Inverse probability weighting to handle attrition in cohort studies: some guidance and a call for caution

BMC Medical Research Methodology (2022)

Pokket Sirisreetreerux

Metten *et al.*
BMC Medical Research Methodology (2022) 22:45
https://doi.org/10.1186/s12874-022-01533-9

BMC Medical Research Methodology

RESEARCH Open Access

Inverse probability weighting to handle attrition in cohort studies: some guidance and a call for caution

Marie-Astrid Metten[1], Nathalie Costet[2], Luc Multigner[2], Jean-François Viel[1] and Guillaume Chauvet[3*]

Introduction

- Cohort studies are essential for investigating associations between exposure and health outcomes.
- The repeated collection of information in successive follow-ups(also called survey waves) allows studying the effects of past exposures on health outcomes occurring at inclusion or thereafter.
- However, such studies are known to be affected by partial and total non-response, which can invalidate the causal inference that can be drawn from them.

Introduction

- Partial non-response
  - refers to missing data that occasionally occurs for certain variables during a survey wave when some individuals fail or refuse to answer some of the questions.
- Total nonresponse (or attrition)
  - occurs when a subset of individuals does not participate in one specific survey wave or quit the study completely

Introduction

- Missing data resulting from non-response can be classified according to their postulated underlying mechanism

  - *missing completely at random* (MCAR), the probability of missing data does not depend on either the observed or unobserved values.
  - *missing at random* (MAR), it depends on the observed data but not the unobserved data.
  - *missing not at random* (MNAR), it depends on the unobserved data.

Introduction

- The simplest and most widely used approach to handle total non-response in cohort studies is complete-case analysis (CCA).

- Several methodological publications have suggested the use of the inverse probability weighting (IPW) method in situations of the MAR mechanism of attrition.
  - recreate a representative sample of the initial cohort by differentially weighting the so-called “complete individuals”.
  - The use of the inverse of this probability implies that a respondent with a high probability of response is given a comparatively lower weight in the analysis.
  - The approach can be summarized as: “the respondents carry the weight of the non-respondents”.

Introduction

- First step - build a response model (logistic regression model) to obtain weights
- Second step – build association model that will use weights from the first step, the method is also called “weighted complete-case analysis” or “inverse probability of *participation/attrition* weighting” (IPPW/IPAW) in the literature

Introduction

- IPW
  - was developed for reducing the effects of confounding in observational studies (propensity score method)
  - was extended to correct for selection biases in situations of attrition.

Aims of the study

- In this article, the author will focus on attrition resulting in a missing outcome of interest.
- This study aimed
  - evaluating through simulations the ability of the IPW method to correct for a selection bias under various missingness mechanisms and specifications of the response model.
  - Response model specifications were compared in terms of bias, variance and mean square error of the association estimates between the xposure and the outcome.

Which variables should be introduced into the response model?

- IPPW method considers the exposure and the outcome, and also the response as a third variable.
- Relatively few authors have addressed the question of which variables should be introduced in the response model from which the weighting is derived.

Which variables should be introduced into the response model?

- None of the proposed strategies in the literature has been tested through simulations and they do not appear to be applied by researchers.
- Therefore, the authors propose here to provide insight on this issue by studying the impact of the type of variables included in the response model on the bias and variance of the exposure regression coefficient in the association model.

Simulation study

- They conducted a Monte-Carlo simulation study under several MAR and MNAR scenarios.
- They aimed to evaluate

i) the relative performance of the IPPW method relative to CCA

ii) how the specification of the response model in the IPPW method affects the bias of the exposure regression coefficient $\hat{\beta}$, its variance and mean square error, and the coverage rate of confidence intervals.

Data-generating process

1. created a sample of size *n* = 1,000, containing seven covariates *z*1, …, *z*7 generated independently according to standard normal distributions.
2. then generated an exposure variable according to the following model:

$$x_i = 1 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \alpha_5 z_{5i} + \alpha_6 z_{6i} + \epsilon_i,$$

