SCM graphs and missing data[1]

Statistical and causal inference involving missing data demands a model. Directed acyclic graphs (DAGs) provide a convenient and powerful frame for this important exercise as few empirical studies are free of missing data.

There are four categories of missing data: (a) no missingness, (b) missing completely at random (MCAR), (c) missing at random (MAR), and (d) missing not at random (MNAR). Consider the data in table 2 to illustrate.

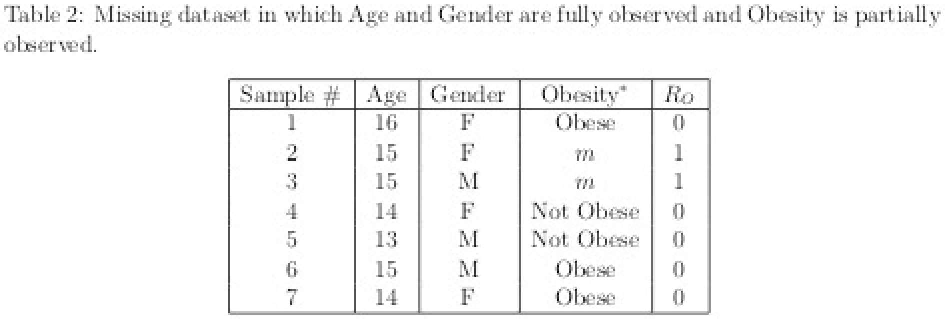


Table 2: Missing dataset in which Age and Gender are fully observed and Obesity is partially observed.

Sample #	Age	Gender	Obesity*	R_O
1	16	F	Obese	0
2	15	F	m	1
3	15	M	m	1
4	14	F	Not Obese	0
5	13	M	Not Obese	0
6	15	M	Obese	0
7	14	F	Obese	0

Graphical representations of the four cases are depicted in figure 1.

[1] This note draws from Mohan and Pearl (2020), *Journal of the American Statistical Association*, "Graphical models for processing missing data." Mohan and Pearl discuss a broad range of issues as indicated in table 1.

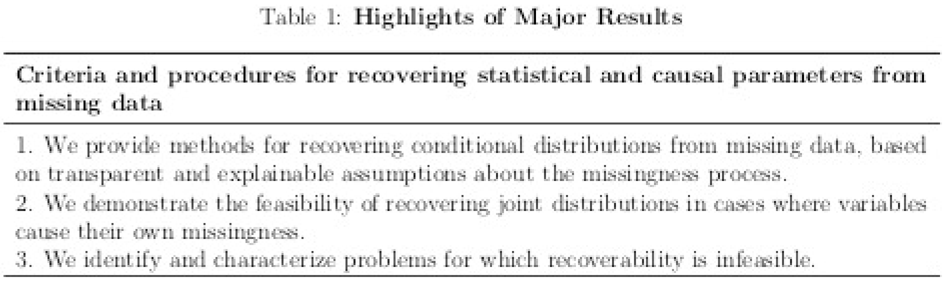


Table 1: **Highlights of Major Results**

Criteria and procedures for recovering statistical and causal parameters from missing data
1. We provide methods for recovering conditional distributions from missing data, based on transparent and explainable assumptions about the missingness process.
2. We demonstrate the feasibility of recovering joint distributions in cases where variables cause their own missingness.
3. We identify and characterize problems for which recoverability is infeasible.

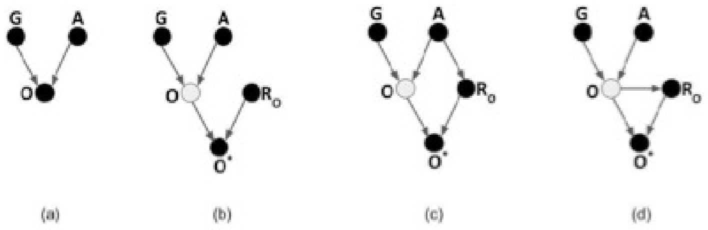


Figure 1: (a) causal graph under no missingness (b), (c) & (d) m-graphs modeling MCAR, MAR and MNAR missingness processes respectively.

Solid (clear) nodes indicate variables involving no missing (some missing) data, R variables represent the status of the causal mechanism for missingness (0 indicates not missing and 1 indicates missing), and variables with an asterisk indicate proxy variables for variables experiencing missingness as defined by equation (1).

