). We thank Geoconcept for allowing us to access to the Universal Geocoder software. We also thank CNAM-TS and the Caisse Primaire d'Assurance Maladie de Paris (CPAM-P, France) for helping make this study possible.

Conflict of interest: none declared.

ABSTRACT

Background: Using empirical data, this article first examined how inverse probability weighting (IPW) and multiple imputation (MI) handled missing outcome data from attrition in the association between individual education and change in body mass index (BMI). Second, simulating additional attrition, we quantified the impact of attrition and assessed how MI performed compared to complete case analysis (CCA) and to a perfectly specified IPW model as gold standards.

Methods: We used data from the two waves of the French RECORD Cohort Study (N = 7,172). After analyzing attrition bias in the observed data (stage 1), we simulated additional missing data in BMI at follow-up under various Missing At Random (MAR) scenarios. IPW and MI analyses in stage 1 and MI in stage 2 were assessed in their ability to account for attrition bias.

Results: With the observed data in stage 1, an inverse association was found between individual education and change in BMI, in CCA as well as with IPW and MI. When additional attrition was simulated under a MAR pattern (stage 2), the bias increased with the magnitude of selective attrition, and MI was useless to address it.

Conclusion: Our simulations revealed that selective attrition in the outcome heavily biased the association of interest. The present article contributes to raise the awareness that for missing outcome data MI does not do better than CCA. More effort is thus needed during the design phase to understand attrition mechanisms by collecting information on the reasons for dropout.

INTRODUCTION

Most studies with a longitudinal design do not address potential biases due to selective attrition.[1] The potential of missing data, in the covariates or in the outcome, to compromise the validity of research results has often been overlooked.[2,3] Little and Rubin[2,4] describe three categories of missing data mechanisms: "Missing Completely At Random" (MCAR), i.e., missing cases are not different than non-missing cases; "Missing At Random" (MAR), i.e., any systematic differences between the missing values and the observed values can be explained by differences in observed data; and "Missing Not At Random" (MNAR), i.e., the probability of missingness depends on an event that the researcher has not measured, for example on the true value of the missing data. Unfortunately, it is impossible when handling observational data to determine with certainty whether the data are MAR or MNAR.[5-7]

When data are missing in the outcome, the most common approach is complete case analysis (CCA), i.e., simply excluding individuals with missing data.[8] In a review of 262 studies published in 2010 in three leading epidemiological journals, Eekhout et al. found that 81% of these studies used CCA.[1] Estimations obtained from CCA are valid when data are MCAR because complete cases are a representative subsample of the sample.[2,9] However, estimates may be biased if excluded individuals are systematically different from those included (MAR or MNAR).[2,6-8]

Multiple Imputation (MI)[9,10] has been proposed as an alternative to CCA for missing covariates or exposures. MI uses data on all subjects (including those with missing data), creating a number of imputed datasets by generating multiple imputed values for each missing data. Each imputed dataset is analyzed separately, and their estimates are combined using Rubin's rules (1987).[10,11] MI yields correct estimators if the imputation model is correctly specified and the

data are MAR.[2,11-13] However, some authors have suggested that when data are missing only in the outcome, MI does not address the bias and can add needless noise to the estimates.[14,15]

In parallel, inverse probability weighting[16] (IPW) has been proposed as a strategy to mitigate attrition biases, when a MAR pattern is assumed. In this method complete cases are weighted by the inverse probability of being a complete case.[8] IPW does not model the distribution of the partially observed variables, but models the determinants of missingness.[8,11]

Considering a mix of empirically observed and simulated attrition, the present article examines the magnitude of bias related to missing outcome data from selective attrition and the performance of approaches to handle missing data. As a case study, we investigated the association between individual education and change in body mass index (BMI) between the first and second waves of the French RECORD Study.[17] A MAR pattern may apply to missing BMI due to material, behavioral, and psychological variables influencing nonparticipation in the follow-up (motivated nonparticipation) that also influence BMI change. In the first stage, using data with the observed attrition, we examined how IPW and MI handled missing outcome data compared to CCA. In a second stage (simulation), we artificially introduced increasing selective attrition following different MAR mechanisms through the manipulation of the empirical data, and examined both the magnitude of the resulting bias and how MI performed compared to CCA.

METHODS

Study population

Data from the first and second waves of the RECORD Cohort Study (www.record-study.org)[17] were used for longitudinal analyses. During the first wave, 7,290 participants aged 30-79 years at their inclusion were recruited without *a priori* sampling in 2007-2008 during free standardized preventive medical checkups conducted by the Centre d'Investigation Préventive et Clinique in

the Paris metropolitan area.[18-20] Only participants residing in 10 (out of 20) administrative districts of Paris or in 111 other municipalities in the region were selected. In the first wave, 83% of the eligible participants at the health centers agreed to participate and completed the data collection.[17] During the second wave, 3,746 participants were reexamined between 2011 and 2013. The overall revisit rate at wave two was 51%. The French Data Protection Authority approved the study protocol.

After excluding participants with missing values for Body Mass Index (BMI) in the first wave, the sample comprised 7,172 participants, of which 3,693 had information on BMI in the second wave.

Measures

Outcome of interest

BMI (kg/m^2)[21] was calculated at each wave, using height (measured with a wall mounted stadiometer) and weight (measured with calibrated scales) recorded by a nurse (standard procedure of the IPC Medical Center[22]). The longitudinal change in BMI from baseline to follow-up was the outcome of interest.

In the first wave, 4.5% of the participants were recruited in other sites than the Paris site of the IPC Medical Center. Nurses in the different sites used the same stadiometers and scales and were trained to apply the same procedure. In the second study wave, all participants were assessed in the Paris site.

Main exposure

As the main exposure, personal education was coded in two classes, high vs. low (self-administered questionnaire). High education corresponded to completing higher secondary school

(i.e., receiving the French baccalaureate, corresponding to 12 years of schooling after preschool) or above.

Covariates

Age at baseline (in years) and sex were considered. Following our previous empirical work,[18,23,24] the socioeconomic status of the neighborhood was assessed with the educational level of residents (percentage of residents with >2 University years), using data from the 2006 population census geocoded at the building address by INSEE (French National Institute of Statistics and Economic Studies). The variable was computed within a street network buffer with a radius of 1000m centered on the participants' residences. ArcInfo 10 (ESRI, Redlands, CA) and its Network analyst were used to derive such buffers based on street network data from the National Geographic Institute. Neighborhood education was used as a continuous variable. Stress, related to whether participants find their lives unpredictable, uncontrollable, and overloaded, was assessed with the Perceived Stress Scale of Cohen.[25] Depressive symptomatology was evaluated with the QD2A scale of Pichot.[26] The perceived stress and depressive symptomatology scores were entered as continuous variables.

Study design and statistical analysis

Stage 1: empirical analysis without added simulated attrition

Three methods (i.e., CCA, MI and IPW) were first applied to the observed dataset. In CCA, only subjects with an observed value for the outcome (change in BMI) were included in the linear regression analysis (the complete dataset comprised 3,693 participants after excluding those with missing BMI in the second wave).

Regarding MI, we created five imputed datasets (N = 7,172),[4,6,27] analyzed them separately (linear regression with change in BMI as the outcome), and combined the results from the different datasets into a single set of parameter estimates and standard errors. The covariates used in the imputation model for change in BMI were age, sex, individual education, residential neighborhood education, baseline BMI, perceived stress, and depressive symptoms (selected on the basis of hypotheses of causal effects on this outcome). MI (by chained equations) was carried out in R using the mi package.[28]

With IPW, the contribution of the complete cases to the regression estimation was weighted by the inverse of their probability of being a complete case. We modeled the probability of being followed in the second wave as a logistic function of the following predictors: age, sex, individual education, residential neighborhood education, baseline BMI, perceived stress, and depressive symptoms (selected on the basis of hypotheses of causal effects on this outcome). The weight for each subject was calculated as the inverse of the predicted probability of participation in the second wave. While 7,172 participants (with BMI at baseline) were considered to produce the weights, 3,693 participants (with BMI also measured in wave 2) were included in the IPW analysis of the relationship of interest.

We could not exactly assess the performance of CCA, MI, and IPW in this first stage, because the true association between individual education and change in BMI was unknown.

Stage 2: simulation of additional attrition

Definition of simulated datasets by manipulation of the empirical dataset. The complete dataset (3,693 participants) was used as a basis to simulate additional attrition. These simulated

incomplete datasets were generated from the complete dataset by randomly simulating the outcome value (BMI in the second wave, thus change in BMI) to be missing. Missingness was simulated according to three MAR mechanisms (see the signed[29] Directed Acyclic Graphs in Figure 1). The attrition rate was defined on the basis of the risk of having missing data at follow up.

Different mechanisms to simulate attrition. We considered two mechanisms of attrition based on observed covariates (individual education and either neighborhood education or baseline BMI) and one mechanism using individual education and a simulated hypothetical covariate. Regarding the first two mechanisms, previous studies have shown that low individual education and low neighborhood education were associated with a higher BMI or a higher change in BMI and that a higher BMI at baseline was associated with a higher change in BMI.[24,30-33] Moreover, as shown with a regression model reported in eAppendix 1, these three baseline characteristics were independently (but weakly) associated with the fact of having a follow-up BMI measurement in the RECORD Study: the odds of participating in wave 2 were higher at high individual education and high neighborhood education levels and among participants who were not obese at baseline. Thus, two types of MAR mechanisms were first generated: missingness in the outcome (1) was set to be conditional on both individual education and neighborhood education and (2) was set to be conditional on both individual education and baseline BMI.

In these two scenarios, the other covariate influencing attrition was weakly associated with the outcome (the correlation between neighborhood education and change in BMI was 0.06 while the correlation between baseline BMI and change in BMI was 0.09). Therefore, in the third mechanism of attrition, we used a simulated hypothetical covariate for which we defined five levels of correlation with BMI change ($r = 0.3; 0.4; 0.5; 0.6; 0.7$).

For each mechanism of attrition, eleven scenarios of attrition level were examined, with an increasing influence of the two selected covariates (individual education and another variable) on the risk of having missing data in wave 2. The eleven scenarios were defined on the basis of eleven odds ratios [OR] ranging from 1.0 to 6.0 (increasing by 0.5) for these associations with missingness (Table 1, see the footnote for details, including a description of the units of variables). In each scenario, the same OR was applied to the relationship of individual education with attrition and to the relationship of the other variable (neighborhood education, baseline BMI, simulated covariate) with attrition. Thus, all the attrition simulated in the data is attributable to increasing selective attrition, while in the initial empirical analysis the overall attrition rate is higher but is likely less attributable to selective attrition. To circumvent random sampling variability, 500 such simulated incomplete datasets were generated for each mechanism and for each scenario of attrition level. Therefore, we obtained 11 scenarios for each of the first two mechanisms of attrition and 11 × 5 correlation levels = 55 scenarios for the third mechanism of attrition (see Table 1).

Quantification of bias and assessment of MI. The CCA analysis was applied to the simulated incomplete datasets with the same adjustment factors than in the first stage. MI was also applied with the following list of predictors in the imputation model: age, sex, individual education, perceived stress, depressive symptoms, and one additional mechanism-dependent variable (neighborhood education with the first, baseline BMI with the second, and the simulated covariate with the third mechanism). It should be noted that IPW relies on a comparable model than the one that we used to simulate additional attrition. The predictors included in the weight model of IPW were those that were used as determinants of missingness with each of the three

distinct MAR mechanisms (e.g., individual education and either neighborhood education, baseline BMI, or the simulated covariate, but without age, sex, stress, and depressive symptoms). Thus in this second stage, we could not evaluate the performance of IPW but used it as a gold standard (as it was perfectly specified) against which to assess the magnitude of bias with CCA and the performance of MI. In this second stage, the CCA estimate in the sample of 3,693 participants (from stage 1) can also be considered as the true estimate (compared to the bias added through the simulations).