Exposure model

Data-generating process

3. Generated an outcome variable according to the following model:

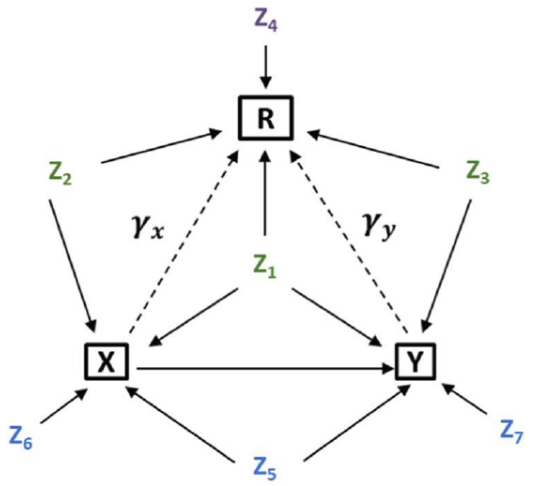
$$y_i = 1 + \beta x_i + \beta_1 z_{1i} + \beta_3 z_{3i} + \beta_5 z_{5i} + \beta_7 z_{7i} + \epsilon_i',$$

Outcome model

4. Generated response probabilities according to the following logistic model:

$$\text{logit}(p_i) = \gamma_0 + \gamma_y y_i + \gamma_x x_i + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i},$$

Data-generating process



R: response variable

Y: outcome variable; **X**: exposure variable; $\mathbf{Z_i}$: covariates

⟶ : correlations

Types of covariates

$\mathbf{Z_1}$: confounding variable associated with the response

$\mathbf{Z_2}$: variable associated with the exposure and the response

$\mathbf{Z_3}$: prognostic variable associated with the response

$\mathbf{Z_4}$: variable only associated with the response

$\mathbf{Z_5}$: confounding variable not associated with the response

$\mathbf{Z_6}$: variable only associated with the exposure

$\mathbf{Z_7}$: prognostic variable not associated with the response

- - - - ▸ Depending on the attrition scenario (MAR, MNAR), the exposure and outcome variables were associated or not with the response variable through γ_x and γ_y

Data-generating process

- The nine response mechanism scenarios are summarized in Table 1.

Table 1 Response mechanism scenarios (data generation)

Scenario	γ_x	γ_y	$\gamma_1, \gamma_2, \gamma_3, \gamma_4$	Description
MAR 1	0.0	0.0	0.1	Response depending only on covariates
MAR 2	0.2	0.0	0.1	Response depending on covariates and exposure
MAR 3	0.5	0.0	0.1	Response depending on covariates and exposure
MNAR 1	0.0	0.2	0.1	Response depending on outcome and covariates
MNAR 2	0.2	0.2	0.1	Response depending on outcome, exposure, and covariates
MNAR 3	0.5	0.2	0.1	Response depending on outcome, exposure, and covariates
MNAR 4	0.0	0.5	0.1	Response depending on outcome and covariates
MNAR 5	0.2	0.5	0.1	Response depending on outcome, exposure, and covariates
MNAR 6	0.5	0.5	0.1	Response depending on outcome, exposure, and covariates

γ_k:regression coefficients of the generated response models ($\text{logit}(p_i) = \gamma_0 + \gamma_y y_i + \gamma_x x_i + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i}$)

Simulation parameters and performance criteria

- Compared the IPPW method to CCA for a parsimonious association model.
- Several response models were tested (see Table 2) to determine the impact of the type of variables included on the β^ regression coefficient of the exposure variable and its variance in the association model.

Table 2 Response models tested

Response model	Set of variables	Description
1	X, Z_1, Z_2, Z_3, Z_4	All variables associated with the response[a]
2	$X, Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7$	The exposure variable X and all covariates
3	X, Z_1, Z_3	The exposure variable X and variables associated with both response and outcome (strategy proposed by Hernan et al. [3])
4	$X, Z_1, Z_2, Z_3, Z_5, Z_7$	All variables associated with the response*, except Z_4 only associated with the response; Adding Z_5 a confounding variable (Z_5) and a prognostic variable (Z_7), neither associated with the response (strategy proposed by Seaman and White [4])
5	X, Z_1	The exposure variable X and the confounding variable associated with the response (Z_1)
6	X, Z_5	The exposure variable X and the confounding variable not associated with the response (Z_5)
7	X, Z_1, Z_5	The exposure variable X and both confounding variables, that associated with the response (Z_1) the other not (Z_5)
8	X, Z_1, Z_5, Z_7	The exposure variable X, both confounding variables (Z_1, Z_5) and a prognostic variable not associated with the response (Z_7)
9	X, Z_5, Z_7	The exposure variable X and a confounding variable and prognostic variable, neither associated with the response (Z_5, Z_7)
10	Z_1, Z_2, Z_3, Z_4	Previous response models without the exposure variable X
11	$Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7$	
12	Z_1, Z_3	
13	Z_1, Z_2, Z_3, Z_5, Z_7	
14	Z_1	
15	Z_5	
16	Z_1, Z_5	
17	Z_1, Z_5, Z_7	
18	Z_5, Z_7	

[a] X was not associated with the response in scenarios MAR 1, MNAR 1, or MNAR 4

Simulation parameters and performance criteria

- Finally, the authors evaluated parsimonious strategies:
  - including only the confounding variable associated with the response,
  - including only the confounding variable not associated with the response,
  - including both
  - including both with the addition of a prognostic variable not associated with the response, and finally,
  - including both confounding and prognostic variables not associated with the response. All
- These response models were then evaluated without the exposure variable X.

