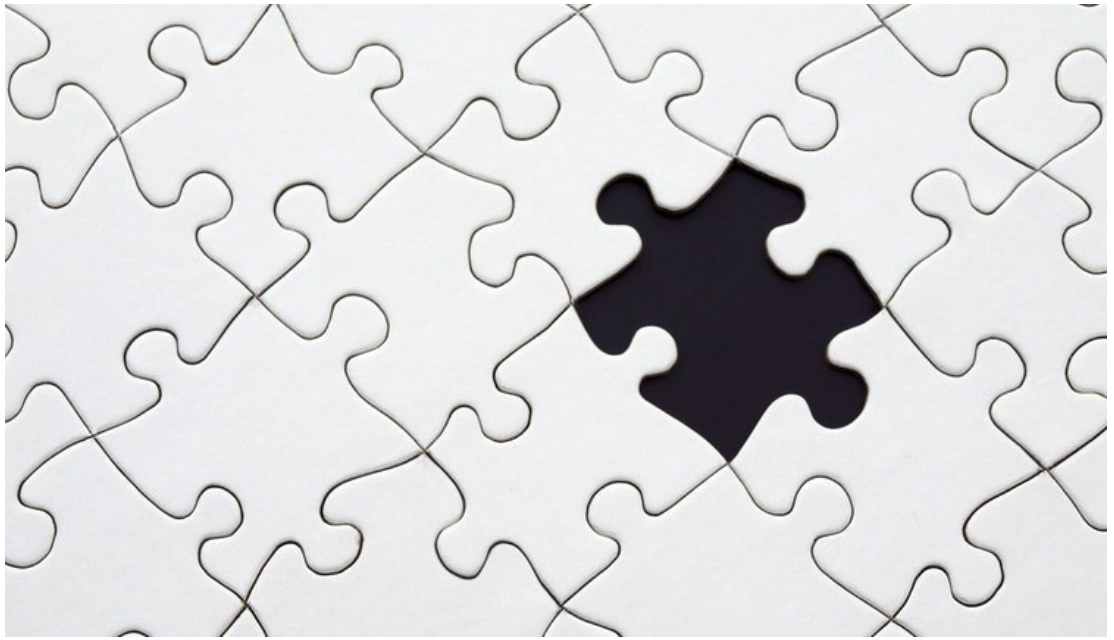
Introduction



- Missing data cause
  - reduce precision
  - induce a large amount of bias

Missing data mechanism

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR) : most common

(believed that) Can be correctly solved by the imputation method

What options for missing data

- Deletion: listwise vs pairwise
- Recover the values
- Imputation (replacement)
  - Education guessing
  - Common-point imputation : replace with grand means
  - Average imputation: replace with group means
  - Regression substitution: replacement with regression predicted values

- Biased variance
- Under-estimated SE
- Ignore natural random valued

- Multiple imputation (MI)
- Maximum likelihood IML)

Limitation :
appropriated for MCAR

- Complete-case analysis
  - Using listwise deletion method
- Single imputation
  - Missing values are imputed by observed sample mean
  - Variance estimates can be underestimate
- Multiple imputation (MI)
  - A simulation-based procedure_prepresents multiple set of possible values for missing data

- Multiple imputation
  - Limitation : produces different estimates every time
  - Requires multiple assumptions
  - Works with MAR
  - Give biased results in MNAR

In practice _still in the developmental process

- mdesc
- If <10% => missing data can be ignored
- If missing >50% => modern methods or delete it

OXFORD JOURNALS
American Journal of Epidemiology

Missing Outcome Data in Epidemiologic Studies

Stephen R Cole, corresponding author Paul N Zivich, Jessie K Edwards, Rachael K Ross, Bonnie E Shook-Sa, Joan T. Price, and Jeffrey S A Stringer

Am J Epidemiol. 2023 Jan; 192(1): 6–10. doi: 10.1093/aje/kwac179

Present by : Chuenkamon Charakorn
Comment by : Monchai Sumtipap
Academic coordinator : Ajarn Atiporn Ingsathit

May 12, 2023

- The authors provided a summary data set taken from a recent randomized trial conducted in Zambia which was not subject to missing data and induced missing outcomes to illustrate 4 scenarios
- Then they analyze the modified data sets using both a naive method and a principled missing-data method.