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases} \qquad (1)$$

Proxy variables (defined by equation (1)) are typically omitted from m-graphs for simplicity and clarity.

1. MCAR data are indicated by $V_0, V_m \perp R$. That is, the causal mechanism for missingness is independent of observed and partially observed data, respectively.
2. MAR data are indicated by $V_m \perp R \mid V_0$. The causal mechanism for missingness is independent of partially observed data conditional on observed data.
3. MNAR data are indicated when neither MCAR or MAR apply.

Example 1 *Consider the problem of recovering the joint distribution given the m-graph in Fig. 1 (c) and dataset in Table 3. Let it be the case that 15-18 year olds were reluctant to reveal their weight, thereby making O a partially observed variable i.e. $V_m = \{O\}$ and $V_o = \{G, A\}$. This is a typical case of v-MAR missingness, since the cause of missingness is the fully observed variable: Age. The following three steps detail the recovery procedure.*

1. Factorization: The joint distribution may be factored as:

$$P(G, O, A) = P(G, O|A)P(A)$$

2. Transformation into observables: G implies the conditional independence $(G, O) \perp\!\!\!\perp R_O|A$ since A d-separates (G, O) from R_O. Thus,

$$P(G, O, A) = P(G, O|A, R_O = 0)P(A)$$

3. Conversion of partially observed variables into proxy variables: $R_O = 0$ implies $O^* = O$ (by eq 1). Therefore,

$$P(G, O, A) = P(G, O^*|A, R_O = 0)P(A) \qquad (2)$$

The RHS of Eq. (2) is expressed in terms of variables in the observed-data distribution. Therefore, $P(G, A, O)$ can be consistently estimated (i.e. recovered) from the available data. The recovered joint distribution is shown in Table 4.

Note that samples in which obesity is missing are not discarded but are used instead to update the weights $p_1, ..., p_{12}$ of the cells in which obesity has a definite value. This can be seen by the presence of probabilities $p_{13}, ..., p_{18}$ in Table 4 and the fact that samples with missing values have been utilized to estimate prior probability $P(A)$ in equation 2. Note also that the joint distribution permits an alternative decomposition:

$$\begin{aligned} P(G, O, A) &= P(O|A, G)P(A, G) \\ &= P(O^*|A, G, R_O = 0)P(A, G) \end{aligned}$$

Table 3: observed-data Distribution $P(G, A, O^*, R_O)$ where Gender (G) and Age (A) are fully observed, Obesity O is corrupted by missing values and Obesity's proxy (O^*) is observed in its place. Age is partitioned into three groups: $[10-13), [13-15), [15-18)$. Gender and Obesity are binary variables and can take values Male (M) and Female (F), and Yes (Y) and No (N), respectively. The probabilities $p_1, p_2, ..., p_{18}$ stand for the (asymptotic) frequencies of the samples falling in the 18 cells (G, A, O^*, R_O).

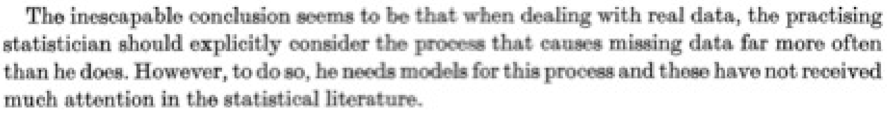
G	A	O^*	R_O	$P(G,A,O^*,R_O)$
M	$10-13$	Y	0	p_1
M	$13-15$	Y	0	p_2
M	$15-18$	Y	0	p_3
M	$10-13$	N	0	p_4
M	$13-15$	N	0	p_5
M	$15-18$	N	0	p_6
F	$10-13$	Y	0	p_7
F	$13-15$	Y	0	p_8
F	$15-18$	Y	0	p_9
F	$10-13$	N	0	p_{10}
F	$13-15$	N	0	p_{11}
F	$15-18$	N	0	p_{12}
M	$10-13$	m	1	p_{13}
M	$13-15$	m	1	p_{14}
M	$15-18$	m	1	p_{15}
F	$10-13$	m	1	p_{16}
F	$13-15$	m	1	p_{17}
F	$15-18$	m	1	p_{18}

Table 4: Recovered joint distribution corresponding to dataset in Table 3 and m-graph in Figure 1(c)

