Title: Zero Inflation as a Missing Data Problem: a Proxy-based Approach

Speakers: Trung Phung

Series: Quantum Foundations, Quantum Information

Date: September 16, 2024 - 2:10 PM

URL: https://pirsa.org/24090108

Zero Inflation as a Missing Data Problem: a Proxy-based Approach

Trung Phung, Jaron J.R. Lee, Opeyemi Oladapo-Shittu, Eili Y. Klein, Ayse Pinar Gurses, Susan M. Hannum, Kimberly Weems, Jill A. Marsteller, Sara E. Cosgrove, Sara C. Keller, Ilya Shpitser

Content

1. Zero Inflation Data
2. Review of missing data
3. Zero Inflation graphical models
4. Non-identifiability
5. Partial Identification
6. Application

What is Zero Inflation?

Age	Therapy	Device	Recorded infection
23	1	3	0
42	2	2	0
67	3	1	1
34	2	3	1

Covariates/Features Outcome

What is Zero Inflation?

			Current test results	If all tests were done
Age	Therapy	Device	Recorded infection	True infection
23	1	3	0	1
42	2	2	0	0
67	3	1	1	1
34	2	3	1	1

Covariates/Features | Outcome | Hidden

Goals in a Zero Inflation problem

Age	Therapy	Device	Recorded infection	True infection
23	1	3	0	1
42	2	2	0	0
67	3	1	1	1
34	2	3	1	1

Goals in a Zero Inflation problem

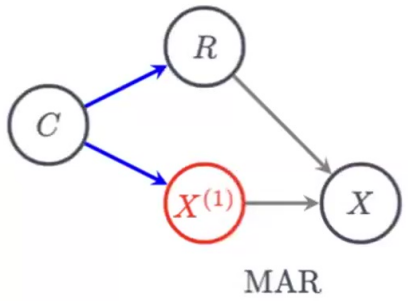
Age	Therapy	Device	Recorded infection	True infection
23	1	3	0	1
42	2	2	0	0
67	3	1	1	1
34	2	3	1	1

Which statistical model fit the data the best?
Pr(Age, Therapy, Device, Recorded Infection)

- Hurdle model
- ⇒ Zero-inflated Poisson ⇐
- …

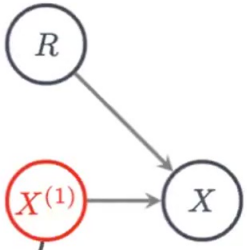
Missing data: the cousin

Age (C1)	Therapy (C2)	Device (C3)	Recorded infection (X)	True infection (X(1))	Indicator (R)
23	1	3	?	1	0
42	2	2	0	0	1
67	3	1	1	1	1
34	2	3	?	0	0



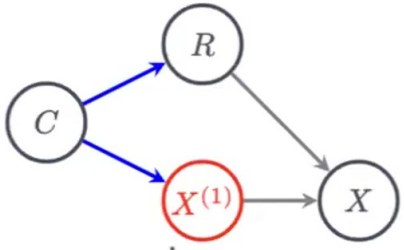
Consistency: when R=1, X = X(1), when R=0, X = ?

Missing data identification for the target law p(X(1))



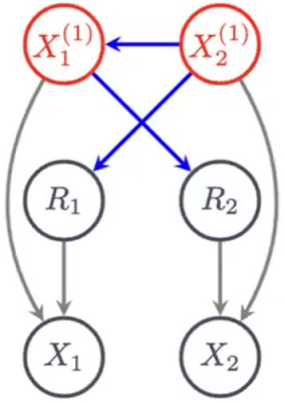
MCAR

$$p(X^{(1)}) = p(X \mid R = 1)$$



MAR

$$p(X^{(1)}) = \sum_c p(X \mid R = 1, c)p(c)$$



MNAR

If $\mathcal{G}$ has no self-censoring edge and no colluder, then $p(X^{(1)})$ is identified given $p(X, R)$

Nabi, Razieh, Rohit Bhattacharya, and Ilya Shpitser. 2020. "Full Law Identification in Graphical Models of Missing Data: Completeness Results." ICML.

Missing Data vs Zero Inflation

Missing data Consistency: when R=1, X = X(1), when R=0, X = ?