Simulation parameters and performance criteria

- The generation of the sample and variables was repeated B = 10,000 times.
- The results were assessed according to the following criteria:
  - The Monte Carlo bias
  - The Monte Carlo variance
  - The mean square error
  - The relative root mean square error
- The author have also computed the coverage rates for the normality-based confidence intervals for $\hat{\beta}_{x'}$ with nominal rates of 2.5% in each tail.

1.Bias in the $\hat{\beta}$ regression coefficient

- No bias with either CCA or the IPPW method for the three MAR scenarios and MNAR scenario 1.
- A bias occurred with both methods for the five other MNAR scenarios, with a greater amplitude for MNAR scenarios 5 (***γx*** = **0.2**, ***γy*** = **0.5**) and 6 (***γx*** = **0.5**, ***γy*** = **0.5**).

Table 3 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MAR 1	0.0	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.670	X	1.686	1.691	1.679	1.683	1.676	1.676	1.678	1.679	1.677
					–	1.684	1.687	1.676	1.680	1.672	1.671	1.674	1.675	1.672
			MSE (10^{-3})	1.670	X	1.686	1.691	1.679	1.683	1.676	1.676	1.678	1.679	1.677
					–	1.684	1.687	1.676	1.680	1.672	1.671	1.674	1.675	1.672
			RRMSE (%)	16.3	X	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4
					–	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4
MAR 2	0.2	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.668	X	1.715	1.720	1.709	1.711	1.704	1.705	1.706	1.707	1.705
					–	1.684	1.689	1.678	1.682	1.673	1.668	1.674	1.676	1.670
			MSE (10^{-3})	1.669	X	1.715	1.720	1.709	1.711	1.705	1.706	1.707	1.708	1.706
					–	1.685	1.689	1.679	1.682	1.674	1.670	1.675	1.677	1.671
			RRMSE (%)	16.3	X	16.6	16.6	16.5	16.5	16.5	16.5	16.5	16.5	16.5
					–	16.4	16.4	16.4	16.4	16.4	16.3	16.4	16.4	16.4
MAR 3	0.5	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.761	X	1.992	2.002	1.981	1.985	1.973	1.977	1.978	1.977	1.976
					–	1.787	1.798	1.777	1.787	1.773	1.763	1.776	1.778	1.765
			MSE (10^{-3})	1.768	X	1.992	2.002	1.981	1.985	1.981	1.984	1.986	1.985	1.983
					–	1.793	1.804	1.784	1.793	1.781	1.771	1.784	1.786	1.773
			RRMSE (%)	16.8	X	17.9	17.9	17.8	17.8	17.8	17.8	17.8	17.8	17.8
					–	16.9	17.0	16.9	16.9	16.9	16.8	16.9	16.9	16.8