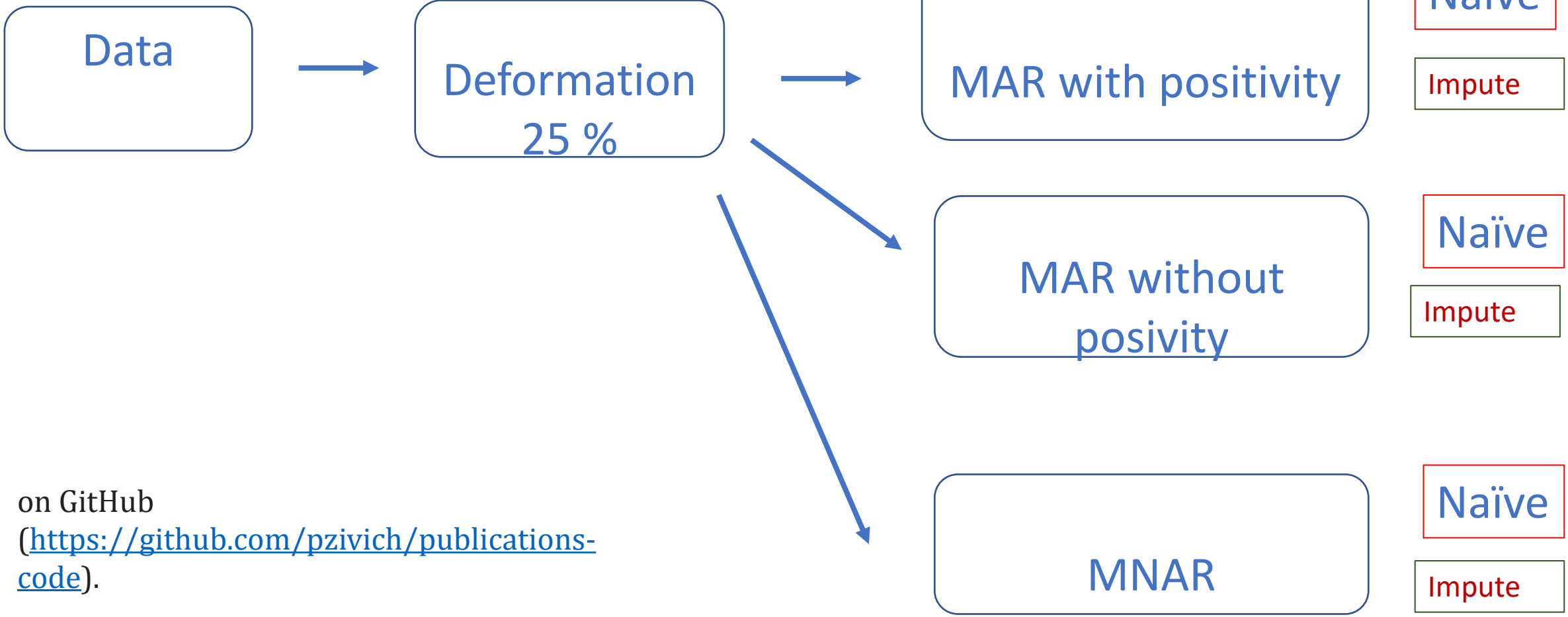
IPOP trial

- The Improving Pregnancy Outcomes With Progesterone (IPOP) Trial.
  - was a double-masked placebo-controlled randomized trial
  - Population : HIV pregnant women, ANC in Lusaka, Zambia
  - Intervention: weekly injections of 17α-hydroxyprogesterone caproate (17p)
  - Control : Placebo
  - Outcomes: composite outcome of preterm birth (birth at <37 weeks' gestation) or stillbirth