G	A	O	$P(G,O,A)$
M	$10-13$	Y	$\frac{p_1*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
M	$13-15$	Y	$\frac{p_2*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
M	$15-18$	Y	$\frac{p_3*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$
M	$10-13$	N	$\frac{p_4*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
M	$13-15$	N	$\frac{p_5*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
M	$15-18$	N	$\frac{p_6*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$
F	$10-13$	Y	$\frac{p_7*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
F	$13-15$	Y	$\frac{p_8*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
F	$15-18$	Y	$\frac{p_9*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$
F	$10-13$	N	$\frac{p_{10}*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
F	$13-15$	N	$\frac{p_{11}*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
F	$15-18$	N	$\frac{p_{12}*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$

Finally we observe that for the MCAR m-graph in Figure 1 (b), a wider spectrum of decompositions is applicable, including:

$$\begin{aligned} P(G,O,A) &= P(O,A,G|R_O=0) \\ &= P(O^*,A,G|R_O=0) \end{aligned}$$



The inescapable conclusion seems to be that when dealing with real data, the practising statistician should explicitly consider the process that causes missing data far more often than he does. However, to do so, he needs models for this process and these have not received much attention in the statistical literature.

Figure 2: Quote from Rubin (1976)

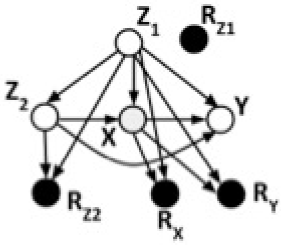


Figure 3: An MNAR m-graph in which joint distribution is not recoverable but $P(Y|X,Z_1,Z_2)$ and $P(Z_1)$ are recoverable. Proxy variables have not been explicitly portrayed, as stated in section 2.1.

A generic example for recoverability under MNAR is presented below.

Example 2 (Recoverability in MNAR m-graphs) *Consider the m-graph G in Figure 3 where all variables are subject to missingness. Y is the outcome of interest, X the exposure of interest and Z_1 and Z_2 are baseline covariates. The target parameter is $P(Y|X,Z_1,Z_2)$, the regression of Y on X given both baseline covariates.*
Since $Y \perp\!\!\!\perp (R_X, R_Y, R_{Z_1}, R_{Z_2})|(X,Z_1,Z_2)$ in G, $P(Y|X,Z_1,Z_2)$ can be recovered as:

$$\begin{aligned} P(Y|X,Z_1,Z_2) &= P(Y|(X,Z_1,Z_2,R_X=0,R_Y=0,R_{Z_1}=0,R_{Z_2}=0)) \\ &= P(Y^*|(X^*,Z_1^*,Z_2^*,R_X=0,R_Y=0,R_{Z_1}=0,R_{Z_2}=0))(\textit{ Using eq 1}) \end{aligned}$$

Though all variables are subject to missingness and missingness is highly dependent on partially observed variables, the graph nevertheless licenses the estimation of the target parameter from samples in which all variables are observed.

The next example illustrates recoverability specific to a causal inquiry even though recoverability is not generally feasible. In this setting, do-calculus is employed to address outcome missingness.

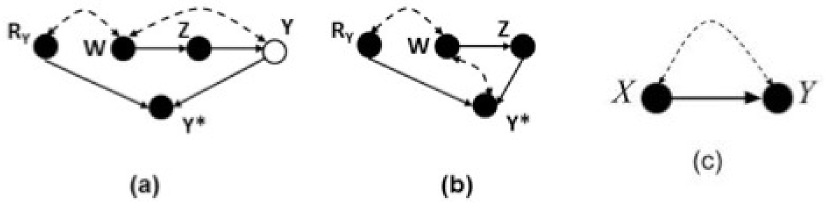


Figure 7: (a) m-graph in which $P(y|do(z))$ is recoverable although Y and R_y are not d-separable. (b) m-graph in which Y is treated as a latent variable and not explicitly portrayed. (c) bow-arc model in which causal effect of X on Y is non-identifiable.

Example 10 *Examine the m-graph in Figure 7(a). Suppose we are interested in the causal effect of Z (treatment) on outcome Y (death) where treatments are conditioned on (observed) X-rays report (W). Suppose that some unobserved factors (say quality of hospital equipment and staff) affect both attrition (R_y) and accuracy of test reports (W). In this setup the causal-effect query $P(y|do(z))$ is identifiable (by adjusting for W) through the estimand:*

$$P(y|do(z)) = \sum_w P(y|z,w)P(w). \tag{7}$$

However, the factor $P(y|z,w)$ is not recoverable (by theorem 3), and one might be tempted to conclude that the causal effect is non-recoverable. We shall now show that it is nevertheless recoverable in three steps.