Table 4 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MNAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MNAR 1	0.0	0.2	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.676	X	1.701	1.707	1.694	1.698	1.688	1.685	1.689	1.690	1.686
					–	1.697	1.703	1.689	1.695	1.682	1.677	1.684	1.685	1.679
			MSE (10^{-3})	1.688	X	1.711	1.715	1.703	1.706	1.700	1.698	1.702	1.702	1.697
					–	1.709	1.714	1.702	1.706	1.695	1.690	1.697	1.698	1.691
			RRMSE (%)	16.4	X	16.5	16.6	16.5	16.5	16.5	16.5	16.5	16.5	16.5
					–	16.5	16.6	16.5	16.5	16.5	16.4	16.5	16.5	16.5
MNAR 2	0.2	0.2	Bias	−0.01	X[b]	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
					–	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
			Variance (10^{-3})	1.668	X	1.746	1.751	1.737	1.740	1.728	1.727	1.730	1.730	1.728
					–	1.695	1.703	1.686	1.694	1.678	1.671	1.681	1.682	1.672
			MSE (10^{-3})	1.881	X	1.907	1.897	1.896	1.885	1.942	1.940	1.943	1.925	1.920
					–	1.904	1.907	1.900	1.900	1.892	1.884	1.895	1.896	1.885
			RRMSE (%)	17.3	X	17.5	17.4	17.4	17.4	17.6	17.6	17.6	17.5	17.5
					–	17.5	17.5	17.4	17.4	17.4	17.4	17.4	17.4	17.4
MNAR 3	0.5	0.2	Bias	−0.03	X[b]	−0.03	−0.02	−0.03	−0.02	−0.03	−0.03	−0.03	−0.03	−0.03
					–	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03
			Variance (10^{-3})	1.756	X	2.044	2.050	2.028	2.032	2.013	2.013	2.014	2.013	2.013
					–	1.793	1.806	1.780	1.794	1.772	1.760	1.777	1.778	1.761
			MSE (10^{-3})	2.572	X	2.708	2.650	2.685	2.630	2.885	2.882	2.885	2.811	2.806
					–	2.611	2.613	2.605	2.607	2.596	2.578	2.604	2.604	2.579
			RRMSE (%)	20.3	X	20.8	20.6	20.7	20.5	21.5	21.5	21.5	21.2	21.2
					–	20.4	20.4	20.4	20.4	20.4	20.3	20.4	20.4	20.3

Table 5 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MNAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MNAR 4	0.0	0.5	Bias	−0.02	X[b]	−0.02	−0.01	−0.01	−0.01	−0.02	−0.02	−0.02	−0.02	−0.02
					–	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02
			Variance (10^{-3})	1.591	X	1.633	1.641	1.625	1.633	1.616	1.611	1.618	1.621	1.614
					–	1.620	1.630	1.613	1.623	1.603	1.594	1.607	1.610	1.597
			MSE (10^{-3})	1.857	X	1.859	1.849	1.849	1.840	1.885	1.878	1.887	1.866	1.852
					–	1.883	1.888	1.883	1.882	1.872	1.861	1.876	1.880	1.864
			RRMSE (%)	17.2	X	17.2	17.2	17.2	17.2	17.4	17.3	17.4	17.3	17.2
					–	17.4	17.4	17.4	17.4	17.3	17.3	17.3	17.3	17.3
MNAR 5	0.2	0.5	Bias	−0.04	X[b]	−0.04	−0.03	−0.04	−0.03	−0.04	−0.04	−0.04	−0.04	−0.04
					–	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04
			Variance (10^{-3})	1.621	X	1.737	1.743	1.727	1.731	1.712	1.708	1.711	1.712	1.709
					–	1.656	1.667	1.648	1.659	1.636	1.624	1.641	1.643	1.626
			MSE (10^{-3})	3.094	X	3.038	2.933	3.013	2.914	3.230	3.219	3.231	3.103	3.081
					–	3.142	3.137	3.144	3.138	3.125	3.104	3.136	3.140	3.108
			RRMSE (%)	22.3	X	22.0	21.7	22.0	21.6	22.7	22.7	22.7	22.3	22.2
					–	22.4	22.4	22.4	22.4	22.4	22.3	22.4	22.4	22.3
MNAR 6	0.5	0.5	Bias	−0.07	X[b]	−0.06	−0.06	−0.06	−0.06	−0.07	−0.07	−0.07	−0.07	−0.07
					–	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07
			Variance (10^{-3})	1.734	X	2.136	2.150	2.118	2.129	2.082	2.079	2.078	2.082	2.084
					–	1.786	1.810	1.770	1.798	1.755	1.742	1.764	1.767	1.744
			MSE (10^{-3})	6.211	X	6.284	5.938	6.229	5.896	6.903	6.887	6.901	6.503	6.473
					–	6.350	6.360	6.341	6.368	6.295	6.251	6.336	6.346	6.262
			RRMSE (%)	31.5	X	31.7	30.8	31.6	30.7	33.2	33.2	33.2	32.3	32.2
					–	31.9	31.9	31.9	31.9	31.7	31.6	31.8	31.9	31.7

Variance of the $\hat{\beta}$ regression coefficient

- The IPPW method was less efficient than CCA for all scenarios.
- The loss of efficiency of the IPPW method was thus particularly pronounced in MAR scenario 3 and MNAR scenarios 3 and 6 (all three characterized by $\gamma_x = \mathbf{0.5}$).