Analysis of the relationship between individual education and change in BMI

For all the approaches enumerated above, a linear model was used to estimate the association between individual education and the change in BMI. Based on the available variables, we could not identify any causal antecedents of individual education (temporally antecedent to individual education) that was also associated with change in BMI (confounder). Parental education and the human development index of the country of birth were tested but did not fulfill the last condition. Thus the association between individual education and BMI change was adjusted only for age and sex, both in the initial empirical and simulation analyses. In the first stage, this model was applied to the complete case database of 3,693 participants for CCA and IPW, while MI was applied to the complemented dataset of 7,172 participants. In the second stage, this model for the individual education – BMI change relationship was applied to simulated complete case databases ($N < 3,693$) for CCA and IPW (even if the weight model of IPW was run among 3,693 participants), while MI was applied to complemented datasets of 3,693 participants. The median of the coefficient over the 500 simulated datasets was used as the final estimate. The uncertainty in this estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

The simulation code is provided in eAppendix 2. All analyses were conducted using R version 3.1.1 (http://www.R-project.org).

RESULTS

Descriptive characteristics for participants with BMI measured at baseline (N = 7,172); participants with BMI also assessed at the follow-up (N = 3,693); and participants who dropped out (N = 3,479) are provided in eAppendix 3. In our sample of 7,172 participants in the first wave, median BMI at baseline was 25.0 (interdecile range: 20.7, 30.7). In the subsample of 3,693 participants followed up in the second wave, 68.6% were men, 70.8% had a high education, the mean age was 51.5 years (SD = 11.3), the median baseline BMI was 25.0 (interdecile range: 20.9, 30.1), and the median change in BMI was 0.1 (interdecile range: -1.4, 1.9).

Stage 1: empirical analysis without added simulated attrition

In the CCA (N = 3,693), after adjustment for age and sex, the change in BMI was lower for participants with a high compared to a low educational level [-0.26, 95% confidence interval (CI): -0.36, -0.15]. As shown in Table 2, IPW and MI also yielded a comparable inverse association between individual education and change in BMI. The final estimate was relatively similar whatever the correction method (IPW or MI).

Stage 2: simulation of additional attrition

The results for the simulated datasets based on three mechanisms of attrition under eleven different scenarios of attrition level (OR = 1.0 to 6 by 0.5) are shown in Figures 2 to 4. Numerical estimates are reported in eAppendix 4. Due to the perfect knowledge of the simulated

missingness mechanisms, the IPW estimates were close from the horizontal reference line, as expected (gold standard).

The first attrition mechanism was grounded on the effect of individual education and neighborhood education on participation in the second wave. Figure 2 shows an inverse association between individual education and change in BMI after adjustment for age and sex, with all methods and under all scenarios of attrition level. The CCA estimate was very close from the thick horizontal line (true estimate), and very close from the IPW estimate (which was still closer from the horizontal line). There was an indication that MI did slightly worst than CCA, with a slight increase of bias with increasing selective attrition. The uncertainty in the estimates increased for CCA, and still more so for MI, with increasing attrition bias.

The second simulated mechanism of attrition resulted from the effects of individual education and baseline BMI on attrition. As shown in Figure 3, the gap between the true coefficient and the CCA coefficient increased weakly but regularly (from -0.26 to -0.22) with the strength of the simulated selective attrition bias. MI could not correct the bias, and on the opposite implied a stronger bias than the CCA itself.

The findings for the third MAR mechanism of attrition using a simulated covariate correlated with the outcome are shown in Figure 4. For CCA as well as for MI, the magnitude of the attrition bias increased both with the strength of the association between the covariates and attrition and with the strength of the correlation between the simulated covariate and BMI change. For example, the coefficient for the association between individual education and BMI change analyzed with CCA varied from -0.26 to -0.06 under the eleven scenarios of attrition level when the correlation between the simulated covariate and BMI change was of 0.3 and from -0.26 to 0.14 when the correlation was of 0.7. MI could not correct this attrition bias.

DISCUSSION

To our knowledge, the present study is the first to rely on a simulation approach applied to longitudinal observed data to quantify the magnitude of bias due to missing information in the outcome at the follow-up and to test the ability of statistical methods to correct this attrition bias.

Stage 1: empirical analysis without added simulated attrition

Previous studies, mostly based on cross-sectional designs,[32] have reported that BMI or body fat increased with decreasing individual/neighborhood socioeconomic levels.[30,31,33,34] A review concluded from studies with measured adiposity and a follow-up of at least 4 years to an inverse association between education and weight gain.[31] Our findings indicate that participants with a high education had a weaker increase in BMI over the follow-up. Our analyses of the empirical sample based on IPW and MI indicate that MAR mechanisms of attrition based on the effects of individual education, neighborhood education, and baseline BMI on participation in the study are unlikely to affect the direction and even the strength of the association of interest. This is likely because there were only weak effects of individual education, neighborhood education, and baseline BMI on participation, and weak effects of neighborhood education and baseline BMI on the change in BMI.

Stage 2: simulation of additional attrition

First and second attrition mechanisms (observed covariates)

There was evidence of a modest attrition bias for the coefficient of interest especially for the second mechanism of attrition (specifying an effect baseline BMI on participation). A likely explanation is that the association of baseline BMI with change in BMI (correlation = 0.09) was stronger than the association of neighborhood education with change in BMI involved in the first

attrition mechanism (correlation = 0.06). The comparison of Figure 2 with Figure 3 shows that the attrition bias was in the opposite direction in these two cases: the covariate that was used as a determinant of attrition was associated in the opposite direction with BMI change (see the signed[29] Directed Acyclic Graphs in Figure 1).

Third attrition mechanism (simulated covariate)

Simulating a covariate allowed us to modulate its correlation with the change in BMI between 0.3 and 0.7. These results emphasize the following important aspects: 1) even in presence of a very strong selective attrition (influence of covariates on dropout), the bias for the association of interest remains of small magnitude if the correlation between the covariates influencing study dropout and the change in BMI is weak; 2) the magnitude of the attrition bias, as expressed by the difference between CCA and the thick horizontal line or the gold standard IPW estimate, increased with the correlation between the simulated covariate and change in BMI; 3) when data are missing in the outcome under a MAR pattern, MI does not do better than CCA.

When data are missing in the covariates, MI is the most commonly advocated method to handle missing data.[2,8,11] However, when data are missing only in the outcome, MI cannot correct the bias even if the imputation model is correct. Compared to CCA, MI can even amplify the observed bias (see the point estimates on Figure 3), which is consistent with previous studies.[35,36] Allison[14] and Hippel[15] showed that if there is no missing data in the independent variables and if there are no strongly correlated auxiliary predictors, then there is no additional information in the imputed values. Hippel[15] and Kullback[37] showed that observations with an imputed outcome contain no information about the regression of Y on X. The approach relies on the available knowledge of how X predicts Y to impute Y, so it does not bring additional information on the

potential influence of X on Y. In this case, using MI is equivalent to performing a CCA, while the imputation of missing outcome values adds noise to these estimates (which is clear from the 95% credible intervals in Figure 4).

The attrition mechanisms examined were based on MAR patterns. eAppendix 5 reports a complementary analysis where a MNAR pattern of missingness was implemented (individual education and BMI change itself influencing study dropout). In this case, the weight model of IPW could not be perfectly specified (i.e., it included individual education, but not the change in BMI, considering that it was inaccessible to the researchers when missing). The findings show that the bias was substantial and that neither MI nor IPW could correct the attrition bias introduced by MNAR outcome data (eAppendix 5).

Strengths and limitations

Regarding study strengths, first, our study combined empirical longitudinal data with simulation approaches: we used the available knowledge on the determinants of attrition in the RECORD Cohort Study to create scenarios of attrition corresponding to realistic situations occurring in practice, allowing us to assess how statistical methods handled missing outcome data in "real life scenarios". Second, in this simulation study, different mechanisms of attrition (including two MAR with observed variables and one MAR with a simulated covariate strongly correlated with the outcome) and different attrition levels were explored.

Regarding limitations, a first shortcoming of the work is related to the fact that in the initial empirical analyses, the true estimate against which to compare the corrected estimates was not known. Second, our simulation study was designed to quantify the magnitude of bias and to illustrate the non-performance of MI to handle missing outcome data; however, this simulation study was not meant to assess the performance of IPW, which by design was perfectly specified.

Third, the present study analyzed only two repeated measures, thus our findings are difficult to generalize to other longitudinal designs with more than two waves (which would likely have to be analyzed with hierarchical linear models or repeated measure models rather than linear models as in the present case). Fourth, explore the few potential confounders that we had for the effect of individual education on change in BMI, but we lacked variables that are difficult to collect (e.g., norms and attitudes prior to the completion of education). Fifth, although we made efforts to correctly specify these models, we acknowledge that the weight and imputation models could be further complemented in the initial empirical analyses. Sixth and finally, our convenience sample, recruited in preventive healthcare centers, was not representative of the Paris Ile-de-France region.[19] However, a large panel of municipalities from the region was *a priori* selected to ensure the presence in the sample of people from all socioeconomic backgrounds.

Conclusion

In conclusion, the results of this study suggest that biases related to selective attrition in longitudinal studies implying missing outcome data can be substantial. Moreover, it was found that when data are missing only in the outcome, MI is not able to correct the bias introduced by MAR patterns of attrition. Thus, although MI is the most commonly advocated method to handle missing data in the covariates in cohort studies, it should not be used in these particular circumstances (outcome missingness under a MAR pattern) where it does not do better and can even do worse than CCA. Overall, our findings therefore emphasize the need to devote more effort during the design phase to plan the collection of relevant information on the cause of study dropout,[38] in order to evaluate the attrition mechanism involved, and to improve the specification of IPW models.

REFERENCES

1. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology*. 2012;23:729-32.
2. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
3. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1:368-76.
4. Little RJA, DB R. *Statistical Analysis with Missing Data*. New Jersey: John Wiley & Sons; 1987.
5. Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63:581-592.
6. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7:147-77.
7. 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.
8. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22:278-95.
9. Little RJA, Rubin DB. *Statistical analysis with missing data*. Vol. 2nd edition. Chichester: Wiley; 2002.
10. Rubin D.B. *Multiple imputation for non response in surveys*. New York: Wiley; 1987.
11. Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics*. 2014;15:719-30.

12. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. 2012;68:129-37.
13. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*. 2010;29:2920-31.
14. Allison PD. *Missing data*. Thousand Oaks, CA: Sage University Papers on Quantitative Application in the Social Sciences; 2001.
15. Hippel PT. Regression with Missing Y's: An Improved Strategy for Analysing Multiply Imputed Data. *Sociological Methodology*. 2007;37:83-117.
16. Hofler M, Pfister H, Lieb R, Wittchen HU. The use of weights to account for non-response and drop-out. *Soc Psychiatry Psychiatr Epidemiol*. 2005;40:291-9.
17. Chaix B, Kestens Y, Bean K, et al. Cohort Profile: Residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases--The RECORD Cohort Study. *Int J Epidemiol*. 2012;41:1283-1292.
18. Chaix B, Bean K, Leal C, et al. Individual/neighborhood social factors and blood pressure in the RECORD Cohort Study: which risk factors explain the associations? *Hypertension*. 2010;55:769-775.
19. Chaix B, Billaudeau N, Thomas F, et al. Neighborhood effects on health: correcting bias from neighborhood effects on participation. *Epidemiology*. 2011;22:18-26.
20. Havard S, Reich BJ, Bean K, Chaix B. Social inequalities in residential exposure to road traffic noise: An environmental justice analysis based on the RECORD Cohort Study. *Occup Environ Med*. 2011;68:366-74.
21. WHO. Obesity: preventing and managing the global epidemic. In: consultation W, ed. Geneva: World Health Organisation, 2000.

22. Thomas F, Bean K, Pannier B, Oppert JM, Guize L, Benetos A. Cardiovascular mortality in overweight subjects: the key role of associated risk factors. *Hypertension*. 2005;46:654-659.

23. Chaix B, Bean K, Daniel M, et al. Associations of supermarket characteristics with weight status and body fat: a multilevel analysis of individuals within supermarkets (RECORD Study). *PLoS One*. 2012;7:e32908.