1) **Know** which values are incorrect (“?”)
2) Don’t know the true values for the incorrects

C1	C2	C3	X	R
23	1	3	**?**	0
42	2	2	0	1
67	3	1	1	1

ZI Consistency: when R=1, X = X(1), when R=0, X = 0

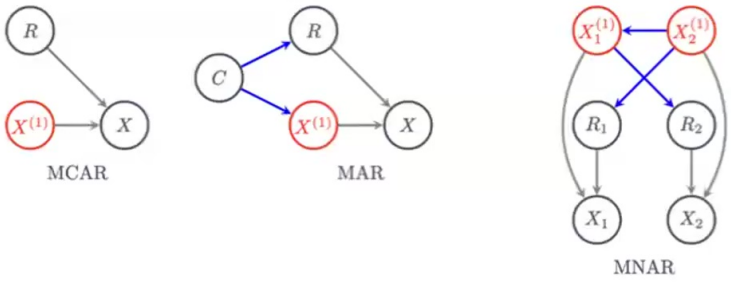
1) **Don’t know** which 0s are incorrect
2) Don’t know the true values for the incorrects

C1	C2	C3	X
23	1	3	**0**
42	2	2	**0**
67	3	1	1

Missing Data vs Zero Inflation

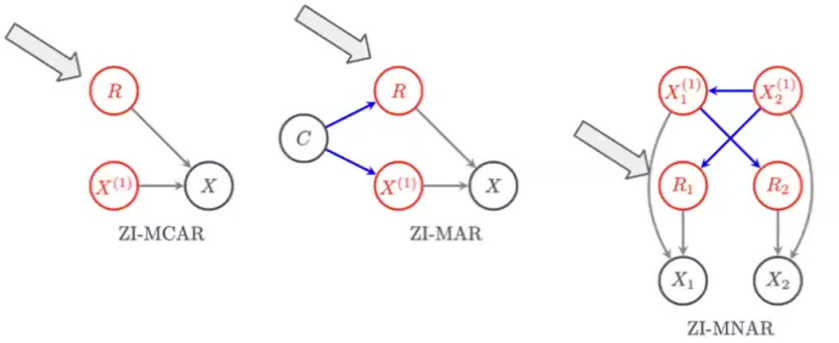
Missing data Consistency: when R=1, X = X(1), when R=0, X = ?

1) **Know** which values are incorrect (“?”)
2) Don’t know the true values for the incorrects



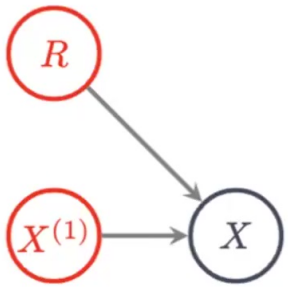
ZI Consistency: when R=1, X = X(1), when R=0, X = 0

1) **Don’t know** which 0s are incorrect
2) Don’t know the true values for the incorrects



Zero Inflation non-id

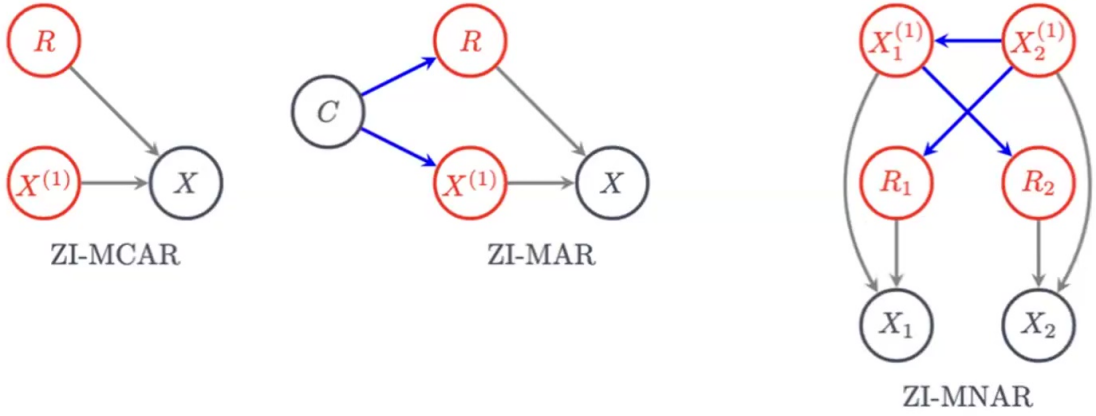
Now R is unobserved, is p(X(1)) identified from p(X)?