Table 3 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MAR 1	0.0	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.670	X	1.686	1.691	1.679	1.683	1.676	1.676	1.678	1.679	1.677
					–	1.684	1.687	1.676	1.680	1.672	1.671	1.674	1.675	1.672
			MSE (10^{-3})	1.670	X	1.686	1.691	1.679	1.683	1.676	1.676	1.678	1.679	1.677
					–	1.684	1.687	1.676	1.680	1.672	1.671	1.674	1.675	1.672
			RRMSE (%)	16.3	X	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4
					–	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4	16.4
MAR 2	0.2	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.668	X	1.715	1.720	1.709	1.711	1.704	1.705	1.706	1.707	1.705
					–	1.684	1.689	1.678	1.682	1.673	1.668	1.674	1.676	1.670
			MSE (10^{-3})	1.669	X	1.715	1.720	1.709	1.711	1.705	1.706	1.707	1.708	1.706
					–	1.685	1.689	1.679	1.682	1.674	1.670	1.675	1.677	1.671
			RRMSE (%)	16.3	X	16.6	16.6	16.5	16.5	16.5	16.5	16.5	16.5	16.5
					–	16.4	16.4	16.4	16.4	16.4	16.3	16.4	16.4	16.4
MAR 3	0.5	0.0	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.761	X	1.992	2.002	1.981	1.985	1.973	1.977	1.978	1.977	1.976
					–	1.787	1.798	1.777	1.787	1.773	1.763	1.776	1.778	1.765
			MSE (10^{-3})	1.768	X	1.992	2.002	1.981	1.985	1.981	1.984	1.986	1.985	1.983
					–	1.793	1.804	1.784	1.793	1.781	1.771	1.784	1.786	1.773
			RRMSE (%)	16.8	X	17.9	17.9	17.8	17.8	17.8	17.8	17.8	17.8	17.8
					–	16.9	17.0	16.9	16.9	16.9	16.8	16.9	16.9	16.8

Table 4 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MNAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MNAR 1	0.0	0.2	Bias	0.00	X[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			Variance (10^{-3})	1.676	X	1.701	1.707	1.694	1.698	1.688	1.685	1.689	1.690	1.686
					–	1.697	1.703	1.689	1.695	1.682	1.677	1.684	1.685	1.679
			MSE (10^{-3})	1.688	X	1.711	1.715	1.703	1.706	1.700	1.698	1.702	1.702	1.697
					–	1.709	1.714	1.702	1.706	1.695	1.690	1.697	1.698	1.691
			RRMSE (%)	16.4	X	16.5	16.6	16.5	16.5	16.5	16.5	16.5	16.5	16.5
					–	16.5	16.6	16.5	16.5	16.5	16.4	16.5	16.5	16.5
MNAR 2	0.2	0.2	Bias	−0.01	X[b]	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
					–	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
			Variance (10^{-3})	1.668	X	1.746	1.751	1.737	1.740	1.728	1.727	1.730	1.730	1.728
					–	1.695	1.703	1.686	1.694	1.678	1.671	1.681	1.682	1.672
			MSE (10^{-3})	1.881	X	1.907	1.897	1.896	1.885	1.942	1.940	1.943	1.925	1.920
					–	1.904	1.907	1.900	1.900	1.892	1.884	1.895	1.896	1.885
			RRMSE (%)	17.3	X	17.5	17.4	17.4	17.4	17.6	17.6	17.6	17.5	17.5
					–	17.5	17.5	17.4	17.4	17.4	17.4	17.4	17.4	17.4
MNAR 3	0.5	0.2	Bias	−0.03	X[b]	−0.03	−0.02	−0.03	−0.02	−0.03	−0.03	−0.03	−0.03	−0.03
					–	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03	−0.03
			Variance (10^{-3})	1.756	X	2.044	2.050	2.028	2.032	2.013	2.013	2.014	2.013	2.013
					–	1.793	1.806	1.780	1.794	1.772	1.760	1.777	1.778	1.761
			MSE (10^{-3})	2.572	X	2.708	2.650	2.685	2.630	2.885	2.882	2.885	2.811	2.806
					–	2.611	2.613	2.605	2.607	2.596	2.578	2.604	2.604	2.579
			RRMSE (%)	20.3	X	20.8	20.6	20.7	20.5	21.5	21.5	21.5	21.2	21.2
					–	20.4	20.4	20.4	20.4	20.4	20.3	20.4	20.4	20.3

Table 5 Simulation study results: bias, variance, mean square error and related root mean square error in the $\hat{\beta}$ regression coefficient for CCA and the IPPW method (18 response models), for three MNAR response mechanism scenarios