24. Leal C, Bean K, Thomas F, Chaix B. Are associations between neighborhood socioeconomic characteristics and body mass index or waist circumference based on model extrapolations? *Epidemiology*. 2011;22:694-703.

25. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav*. 1983;24:385-396.

26. Pichot P, Boyer P, Pull CB, Rein W, Simon M, Thibault A. Le questionnaire QD 2. La forme abrégée QD 2A. *Rev Psychol Appl*. 1984;4:323-340.

27. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8:3-15.

28. Su Y.S., Gelman A., Hill J., Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*. 2011;45:6-15.

29. VanderWeele TJ, Robins JM. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society. Series B, Statistical methodology*. 2010;72:111-127.

30. Lewin A, Pannier B, Meline J, Karusisi N, Thomas F, Chaix B. Residential neighborhood, geographic work environment, and work economic sector: associations with body fat measured by bioelectrical impedance in the RECORD Study. *Ann Epidemiol*. 2014;24:180-6.

31. Ball K, Crawford D. Socioeconomic status and weight change in adults: a review. *Soc Sci Med*. 2005;60:1987-2010.

32. Mujahid MS, Diez Roux AV, Borrell LN, Nieto FJ. Cross-sectional and longitudinal associations of BMI with socioeconomic characteristics. *Obes Res*. 2005;13:1412-1421.

33. McLaren L. Socioeconomic status and obesity. *Epidemiol Rev*. 2007;29:29-48.

34. Kershaw KN, Albrecht SS, Carnethon MR. Racial and ethnic residential segregation, the neighborhood socioeconomic environment, and obesity among Blacks and Mexican Americans. *Am J Epidemiol*. 2013;177:299-309.

35. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol*. 2002;55:329-37.

36. Kristman VL, Manno M, Cote P. Methods to account for attrition in longitudinal data: do they work? A simulation study. *Eur J Epidemiol*. 2005;20:657-62.

37. Kullback S. *Information Theory and Statistics.* New York: Wiley; 1959.

38. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc*. 1995;90:1112-1121.

FIGURE LEGENDS

FIGURE 1. Signed directed acyclic graphs depicting the different selection biases that were introduced with the three mechanisms of attrition examined [following Missing At Random (MAR) patterns]. These mechanisms were based on the effects of the following variables on participation: individual education and neighborhood education in mechanism 1 (1A); individual education and baseline BMI in mechanism 2 (1B); individual education and a simulated covariate in mechanism 3 (1C). Signed graphs were established based on the assumption of so-called weak monotonic effects.[29] Participants with a high level of neighborhood education or a low baseline BMI (respectively in the first and second MAR patterns) and participants with high individual level of education are more likely to participate in the second wave. The attrition bias in the estimate of interest is introduced by conditioning on participation (P = 1), which is a common effect of the variable of interest (individual education) and of another covariate that also influences the change in BMI

FIGURE 2. Associations between individual education and change in BMI in simulated datasets based on the first attrition mechanism (MAR, individual education, neighborhood education) under eleven scenarios of attrition level. Point estimates and 95% credible intervals (median and 2.5^{th} and 97.5^{th} percentiles over 500 realizations of each scenario) for each method are represented as continuous lines to ease the reading but refer to discrete estimates for each attrition level. The horizontal thick line represents the true coefficient of reference for the simulation work, from the analysis without missing data (N = 3,693). As IPW was considered as a gold standard for the point estimate (due to the perfectly specified weight model), only the point estimate but not its 95% credible interval is reported.

FIGURE 3. Associations between individual education and change in BMI in simulated datasets based on the second attrition mechanism (MAR, individual education, baseline BMI) under eleven scenarios of attrition level. Point estimates and 95% credible intervals (median and 2.5th and 97.5th percentiles over 500 realizations of each scenario) for each method are represented as continuous lines to ease the reading but refer to discrete estimates for each attrition level. The horizontal thick line represents the true coefficient of reference for the simulation work, from the analysis without missing data (N = 3,693). As IPW was considered as a gold standard for the point estimate (due to the perfectly specified weight model), only the point estimate but not its 95% credible interval is reported.

FIGURE 4. Associations between individual education and change in BMI in simulated datasets based on the third attrition mechanism (MAR, individual education, simulated covariate with five levels of correlation with the outcome: r = 0.3 to 0.7 by 0.1) under eleven scenarios of attrition level. Point estimates and 95% credible intervals (median and 2.5th and 97.5th percentiles over 500 realizations of each scenario) for each method are represented as continuous lines to ease the reading but refer to discrete estimates for each attrition level. The horizontal thick line represents the true coefficient of reference for the simulation work, from the analysis without missing data (N = 3,693). As IPW was considered as a gold standard for the point estimate (due to the perfectly specified weight model), only the point estimate but not its 95% credible interval is reported.

TABLE 1. Number of observations with a missing outcome at the follow-up (and corresponding percentage) for each mechanism of attrition under each attrition level (the initial database in which attrition is simulated comprises 3,693 participants)

Attrition level	First attrition mechanism[a]	Second attrition mechanism[b]	Third attrition mechanism with a simulated covariate[c]				
			r = 0.3	r = 0.4	r = 0.5	r = 0.6	r = 0.7
OR = 1.0[d]	67 (1.8)	66 (1.8)	66 (1.8)	67 (1.8)	66 (1.8)	66 (1.8)	66 (1.8)
OR = 1.5	141 (3.8)	140 (3.8)	140 (3.8)	141 (3.8)	139 (3.8)	140 (3.8)	140 (3.8)
OR = 2.0	248 (6.7)	241 (6.5)	241 (6.5)	248 (6.7)	242 (6.5)	242 (6.5)	242 (6.5)
OR = 2.5	385 (10.4)	376 (10.2)	368 (10.0)	385 (10.4)	368 (10.0)	369 (10.0)	368 (10.0)
OR = 3.0	537 (14.5)	509 (13.8)	506 (13.7)	537 (14.5)	506 (13.7)	510 (13.8)	509 (13.8)
OR = 3.5	687 (18.6)	654 (17.7)	650 (17.6)	687 (18.6)	651 (17.6)	655 (17.7)	653 (17.7)
OR = 4.0	832 (22.5)	795 (21.5)	794 (21.5)	832 (22.5)	797 (21.6)	797 (21.6)	798 (21.6)
OR = 4.5	962 (26.0)	930 (25.2)	932 (25.2)	962 (26.0)	933 (25.3)	934 (25.3)	934 (25.3)
OR = 5.0	1083 (29.3)	1060 (28.7)	1060 (28.7)	1083 (29.3)	1060 (28.7)	1063 (28.8)	1060 (28.7)
OR = 5.5	1192 (32.3)	1177 (31.9)	1176 (31.8)	1192 (32.3)	1179 (31.9)	1179 (31.9)	1178 (31.9)
OR = 6.0	1293 (35.0)	1285 (34.8)	1285 (34.8)	1293 (35.0)	1289 (34.9)	1286 (34.8)	1287 (34.8)

[a] MAR mechanism: individual education and neighborhood education influence dropout.

[b] MAR mechanism: individual education and baseline BMI influence dropout.

[c] MAR mechanism: individual education and a simulated variable strongly correlated with the change in BMI influence dropout. Several degrees of correlation of this variable with dropout are considered.

[d] The OR refers to a multiplicative coefficient for the odds of having a missing outcome in wave 2 for having a low rather than a high individual education level (binary variable), for a 2-standard deviation decrease in the proportion of high educated residents in the neighborhood, for a 2-standard deviation increase in baseline BMI (a 2-standard deviation was chosen to mimic the effect of a binary variable with approximately half of the population in each group), and for a 2-standard deviation increase in the simulated covariate strongly correlated with the change in BMI (r = 0.3 to 0.7 by 0.1). The distribution of the standardized baseline BMI and standardized change in BMI were shifted towards the positive, so as to have a minimum value equal to 0, yielding the minimal (and non-negative) probability of attrition when these recoded variables were equal to 0. The distribution of standardized neighborhood education was shifted towards the negative, so as to have a maximum value equal to 0, yielding (due to a negative coefficient in our model) the minimal probability of attrition when this recoded variable was equal to 0 (i.e., when neighborhood education was the highest) and an increase in the odds of attrition with more negative values of this variable. The model to define the probability of attrition for each participant was a logistic model with an intercept equal to –4.

TABLE 2. Empirical associations[a] between individual education (high vs. low) and change in BMI (adjusted for age and sex), either from the complete case analysis or corrected from the observed attrition bias through IPW or MI[b]

	β (95% CI)[c]
CCA (N = 3,693)	-0.26 (-0.37 ; -0.15)
IPW[d] (N = 3,693)	-0.25 (-0.36 ; -0.14)
MI[d] (N = 7,172)	-0.26 (-0.37 ; -0.14)

[a] The associations reported here are based on the original RECORD sample, without any modification through simulations.

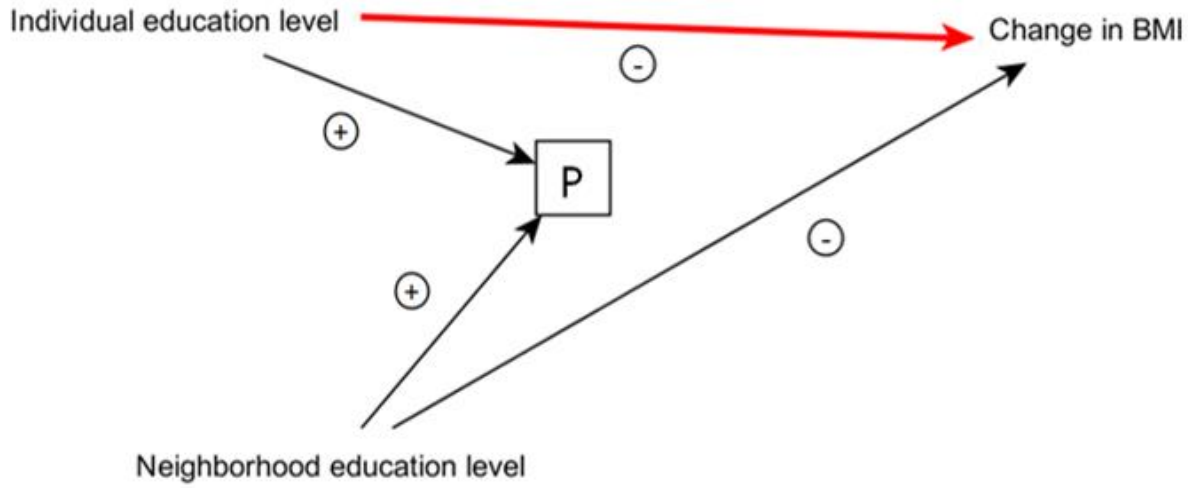
[b] The average of the predicted change in BMI in the low education group was 0.34 in the CCA model, 0.34 in the IPW model, and 0.25 in the model based on MI.

[c] β, beta coefficient; CI, confidence interval.

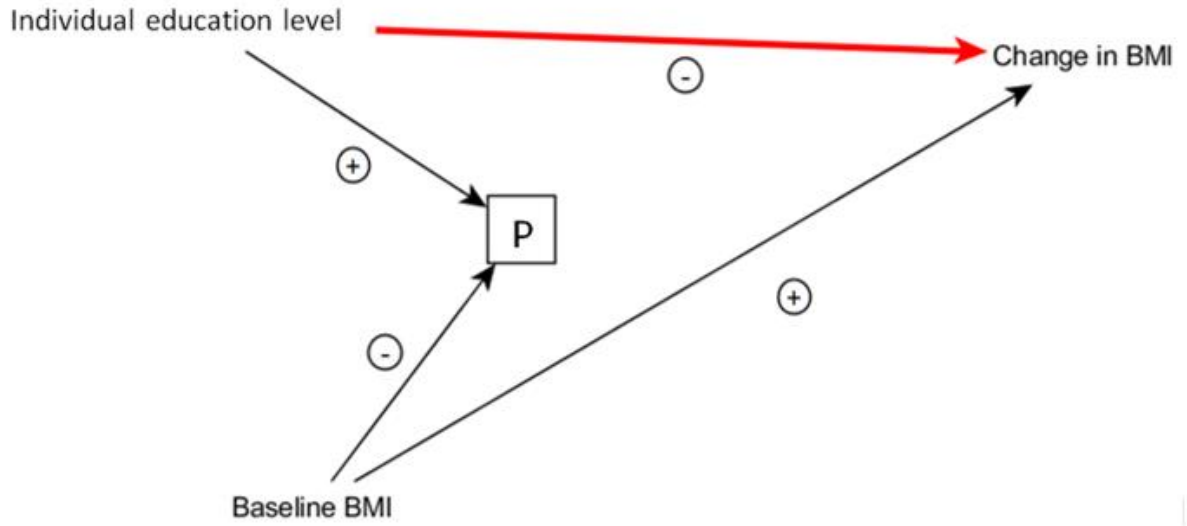
[d] The covariates used in the weight model of IPW (N = 7,172) and in the imputation model of MI were age, sex, individual education, residential neighborhood education, baseline BMI, perceived stress, and depressive symptoms.

