Compensating for Non-response in Biosocial Research: Simulation Study from a Cross-sectional Analysis

Authors: Tina Hannemann[1], Natalie Shlomo[1], Tarani Chandola[1] and Georgia Chatzi[1]

[1] Social Statistics Department, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL United Kingdom

Abstract: We present a simulation study from real data of Wave 2 of the UK English Longitudinal Study of Ageing. In this study there are 4 stages where potential missing data can arise: questionnaire stage; nurse visit stage, blood collection stage and having a valid value for the C-reactive protein (CRP), a stress biomarker that will be used as the dependent variable in a substantive analysis where we research the effects of socio-economic indicators and other control variables. Generating a full sample, we induce missing data on variables according to the stages of missingness with 200 repetitions under three missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). We then carried out a complete case analysis and four compensation methods: inverse propensity weighting, a selection model according to the 2-stage Heckman approach, multiple imputation and a new proposed approach combining the selection model with the multiple imputation. In this study, we found that overall the multiple imputation proved to be the best compensation method for all missing data mechanisms.

Keywords: Missing data mechanisms, Inverse propensity weighting, Selection model, Multiple imputation, C-Reactive Protein biomarker

1. Introduction

In recent years, a greater number of large-scale socio-demographic surveys now include biomarker information from blood, hair or saliva samples. In the UK, several national longitudinal surveys collect this biomarker data such as the National Child Development Study, Understanding Society - The UK Household Longitudinal Study and the English Longitudinal Study of Aging (ELSA).

In the past, collection of such data, which often requires additional contact with participants and the use of clinical personnel and lab-technology for analysis and documentation, was too costly and laborious to be performed on large samples in national surveys. However, recent advancements in medical analysis methods have made it possible to collect biomarker data from large representative samples. Data from inflammatory and stress-related biomarkers collected together with sociodemographic and health-related characteristics allow for the comprehensive investigation of the social effects on biological mechanisms. In particular, the analysis of C-reactive proteins (CRP) in the blood as early indicator of chronic inflammation, the underlying cause for many cardio-vascular diseases, is a focus of many studies examining associations with socioeconomic position. Therefore, with the increase in research on biomarkers and the rising availability of such data in surveys, social research using biomarker data has experienced a boost in popularity, which can only be expected to continue to rise in the near future. Examples of biomarker socio-studies can be found in Fraga et al. 2015; Loucks et al. 2010; Maharani 2019; Muennig et al. 2007; Pollitt et al. 2007; Stringhini et al. 2013.

This new development in the collection of biomarker data does not come without its drawbacks. With the invasive nature of the biomarker sample collection, missing data among biomarkers is more frequent than for standard socio-economic questions and has potentially larger bias. This missing data could be a potential threat for the validity of biomarker research and its results and conclusions. For example, during data collection for the ELSA dataset, participants' health determines the further collection of blood samples. Only those participants in good physical health are asked for consent to collect blood samples. Specific health conditions and poor general health, together with refusal of a blood sample and technical problems during the collection and analysis of the blood samples in the laboratory, reduce the sample size. In wave 2 of ELSA, the original sample size of 8,780 was reduced by 32% with eligible blood samples available for only 5,899 participants, which is further reduced to 5,821 when accounting for missing item responses.

The pattern of missingness among biomarker data cannot be assumed to be random due to the complexities of the necessary good general health conditions and the need to consent at three different stages during data collection: i) main interview, ii) nurse visit, and iii) blood collection. This leaves the question of how representative is the analysis of only those cases which have valid biomarker data compared to the well-structured representative original sample. Indeed, many biomarker research studies are carried out using a complete case analyses where missing data is simply deleted from the study. Ignoring those participants who either are not fit enough for the blood sample or refuse to take part in the nurse visit or the blood sample, includes potentially large bias to the analysis. This study investigates the potential impact of missing data on the relationship between biomarkers and socio-economic position. In order to analyse the effect of missing data, different types of missingness have to be considered. Rubin (1976) classified three types of missing data mechanisms:

- Missing Completely at Random (MCAR) – the probability of missingness is unrelated to the set of observed responses and unobserved target variables, assuming the missing data is a random sub-sample of the full sample;
- Missing at Random (MAR) – the probability of missingness is related to the set of observed instrumental and auxiliary variables but unrelated to the unobserved target variables. Therefore, the missing data is assumed MCAR within strata defined by cross-classifying the observed variables;
- Missing Not at Random (MNAR) – the probability of missingness is directly related to the unobserved target variables, making the available biomarker sub sample not representative of the full sample.

The analysis method of *complete case* is only valid for the MCAR mechanism. Given the selectivity of the missing biomarker data, the assumptions for the MCAR mechanisms are not easily found in reality. As mentioned, missing information can occur at several stages of data collection and due to different reasons. For example, people with poor health might be excluded from blood sample collection. If this group of individuals has higher CRP levels than those that are of good health, due to their health condition, which excluded them from the biomarker sample, the remaining sample is no longer representative. For this case the assumption of MNAR is more credible. On the other hand, if the analyses controls for observed auxiliary variables such as poor health and age which are correlated with the missing CRP, this could lead to the assumption of MAR. Under the MAR mechanism, we assume that the missing data bias is negligible as long as all characteristics which determine missingness are controlled for in the model. This is a strong assumption and often impossible to test with survey data.