IPOP trial

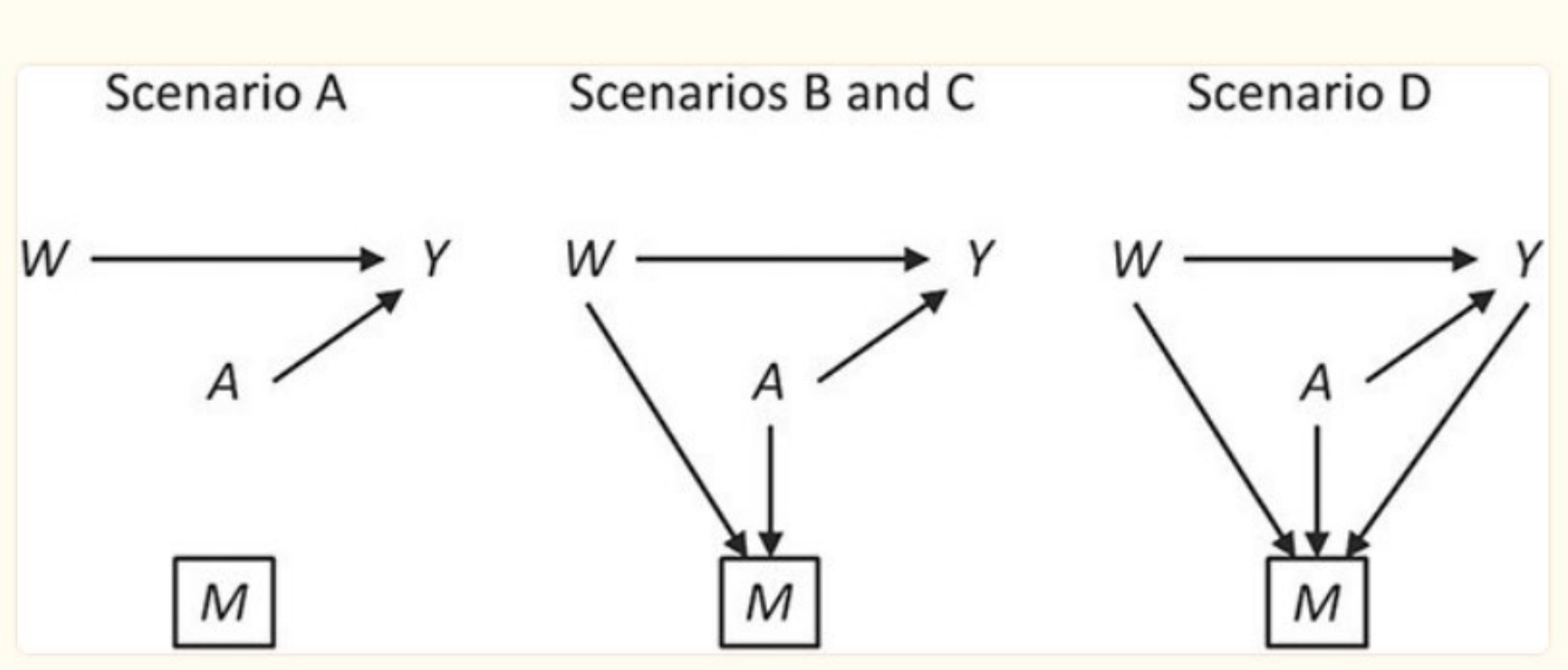
- Eligible women were aged 18 years or older, had a viable singleton pregnancy at less than 24 weeks' gestation, had confirmed HIV infection, and were currently receiving or intended to commence the use of antiretroviral therapy.
- Exclude : prior spontaneous preterm birth

- N = 800
- Weekly injection 17P vs placebo
- Cervical length 4 cm => effect the outcome in general (here is equal in both group)



on GitHub (https://github.com/pzivich/publications-code).