Figure 1.

1A.



1B.



1C.

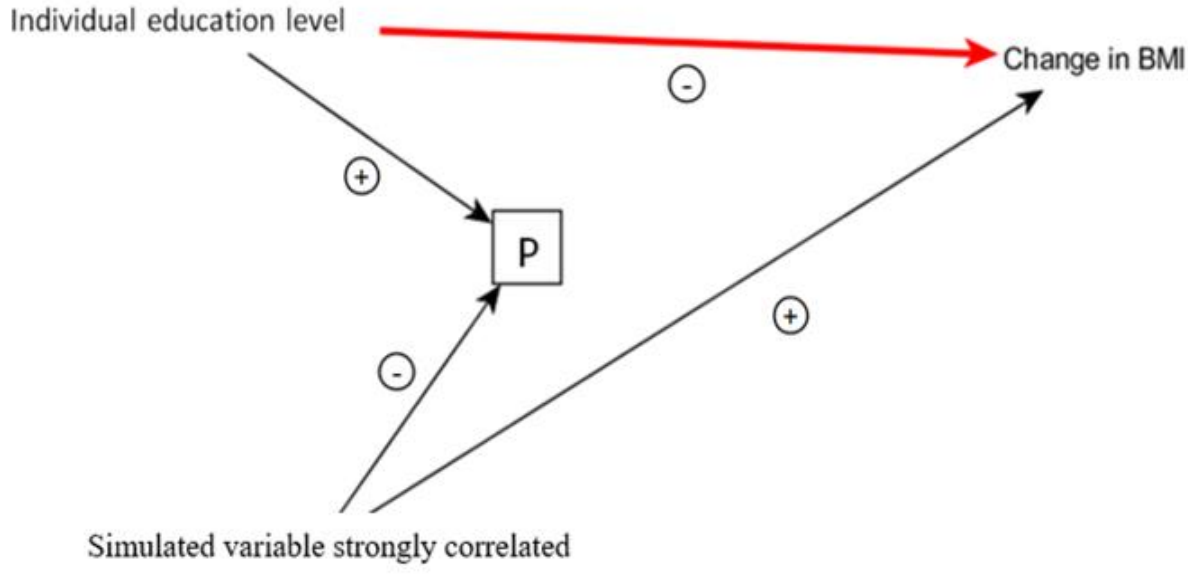


Figure 2

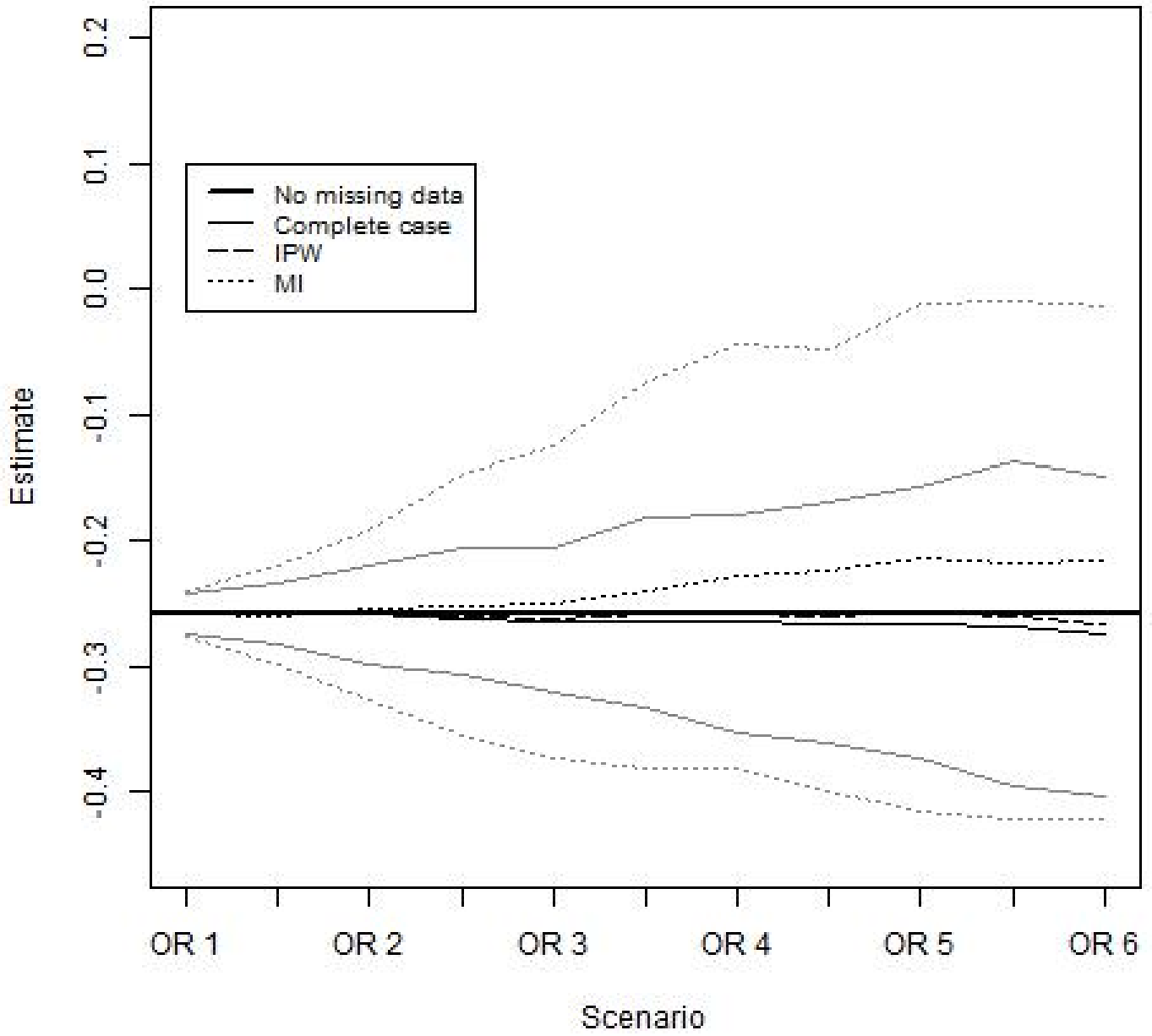
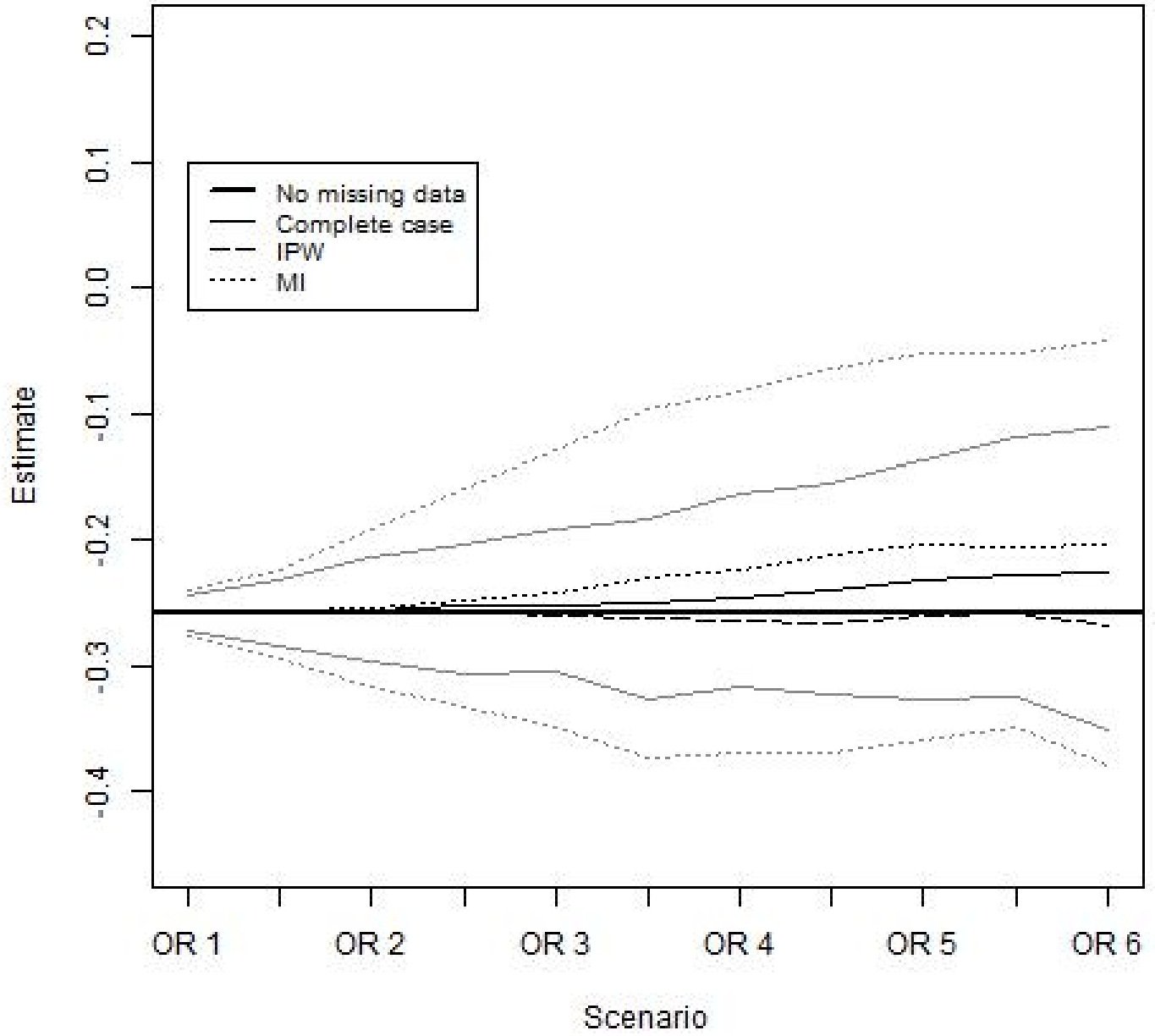

No missing data
Complete case
IPW
MI
Estimate
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
OR 1
OR 2
OR 3
OR 4
OR 5
OR 6
Scenario

Figure 3


No missing data
Complete case
IPW
MI
Estimate
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
OR 1
OR 2
OR 3
OR 4
OR 5
OR 6
Scenario

