Missing Data Part 3

Jarvis T. Chen

17 April 2023

Department of Social and Behavioral Sciences
Harvard T. H. Chan School of Public Health

Causal inference as a missing data problem

Causal inference as a missing data problem

Assume that we are interested in the **causal effect** of a binary exposure A on a binary outcome Y. We can conceive of two **potential outcomes** for each individual in our study population, $Y_{a=1}$ and $Y_{a=0}$, depending on whether the individual is exposed to $a = 1$ or $a = 0$.

Individual	$Y_{a=1}$	$Y_{a=0}$
1	0	1
2	1	0
3	0	0
4	0	0
5	1	1
6	0	0

Causal inference as a missing data problem

We only get to observe each individual's exposure A and outcome Y.

Individual	A	Y
1	1	0
2	0	0
3	0	0
4	0	0
5	1	1
6	0	0

Causal inference as a missing data problem

This is the **fundamental problem of causal inference**: we only ever observe one outcome for each individual, i.e. the one corresponding to the treatment (exposure) that they did receive.

Individual	A	$Y_{a=1}$	$Y_{a=0}$
1	1	0	?
2	0	?	0
3	0	?	0
4	0	?	0
5	1	1	?
6	0	?	0

Causal inference as a missing data problem

- In an observational setting, we may also have measured a vector of covariates C_i for each individual.
- You're familiar with the notion of conditional exchangeability that allows us to draw valid causal inferences from observational data, i.e. $Y_a \perp\!\!\!\perp A|C$
- This is equivalent to a statement about ignorability (unconfoundedness) of the **treatment assignment mechanism**,

$$P(A|Y_{a=0}, Y_{a=1}, C) = P(A|C)$$

- We can draw a parallel between the concept of the **treatment assignment mechanism** and the **missing data mechanism**.
- For each individual, the observed exposure A determines whether we get to observe $Y_{a=1}$ or $Y_{a=0}$, we can think of this as

$$P(A|Y^{obs}, Y^{mis}, C) = P(A|C)$$

- Recall the definition of MAR from last week:

$$P(\mathbf{R}|\mathbf{Z}^{obs}, \mathbf{Z}^{mis}) = P(\mathbf{R}|\mathbf{Z}^{obs}, \boldsymbol{\theta})$$

Individual	A	$Y_{a=1}$	$Y_{a=0}$	C
1	1	0	?	c_i
2	0	?	0	c_i
3	0	?	0	c_i
4	0	?	0	c_i
5	1	1	?	c_i
6	0	?	0	c_i

Supplemental Slides: Complete case analysis vs. multiple imputation

Data on exposure are MCAR

Your experience with DAGs and selection bias can help you think through the consequences of doing a **complete case analysis** in the presence of missing data.

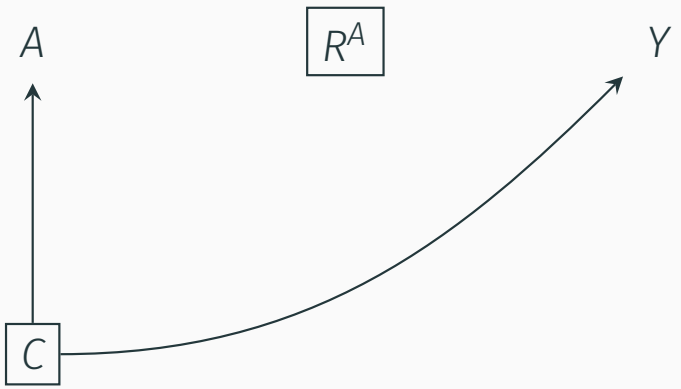


Figure 1: DAG drawn under the null. Data on exposure are MCAR. Complete case analysis yields unbiased estimate of the effect of *A* on *Y* as long as we control for *C*.

Data on exposure are MAR

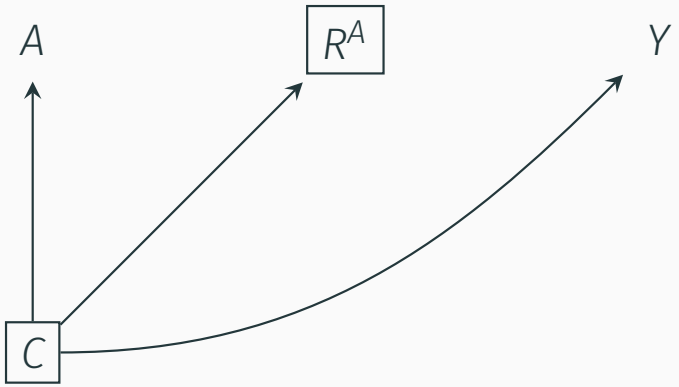


Figure 2: DAG drawn under the null. Data on exposure are MAR. Complete case analysis yields unbiased estimate of the effect of *A* on *Y* as long as we control for *C*. Multiple imputation using information on *C* to predict missing *A* helps with efficiency.