Causal diagrams for possible missing-data scenarios in the IPOP Trial



W denotes the covariate short cervix,
A denotes treatment with 17α-hydroxyprogesterone caproate,
Y denotes preterm birth outcome, and
M denotes a missing value for the outcome.
Boxes denote restriction to observed data.

Imputation

- IPW
- MI

The 4 missing-data mechanisms

1. *Missing completely at random*: 25% of patients had their outcome set to missing, independent of 17p treatment, short cervix, or the value of the outcome itself. Therefore, no bias should be incurred even with a naive analysis, but a loss in precision is to be expected.

2. *Missing at random with positivity*: 50% of patients with both 17p treatment and a short cervix and 50% of patients with neither 17p treatment nor a short cervix had their outcomes set to missing. Other patients had complete data.

The 4 missing-data mechanisms

2. *Missing at random with positivity*:

- Therefore, among the patients with the outcome observed, the odds ratio for the association between short cervix and no 17p treatment was 4.3.
- This relationship is expected to cause positive bias because the no-17p treatment group is enriched with patients with a short cervix, and short cervix was associated with a nearly 2-fold increased risk of preterm birth.
- Regarding missing data, positivity is the condition that each woman has a positive probability of having observed data given measured covariates.

The 4 missing-data mechanisms

3. *Missing at random without positivity*: All patients with both 17p treatment and a short cervix had their outcomes set to missing. Other patients had complete data.

- Therefore, the probability of being observed was 0 (nonpositive) for patients with both 17p treatment and a short cervix.

4. *Missing not at random*: Among women who did not have a preterm birth, 50% with both 17p treatment and a short cervix and 50% with neither 17p treatment nor a short cervix had their outcomes set to missing. Additionally, 50% of women with preterm birth who were treated with 17p had their outcomes set to missing.

- Therefore, a bias is induced which cannot be removed without knowledge of the data that are missing.

Distribution of Data From the IPOP Trial and Possible Missing-Data Scenarios (n = 800), Zambia, 2018–2020

Data Set	Total No.	Cervix Length and 17p Treatment Arm			
		Cervix ≥4 cm		Cervix <4 cm	
		No 17p	17p	No 17p	17p
Observed IPOP data					
No. of women	800	215	222	186	177
No. of preterm births	72	15	13	21	23
No. of missing data points	0	0	0	0	0
Missing-data mechanism					
MCAR[a]					
No. of women	601	161	167	140	133
No. of preterm births	54	11	10	16	17
No. of missing data points	199	54	55	46	44
MAR with positivity[b]					
No. of women	605	108	222	186	89
No. of preterm births	54	8	13	21	12
No. of missing data points	195	107	0	0	88
MAR without positivity[c]					
No. of women	623	215	222	186	0
No. of preterm births	49	15	13	21	0
No. of missing data points	177	0	0	0	177
MNAR[d]					
No. of women	583	100	216	186	81
No. of preterm births	55	15	7	21	12
No. of missing data points	217	115	6	0	96

MAR, missing at random;
MCAR, missing completely at random;
MNAR, missing not at random.

Approximately 25% missing data from each stratum.

[b] Approximately 50% missing data from the first and fourth strata.

[c] All women from the fourth stratum were missing outcome data.

[d] Approximately 50% of term births were missing from the first and fourth strata and 50% of preterm births were missing from the second and fourth strata.

Statistical methods

- For the naive method, risk ratios were estimated using a log binomial model fitted to the complete records by maximum likelihood, with Wald-type 95% confidence intervals computed using the model-based standard error.

- Principled approaches with which to account for missing data include imputation, weighting, or direct maximum likelihood.

- With only the outcome missing, they accounted for missing data using a direct maximum likelihood approach, specifically, generalized computation (g-computation) to estimate the treatment effect accounting for the missing outcome data.

- The generalized formula can be used to construct a g-computation algorithm that provides a maximum likelihood estimator of the risk under binary treatment *a*, given as

$$n^{-1}\sum_{i=1}^{n} m(a, W_i; \hat{\beta}),$$

$$n^{-1} \sum_{i=1}^{n} m(a, W_i; \hat{\beta}),$$

where m(a, W_i; $\hat{\beta}$) is the probability of the potential outcome Y_i^a estimated using the observed data, W_i is a set of covariates (where *i* indexes the n participants), and β is a set of parameters from the model m.