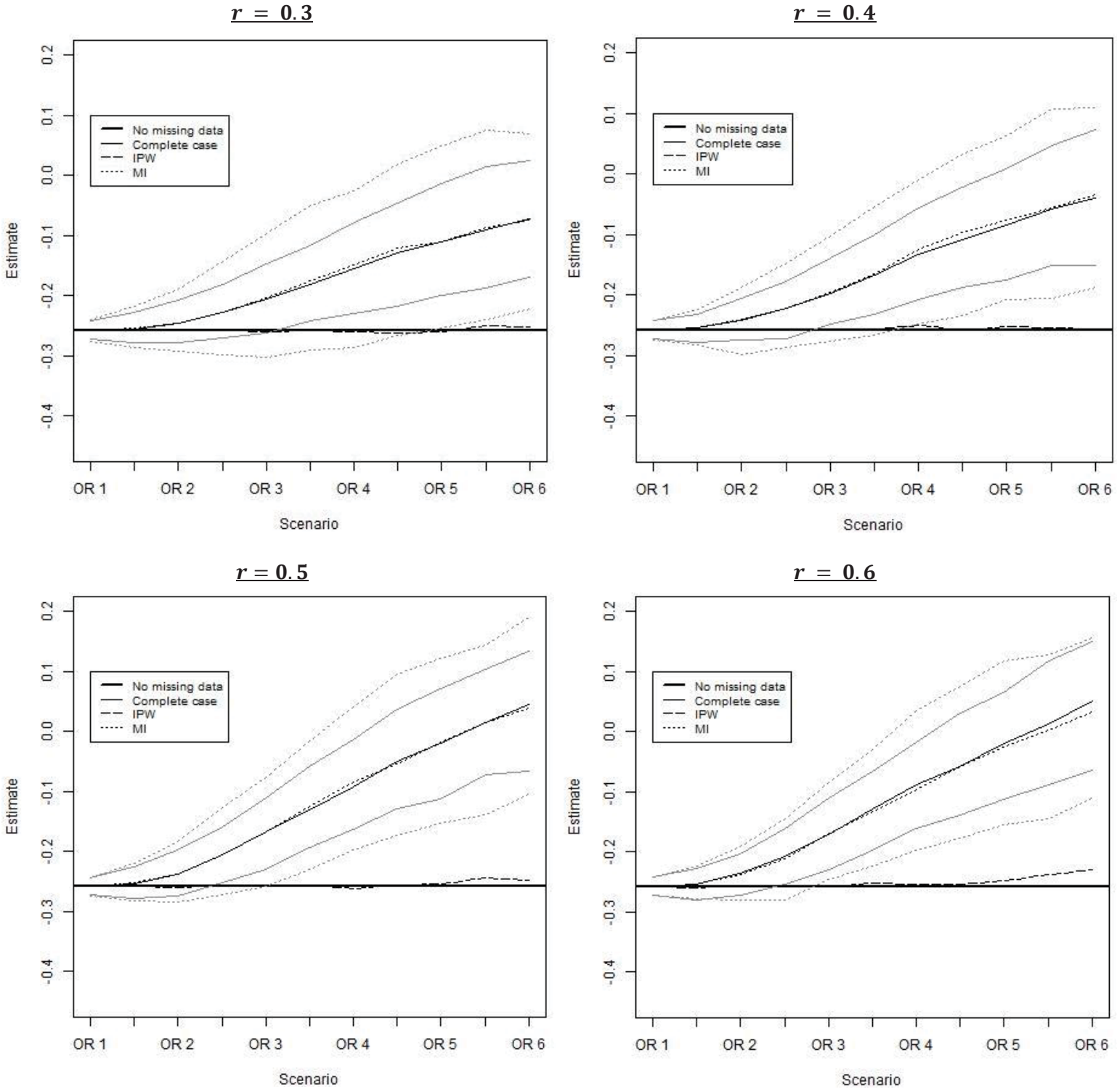

$r = 0.3$
$r = 0.4$
$r = 0.5$
$r = 0.6$
$r = 0.7$
No missing data
Complete case
IPW
MI
Estimate
Scenario
OR 1
OR 2
OR 3
OR 4
OR 5
OR 6

Figure 4

Appendix 1. The determinants of participation in wave 2

Appendix Table 1. Logistic model for the determinants of participation in wave 2 (N = 7,172)

	Participation in wave 2 β (95% CI)
Low individual education (vs. high)	-0.069 (-0.09 ; -0.04)
Obese in wave 1 (vs. normal weight)	-0.067 (-0.10 ; -0.03)
Neighborhood proportion of high educated[a]	0.098 (0.01 ; 0.18)

[a] While the two other explanatory variables were coded as categorical variables, neighborhood education was expressed as a continuous variable comprised between 0 and 1.