We present here a large-scale simulation study based on wave 2 of the ELSA study. The simulation is based on artificially introducing missing biomarker data according to the three missing data mechanisms on a sample dataset of size 10,000 generated from the complete cases (5,821 respondents) of the blood sample. We generate missing data using the three missing data mechanisms according to the true response rates found in ELSA wave 2 and generate 200 datasets each. We then apply five different compensation methods for the missing data and compare their impact on effect sizes of a regression model examining the association between socioeconomic position and inflammation using logged CRP as the response variable with those of the baseline model from the full dataset.

Section 2 describes the data and the substantive model that will be used in the simulation study. Section 3 explains how the missing data mechanisms are introduced into the data and Section 4 the methods that will be used to compensate for the missing data. Section 5 presents the results of the simulation study. We conclude in Section 6 with a discussion and future work.

2. Data and Substantive Model as Baseline for the Simulation Study

ELSA is a longitudinal study that collects multidisciplinary data representative of people ages 50 and over. The original sample was drawn from the Health Survey for England (HSE), which collected cross-sectional data of the general population annually. The sample was drawn by a multistage stratified random probability design from years 1998, 1999 and 2001 of HSE. The first wave of ELSA took place in 2002/2003 and the participants were interviewed every two years. In every wave, the data collection is comprised of a computer-assisted personal interview (CAPI) and a self-completion questionnaire. As of 2019, there are 8 waves (wave 2, 4, 6, and 8) of data which include biomarker from blood samples, collected in an additional nurse visit. For all three stages of data collection – main interview, nurse visit and blood sample – separate consent had to be given by the participants. If the nurse decided that a participant was not fit enough for a blood sample, due to medical pre-conditions or health data obtained during the nurse visit, the request for a blood sample was not extended to the participant.

This simulation study will use the complete set of cases with eligible biomarker data from wave 2 of ELSA to generate a sample size of 10,000 which will be denoted as the ‘simulation dataset’. The full biomarker sample from wave 2 without item missing data for variables in our substantive model includes 5,821 complete cases. We replicate the sample and then randomly delete 1,642 of the duplicates to obtain a sample size of 10,000. Thus, we can run a baseline model on the simulation dataset to obtain the ‘true values’ of the effect sizes on the effects of socio-economic status on the biomarker log(CRP). Note that for real survey data where the missing information remains unknown, we would not have knowledge of the ‘true’ effect sizes.

The substantive model in our simulation study uses log(CRP) values as the response variable. We then run three separate models using a set of socio-economic indicators (education level, occupational class and wealth quintile) and also include control variables: age, sex, general health, employment status at time of the interview. We keep the substantive model simple and concentrate on the complexity of the simulation study assessing compensation methods for missing data.

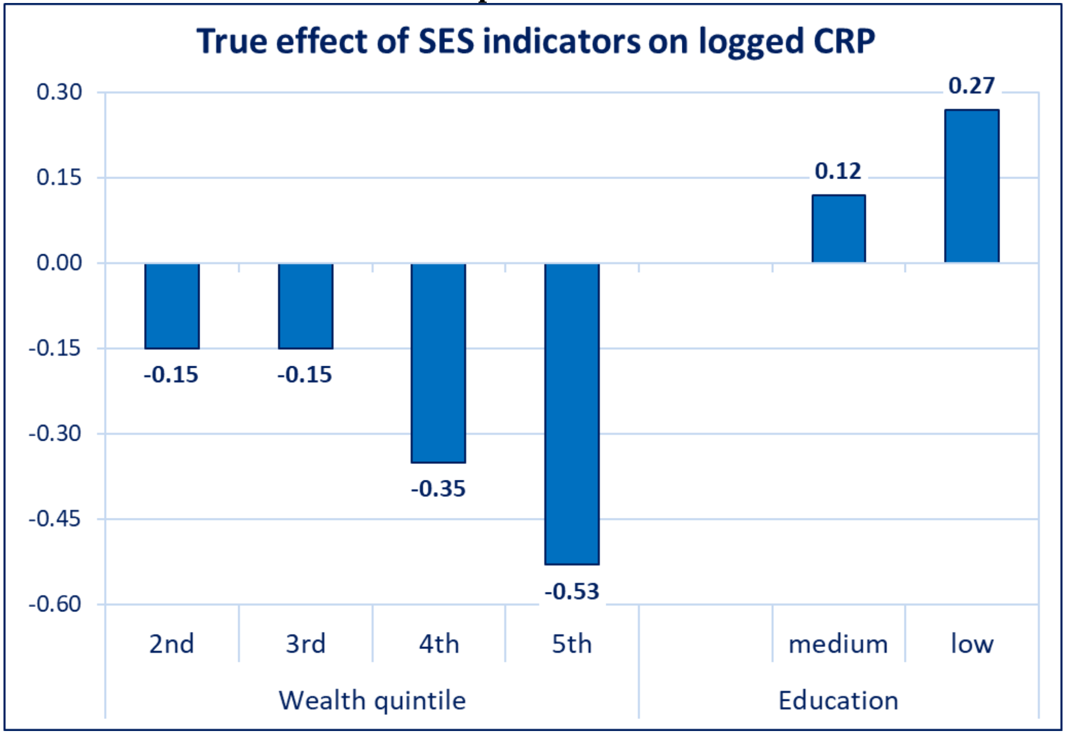
The socio-economic status variables are derived from the main questionnaire and defined as follows: Education level was coded originally in seven 7 valid instances, which were regrouped for this study into three categories: low education (no or foreign/other qualification), medium education (A-level, O-level or other graded equivalent qualification) and high education (higher education below degree or degree and equivalent qualification). Occupational class is used with five ordinal classifications as i) managerial and professional, ii) intermediate, iii) small employers and own accountant workers and lower supervisory and technical, iv) semi-routine and other occupations, and v) no information. The latter was created in order to preserve the sample size as it is the largest group for this variable. The wealth variable is derived from the continuous wealth variable in the main questionnaire and used as a quintile categorical variable for this study.