WEB TABLE 1: Inverse probability weighted and Multiply Imputed IPOP Trial Results for Effect of No 17p on Preterm Birth Under Various Missing Data Mechanisms, 2018 to 2020, N = 800

Missing Data Mechanism	Naive		Inverse Probability Weighted [a]		Multiple Imputation [b]	
	Risk Ratio (95% CI)	RMSE (SE) [c]	Risk Ratio (95% CI)	RMSE (SE) [c]	Risk Ratio (95% CI)	RMSE (SE) [c]
No Missing	1.00 (0.65, 1.56)	0.225 (0.225)	NA	NA	NA	NA
A. MCAR	1.00 (0.60, 1.66)	0.260 (0.260)	1.00 (0.60, 1.66)	0.260 (0.260)	0.99 (0.60, 1.65)	0.259 (0.259)
B. MAR positive	1.23 (0.74, 2.04)	0.339 (0.261)	1.00 (0.58, 1.71)	0.274 (0.274)	1.01 (0.60, 1.70)	0.266 (0.266)
C. MAR nonpositive	1.53 (0.83, 2.83)	0.536 (0.313)	1.53 (0.83, 2.83)	0.533 (0.313)	1.21 (0.65, 2.26)	0.370 (0.317)
D. MNAR	1.97 (1.16, 3.35)	0.740 (0.271)	1.59 (0.91, 2.76)	0.543 (0.282)	1.58 (0.92, 2.71)	0.533 (0.274)

Abbreviations: CI, confidence interval; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; RMSE, root mean squared error; SE, standard error; NA, not applicable.

[a] Weights were estimated as the inverse of the predicted probability of being observed from a

- To implement the g-computation approach.
- Construct the 2 potential outcomes and add them to the data set.

  set $Y_i^a = Y_i$ when $A_i = a$, by invoking the causal consistency assumption

  When Ai ≠ a, the constructed potential outcome Y_i^a is missing.

  - When the observed outcome Y_i is missing, then both constructed potential outcomes are Y_i^a missing. The data set with the 2 constructed potential outcomes is illustrated.

WED TABLE 2: Structure of Missing Data [a]

Short Cervix, W	17p, A	Preterm, Y	Y^0	Y^1
0	0	0	0	M
0	0	1	1	M
0	0	M	M	M
0	1	0	M	0
0	1	1	M	1
0	1	M	M	M
1	0	0	0	M
1	0	1	1	M
1	0	M	M	M
1	1	0	M	0
1	1	1	M	1
1	1	M	M	M

[a] Missing value denoted by "M"

- Fit a pair of logistic regression models, one with each potential outcome as the outcome, both conditional on short cervix status.
- The fitted logistic regression models are used to predict the probability of the potential outcome under plan a, which is $m(a, W_i; \hat{\beta})$.
- Estimate the preterm birth risk under treatment a by taking the average of the predicted values $m(a, W_i; \hat{\beta})$.

Assumptions

- First, women treated with 17p are assumed to be marginally exchangeable with women treated with placebo given the randomized design.
- Second, women who are missing data are assumed to be exchangeable with women with observed data, conditional on short cervix and 17p treatment. Conceptually, g-computation imputes missing potential outcome data, whether those data are missing because the outcome is missing or missing because the woman received the alternate treatment (i.e.,$A_i \neq a$)

- Wald-type 95% confidence intervals were computed with the bootstrap standard error—that is, the standard deviation of 500 bootstrap random samples, each of size *n*, taken with replacement from the observed data.
- The authors compared the risk ratios, both naive and accounting for missing data in a principled manner, using the IPOP full-data estimate as a gold standard.
- They also calculated the root mean squared error (i.e., the square root of the sum of squared bias and variance).