Recovering $P(y|do(z)$ **given the m-graph in Figure 7(a)** *The first step is to transform the query (using the rules of do-calculus) into an equivalent expression such that no partially observed variables resides outside the do-operator.*

$$\begin{aligned} P(y|do(z)) &= P(y|do(z), R_y = 0) \textit{ (follows from rule 1 of do-calculus)} \\ &= P(y^*|do(z), R_y = 0) \textit{ (using eq 1)} \end{aligned} \tag{8}$$

The second step is to simplify the m-graph by removing superfluous variables, still retaining all relevant functional relationships. In our example, Y *is irrelevant once we treat* Y^* *as an outcome. The reduced m-graph is shown in Figure 7(b). The third step is to apply the do-calculus (Pearl (2009b)) to the reduced graph (7(b)), and identify the modified query* $P(y^*|do(z), R_y = 0)$.

$$P(y^*|do(z), R_y = 0) = \sum_w P(y^*|do(z), w, R_y = 0)P(w|do(z), R_y = 0) \tag{9}$$

$$P(y^*|do(z), w, R_y = 0) = P(y^*|z, w, R_y = 0) \textit{ (by Rule-2 of do-calculus)} \tag{10}$$

$$P(w|do(z), R_y = 0) = P(w|R_y = 0) \textit{ (by Rule-3 of do-calculus).} \tag{11}$$

Substituting (10) and (11) in (9) the causal effect becomes

$$P(y|do(z)) = \sum_w P(y^*|z, w, R_y = 0)P(w|R_y = 0), \tag{12}$$

which permits us to estimate our query from complete cases only. While in this case we were able to recover the causal effect using one pass over the three steps, in more complex cases we might need to repeatedly apply these steps in order to recover the query.

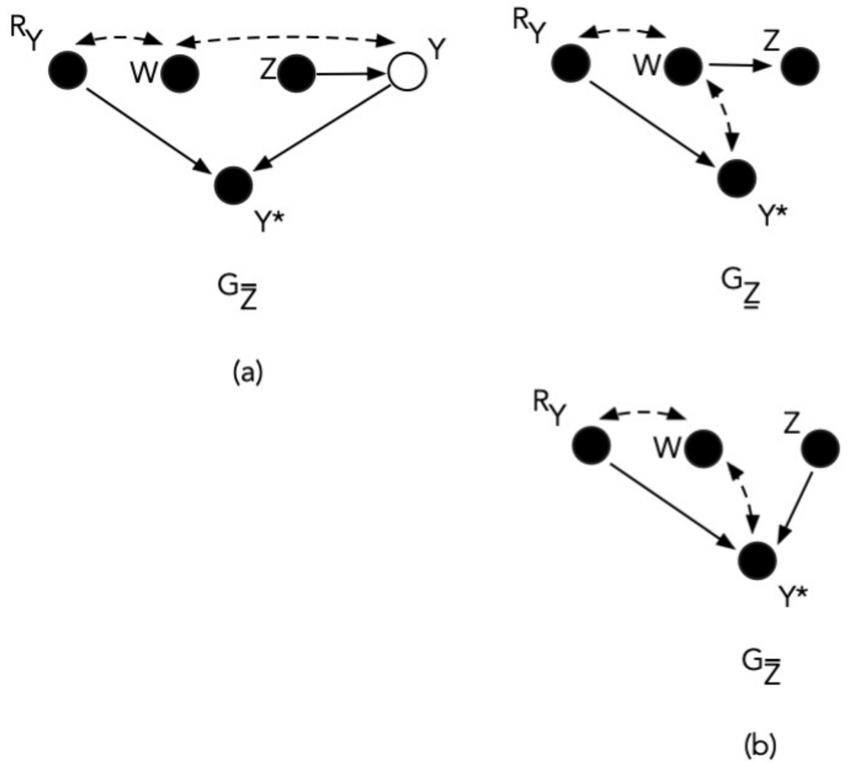


Figure 7 (a) and (b) subgraphs