The controls are commonly found in inflammation research using log(CRP) as the response variable. Age is used as a categorical variable with possible values of: i) 51-55, ii) 56-60, iii) 61-65, iv) 66-70, v) 71-75, vi) 76-80 and vii) 81+. Sex, general health and employment are all measured as binary variables depending if the participant is male or female; general in good or bad heath; and in paid labour or not during data collection.

As mentioned, the models are run separately for each of the socioeconomic status variables along with the control variables on the simulation dataset to form the baseline for the simulation study. All results from the simulation study on datasets with missing data will be compared to this baseline, in order to establish the impact and variation caused by the compensation methods for missing data depending on the missing data mechanism. The variable occupation behaves similarly to the other socioeconomic variables, and hence we present here the results for the wealth and education level socioeconomic variables.

Figure 1 presents the coefficients for the baseline models for the education level and wealth variables on the simulation dataset and including the control variables in each model. The results for the baseline models for each of the socioeconomic status variables are as expected from previous literature. Medium and low education show significant higher levels of CRP compared to the high education reference category indicating the protective effect of higher education towards chronic stress. There is a clear wealth gradient for the analysed wealth quintiles with the wealthiest people showing significantly lower CRP levels.

Figure 1: Coefficients for two socioeconomic status indicators on the level of log(CRP) using ELSA simulation dataset with complete biomarker information



Note: Each socioeconomic status variable was analysed in a separate model to avoid cross-contamination of the effect sizes. Control variables include age groups, sex, general health and employment situation.

3. Simulating missing data patterns

In order to simulate missing data mechanisms, we use the true response rates from ELSA wave 2. We assume a monotonic missing data pattern where missing data occurs monotonically from the main questionnaire, then nurse visit and then blood sample and finally the value of CRP. Compared to the 11,391 wave 1 core members, wave 2 experienced attrition by 23% resulting in 8,780 core members. This sample is further reduced by 12.7% missing data due to a refusal of the nurse visit, then 13.2% missing data due to refusal of a blood sample during the nurse visit, and finally 11.3% cases that did not have a valid blood sample due to various reasons (Scholes et al. 2008). The latter case could be due to late analysis of spoiled blood samples or other technical issues and likely unrelated to the health condition of the participant. Based on these percentages, we simulate 200 datasets for each of the MCAR, MAR and MNAR missing data mechanisms as follows:

For the case of MCAR missing data mechanism, we draw random samples at each stage of the missing data (questionnaire, nurse visit and blood sample, valid CRP) according to the overall percentages of missingness above and denote the relevant variables in the selected records as missing. Specifically, in order to obtain a sample of 8,780 out of the 10,000 in our simulation dataset, we randomly select 12.2% of the records and denote the variables obtained at the questionnaire stage as missing. Note that due to the monotonic pattern of missingness, all the variables in the subsequent stages of nurse visit, blood sample and valid CRP are also denoted as missing. Out of the 8,780 records not missing, we randomly select 12.7% and denote the variables associated to the nurse visit as missing as well as those variables in the remaining stages. Out of those not missing, we randomly select 13.2% and denote the variables associated to the blood sample as missing as well as the valid CRP. Finally, out of those not missing, we randomly select 11.3% and denote them as having missing data for the CRP.

For the case of MAR missing data mechanism, where missing data is dependent on covariates but not on the outcome variables, we simulated missing data similar to the method for MCAR mechanism but separately within 16 strata built by four covariates in the baseline model in dichotomized form: age (60 and below or above 60), sex, health (ill heath, other) and wealth (1^{st}-2^{nd} quintile or 3^{rd} -5^{th} quintile). The percentages for generating the missing data in each of the 16 strata are based on the true response rates from ELSA Wave 2 and assuming again the monotonic missing data pattern. The percentages are presented in the appendix.

For the case of MNAR missing data mechanism, we used the approach taken for the MAR mechanism, however in the final stage of drawing records for the missing data of CRP, we disproportionately sampled more records to denote as missing values from among those cases with high CRP having a value over 3 (90% of the cases to be declared missing) compared to the cases with low CRP having a value less than 3 (10% of the cases to be declared missing).

Compensation Methods for Missing Data

In addition to the complete case scenario where only cases with complete information are analysed while cases with missing information are disregarded, the study considers four statistical approaches for compensating for missing data.