Results

- While accounting for data that are missing completely at random can improve precision in comparison with a complete-case analysis, a reduction in the standard error upon accounting for the missing data was not seen in this simple example.

Results

Table 2. Effect of No 17p Use on Preterm Birth Under Various Missing-Data Mechanisms in the IPOP Trial (n = 800), Zambia, 2018–2020

Missing-Data Mechanism	Analysis							
	Naive				Imputed[a]			
	RR	95% CI	RMSE	SE for Log RR	RR	95% CI	RMSE	SE for Log RR
No missing data	1.00	0.65, 1.56	0.225	0.225	N/A	N/A	N/A	N/A
MCAR	1.00	0.60, 1.66	0.260	0.260	0.98	0.58, 1.67	0.273	0.273
MAR with positivity	1.23	0.74, 2.04	0.339	0.261	0.98	0.57, 1.70	0.280	0.280
MAR without positivity	1.53	0.83, 2.83	0.536	0.313	1.52	0.78, 2.97	0.547	0.340
MNAR	1.97	1.16, 3.35	0.740	0.271	1.56	0.88, 2.78	0.540	0.292

Abbreviations: 17p, 17α-hydroxyprogesterone caproate; CI, confidence interval; IPOP, Improving Pregnancy Outcomes With Progesterone; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; N/A, not applicable; RMSE, root mean squared error; RR, risk ratio; SE, standard error.

[a] Generalized computation accounting for cervix length <4 cm, with the SE estimated by the standard deviation of 500 bootstrap samples.

Results

Table 2. Effect of No 17p Use on Preterm Birth Under Various Missing-Data Mechanisms in the IPOP Trial (n = 800), Zambia, 2018–2020

Missing-Data Mechanism	Analysis							
	Naive				Imputed[a]			
	RR	95% CI	RMSE	SE for Log RR	RR	95% CI	RMSE	SE for Log RR
No missing data	1.00	0.65, 1.56	0.225	0.225	N/A	N/A	N/A	N/A
MCAR	1.00	0.60, 1.66	0.260	0.260	0.98	0.58, 1.67	0.273	0.273

When data were MCAR, the naive complete-case analysis had no bias but there was a loss of precision, with the standard error for the log risk ratio being 1.16 (= 0.260/0.225) times larger than the full-data standard error. While accounting for data that are MCAR can improve precision in comparison with a complete-case analysis, a reduction in the standard error upon accounting for the missing data was not seen in this simple example.

Results

Table 2. Effect of No 17p Use on Preterm Birth Under Various Missing-Data Mechanisms in the IPOP Trial (n = 800), Zambia, 2018–2020

Missing-Data Mechanism	Analysis							
	Naive				Imputed[a]			
	RR	95% CI	RMSE	SE for Log RR	RR	95% CI	RMSE	SE for Log RR
No missing data	1.00	0.65, 1.56	0.225	0.225	N/A	N/A	N/A	N/A
MCAR	1.00	0.66, 1.66	0.260	0.260	0.98	0.58, 1.67	0.273	0.273
MAR with positivity	1.23	0.74, 2.04	0.339	0.261	0.98	0.57, 1.70	0.280	0.280

When data were MAR with positivity, there was notable bias. The bias was ameliorated upon accounting for the missing data, with some cost in precision, as the standard error for the log risk ratio was 1.07 (= 0.280/0.261) times larger than it was without bias correction.

Taking the estimated risk ratio from the full data set as the truth, the root mean squared error was 0.339 for the naive log risk ratio and 0.280 for the imputed risk ratio, suggesting that here the reduction in bias outweighed any loss of precision in terms of squared error.