Appendix 2. Simulation code

```
#########################################################
# A script to accompagny
#
# Lewin, Brondeel, Benmarhnia, Thomas, Chaix. Attrition bias related to missing
# outcome data: a longitudinal simulation study.
# Epidemiology.
#
# This script describes the code used in this simulation study.
# At the end of this document, there is some extra script to make up
# your own dataset, enabling interested readers to test the code.
# For questions, feel free to contact Ruben Brondeel
# e-mail: Ruben.Brondeel@gmail.com
#########################################################

#########################################################
# Script created under
# R version 3.3.0
# mi-package version 1.0
#########################################################

#########################################################
# Step 1: Construction of simulation datasets.
# 500 datasets are constructed for 11 scenarios.
# Each dataset has the analyses variables in common,
# they differ in the missingness indicators (0/1) (named 'filters' in script).
#########################################################

# 1.0 Creation of a file directory
scenario <- c("scen10", "scen15", "scen20", "scen25", "scen30", "scen35", "scen40",
 "scen45", "scen50", "scen55", "scen60")
for(i in scenario){
 path <- paste('my data', i,sep="")
 dir.create(path, showWarnings = TRUE, recursive = TRUE)
}

# 1.1 Construction of a 'simulated' variables to test variables with a
# stronger link to the outcome than the original independent variables.
# The variable newVar30 is a continuous variable with N(0,1).
# The Pearson correlation with the outcome (bmi_change) is set to 0.30.
# We look for a value x for which the squared difference
# of 0.30 and the correlation bmi_change and newvar is minimal (i.e. ~= 0).



fr <- function(x, y) {
  set.seed(04072015)
  newvar <-  x*set$bmi_change + rnorm(nrow(set), 0, 1)
  (y - cor(set$bmi_change, newvar))^2
}

a <- optimize(f = fr, interval=c(0,1), y=0.30)$minimum

set.seed(04072015)
set$newVar30 <-  a*set$bmi_change + rnorm(nrow(set), 0, 1)
```

```
# 1.2 Creation of the missingness indicators.
# We create a function 'miss' that will create 4 missingness indicators,
# one for each model (Model MAR 1, ...) given a certain level of association
# between the respective variables and the probability of having a missing value.
# This process is repeated 500 times. The results is 500 datasets with the
# original data and 4 indicators (0/1) of missingness.


nsim <- 500   # number of simulations. Lower this number for a test run.

miss <- function(scenario, OR, data){
setwd(paste(path.main,'1. data/miss/',scenario, sep=""))
 var <- c("ID", "bmi_change", "educ_indiv", "educ_neigh", "bmi_t1",
        "age", "male", "stress", "depression", "newVar30")

 # 1.2.1: we standardize the continuous variables.
 # So the associations with the missingness are comparable.
 # This is for the creation of missingness indicators only, not for later analyses.
 x  <- data$educ_indiv
 y1 <- as.numeric(scale(data$educ_neigh))
 y2 <- as.numeric(scale(data$bmi_t1))
 y3 <- as.numeric(scale(data$newVar30))
 y4 <- as.numeric(scale(data$bmi_change))


 # 1.2.2: A shift of the variables with a negative association to the missingness
 # towards negative values; and the variables with a positive association towards
 # positive values.
 # This allows for strictly positive additions in the formula of the logodds.
 # Therefore, higher associations with the missingness,
 # will always lead to a higher dropout.
 # This is not necessary from a mathematical point of view,
 # but it seems reasonable from an epidemiological point of view.
 # In real cases of dropout, we can assume that higher levels of dropout
 # are more probable to lead to higher associations and therefore biases, even
 # though this is not a hard rule.
 x  <- x  - 1
 y1 <- y1 - 3 # the three in these formulas are used not to depend
 y2 <- y2 + 3 #  on the randomness of minimum and maximum values.
 y3 <- y3 + 3
 y4 <- y4 + 3

 # 1.2.3 We divide the continuous variables by two: for more interpretable OR's,
 # and to compare more easily with the dichotomous individual education variable
 # (=x).
 y1 <- y1/2
 y2 <- y2/2
 y3 <- y3/2
 y4 <- y4/2

 # 1.2.4 We calculate the probabilities of the dropout for a given odds ratio for #
 the 4 models.
 # pr1  = Model MAR 1 [individual level education (=x) and neighborhood education
 (=y1)]
 # pr2  = Model MAR 2 [individual level education (=x) and BMI at time 1 (=y2)]
```

```
 # pr3  = Model MAR 3 simulated variable [individual level education (=x) and new
 variable (=y3)]
 # pr4  = Model MNAR  [individual level education (=x) and change of BMI between
 time 1 and time 2(=y4)]

 z <-  - 4 - log(OR)*x - log(OR)*y1
 pr1 <- exp(z)/(1+exp(z))

 z <-  - 4 - log(OR)*x + log(OR)*y2
 pr2 <- exp(z)/(1+exp(z))

  z <-  - 4 - log(OR)*x + log(OR)*y3
 pr3 <- exp(z)/(1+exp(z))

 z <-  - 4 - log(OR)*x + log(OR)*y4
 pr4 <- exp(z)/(1+exp(z))

  # 1.2.5. Probability of participation.
 # Probability participation = 1 - probability of dropout.
 # This is an arbitrary choice, but has no influence on the results.

 pr1 <- 1 - pr1
 pr2 <- 1 - pr2
 pr3 <- 1 - pr3
 pr4 <- 1 - pr4

 # 1.2.6 Creation of simulation datasets.
 # each time a different random sample of 0's and 1's,
 # given the binomial distribution and the above calculated probability.
 for(i in 1:500){
    data1 <- data

    data1$filterMAR1 <- rbinom(nrow(data1), 1, pr1)
    data1$filterMAR2 <- rbinom(nrow(data1), 1, pr2)
    data1$filterMAR3 <- rbinom(nrow(data1), 1, pr3)
    data1$filterMNAR <- rbinom(nrow(data1), 1, pr4)


    set1 <- set1[,var]
    name.set <- paste("set", i, sep=".")
    write.csv(set1, name.set, row.names=FALSE)
 }
}

var00 <- c("ID", "bmi_change", "educ_indiv", "educ_neigh", "bmi_t1", "age", "male",
         "stress", "depression", "newVar30")


 setwd(paste(path.main,'1. data/miss/',scenario, sep=""))
 write.csv(set[,var00], 'set00', row.names=FALSE)

 # we apply this function to the 11 selected odd's ratios.

 miss("scen10", OR=1.0, set)
 miss("scen15", OR=1.5, set)
 miss("scen20", OR=2.0, set)
```

```
miss("scen25", OR=2.5, set)
miss("scen30", OR=3.0, set)
miss("scen35", OR=3.5, set)
miss("scen40", OR=4.0, set)
miss("scen45", OR=4.5, set)
miss("scen50", OR=5.0, set)
miss("scen55", OR=5.5, set)
miss("scen60", OR=6.0, set)



##########################################################
# Step 2: multiple imputation.
# !!!! Note: this step might be very long in computation time.
# When testing, you might want to choose a low number of simulations
# Only the code for Model MAR 1 is presented below.
# However, the code for the other models is basically a copy of this code.
# The package mi Version 1.0.
##########################################################
library(mi)


# 2.0 Creation of a file directory
scenario <- c("scen10", "scen15", "scen20", "scen25", "scen30", "scen35", "scen40",
 "scen45", "scen50", "scen55", "scen60")
for(i in scenario){
 path <- paste('my data', i,sep="")
 dir.create(path, showWarnings = TRUE, recursive = TRUE)
}


# 2.1 A function is created to perform multiple imputation over all simulation sets
# per scenario.
final.mi1<- function(scenario){

var <- c("ID", "bmi_change", "educ_indiv", "age", "male", "educ_neigh", "stress",
 "depression")

for(i in 1:500){
    flush.console()

    # 2.1.1 A simulation dataset is read.

    set1 <- read.csv(paste(path.main, '1. simulation datasets/', scenario,
 '/set.',i,sep=""))


    # 2.1.2 The dependent variable bmi_change is set to NA if the participation
    # indicator = 0.
    set1$bmi_change[which(set1$filterMAR1 == 0)] <- NA

    # 2.1.3 Only the varaibles needed for model MAR 1 are selected
    # For the different models, different variables were used instead of
    # 'educ_neigh'
    # (i.e.  model MAR 2 : 'bmi_t1', model MAR 3 : 'newVAR30' and model MNAR : no
    # extra variable).
```

```
    set1 <- set1[,var]
    set1 <- missing_data.frame(set1)
    set1 <- change(set1, y = c("ID"), what = "type", to = c("irrelevant"))

    # 2.1.5 Actual imputation, with 5 impuation datasets.
    mi.data <- mi(set1, n.chain=5, max.minutes=10000, n.iter=50)

    # 2.1.6 The dataset is saved in an .Rdata workspace image.
    image <- paste("my data", , scenario,"\\set.",i,".Rdata",sep="")
    save(mi.data,file=image)
    }
}

# 2.2 The function is applied over the 11 chosen odds ratios (e.g. scen10
# corresponds to OR=1.0).
# This step might take some time. Therefore, the code can be run for 'complete
# case' and 'IPW' only.


scenario = 'scen10'
i = 1

final.mi1('scen10')
final.mi1('scen15')
final.mi1('scen20')
final.mi1('scen25')
final.mi1('scen30')
final.mi1('scen35')
final.mi1('scen40')
final.mi1('scen45')
final.mi1('scen50')
final.mi1('scen55')
final.mi1('scen60')

############################################################
# Step 3: 'Complete case' and 'Inverse Probability weighting' methods performed
# simultaneously.
############################################################
# rep.glm is the function that repeats 500 times the general linear models
# with or without weighing, given the level of association with the missingness
# (='scenario').

rep.glm <- function(scenario){
 # 3.2.1 we initiate a results table 'res.tab'.
 res.tab <- data.frame(set = seq(1, nsample, 1), nmiss = rep(NA, nsample))

 for(i in 1:nsample){
    # 3.2.2 One simulation dataset read
    path <- paste(path.main,'1. simulation datasets/', scenario, sep="")
    db <- read.csv(paste(path, '/set.', i, sep=""))

    # 3.2.3 Missings values in 'bmi_change' are set.
    idx <- which(db$filterMAR1 == 0); db$bmi_change[idx] <- NA

    # 3.2.4 Complete case regression.
    fitc <- summary(glm(bmi_change ~ educ_indiv + age + male ,data=db))
```

```
    # 3.2.5 Inverse probability weighting.
    db$plog <- predict(glm(filterMAR1 ~ educ_indiv, educ_neigh,
                      data=db, family='binomial'), type='response')

    fiti <- summary(glm(bmi_change ~ educ_indiv + age + male,
                   weight=1/plog,data=db))

    # 3.2.6 Number of missing values.
    idx <- which(db$filterMAR1 == 0)
    res.tab[i,'nmiss'] <- length(idx)

    # 3.2.7 The coefficients from the models extracted and save in table res.tab.
    res.tab[i,'cce']    <- fitc$coef['educ_indiv','Estimate']
    res.tab[i,'cwe']    <- fiti$coef['educ_indiv','Estimate']

 }
 res.tab
}



# the function is applied to all odds ratios.
gl10 <- rep.glm('scen10')
gl15 <- rep.glm('scen15')
gl20 <- rep.glm('scen20')
gl25 <- rep.glm('scen25')
gl30 <- rep.glm('scen30')
gl35 <- rep.glm('scen35')
gl40 <- rep.glm('scen40')
gl45 <- rep.glm('scen45')
gl50 <- rep.glm('scen50')
gl55 <- rep.glm('scen55')
gl60 <- rep.glm('scen60')

##########################################################
# step 4: Multiple imputation methods.
##########################################################
# 4.1 Regression function for Multiple Imputation datasets.
rep.mi <- function(scenario){

 # 4.1.1 We initiate a results table 'tabel'.
tabel <- data.frame(set=1:500, coef = rep(NA,500))
 var <- c("ID", "bmi_y2_Y1", "nivetude_h", "Age", "homme")

 for(i in 1:500){

    # 4.1.2 Previously saved Rdata workspace image is loaded.
    load(paste('set.', scenario, i,".Rdata",sep=""))

    # 4.1.3 Regression on 5 datasets and pooling.
    sum.fit <- summary (pool(bmi_change ~ educ_indiv  + age + male, data=mi.data, m
= 5, FUN=glm))

    # 4.1.4 Extraction of the coefficients.
    res.tab[i,'coef']   <- sum.fit$coef['educ_indiv','Estimate']
```

```
 }
 res.tab
}

# 4.2 The function is applied to all scenarios of different levels
# of association with the missingness.
mi10 <- rep.mi('scen10')
mi15 <- rep.mi('scen15')
mi20 <- rep.mi('scen20')
mi25 <- rep.mi('scen25')
mi30 <- rep.mi('scen30')
mi35 <- rep.mi('scen35')
mi40 <- rep.mi('scen40')
mi45 <- rep.mi('scen45')
mi50 <- rep.mi('scen50')
mi55 <- rep.mi('scen55')
mi60 <- rep.mi('scen60')

# When doing the test, it is adviced to save the image at this point,
# especially with a high number of simulations.
# If you wish to do so, uncomment the 3 lines below.
# setwd(path.main)

save.image("model 1.RData")
load("model 1.RData")




############################################################
# step 5: summary of the coefficients of the simulations.
############################################################
# 5.1 some simple summary functions.
# 5.1.1 The means of the coefficients.
mean.coef <- function(coef){
  mcoef <- c( mean(gl10[,coef]), mean(gl15[,coef]), mean(gl20[,coef]),
      mean(gl25[,coef]), mean(gl30[,coef]), mean(gl35[,coef]), mean(gl40[,coef]),
      mean(gl45[,coef]), mean(gl50[,coef]), mean(gl55[,coef]), mean(gl60[,coef]))
  mcoef
}

mean.coef.mi <- function(coef){
  mcoef <- c( mean(mi10[,coef]), mean(mi15[,coef]), mean(mi20[,coef]),
      mean(mi25[,coef]), mean(mi30[,coef]), mean(mi35[,coef]),mean(mi40[,coef]),
      mean(mi45[,coef]), mean(mi50[,coef]), mean(mi55[,coef]),mean(mi60[,coef]))
  mcoef
}

# 5.1.2 The median.
medi.coef <- function(coef){
  mcoef <- c( median(gl10[,coef]), median(gl15[,coef]), median(gl20[,coef]),
       median(gl25[,coef]), median(gl30[,coef]), median(gl35[,coef]),
```

```
      median(gl40[,coef]), median(gl45[,coef]), median(gl50[,coef]),
      median(gl55[,coef]), median(gl60[,coef]))
  mcoef
}

medi.coef.mi <- function(coef){
  mcoef <- c( median(mi10[,coef]), median(mi15[,coef]), median(mi20[,coef]),
      median(mi25[,coef]), median(mi30[,coef]), median(mi35[,coef]),
      median(mi40[,coef]), median(mi45[,coef]), median(mi50[,coef]),
      median(mi55[,coef]), median(mi60[,coef]))
  mcoef
}

# 5.1.3 The upper limit, i.e 97.5 percentile.
ul.coef <- function(coef){
  mcoef <- c( quantile(gl10[,coef],0.975),
              quantile(gl15[,coef],0.975), quantile(gl20[,coef],0.975),
              quantile(gl25[,coef],0.975), quantile(gl30[,coef],0.975),
              quantile(gl35[,coef],0.975), quantile(gl40[,coef],0.975),
              quantile(gl45[,coef],0.975), quantile(gl50[,coef],0.975),
              quantile(gl55[,coef],0.975), quantile(gl60[,coef],0.975))
  mcoef
}

ul.coef.mi <- function(coef){
  mcoef <- c( quantile(mi10[,coef],0.975),
              quantile(mi15[,coef],0.975), quantile(mi20[,coef],0.975),
              quantile(mi25[,coef],0.975), quantile(mi30[,coef],0.975),
              quantile(mi35[,coef],0.975), quantile(mi40[,coef],0.975),
              quantile(mi45[,coef],0.975), quantile(mi50[,coef],0.975),
              quantile(mi55[,coef],0.975), quantile(mi60[,coef],0.975))
  mcoef
}

# 5.1.4 The lower limit, i.e 2.5 percentile.
ll.coef <- function(coef){
  mcoef <- c( quantile(gl10[,coef],0.025),
              quantile(gl15[,coef],0.025), quantile(gl20[,coef],0.025),
              quantile(gl25[,coef],0.025), quantile(gl30[,coef],0.025),
              quantile(gl35[,coef],0.025), quantile(gl40[,coef],0.025),
              quantile(gl45[,coef],0.025), quantile(gl50[,coef],0.025),
              quantile(gl55[,coef],0.025), quantile(gl60[,coef],0.025))
  mcoef
}

ll.coef.mi <- function(coef){
  mcoef <- c( quantile(mi10[,coef],0.025),
              quantile(mi15[,coef],0.025), quantile(mi20[,coef],0.025),
              quantile(mi25[,coef],0.025), quantile(mi30[,coef],0.025),
              quantile(mi35[,coef],0.025), quantile(mi40[,coef],0.025),
              quantile(mi45[,coef],0.025), quantile(mi50[,coef],0.025),
              quantile(mi55[,coef],0.025), quantile(mi60[,coef],0.025))
  mcoef
}

# 5.2 The number of missings for each OR.
```

```
nmiss <- mean.coef('nmiss')

# 5.3 The median, 0.025 and 0.975 percentile of the coefficients, for each OR.
# 5.3.1 Complete case coefficients.
ccm <- medi.coef('cce')
ccl <- ll.coef('cce')
ccu <- ul.coef('cce')

# 5.3.2 Inverse probability weighting coefficients
ipm <- medi.coef('cwe')
ipl <- ll.coef('cwe')
ipu <- ul.coef('cwe')

# 5.3.3 Multiple imputation coefficients
mim <- medi.coef.mi('coef')
mil <- ll.coef.mi('coef')
miu <- ul.coef.mi('coef')


##########################################
# 6. Graph.
##########################################

# 6.1 regression in case of no missing data, the coefficient is used as the
# reference.
fit0  <- summary(glm(bmi_change ~ educ_indiv + age + male, data=set))
coef0 <- fit0$coef['educ_indiv','Estimate']

# 6.2 A character strings used in the plot below.
ORlevel <- as.factor(paste("OR", seq(1.0, 6, 0.5), sep="."))

# 6.3 Uncomment the 2 lines 'jpeg' and the line 'dev.off' at the end to save the
# graph.
# jpeg(filename = paste(path.main,'/plot model MAR 1.jpg', sep=""),

# 6.4 Actual plot.
# Change the ylim if it's not adapted to your results.
plot.default(ORlevel, cce, lim=c(-0.45,0.2),  type="n",xaxt="n",
ylab='Estimate', xlab='Scenario')
 axis(1, at = seq(1,11,1), labels = ORlevel)

# 6.4.1 The reference regression coefficient.
abline(coef0,0,lwd=2.5)

# 6.4.2 The complete case medians, lower and upper levels for each scenario of
# association.
lines(ORlevel, ccm,)
lines(ORlevel, ccl, lty=1)
lines(ORlevel, ccu, lty=1)

# 6.4.3 The IPW results.
lines(ORlevel, ipm, lty=5)

# 6.4.4 The Multiple Imputation results.
lines(ORlevel, mim, lty=3)
lines(ORlevel, mil, lty=3)
```

```
lines(ORlevel, miu, lty=3)

legend(1,0.1, legend=c('No missing data', 'Complete case', 'IPW', 'MI'),
    cex=0.8, lty=c(1,1,5,3), lwd=c(2,1,1,1))

dev.off()
```

```
#########################################################
# 0. Extra: A part of code to simulate a dataset similar to the dataset used in the
# article.
# The results will be different from any of the presented models.
# This is because the construction of this dataset only includes
# the correlations between the variables,
# but not the mean values nor the variances of the variables.
# Also, the dichotomous variables are replaced by continuous variables.
# However, it does enable the reader to test the code,
# and it will give similar results on
# the comparison of the 3 methods (i.e. complete case, IPW and MI).
#########################################################

# 0.1. Set and create a directory where you want to save the simulation datasets.
# Change the directory name in the line below to your personal preference.
path.main  <- "C:/your_file/"
dir.create(path.main, showWarnings = FALSE, recursive = TRUE)

# 0.2. Create subdirectory needed for the data storage.
models <- c('modelMAR1', 'modelMAR2', 'modelMAR3', 'modelMNAR')

scenarios <- c("scen10", "scen15", "scen20", "scen25", "scen30", "scen35",
      "scen40", "scen45", "scen50", "scen55", "scen60")

for(i in models){
for(j in scenarios){
    path <- paste(path.main, '1. simulation datasets/', j, sep="")
    dir.create(path, showWarnings = FALSE, recursive = TRUE)

    path <- paste(path.main, '2. multiple imputed datasets/', i, '/', j, sep="")
    dir.create(path, showWarnings = FALSE, recursive = TRUE)
 }
}

# 0.3. Construction of the dataset.
# 0.3.1 We define a correlation matrix.
cor.matrix <-
 matrix(cbind(   1.000, -0.071, -0.066, -0.097, -0.087, -0.039,  0.300,
                -0.071,  1.000,  0.308, -0.128, -0.100,  0.081,  0.009,
                -0.066,  0.308,  1.000, -0.150,  0.150,  0.020, -0.025,
                -0.097, -0.128, -0.150,  1.000,  0.094,  0.116, -0.023,
                -0.087, -0.100,  0.150,  0.094,  1.000, -0.105, -0.023,
                -0.039,  0.081,  0.020,  0.116, -0.105,  1.000, -0.011,
                 0.300,  0.009, -0.025, -0.023, -0.023, -0.011,  1.000),
        nrow=7)

# 0.3.2 Cholensky decomposition of the correlation matrix
chol.matrix <- t(chol(cor.matrix))

# 0.3.3 Creation of a dataset with 7 variables and 500 participants.
# The number of variables is fixed but the number of participants can be augmented,
# if you wish the results to be less depend of the random sample simulated by the
# cholesky decomposition.
# In this dataset, all variables are uncorrelated.
```

```
number.var <- 7
number.obs <- 500
uncorr.set = matrix(rnorm(number.var * number.obs,0,1), nrow=number.var,
ncol=number.obs);

# 0.3.4 The uncorrelated dataset is multiplied with the result of the Cholensky
# decomposition.
# The result is a 'mock' dataset with similar characteristics to the dataset used
# in the article. This set can be used to test the code above.
set <- as.data.frame(t(chol.matrix %*% uncorr.set))
names(set) <- c("bmi_change", "educ_indiv", "educ_neigh", "bmi_t1" ,
        "age", "male", "newVar30")

##########################################################
```