ZI-MCAR

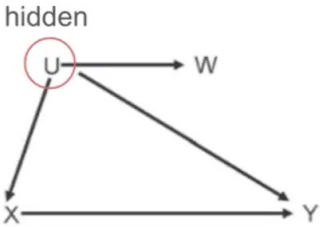
No. p(X(1)) in any ZI m-DAG is non-id non-parametrically.

Identification: what's next?



Recall: if we know p(X, R), we get identification ⇒ Find p(X,R)

Kuroki Pearl method



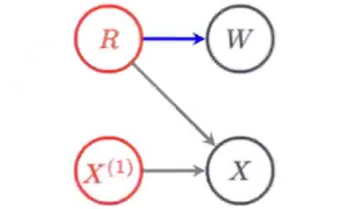
Question: Identify p(y|x,u) from p(x,y,w) and p(w|u)?

Observed proxy W of hidden U satisfying

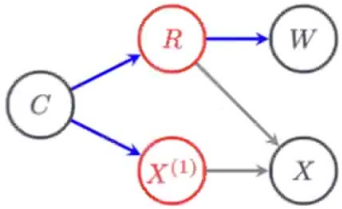
1. W is independent of everything given U
2. p(W|U) forms an invertible matrix
3. p(W|U) is given

Kuroki, Manabu, and Judea Pearl. 2014. “Measurement Bias and Effect Restoration in Causal Inference.” *Biometrika* 101 (2): 423–37.

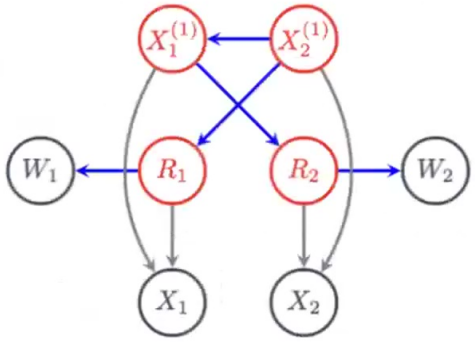
Kuroki Pearl method in ZI



Proxy-augmented ZI MCAR



Proxy-augmented ZI MAR



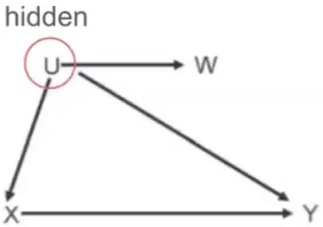
Proxy-augmented ZI MNAR

Question: Identify p(X, R) from p(X,W) and p(W|R)?

Conditions:

1) W indep others given R.
2) p(W|R) form an invertible matrix
3) p(W|R) is known.

Kuroki Pearl method



Question: Identify p(y|x,u) from p(x,y,w) and p(w|u)?

$$\mathrm{pr}(y, w \mid x) = \sum_{i=1}^{k} \mathrm{pr}(y, u_i \mid x)\mathrm{pr}(w \mid u_i).$$

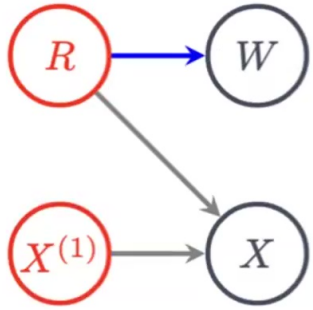
$$\underbrace{\begin{pmatrix} \mathrm{pr}(y, w_1 \mid x) \\ \vdots \\ \mathrm{pr}(y, w_k \mid x) \end{pmatrix}}_{V_{xy}(w)} = \underbrace{\begin{pmatrix} \mathrm{pr}(w_1 \mid u_1) & \cdots & \mathrm{pr}(w_1 \mid u_k) \\ \vdots & \ddots & \vdots \\ \mathrm{pr}(w_k \mid u_1) & \cdots & \mathrm{pr}(w_k \mid u_k) \end{pmatrix}}_{M(w,u)} \underbrace{\begin{pmatrix} \mathrm{pr}(y, u_1 \mid x) \\ \vdots \\ \mathrm{pr}(y, u_k \mid x) \end{pmatrix}}_{V_{xy}(u)}$$

$V_{xy}(w)$ (known) — $M(w,u)$ (known) — $V_{xy}(u)$ (unknown)

$$V_{xy}(u) = M(w, u)^{-1} V_{xy}(w).$$

Kuroki, Manabu, and Judea Pearl. 2014. “Measurement Bias and Effect Restoration in Causal Inference.” *Biometrika* 101 (2): 423–37.