The approaches are:

i. **Inverse propensity weights:** Based on a response model using a logistic regression, we estimate the propensity of missingness of the CRP variable according to the following explanatory variables: BMI, blood pressure, age groups, wealth quintiles, alcohol consumption, occupational group, marital status, geographical region (GOR), sex, general health, physical activity level, the presences of longstanding illness, permanent pain, coronary vascular disease or regular medication. The complete cases are then weighted by the inverse of the estimated response propensity and used in the substantive modelling stage as a weight (these are defined as the ‘pweight’ in STATA under the svy commands).

ii. **Inverse Mills' Ratio (Heckman) selection model:** Heckman (1976) proposed a two-step procedure to deal with censored data which has been shown to be useful for compensating for selective nonresponse (Little 1982). The first step is to develop a response model based on a probit regression, using the same explanatory variables as for the inverse propensity weights described above. From these model predictions, a Mills Ratio is calculated and its inverse is then used as an additional covariate in the substantive modelling stage. The coefficients for the socioeconomic variables are then adjusted to account for the additional explanatory variable of the inverse Mills Ratio to facilitate comparisons to other approaches.

iii. **Multiple imputation:** Multiple datasets with complete information are created from imputed values for each missing item which represents a distribution of possibilities (Rubin 1987). In our simulation study, each value is imputed through a regression model under the MICE (multiple imputation chained equations) approach in STATA, where each variable with missing data is modelled conditionally upon the other variables in the data and its data type (Raghunathan et al. 2001; van Buuren 2007). The substantive model is then run on each complete dataset and the estimates are combined according to formulae given by Rubin (1987). Our estimates are based on imputations for 15 complete datasets.

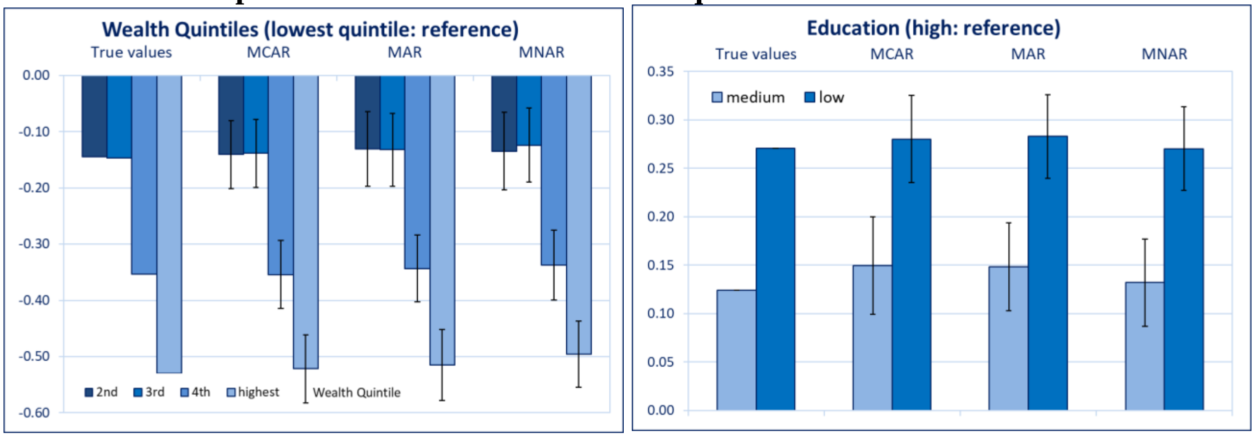
iv. **Combined method of multiple imputation and the Inverse Mills Ratio selection model:** We propose here a new approach where we include the inverse Mills Ratio into the multiple imputation chained equations method that are used to carry out the multiple imputations. The aim is to compensate for the missing data at the sage of imputation and not at the stage of the substantive analysis. We first carried out standard multiple imputation using the MICE procedure to impute variables that are to be used in the probit response model (not including the CRP value). This provided 3 complete datasets. Then, on each dataset we run the probit response model and calculate the inverse Mills ratio. The inverse Mills ratios are then added as an additional covariate in another set of MICE to impute the CRP values on 3 additional datasets. In total there are 9 imputed datasets. The substantive model is run on each dataset and the results combined according to formulae given by Rubin (1987).

Each of the compensation mechanisms and the complete case analysis are performed separately for the socio-economic indicators - education, wealth quintile and occupational group (not shown here) - on each of the 200 datasets generated for each of the missing data mechanisms MCAR, MAR and MNAR. Results are presented as averages over all 200 simulations of each combination of missing data mechanism, socioeconomic indicator and compensation methods.

4. Results

We present in Figures 2 through 6 the effect sizes (coefficients) from the substantive model for the wealth and education level variable (the occupation variable performed similarly). Each of the figures show first the baseline model effect sizes (denoted ‘True values’) alongside the average effects sizes across the 200 datasets for each of the three missing data mechanisms, including the confidence interval obtained from the simulation standard error. We start with Figure 2 which shows the average effect sizes and confidence intervals across the 200 samples for the complete case analysis.

Figure 2: **Complete Case Analysis of effect sizes for MCAR, MAR and MNAR missing data mechanisms compared to 'true' values for wealth quintiles and education level**