Scenario[a]	γ_x	γ_y		CCA		IPPW method								
						Response models								
						(X), Z_1, Z_2, Z_3, Z_4	(X), Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7	(X), Z_1, Z_3	(X), Z_1, Z_2, Z_3, Z_5, Z_7	(X), Z_1	(X), Z_5	(X), Z_1, Z_5	(X), Z_1, Z_5, Z_7	(X), Z_5, Z_7
MNAR 4	0.0	0.5	Bias	−0.02	X[b]	−0.02	−0.01	−0.01	−0.01	−0.02	−0.02	−0.02	−0.02	−0.02
					–	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02
			Variance (10^{-3})	1.591	X	1.633	1.641	1.625	1.633	1.616	1.611	1.618	1.621	1.614
					–	1.620	1.630	1.613	1.623	1.603	1.594	1.607	1.610	1.597
			MSE (10^{-3})	1.857	X	1.859	1.849	1.849	1.840	1.885	1.878	1.887	1.866	1.852
					–	1.883	1.888	1.883	1.882	1.872	1.861	1.876	1.880	1.864
			RRMSE (%)	17.2	X	17.2	17.2	17.2	17.2	17.4	17.3	17.4	17.3	17.2
					–	17.4	17.4	17.4	17.4	17.3	17.3	17.3	17.3	17.3
MNAR 5	0.2	0.5	Bias	−0.04	X[b]	−0.04	−0.03	−0.04	−0.03	−0.04	−0.04	−0.04	−0.04	−0.04
					–	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04	−0.04
			Variance (10^{-3})	1.621	X	1.737	1.743	1.727	1.731	1.712	1.708	1.711	1.712	1.709
					–	1.656	1.667	1.648	1.659	1.636	1.624	1.641	1.643	1.626
			MSE (10^{-3})	3.094	X	3.038	2.933	3.013	2.914	3.230	3.219	3.231	3.103	3.081
					–	3.142	3.137	3.144	3.138	3.125	3.104	3.136	3.140	3.108
			RRMSE (%)	22.3	X	22.0	21.7	22.0	21.6	22.7	22.7	22.7	22.3	22.2
					–	22.4	22.4	22.4	22.4	22.4	22.3	22.4	22.4	22.3
MNAR 6	0.5	0.5	Bias	−0.07	X[b]	−0.06	−0.06	−0.06	−0.06	−0.07	−0.07	−0.07	−0.07	−0.07
					–	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07	−0.07
			Variance (10^{-3})	1.734	X	2.136	2.150	2.118	2.129	2.082	2.079	2.078	2.082	2.084
					–	1.786	1.810	1.770	1.798	1.755	1.742	1.764	1.767	1.744
			MSE (10^{-3})	6.211	X	6.284	5.938	6.229	5.896	6.903	6.887	6.901	6.503	6.473
					–	6.350	6.360	6.341	6.368	6.295	6.251	6.336	6.346	6.262
			RRMSE (%)	31.5	X	31.7	30.8	31.6	30.7	33.2	33.2	33.2	32.3	32.2
					–	31.9	31.9	31.9	31.9	31.7	31.6	31.8	31.9	31.7

Mean square error of the $\hat{\beta}$ regression coefficient

- For the MAR scenarios, all the tested estimators are unbiased and there is therefore no difference between the variance and the mean square error.
- For MNAR scenarios 1 to 3, the mean square error increases with γ_x, i.e. when the correlation between the exposure variable and the response increases.
- This also holds true for MNAR scenarios 4 to 6.

Illustrative example

- The author analyzed the association between prepregnancy maternal BMI with the child's BMI at age 7 in TIMOUN, a prospective mother-child cohort study conducted in the Guadeloupe archipelago (French West Indies)

Illustrative example

- Between November 2004 and December 2007, 1068 pregnant women were enrolled in TIMOUN.
- At inclusion, women were interviewed by trained midwives
- In total, 1033 single live births were registered.
- When the children were 7 years of age, among the 1033 mother-child couples initially included, 592 participated in this second wave, representing 57% of the initial sample.
- Final population of 590 for the association studied (exclude 2 children because weight was not measured)

Outcome and exposure

- The exposure of interest was the pre-pregnancy maternal BMI (kg/m^2).
  - calculated from the mothers' self reported weight and height before pregnancy at inclusion
- The outcome of interest was the child's BMI at 7 years.
  - calculated from measurements performed during a medical examination at 7 years.