Kuroki Pearl method in ZI



Proxy-augmented ZI MCAR

$$p(W, X) = \sum_r p(W \mid r) p(r, X)$$

$$\underbrace{\begin{pmatrix} p_{w_0,x_0} & p_{w_0,x_1} \\ p_{w_1,x_0} & p_{w_1,x_1} \end{pmatrix}}_{\mathbf{p}_{W,X}} = \underbrace{\begin{pmatrix} p_{w_0|r_0} & p_{w_0|r_1} \\ p_{w_1|r_0} & p_{w_1|r_1} \end{pmatrix}}_{\mathbf{p}_{W|R}} \underbrace{\begin{pmatrix} p_{r_0,x_0} & p_{r_0,x_1} \\ p_{r_1,x_0} & p_{r_1,x_1} \end{pmatrix}}_{\mathbf{p}_{R,X}}$$

$$\mathbf{p}_{R,X} = \mathbf{p}_{W|R}^{-1} \mathbf{p}_{W,X}$$

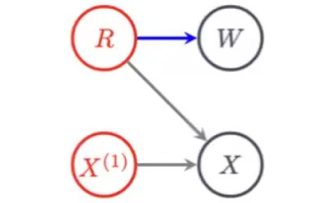
Kuroki-Pearl Step

Finally, identify p(X(1)) from p(x,w) and p(w|r)?

Know p(R, X) ⇒ know p(X(1)) = p(X|R=1)

missing data ID step

Kuroki Pearl method in ZI: Compatibility Problem



Proxy-augmented ZI MCAR

Conditions:

1) W indep others given R.
2) p(W|R) form an invertible matrix
3) p(W|R) is known.

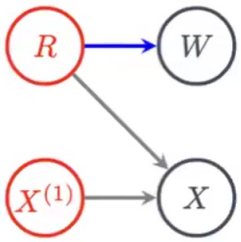
In ZI: p(W|R) is **not given** ⇒ guess

⇒ Not all choices of p(W|R) are compatible to the given p(W, X)
⇒ **Task:** Find compatible p(W|R)

Compatibility: There is some p(X(1), R, X, W) in the ZI model yielding both observed p(W,X) and the guess p(W|R).

Find compatibility set

Full model for p(X(1), X, R, W)



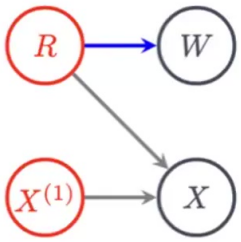
Proxy-augmented ZI MCAR

Model for marginal p(X, R, W)

$$\mathscr{P} = \left\{ (\mathbf{q}_{W|R}, \mathbf{q}_{RX}) \;\middle|\; \begin{array}{l} \mathbf{q}_{W|R} \geq 0, \forall r \left(\sum_w q_{w|r} = 1 \right), q_{w_0|r_0} \neq q_{w_0|r_1}, \\ \mathbf{q}_{RX} \geq 0, \sum_{rx} q_{rx} = 1, \forall x \neq 0 (q_{r_0 x} = 0) \end{array} \right\}.$$

⇒ Compatibility: There is some p(R, X, W) in the ZI model yielding both p(W,X) and p(W|R).

Compatibility bound



Proxy-augmented ZI MCAR

$$\underbrace{\begin{pmatrix} p_{w_0|r_0} & p_{w_0|r_1} \\ p_{w_1|r_0} & p_{w_1|r_1} \end{pmatrix}}_{\mathbf{p}_{W|R}}$$

identified

$$p_{w_0|r_1} = p_{w_0|x_1}$$

Find the compatible set by solving

$$\begin{aligned} \max_{q_{w_0|r_0}} \quad & \pm q_{w_0|r_0} \\ \text{s.t.} \quad & \mathbf{q}_{W|R}\mathbf{q}_{RX} = \mathbf{p}_{WX}, \\ & \mathbf{q}_{W|R} \geq 0, \forall r(\textstyle\sum_w q_{w|r} = 1), q_{w_0|r_0} \neq q_{w_0|r_1}, \\ & \mathbf{q}_{RX} \geq 0, \textstyle\sum_{rx} q_{rx} = 1, q_{w_0|r_1} = p_{w_0|x_1}. \end{aligned}$$