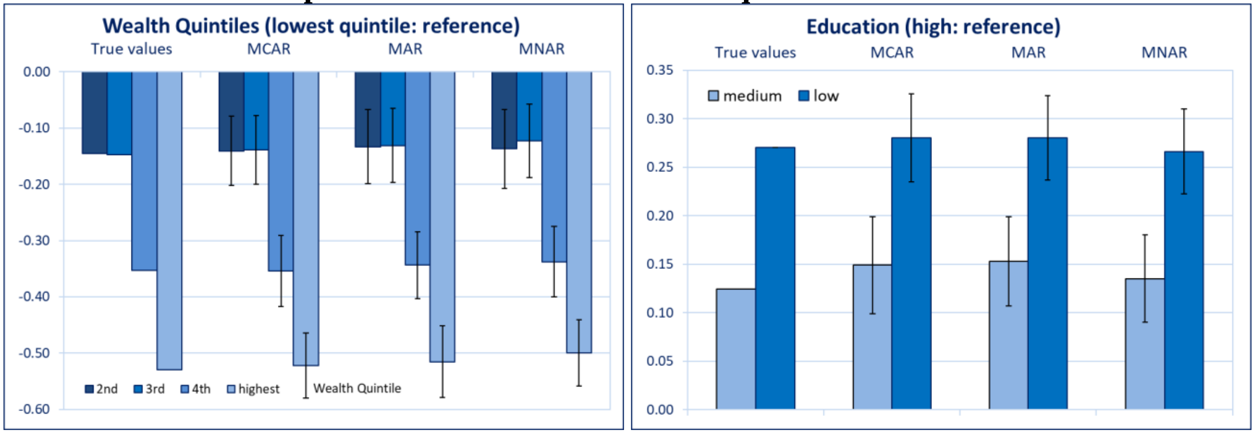


Note: Effects for each of the three missing patterns are averages over 200 simulations. All models control for age group, sex, general health and employment status at time of the interview.

For the complete case analysis in Figure 2, the effect size among the MCAR simulated datasets showed the smallest deviations from the 'true' effect sizes for the wealth variable as expected. For the MAR, and even stronger for the MNAR missing pattern, there is underestimation of the effect sizes for the wealth variable. MAR and MNAR simulated datasets demonstrate more realistic scenarios and therefore researchers would underestimate the predicted CRP value and the association between socioeconomic position and inflammation using only those cases with complete information. For the education variable, the picture is different. For all missing data mechanisms there is overestimation of the effect sizes for medium and lower education level compared to high education level. The overestimation is, surprisingly, highest for the MCAR and MAR scenario and lowest for the MNAR case. Thus, we may be overemphasising the effect size of education on predicted CRP values in the complete case analysis without compensating for the missing information. We note however, that the confidence intervals are wide and generally include the 'true' effect sizes.

Next, we show in Figure 3 the average effect sizes and confidence intervals across the 200 samples for the wealth and education level variables under the compensation method of inverse propensity weights on the complete case sample, according to the three missing data mechanisms.

Figure 3: Inverse Propensity Weighting effect sizes for MCAR, MAR and MNAR missing data mechanisms compared to true values for wealth quintiles and education level

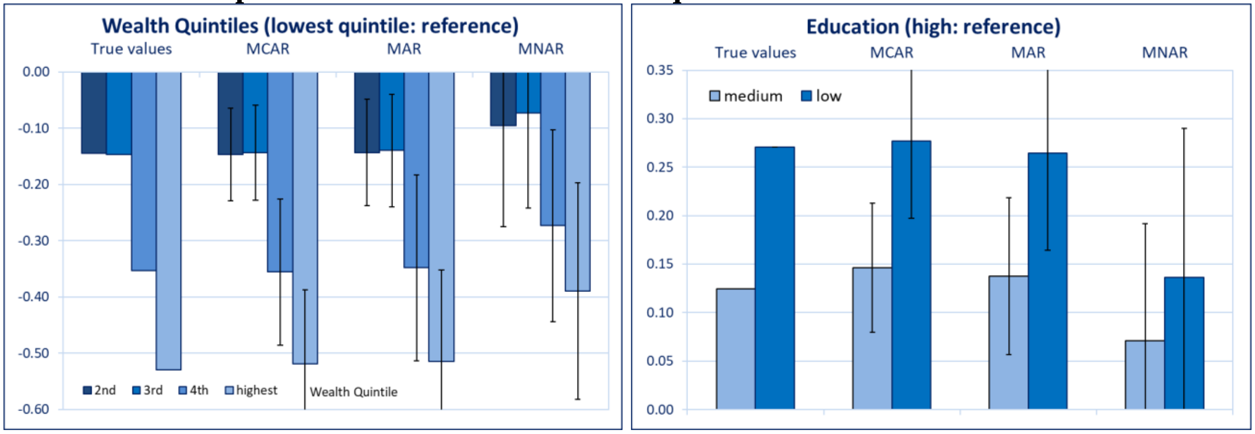


Note: Effects for each of the three missing patterns are averages over 200 simulations. All models control for age group, sex, general health and employment status at time of the interview.

In Figure 3, the results for the inverse propensity weighting compensating method shows very similar results to the simple complete case method shown in Figure 2. In general, we see the same pattern of effect sizes for the socio-economic indicators for all missing data patterns, with slight underestimation of effects for wealth quintiles, especially for the MNAR case which was expected. We also see overestimation of effect sizes for education, again mostly for MCAR and MAR and less for MNAR, which is unexpected. The conclusion would be, that inverse propensity weighting failed to compensate for all of the missing data bias and hardly showed any advantage over the complete case analysis. It appears that the impact on the effect sizes may be dependent on the socioeconomic indicator (similar results were observed for occupation as well). Again, we observe in Figure 3 that the 'true' effect sizes are included within the confidence intervals.

In Figure 4, we demonstrate the next compensation method based on including the inverse Mills Ratio in the substantive model, also called the 2-stage Heckman selection model, where in the first stage we use the probit model for estimating the Mills Ratio. Figure 4 shows the average effect sizes (after adjusting effect sizes to account for the first stage coefficients from the probit model) and confidence intervals across the 200 samples for the wealth quintiles and education level variables according to the three missing data mechanisms.