Data on exposure are MAR

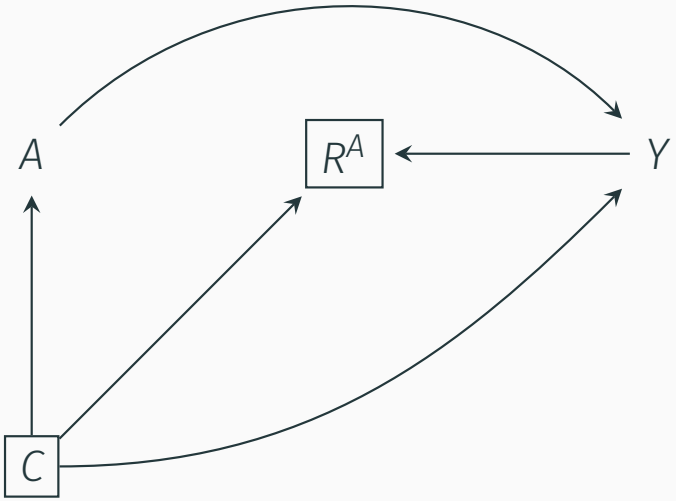
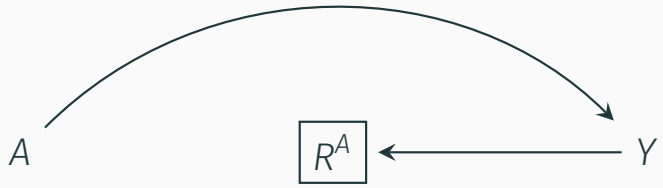


Figure 3: DAG drawn **off null**. Data on exposure are MAR. Complete case analysis **does not yield unbiased estimate** of the effect of A on Y (because missingness depends on observed values of the outcome) **unless** the true relationship between exposure and outcome is null. Multiple imputation using information on $f(A|Y, C)$ allows us to recover an unbiased estimate of the effect of A on Y.

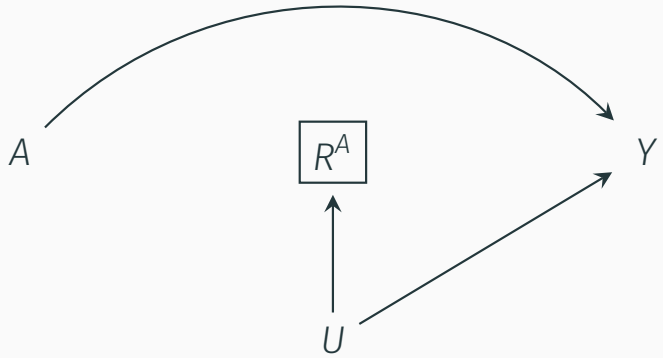
Note that the DAG does not capture the possibility of bias – we have selection bias without collider stratification.

Selection bias without colliders

We can draw an even simpler version with exposure data MAR:



This similar in structure to Figure 1 of Hernán 2017, in which *U* is a common cause of selection and *Y* and there is effect measure modification of the *AY* relationship across levels of *U*.



Hernán notes: "Causal directed acyclic graphs are nonparametric and thus cannot generally encode biases that depend on a particular parameterization of the effect. This is also the reason why the distinction between bias under the null and bias off the null is important for selection bias but not for confounding." Hernán M, Invited commentary: selection bias without colliders, *Am J Epidemiol* 2017;185(11):1048–1050.

Data on exposure are MAR

What if missingness depends on C and M, but we are interested in the total effect of A on Y?

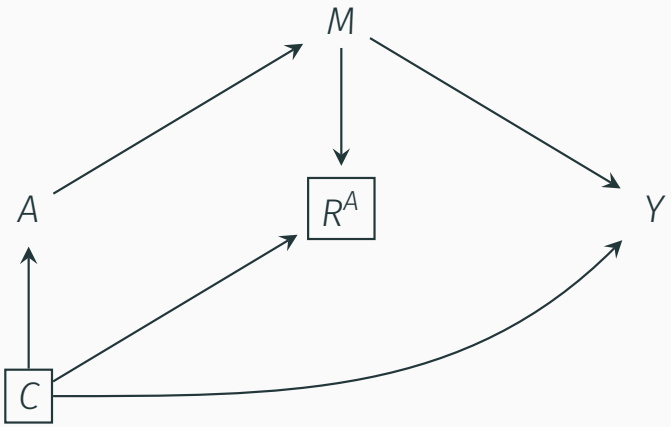


Figure 4: DAG drawn under the null. Data on exposure is MAR. Complete case analysis **does not yield unbiased estimate** of the total effect of A on Y because missingness depends on C and M but we don't want to adjust for M in the model for $Y|A, C$. Multiple imputation sampling from $f(A|M, C)$ allows us to fit the model we want in the imputed datasets. Note that if we adjust for M in the analytic model, our estimate the controlled direct effect of A on Y is unbiased in the complete case analysis.