Results

Table 2. Effect of No 17p Use on Preterm Birth Under Various Missing-Data Mechanisms in the IPOP Trial (n = 800), Zambia, 2018–2020

Missing-Data Mechanism	Analysis							
	Naive				Imputed[a]			
	RR	95% CI	RMSE	SE for Log RR	RR	95% CI	RMSE	SE for Log RR
No missing data	1.00	0.65, 1.56	0.225	0.225	N/A	N/A	N/A	N/A
MCAR	1.00	0.60, 1.66	0.260	0.260	0.98	0.58, 1.67	0.273	0.273
MAR with positivity	1.23	0.74, 2.04	0.339	0.261	0.98	0.57, 1.70	0.280	0.280
MAR without positivity	1.53	0.83, 2.83	0.536	0.313	1.52	0.78, 2.97	0.547	0.340

When data were MAR without positivity, there was notable bias. Here the bias was not ameliorated upon accounting for the missing data. In this example, the effect of no 17p treatment on preterm birth is homogeneous on the ratio scale (as well as the difference scale, since there is no effect) for women with and without a short cervix, as can be verified in the full data. Therefore, we can restrict analysis to the subset in which we have positivity (i.e., where the cervix is ≥4 cm long) and obtain an unbiased estimate of the effect of no 17p treatment on preterm birth, albeit with loss of precision.

Results

Table 2. Effect of No 17p Use on Preterm Birth Under Various Missing-Data Mechanisms in the IPOP Trial (n = 800), Zambia, 2018–2020

Missing-Data Mechanism	Analysis							
	Naive				Imputed[a]			
	RR	95% CI	RMSE	SE for Log RR	RR	95% CI	RMSE	SE for Log RR
No missing data	1.00	0.65, 1.56	0.225	0.225	N/A	N/A	N/A	N/A
MCAR	1.00	0.60, 1.66	0.260	0.260	0.98	0.58, 1.67	0.273	0.273
MAR with positivity	1.23	0.74, 2.04	0.339	0.261	0.98	0.57, 1.70	0.280	0.280
MAR without positivity	1.53	0.83, 2.83	0.536	0.313	1.52	0.78, 2.97	0.547	0.340
MNAR	1.97	1.16, 3.35	0.740	0.271	1.56	0.88, 2.78	0.540	0.292

When data were missing not at random (i.e., depended on values of the missing variables themselves), there was again notable bias when using the naive complete-case estimator. Here, accounting for the missing data reduced but did not eliminate bias.

rone;
uared
ples.

Discussion

- Missing data were an important problem 50 years ago, and was suspected they will remain so.
- Why are missing data so important? Everything you don't know is missing data; and much of what you think you know is affected by missing data.
- If we don't have a formal way to represent and analyze missing data, we may not be able to even recognize what is missing, let alone make accurate inferences when there are missing data.
- Missing data are arguably the central analytical problem for epidemiology, because both confounding and measurement error may be framed as implicit missing-data problems.
- Ignoring missing data stubbornly remains standard practice in epidemiology.

Discussion

- There are limitations to this illustration.

1. The g-computation approach does not easily allow one to have different covariate sets for the missing data and treatment exchangeability assumptions.

2. The inverse probability weighting can be used instead.

3. The variance for the imputation estimator was estimated using the bootstrap, but M-estimation could have been used instead, which avoids the computationally intensive resampling procedure.

Discussion

- Missing data come in many forms. One way to classify missing data is as data missing completely at random, missing at random, or missing not at random.
- In empirical work, we rarely know which form of missingness is operating.
- If data are missing completely at random, then accounting for missing data may improve precision, though this is not guaranteed, as is seen in the example.
- If data are missing at random with positivity, then accounting for missing data can remove bias.
- However, data can be missing at random without positivity, we may not be able to obtain unbiased estimates of the parameter of interest.

Discussion

- In this example, the bias induced when data were missing not at random was reduced by accounting for missing data in the analysis. One can envision scenarios where accounting for data missing not at random increases bias.

In summary

- MCAR
  - Ignorable
  - Complete-case analysis (Naïve)
  - MI
- MAR
  - MI
- MNAR
  - Naïve method and acknowledging limitations

To summarize, it seems better to account for, rather than ignore, missing data.

Main objective of missing data replacement
1. minimize bias
2. maximize use of available information
3. get good estimates of uncertainty