Figure 4: **Inverse Mills Ratio effect sizes for MCAR, MAR and MNAR missing data mechanisms compared to true values for wealth quintiles and education level**

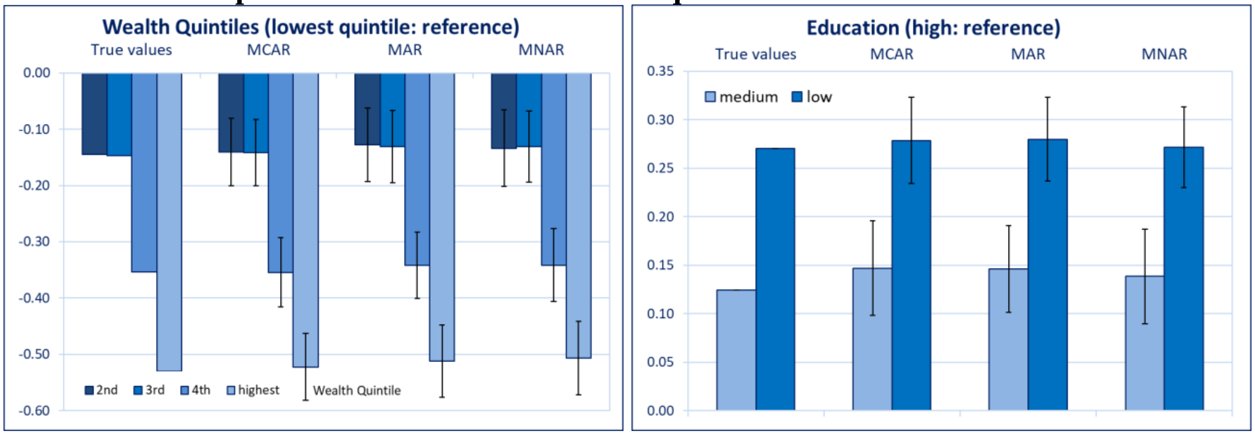


Note: Effects for each of the three missing patterns are averages over 200 simulations. All models control for age group, sex, general health and employment status at time of the interview.

In Figure 4, for the wealth quintiles and education level, we see similar results for MCAR and MAR missing patterns compared to the complete case analysis. The inverse Mills Ratio selection model was expected to perform best for the MNAR missing data mechanism; however, we see a large underestimation of the wealth quintile and education effects. For education, the effect size is almost halved from 0.12 (medium) and 0.27 (low) according to the 'true' effect size values to 0.07 and 0.14 respectively under the MNAR missing pattern using the inverse Mills Ratio. For the wealth variable the underestimation is less in magnitude but still about one third of the true effect size. On the other hand, we see very large confidence intervals under the inverse Mills Ratio approach and the 'true' effect sizes are included within the confidence intervals.

In Figure 5, we test the compensation method of multiple imputation according to the three missing data mechanisms and show the average effect sizes and confidence intervals across the 200 samples for the wealth quintiles and education level alongside the 'true' effect sizes.

Figure 5: Multiple imputation effect sizes for MCAR, MAR and MNAR missing data mechanisms compared to true values for wealth quintiles and education level

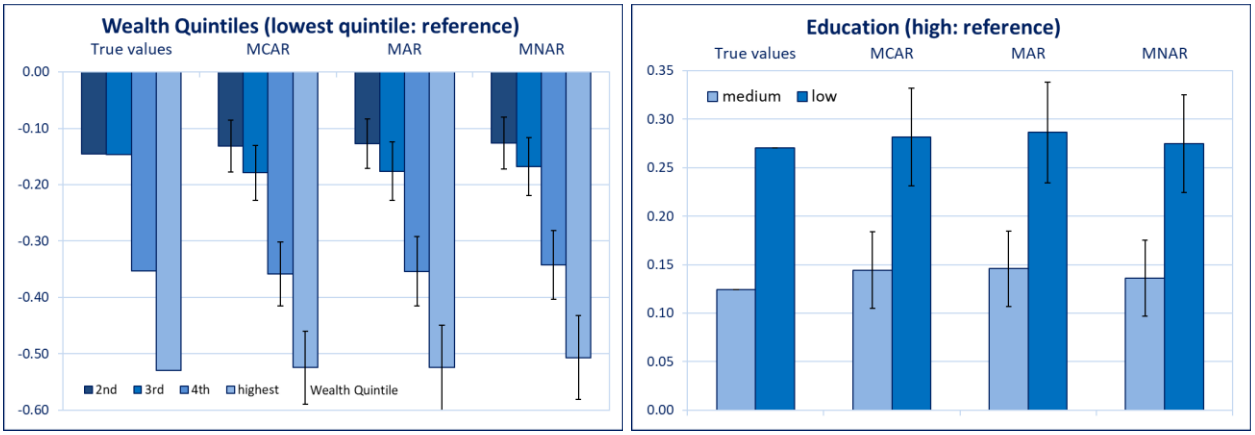


Note: Effects for each of the three missing patterns are averages over 200 simulations. All models control for age group, sex, general health and employment status at time of the interview.

In Figure 5, the results for multiple imputation as a compensation method for missing data looks promising. Among all tested compensation methods, multiple imputation produced results that are closest to the true values of the baseline model, across all missing data patterns. There is only minimal underestimation for wealth quintiles for the MAR and MNAR scenario. However, we do see similar overestimation of effects size for education as in the complete case analysis, which is highest for the MCAR and MAR scenario. We also observe as in the previous figures that the ‘true’ effect sizes are included within the confidence intervals.