p(w=0|r=0) and p(RX) are free variables

⇒ quadratic program

⇒ Can be solve numerically

Compatibility bound: linearize

$$\begin{aligned}\max_{q_{w_0|r_0}} \quad & \pm q_{w_0|r_0} \\ \text{s.t.} \quad & \mathbf{q}_{W|R}\mathbf{q}_{RX} = \mathbf{p}_{WX}, \\ & \mathbf{q}_{W|R} \geq 0,\ \forall r\left(\textstyle\sum_w q_{w|r} = 1\right),\ q_{w_0|r_0} \neq q_{w_0|r_1}, \\ & \mathbf{q}_{RX} \geq 0,\ \textstyle\sum_{rx} q_{rx} = 1,\ q_{w_0|r_1} = p_{w_0|x_1}.\end{aligned}$$

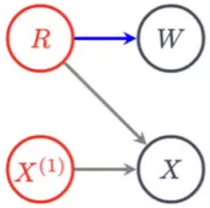
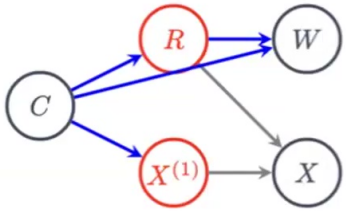
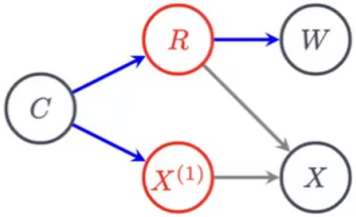
Just need to check it is a probability

$$\mathbf{q}_{RX} = [\mathbf{q}_{W|R}]^{-1}\mathbf{p}_{WX}.$$

$$\frac{1}{q_{w_0|r_0} - q_{w_0|r_1}} \begin{pmatrix} 1 - q_{w_0|r_1} & -q_{w_0|r_1} \\ q_{w_0|r_0} - 1 & q_{w_0|r_0} \end{pmatrix}$$

Linear program with only p(w=0|R=0) is the variable

$$\begin{aligned}\max_{q_{w_0|r_0}} \quad & \pm q_{w_0|r_0} \\ \text{s.t.} \quad & s \cdot \begin{pmatrix} 1 - q_{w_0|r_1} & -q_{w_0|r_1} \\ q_{w_0|r_0} - 1 & q_{w_0|r_0} \end{pmatrix} \mathbf{p}_{WX} \geq \mathbf{0}, \\ & s \cdot q_{w_0|r_0} > s \cdot q_{w_0|r_1}, 0 \leq q_{w_0|r_0} \leq 1, \\ & q_{w_0|r_1} = p_{w_0|x_1}.\end{aligned}$$