Covariates

The covariates considered in the analysis were

- Maternal age at birth (continuous)
- maternal educational level (< 5 years, 5–12 years, > 12 years)
- maternal place of birth (French West Indies, other Caribbean island, Europe)
- non-gestational maternal diabetes (yes, no)
- Enrollment site (university hospital, local hospital, antenatal care dispensary)
- maternal alcohol consumption during pregnancy (yes, no)
- maternal smoking during pregnancy (yes, no)
- sex of the child (boy, girl)

Covariates

- The proportion of missing data within these covariates did not exceed 3%, except for maternal alcohol consumption during pregnancy (5.6%).
- For the variables with missing values, a single imputation by the modal value was previously performed.

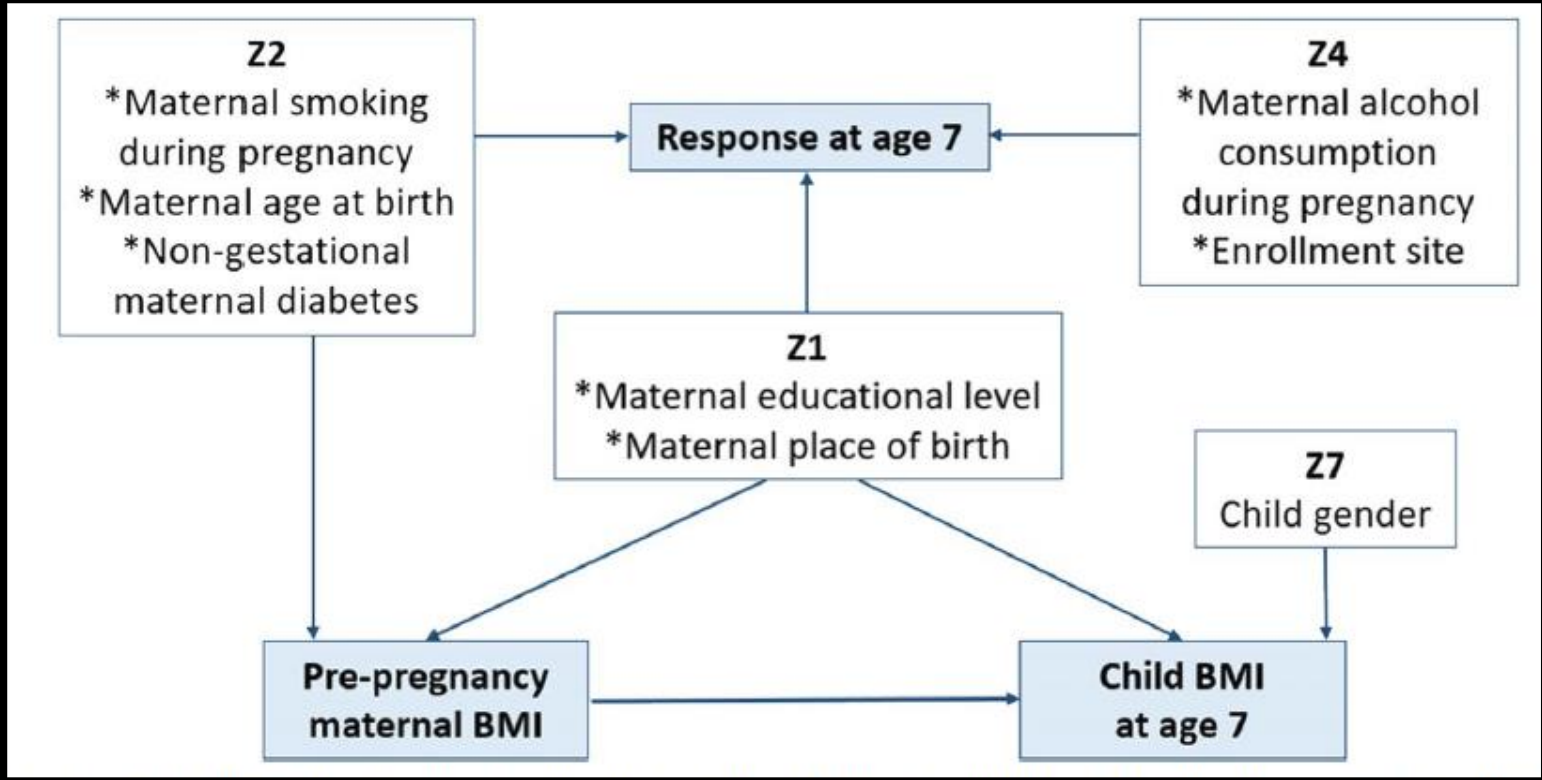
Directed Acyclic Graph (DAG)