For the last compensation method in this simulation study, we show in Figure 6 the average effect sizes and confidence intervals across the 200 samples for the wealth quintiles and education level under the new approach of including the inverse Mills Ratio in the multiple imputation method of chained equations (MICE) according to the three missing data patterns.

Figure 6: Combined inverse Mills Ratio and multiple imputation effect sizes for MCAR, MAR and MNAR missing data mechanisms compared to true values for wealth quintiles and education level



Note: Effects for each of the three missing patterns are averages over 200 simulations. All models control for age group, sex, general health and employment status at time of the interview.

In Figure 6, we see that the proposed approach of combining the inverse Mills Ratio with the multiple imputation technique had relatively close results to the baseline model for all missing data patterns and in particular performed well for the MNAR scenario for the education variable. The results are comparable with the results from the multiple imputation compensation method in Figure 5. However, for the wealth quintiles, particularly the 3rd wealth quintile, we see larger deviation from the 'true' effect size compared to the multiple imputation in Figure 5. The confidence intervals all include the 'true' effect sizes.

As an overall observation from Figures 2 to 6, we observe that the structure of the effect sizes of education level and wealth quintiles on the level of log(CRP) are rather constant and similar for all compensation approaches across the 200 simulations. For instance, the education and wealth gradient (lower education and lower wealth showed highest CRP values) on the inflammation marker, as demonstrated in the baseline model in Figure 1, remained stable, independent from the missing data mechanism or the compensation approach, which is reassuring. The magnitude of the effect sizes and therefore the closeness to the 'truth' on the other hand differed, sometimes to a large extent from the baseline model shown in Figure 1. The large deviations occurred under the inverse Mills Ratio compensation method where we expected to see improvements under the MNAR missing data pattern. However, the effect sizes were underestimated for both socioeconomic indicators. We also found that all 'true' effect sizes were included in the simulation confidence intervals under all compensation methods and that the confidence intervals were relatively large, particularly for the case of the inverse Mills Ratio.

5. Conclusions and Discussion

In this simulation study we tested the impact of the complete case analysis and four different missing data compensation methods on a simulation dataset created from wave 2 of the ELSA data. We also used the same missing data proportions as those observed in the ELSA data to create three missing data mechanisms of MCAR, MAR and MNAR. This was done across 200 iterations for each combination of missing data mechanism and compensation method. The impact of the compensation methods were tested by comparing their resulting effect sizes on a substantive model measuring the effect of socioeconomic indicators (education, wealth quintile and occupation (not shown here)) on the level of log(CRP) to the 'true' effect sizes obtained from the simulation dataset.

The results suggest that a nonresponse bias, due to missing information in biomarker information, does exist and that without acknowledging their existence in the analysis stage, effect size under- or overestimation cannot be avoided. Reassuringly, the pattern of effect sizes did not change drastically for all socioeconomic indicators: e.g. lowest wealth quintile always showed highest log(CRP) value and highest wealth quintile always showed the lowest log(CRP) value, as expected. The magnitude of effect sizes, however, varied depending on the combination of missing data mechanism and compensation method.

The bias in the MCAR scenario was negligible independent of the compensation method used. The complete case analyses performed well as expected except for the slightly higher effect size for the medium education level (see Figure 2). None of the compensation methods corrected this overestimation in the effect size. We also see in Figure 6 that using the approach of the inverse Mills ratio combined with the multiple imputation made the estimation of effect sizes for the wealth quintiles worse than the complete case analysis.

The more realistic scenario of the MAR missing data mechanism depending on other covariates showed more divergence from the true values than MCAR. The bias in the effect sizes was mainly consistent across compensation methods and socioeconomic indicators. For the complete case analysis there is underestimation of effect sizes for the high wealth quintiles and overestimation of the medium and low education levels. There was a slight improvement in the education and wealth quintile variables under the inverse Mills Ratio but the other compensation methods did not show any clearer improvement of the bias.

The largest nonresponse bias was observed for the MNAR scenario, as expected. The complete case analysis underestimated wealth quintiles although there seemed to be little impact on the effect from education level. A similar result was found under the inverse propensity weighting, which is more commonly used in the ELSA study for the biomarker sample (Gale et al. 2013; Jackowska et al. 2013) Whereas we expected the inverse Mills Ratio to perform well for the case of the MNAR mechanism, it performed relatively poor with large under estimation of the effect sizes of the socioeconomic indicators albeit with large confidence intervals that covered the 'true' effect sizes. Further investigation is needed to understand the reason for this outcome, and in particular, the adjustment approach to the effect sizes under the 2-stage Heckman selection model. The multiple imputation method provided a slight under estimation of the effect sizes for high wealth quintiles and over estimation of the medium education level as can also be seen under the proposed approach with the combined inverse Mills Ratio and multiple imputation.

As the specific combination of missing data pattern and compensation mechanisms seems to be important, it is crucial to know the nature of the missing data in a dataset before deciding on a compensation method to use. However, it is often impossible outside simulation studies, to determine with certainty if data is MCAR, MAR or MNAR. We found that multiple imputation performed relatively well across all three missing mechanisms, which makes it a safe choice to use. It provided the best compensation method to mitigate nonresponse bias for analysis with missing data in biomarkers. However, the fact that it performed well for the case of NMAR may not be so surprising given the way the missing data for the MNAR mechanism was generated. In the first step, the MNAR missing data was generated within the 16 strata similar to the MAR mechanism and in the second step, differential missingness was created for the CRP value according to high/low CRP. Nevertheless, using compensation methods that are typically used under the MAR mechanism also helped for the case of the MNAR mechanism in this simulation study. Our proposed approach combining the inverse Mills Ratio and the multiple imputation method performed almost as well as multiple imputation on its own. However, given the additional analytical steps and necessary computing power, it does not show any major advantage over multiple imputation, an established statistical approach for compensating for nonresponse bias, which has been integrated in many statistical software packages already.