Hughes et al. 2019 note "Efficiency gains of MI over CCA are greatest when there are small amounts of missing data on many variables and/or auxiliary variables that provide information about the missing values."

Data on exposure are MNAR

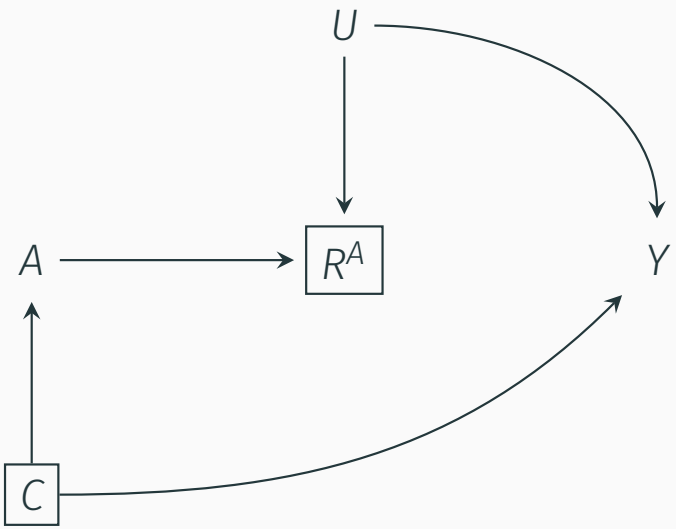


Figure 5: DAG drawn under the null. Data on exposure is MNAR. Complete case analysis **does not yield unbiased estimate** of the effect of A on Y even when controlling for C. Here, the DAG does show us that collider stratification leads to biased estimation of the $A \rightarrow Y$ effect in the complete case analysis. Multiple imputation cannot help us recover the full joint distribution in the population where missingness had not occurred.

Data on exposure are MNAR

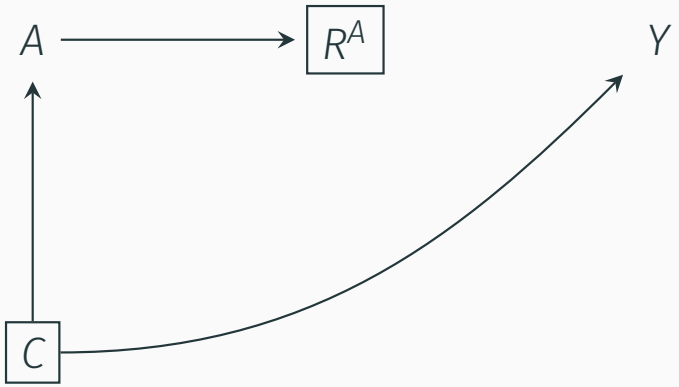


Figure 6: DAG drawn under the null. Data on exposure are MNAR. Contrary to popular belief, complete case analysis yields an unbiased estimate of the effect of *A* on *Y* provided we control for *C*. However, multiple imputation cannot help us because the data on exposure are MNAR!

Linear vs. logistic regression

Though noted in the literature, it is underappreciated that the potential bias of the exposure regression coefficient in complete case analysis depends on whether the analytic model is a linear or logistic regression model.

	Exposure regression coefficient	
Variable missingness is dependent on	Linear	Logistic
None (i.e. Missing Completely At Random	Unbiased	Unbiased
Outcome	Biased*	Unbiased
Exposure (and possible confounders)	Unbiased	Unbiased
Outcome and confounders	Biased	Unbiased
Outcome and exposure (and possible confounders)	Biased	Biased †

Table 1 from Hughes et al. 2019

* Biased in general, except when in truth there is no association between the outcome and the exposure (i.e. the true value of the exposure regression coefficient is zero).

† Biased in general, except when missingness depends on the outcome and exposure independently.

Summary

- DAGs are helpful for thinking about whether unbiased estimates can be obtained from a complete case analysis, but note that they don't encode biases that depend on a particular parameterization of the effect estimate (e.g. the off null bias due to missingness that depends on the outcome in Figure 3).
- Thinking about the missing data mechanism (MCAR or MAR vs. MNAR) can help you think about whether multiple imputation can help you recover the joint distribution of the variables that you would have observed in a population where missingness had not occurred.
- Multiple imputation is a valid approach for all MAR mechanisms, while complete case analysis may give biased results when the chance of being a complete case depends on the observed value of the outcome.
- Multiple imputation can use information from auxiliary variables (not included in the main analysis) that provide information about the missing values.
- When the exposure and/or confounders in the main analysis are MNAR, complete case analysis is a valid approach but multiple imputation may give biased results!

---

† See Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* 2019, 1294-1304. doi: 10.1093/ije/dyz032