Appendix 3. Descriptive characteristics of the RECORD participants.

Appendix Table 2. Descriptive characteristics for the RECORD participants with BMI measured at baseline (N = 7,172); for the participants with BMI also assessed in the second wave (N=3,693); and for participants with BMI at baseline but who dropped out in the second wave (N = 3,479)

Variables	Participants with baseline BMI (N = 7,172)	Participants with BMI also in the second wave (N = 3,693)	Participants who dropped out in the second wave (N=3,479)
BMI (kg/m^2); mean (SD)	25.4 (4.2)	25.4 (3.9)	25.5 (4.4)
Age; mean (SD)	50.3 (11.7)	51.5 (11.3)	48.9 (11.9)
Male (%)	65.6	68.6	62.4
Individual education			
Low (%)	32.2	29.2	35.4
High (%)	67.8	70.8	64.6
Proportion of residents with >2 University years; mean (SD)[a]	0.41 (0.14)	0.42 (0.14)	0.40 (0.15)
Perceived stress; mean (SD)[b]	4.10 (3.0)	3.9 (2.9)	4.30 (3.1)
Depressive symptoms; mean (SD)[c]	1.64 (2.7)	1.52 (2.5)	1.78 (2.8)

[a] This variable was computed using data from the 2006 population census geocoded at the building address, within a street network buffer with a radius of 1000m centered on the participants' residences.

[b] Assessed with the Perceived Stress Scale of Cohen.

[c] Evaluated with the QD2A scale of Pichot.

Appendix 4. Numerical findings corresponding to Figures 2 – 4 in the main text.

Appendix Table 3. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the first attrition mechanism (MAR, individual education, neighborhood education) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.258	(-0.274; -0.241)	-0.260	(-0.279; -0.241)	-0.258	(-0.274; -0.241)
OR = 1.5	-0.257	(-0.282; -0.233)	-0.250	(-0.288; -0.230)	-0.257	(-0.281; -0.233)
OR = 2.0	-0.258	(-0.298; -0.220)	-0.247	(-0.326; -0.202)	-0.257	(-0.297; -0.218)
OR = 2.5	-0.262	(-0.306; -0.206)	-0.266	(-0.344; -0.143)	-0.259	(-0.307; -0.199)
OR = 3.0	-0.264	(-0.320; -0.205)	-0.270	(-0.316; -0.187)	-0.262	(-0.321; -0.197)
OR = 3.5	-0.264	(-0.333; -0.182)	-0.269	(-0.431; -0.141)	-0.258	(-0.343; -0.166)
OR = 4.0	-0.264	(-0.353; -0.179)	-0.255	(-0.347; -0.063)	-0.257	(-0.372; -0.145)
OR = 4.5	-0.267	(-0.360; -0.168)	-0.210	(-0.384; -0.071)	-0.259	(-0.391; -0.126)
OR = 5.0	-0.266	(-0.372; -0.156)	-0.205	(-0.401; -0.069)	-0.256	(-0.413; -0.106)
OR = 5.5	-0.268	(-0.396; -0.137)	-0.212	(-0.388; -0.073)	-0.260	(-0.439; -0.065)
OR = 6.0	-0.274	(-0.402; -0.148)	-0.201	(-0.401; -0.077)	-0.267	(-0.474; -0.073)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix Table 4. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the second attrition mechanism (MAR, individual education, baseline BMI) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.258	(-0.272; -0.243)	-0.258	(-0.276; -0.239)	-0.258	(-0.272; -0.243)
OR = 1.5	-0.257	(-0.285; -0.232)	-0.257	(-0.294; -0.223)	-0.258	(-0.286; -0.232)
OR = 2.0	-0.256	(-0.296; -0.214)	-0.253	(-0.316; -0.191)	-0.258	(-0.301; -0.211)
OR = 2.5	-0.253	(-0.306; -0.202)	-0.247	(-0.333; -0.159)	-0.256	(-0.322; -0.191)
OR = 3.0	-0.252	(-0.305; -0.191)	-0.241	(-0.349; -0.128)	-0.260	(-0.335; -0.173)
OR = 3.5	-0.250	(-0.326; -0.182)	-0.230	(-0.373; -0.096)	-0.263	(-0.375; -0.134)
OR = 4.0	-0.245	(-0.317; -0.164)	-0.223	(-0.368; -0.082)	-0.264	(-0.383; -0.096)
OR = 4.5	-0.239	(-0.322; -0.155)	-0.211	(-0.369; -0.063)	-0.266	(-0.409; -0.056)
OR = 5.0	-0.232	(-0.327; -0.137)	-0.203	(-0.360; -0.052)	-0.259	(-0.454; 0.055)
OR = 5.5	-0.227	(-0.323; -0.119)	-0.206	(-0.349; -0.051)	-0.257	(-0.477; 0.083)
OR = 6.0	-0.225	(-0.350; -0.110)	-0.203	(-0.379; -0.041)	-0.269	(-0.531; -0.010)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix Table 5. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the attrition mechanisms with simulated correlation (MAR, individual education, simulated variable: r = 0.3) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.257	(-0.272; -0.241)	-0.257	(-0.276; -0.240)	-0.258	(-0.272; -0.241)
OR = 1.5	-0.255	(-0.277; -0.228)	-0.255	(-0.286; -0.217)	-0.258	(-0.282; -0.232)
OR = 2.0	-0.245	(-0.278; -0.206)	-0.246	(-0.293; -0.189)	-0.258	(-0.292; -0.223)
OR = 2.5	-0.228	(-0.270; -0.182)	-0.228	(-0.299; -0.142)	-0.258	(-0.307; -0.212)
OR = 3.0	-0.205	(-0.262; -0.147)	-0.203	(-0.302; -0.096)	-0.259	(-0.312; -0.197)
OR = 3.5	-0.180	(-0.242; -0.117)	-0.176	(-0.291; -0.049)	-0.257	(-0.339; -0.185)
OR = 4.0	-0.156	(-0.230; -0.078)	-0.149	(-0.287; -0.026)	-0.260	(-0.347; -0.166)
OR = 4.5	-0.129	(-0.218; -0.046)	-0.121	(-0.265; 0.019)	-0.261	(-0.366; -0.145)
OR = 5.0	-0.110	(-0.199; -0.014)	-0.110	(-0.253; 0.049)	-0.259	(-0.390; -0.134)
OR = 5.5	-0.089	(-0.187; 0.014)	-0.086	(-0.240; 0.075)	-0.249	(-0.416; -0.112)
OR = 6.0	-0.072	(-0.170; 0.026)	-0.075	(-0.220; 0.069)	-0.251	(-0.432; -0.102)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix Table 6. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the attrition mechanisms with simulated correlation (MAR, individual education, simulated variable: r = 0.4) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.258	(-0.273; -0.241)	-0.258	(-0.274; -0.241)	-0.257	(-0.272; -0.242)
OR = 1.5	-0.254	(-0.278; -0.231)	-0.254	(-0.282; -0.223)	-0.258	(-0.283; -0.234)
OR = 2.0	-0.241	(-0.274; -0.206)	-0.239	(-0.298; -0.187)	-0.256	(-0.291; -0.221)
OR = 2.5	-0.221	(-0.272; -0.176)	-0.220	(-0.285; -0.147)	-0.257	(-0.308; -0.211)
OR = 3.0	-0.196	(-0.247; -0.139)	-0.194	(-0.276; -0.102)	-0.257	(-0.315; -0.193)
OR = 3.5	-0.167	(-0.232; -0.100)	-0.165	(-0.266; -0.054)	-0.256	(-0.341; -0.176)
OR = 4.0	-0.132	(-0.208; -0.056)	-0.125	(-0.248; 0.008)	-0.249	(-0.352; -0.162)
OR = 4.5	-0.108	(-0.187; -0.021)	-0.097	(-0.234; 0.033)	-0.258	(-0.368; -0.142)
OR = 5.0	-0.084	(-0.176; 0.009)	-0.075	(-0.206; 0.064)	-0.252	(-0.387; -0.123)
OR = 5.5	-0.057	(-0.152; 0.047)	-0.055	(-0.204; 0.109)	-0.253	(-0.420; -0.095)
OR = 6.0	-0.040	(-0.151; 0.073)	-0.034	(-0.187; 0.109)	-0.255	(-0.474; -0.078)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix Table 7. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the attrition mechanisms with simulated correlation (MAR, individual education, simulated variable: r = 0.5) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.257	(-0.273; -0.243)	-0.258	(-0.273; -0.243)	-0.257	(-0.273; -0.243)
OR = 1.5	-0.253	(-0.278; -0.226)	-0.252	(-0.281; -0.220)	-0.258	(-0.284; -0.232)
OR = 2.0	-0.237	(-0.274; -0.197)	-0.238	(-0.285; -0.182)	-0.260	(-0.294; -0.223)
OR = 2.5	-0.205	(-0.253; -0.159)	-0.205	(-0.271; -0.126)	-0.256	(-0.313; -0.208)
OR = 3.0	-0.168	(-0.230; -0.110)	-0.167	(-0.257; -0.075)	-0.258	(-0.321; -0.188)
OR = 3.5	-0.130	(-0.192; -0.058)	-0.125	(-0.230; -0.015)	-0.256	(-0.336; -0.171)
OR = 4.0	-0.092	(-0.162; -0.014)	-0.084	(-0.197; 0.042)	-0.262	(-0.355; -0.161)
OR = 4.5	-0.050	(-0.128; 0.038	-0.053	(-0.174; 0.095)	-0.255	(-0.397; -0.137)
OR = 5.0	-0.019	(-0.112; 0.073)	-0.017	(-0.153; 0.123)	-0.254	(-0.424; -0.113)
OR = 5.5	0.016	(-0.071; 0.104)	0.016	(-0.138; 0.144)	-0.243	(-0.426; -0.090)
OR = 6.0	0.046	(-0.066; 0.135)	0.040	(-0.104; 0.190)	-0.247	(-0.493; -0.064)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix Table 8. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the attrition mechanisms with simulated correlation (MAR, individual education, simulated variable: r = 0.6) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.257	(-0.272; -0.241)	-0.257	(-0.272; -0.243)	-0.257	(-0.272; -0.242)
OR = 1.5	-0.254	(-0.281; -0.228)	-0.254	(-0.279; -0.222)	-0.259	(-0.285; -0.232)
OR = 2.0	-0.235	(-0.272; -0.202)	-0.237	(-0.281; -0.191)	-0.256	(-0.293; -0.222)
OR = 2.5	-0.207	(-0.253; -0.160)	-0.212	(-0.280; -0.145)	-0.256	(-0.307; -0.208)
OR = 3.0	-0.170	(-0.229; -0.110)	-0.170	(-0.246; -0.084)	-0.259	(-0.327; -0.186)
OR = 3.5	-0.129	(-0.196; -0.065)	-0.132	(-0.124; -0.029)	-0.253	(-0.345; -0.162)
OR = 4.0	-0.089	(-0.161; -0.017)	-0.096	(-0.197; 0.036)	-0.254	(-0.373; -0.154)
OR = 4.5	-0.059	(-0.139; 0.032)	-0.057	(-0.177; 0.075)	-0.254	(-0.396; -0.131)
OR = 5.0	-0.019	(-0.113; 0.066)	-0.025	(-0.154; 0.118)	-0.247	(-0.425; -0.112)
OR = 5.5	0.012	(-0.088; 0.118)	0.002	(-0.145; 0.127)	-0.237	(-0.529; -0.086)
OR = 6.0	0.051	(-0.063; 0.151)	0.032	(-0.111; 0.156)	-0.229	(-0.503; -0.069)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