We note that in this simulation study, we conducted 200 iterations for each missing data mechanism and compensation method and we obtained large simulation confidence intervals, especially for the inverse Mills Ratio method. Therefore, future work will be to expand the simulation study to a larger scale with more iterations in order to narrow the confidence intervals and achieve more accurate results. In addition, other future work will be to study the impact of missing data over time in a longitudinal study, such as ELSA, using a substantive analysis of a growth curve model for CRP.

Acknowledgement: This work was supported by the United Kingdom Economic and Social Research Council, National Centre for Research Methods [ES/L008351/1]

References:

1.Fraga, S., Marques-Vidal, P., Vollenweider, P., Waeber, G., Guessous, I., Paccaud, F., et al. (2015). Association of socioeconomic status with inflammatory markers: A two cohort comparison. *Preventive Medicine*. 2015; 71: 12–19.

2.Loucks, E. B., Pilote, L., Lynch, J. W., Richard, H., Almeida, N. D., Benjamin, E. J., & Murabito, J. M. Life course socioeconomic position is associated with inflammatory markers: The Framingham Offspring Study. *Social Science and Medicine*. 2010; *71*(1): 187–195.

3.Maharani, A. Socio-economic inequalities in C-reactive protein levels: Evidence from longitudinal studies in England and Indonesia. *Brain, Behavior, and Immunity*. 2019; *82*(July): 122–128.

4.Muennig, P., Sohler, N., & Mahato, B. Socioeconomic status as an independent predictor of physiological biomarkers of cardiovascular disease: Evidence from NHANES. *Preventive Medicine*. 2007; *45*(1): 35–40.

5.Pollitt, R. A., Kaufman, J. S., Rose, K. M., Diez-Roux, A. V, Zeng, D., & Heiss, G. (2007). Early-life and adult socioeconomic status and inflammatory risk markers in adulthood. *European Journal of Epidemiology*. 2007; *22*(1): 55–66.

6.Stringhini, S., da Batty, G. D., Bovet, P., Shipley, M. J., Marmot, M. G., Kumari, M., et al. Association of Lifecourse Socioeconomic Status with Chronic Inflammation and Type 2 Diabetes Risk: The Whitehall II Prospective Cohort Study. *PLoS Medicine*. 2013; *10*(7): 1–15.

7.Rubin, D. B. Inference and missing data. *Biometrika*. 1976; *63*(3): 581–592.

8.Scholes, S., Taylor, R., Cheshire, H., Cox, K., & Lessof, C. Retirement, health and relationships of the older population in England: The 2004 English Longitudinal Study of Ageing Technical Report. *London, UK National Centre for Social Research*, (November). 2008.

9.Heckman, J. J. The common structure of statistical models of trancation, sample selection and mimited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*. 1976; *5*(4): 475–492.

10.Little, R. J. A. Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*. 1982; 77(378), 237–250.

11.Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons Inc.; 1987.

12.Raghunathan, T., Lepkowski, J., Van Hoewyk, J., & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology. 2001; 27*(1): 85–96.

13. van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007; *16*(3): 219–242.

14. Gale, C. R., Baylis, D., Cooper, C., & Sayer, A. A. Inflammatory markers and incident frailty in men and women: The English Longitudinal Study of Ageing. AGE. 2013; 35(6): 2493–2501.

15. Jackowska, M., Kumari, M., & Steptoe, A. (2013). Sleep and biomarkers in the English Longitudinal Study of Ageing: Associations with C-reactive protein, fibrinogen, dehydroepiandrosterone sulfate and hemoglobin. Psychoneuroendocrinology. 2013; 38(9): 1484–1493.

Appendix:

Percentage Missing Data Within Strata ELSA Wave 2

Strata	Age group Sex Health Wealth	Sample Size	Percent Attrition	Percent Missing Nurse Visit	Percent Missing Blood Collection	Percent Missing CRP
1	0000	770	18.7%	10.3%	8.5%	9.1%
2	0001	1002	14.5%	8.2%	8.8%	10.3%
3	0010	180	5.2%	19.7%	16.0%	15.0%
4	0011	96	7.8%	13.3%	16.7%	16.7%
5	0100	911	14.8%	14.7%	12.9%	10.1%
6	0101	1038	13.4%	9.6%	9.9%	9.6%
7	0110	238	7.0%	17.0%	21.9%	13.6%
8	0111	150	7.8%	12.7%	22.2%	16.0%
9	1000	941	12.9%	11.8%	7.4%	10.0%
10	1001	1056	19.4%	7.4%	6.6%	16.4%
11	1010	329	17.4%	12.1%	13.0%	18.6%
12	1011	160	9.5%	15.1%	17.8%	8.3%
13	1100	1297	16.1%	14.1%	14.0%	14.3%
14	1101	1236	3.8%	10.3%	10.5%	12.3%
15	1110	446	3.5%	18.0%	25.0%	14.7%
16	1111	150	3.9%	13.6%	18.8%	12.3%