- no variables of type Z3, Z5, or Z6 were present in this example.
- Assume a situation equivalent to the MAR1 scenario in the simulation study

Table 1 Response mechanism scenarios (data generation)

Scenario	γ_x	γ_y	$\gamma_1, \gamma_2, \gamma_3, \gamma_4$	Description
MAR 1	0.0	0.0	0.1	Response depending only on covariates
MAR 2	0.2	0.0	0.1	Response depending on covariates and exposure
MAR 3	0.5	0.0	0.1	Response depending on covariates and exposure
MNAR 1	0.0	0.2	0.1	Response depending on outcome and covariates
MNAR 2	0.2	0.2	0.1	Response depending on outcome, exposure, and covariates
MNAR 3	0.5	0.2	0.1	Response depending on outcome, exposure, and covariates
MNAR 4	0.0	0.5	0.1	Response depending on outcome and covariates
MNAR 5	0.2	0.5	0.1	Response depending on outcome, exposure, and covariates
MNAR 6	0.5	0.5	0.1	Response depending on outcome, exposure, and covariates

γ_k:regression coefficients of the generated response models ($\text{logit}(p_i) = \gamma_0 + \gamma_y y_i + \gamma_x x_i + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i}$)



Analysis

- A linear regression model was fitted with confounding variables.
- Both CCA and the IPPW method were applied.

Results

- The β coefficients related to the exposure of interest were very similar between CCA and the IPPW method.
- Within the IPPW results, the most effective response model strategy was the one including only Z1 variables (maternal educational level and maternal place of birth).

Table 7 Adjusted association between pre-pregnancy maternal BMI and child BMI at age 7 (CCA and IPPW method)

	β	SE
CCA ($N = 590$)[a]	0.142	0.0197
IPPW method ($N = 590$)[a]		
Response models		
Z_1, Z_2, Z_4	0.137	0.0229
Z_1, Z_2, Z_4, Z_7	0.138	0.0231
Z_1, Z_2, Z_7	0.140	0.0234
Z_1	0.140	0.0224

Discussion

- Attrition is a major methodological issue in cohort studies.
- It challenges the validity of association analyses because its occurrence is generally not completely at random.

- Our simulation study showed no superiority of method over CCA in terms of bias, and it even led to a loss of efficiency.
- Both were similarly unbiased in the MAR scenarios and similarly biased in most MNAR scenarios

Discussion

- For the MNAR scenarios, the absolute bias increased as the correlation between the exposure and the response increased.
- As a result, the mean square error is high for these scenarios when

$\gamma_x = \mathbf{0.5}$.

- In addition, because the bias is negative, the confidence intervals are shifted to the left and the nominal error rates are poorly respected.

- One explanation for the loss of efficiency observed with the IPPW method lies in the fact that adding covariates in the response model tends to increase the variability of estimated weights.

Discussion

- Our study aimed also to assess the impact of the choice of the variables included in the response model on the bias of the exposure regression coefficient and its variance.
- It is preferable not to include the exposure variable in the response model, otherwise variance inflation would be observed.
- The strategy for constructing the response model requires clear identification of the role played by the variables.
- This can be based on a structural approach using DAGs.

Strengths

- Firstly, we tested through simulations nine response mechanism scenarios, corresponding to three degrees of correlation between the response variable and our interest variables (exposure, outcome)
- Secondly, we evaluated the impact of several response models on the estimated exposure effect.

Limitations

- First, our simulation framework did not consider binary outcomes, although this is a common situation in epidemiology.
- Second, the level of attrition was kept constant, at a quite high but realistic level (40%) for cohort studies.
- Third, we did not vary the degree of correlation between the covariates Z and the response or our variables of interest (exposure, outcome).
  - strong correlation between the outcome and a variable of type Z3 (associated with the outcome and response) could increase the magnitude of the bias

Conclusion

- This study suggests that using IPPW to handle attrition in cohort studies does not reduce bias and may result in a loss of efficiency.
- These results therefore raise questions about the contribution of the IPW method to correcting possible selection bias that occurs in situations of attrition that lead to a missing outcome in association analyses.
- If the method is to be used, we encourage use of only the confounding variables of the association of interest in the response model.

Thank you