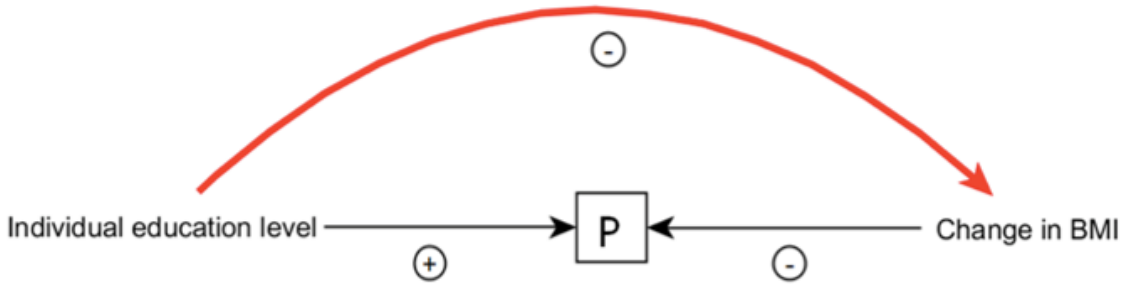
Appendix Table 9. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with the attrition mechanisms with simulated correlation (MAR, individual education, simulated variable: r = 0.7) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.257	(-0.273; -0.242)	-0.257	(-0.271; -0.244)	-0.258	(-0.273; -0.243)
OR = 1.5	-0.249	(-0.274; -0.223)	-0.251	(-0.275; -0.225)	-0.257	(-0.281; -0.231)
OR = 2.0	-0.227	(-0.264; -0.188)	-0.232	(-0.274; -0.190)	-0.257	(-0.294; -0.223)
OR = 2.5	-0.188	(-0.234; -0.141)	-0.198	(-0.255; -0.124)	-0.258	(-0.308; -0.209)
OR = 3.0	-0.136	(-0.198; -0.080)	-0.150	(-0.225; -0.065)	-0.258	(-0.332; -0.195)
OR = 3.5	-0.084	(-0.146; -0.025)	-0.097	(-0.188; -0.011)	-0.254	(-0.350; -0.173)
OR = 4.0	-0.031	(-0.105; 0.035)	-0.049	(-0.147; 0.045)	-0.251	(-0.388; -0.151)
OR = 4.5	0.018	(-0.066; 0.098)	-0.013	(-0.136; 0.085)	-0.251	(-0.465; -0.129)
OR = 5.0	0.063	(-0.026; 0.149)	0.018	(-0.089; 0.144)	-0.241	(-0.469; -0.100)
OR = 5.5	0.098	(0.000; 0.199)	0.046	(-0.068; 0.163)	-0.243	(-0.518; -0.087)
OR = 6.0	0.135	(0.021; 0.240)	0.077	(-0.051; 0.219)	-0.238	(-0.546; -0.062)

[1] The true coefficient was considered to be β = -0.26.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5th and 97.5th percentiles over the simulated datasets.

Appendix 5. Exploration of a Missing Not At Random pattern of attrition for the association between individual education and change in BMI



Appendix Figure 1. This Figure reports a signed Directed Acyclic Graph of the attrition bias following a Missing Not At Random (MNAR) mechanism, for the effect of individual education on change of BMI. Invidual education and change in BMI are causes of non-participation in the second wave (P = 0). Participants with a high individual education level and a lower increase in BMI are more likely to participate. The MNAR attrition bias in the estimate of interest is introduced by conditioning on participation (P = 1), which is caused by both the explanatory variable of interest and the outcome. Signed graphs were established based on the assumption of so-called weak monotonic effects.

Appendix Table 10. Number of observations with a missing outcome at the follow-up (and corresponding percentage) for the MNAR mechanism of attrition under each attrition level (the initial database in which attrition is simulated comprises 3,693 participants)

Attrition level	MNAR attrition mechanism[a]
OR = 1.0[b]	66 (1.8)
OR = 1.5	140 (3.8)
OR = 2.0	243 (6.6)
OR = 2.5	370 (10)
OR = 3.0	510 (13.8)
OR = 3.5	655 (17.8)
OR = 4.0	785 (21.2)
OR = 4.5	930 (25.2)
OR = 5.0	1054 (28.5)
OR = 5.5	1171 (31.7)
OR = 6.0	1277 (34.5)

[a] MNAR mechanism: individual education and change in BMI influence dropout.

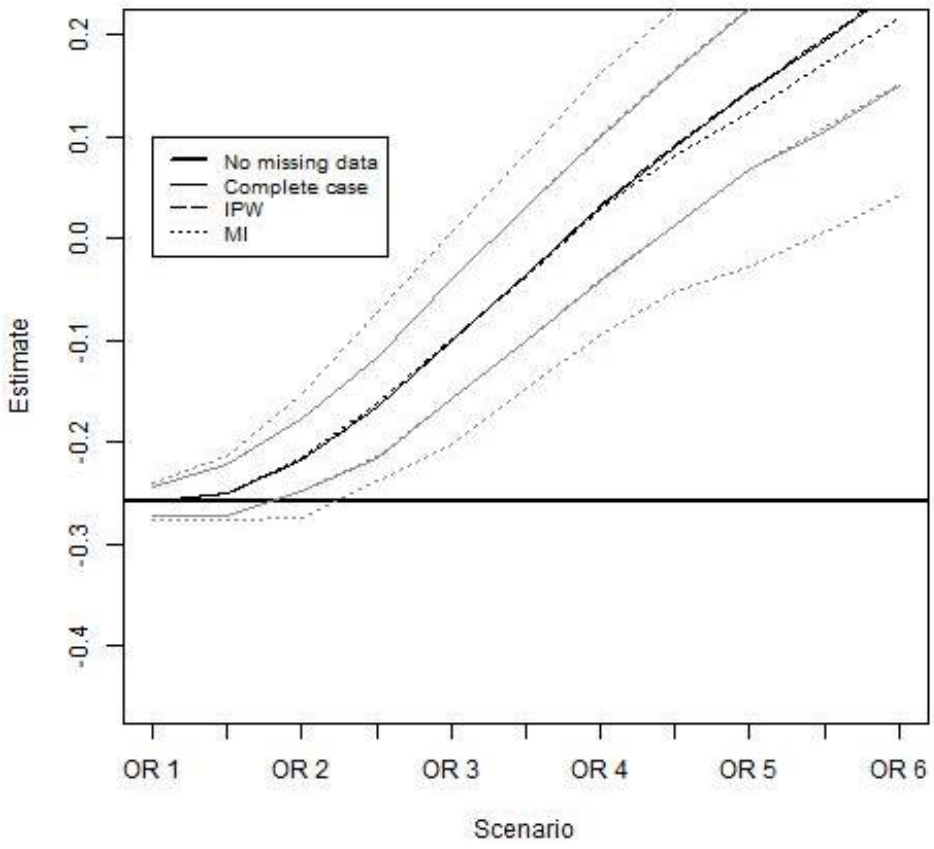
[b] The OR refers to a multiplicative coefficient for the odds of having a missing outcome in wave 2 for having a low rather than a high individual education level (binary variable) and for a 2-standard deviation increase in the change in BMI (a 2-standard deviation was chosen to mimic the effect of a binary variable with approximately half of the population in each group). The distribution of the standardized change in BMI was shifted towards the positive, so as to have a minimum value equal to 0, yielding the minimal probability of attrition when this recoded variable was equal to 0. The model to define the probability of attrition for each participant was a logistic model with an intercept equal to –4.

Appendix Table 11. Median beta coefficients[1] over the 500 simulated databases and 95% credible intervals for the association between individual education and change in BMI with a MNAR attrition mechanism (based on influences of individual education and change in BMI on participation) from complete case analysis, inverse probability weighting and multiple imputation

Attrition level	CCA		MI		IPW	
	β (95% credible interval)[2]		β (95% credible interval)[2]		β (95% credible interval)[2]	
OR = 1.0	-0.257	(-0.272; -0.243)	-0.257	(-0.275 – -0.240)	-0.257	(-0.272 ; -0.243)
OR = 1.5	-0.249	(-0.273; -0.222)	-0.250	(-0.277 – -0.213)	-0.249	(-0.273 ; -0.222)
OR = 2.0	-0.217	(-0.249; -0.176)	-0.215	(-0.275 – -0.153)	-0.217	(-0.249 ; -0.176)
OR = 2.5	-0.165	(-0.214; -0.116)	-0.160	(-0.237 – -0.072)	-0.165	(-0.214 ; -0.116)
OR = 3.0	-0.101	(-0.157; -0.040)	-0.097	(-0.202 – 0.007)	-0.100	(-0.156 ; -0.039)
OR = 3.5	-0.036	(-0.100; 0.030)	-0.037	(-0.146 – 0.085)	-0.035	(-0.099 ; 0.031)
OR = 4.0	0.032	(-0.041 – 0.100)	0.029	(-0.094 – 0.163)	0.033	(-0.039 ; 0.101)
OR = 4.5	0.090	(0.013 – 0.164)	0.082	(-0.052 – 0.224)	0.091	(0.013 ; 0.166)
OR = 5.0	0.145	(0.067 – 0.225)	0.124	(-0.028 – 0.278)	0.147	(0.068 ; 0.226)
OR = 5.5	0.195	(0.107 – 0.280)	0.173	(0.006 – 0.322)	0.198	(0.111 ; 0.282)
OR = 6.0	0.246	(0.150 – 0.342)	0.218	(0.043 – 0.394)	0.249	(0.153 ; 0.346)

[1] The true coefficient was considered to be $\beta = -0.26$.

[2] β was estimated as the median over the simulated datasets. The uncertainty in the estimate was assessed with the 2.5^{th} and 97.5^{th} percentiles over the simulated datasets.



Appendix Figure 2. Associations between individual education and change in BMI in simulated datasets with a MNAR attrition mechanism (based on influences of individual education and change in BMI on participation) under eleven scenarios of attrition level