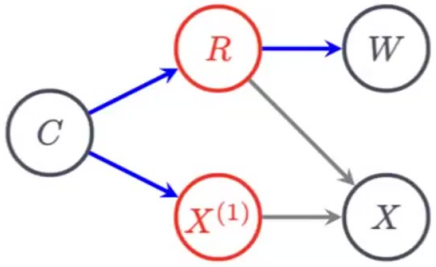
 Proxy-augmented ZI MCAR	 Proxy-augmented ZI MAR *	 Proxy-augmented ZI MAR
(A1) $W \perp\!\!\!\perp X^{(1)} \mid R$, **(A2)** The matrix $\mathbf{p}_{W\mid R}$ is invertible.	**(A1†)** $\forall i, W_i \perp\!\!\!\perp X^{(1)}, R_{-i} \mid C, R_i$. **(A2†)** The matrix $\mathbf{p}_{W\mid R,c}$ is invertible for every value c.	**(A1*)** $\forall i, W_i \perp\!\!\!\perp X^{(1)}, C, R_{-i} \mid R_i$. **(A2*)** The matrix $\mathbf{p}_{W\mid R}$ is invertible.
$\forall x \neq 0, p_{w_0\mid x} = p_{w_0\mid x_1}$	$\forall c, \forall x \neq 0, p_{w_0\mid x,c} = p_{w_0\mid x_1,c}$.	$\forall c, \forall x \neq 0, p_{w_0\mid x,c} = p_{w_0\mid x_1}$, *either* $\forall c \left(p_{w_0\mid x_0,c} \leq p_{w_0\mid x_1}\right)$ *or* $\forall c \left(p_{w_0\mid x_0,c} \geq p_{w_0\mid x_1}\right)$
$q_{w_0\mid r_1} = p_{w_0\mid x_1}$ $q_{w_0\mid r_0} \in \begin{cases} [p_{w_0\mid x_0}, 1] \text{ if } p_{w_0\mid x_0} > p_{w_0\mid x_1} \\ [0, p_{w_0\mid x_0}] \text{ if } p_{w_0\mid x_0} < p_{w_0\mid x_1} \\ (0,1) \setminus \{p_{w_0\mid x_0}\} \text{ if } p_{w_0\mid x_0} = p_{w_0\mid x_1} \end{cases}$	$q_{w_0\mid r_1,c} = p_{w_0\mid x_1,c}$ $q_{w_0\mid r_0,c} \in \begin{cases} [p_{w_0\mid x_0,c}, 1] \text{ if } p_{w_0\mid x_0,c} > p_{w_0\mid x_1,c} \\ [0, p_{w_0\mid x_0,c}] \text{ if } p_{w_0\mid x_0,c} < p_{w_0\mid x_1,c} \\ (0,1) \setminus \{p_{w_0,\mid x_0,c}\} \text{ if } p_{w_0\mid x_0,c} = p_{w_0\mid x_1,c} \end{cases}$	$q_{w_0\mid r_1} = p_{w_0\mid x_1}$ $q_{w_0\mid r_0} \in \begin{cases} [\max_c p_{w_0\mid x_0,c}, 1] \text{ if } \exists \tilde{c}, p_{w_0\mid x_0,\tilde{c}} > p_{w_0\mid x_1}, \\ [0, \min_c p_{w_0\mid x_0,c}] \text{ if } \exists \tilde{c}, p_{w_0\mid x_0,\tilde{c}} < p_{w_0\mid x_1}, \\ (0,1) \setminus \{p_{w_0\mid x_1}\} \text{ if } \forall c, p_{w_0\mid x_0,c} = p_{w_0\mid x_1}. \end{cases}$

CLABSI

Covariates C | Outcome (Y) | Proxy (W)

Peds	Chemotherapy	OPAT	TPN	OtherTherapy	PICC	Port	TunneledCVC	NHSN_CLABSI	EPIC
0	0	1	0	0	1	0	0	0	0
1	0	0	1	0	0	0	1	1	0
0	1	0	0	0	0	1	0	1	0
0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	1	0	0	1	0
0	0	0	1	0	0	1	0	1	0

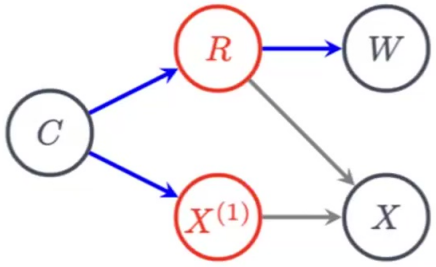
CLABSI



Proxy-augmented ZI MAR

Why R → W not W → R?

CLABSI

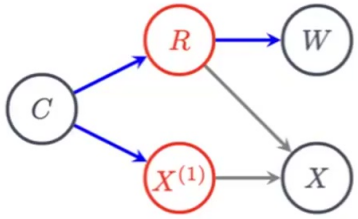


Proxy-augmented ZI MAR

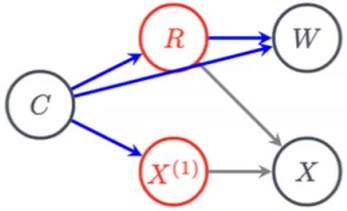
1. Full law model $p(X^{(1)}, X, R, W, C)$ Markov to the graph.
2. Fit $\hat{p}(X, W, C)$ using EM.
3. Compatible bound $p(W \mid R)$.
4. Grid search the bound
    1. for each $p(W \mid R)$ get a $p(R, X, C)$.
    2. True rate $p(X^{(1)}) = \sum_c p(X \mid R = 1, c)p(c)$.

$$q_{w_0|r_1} = p_{w_0|x_1}$$

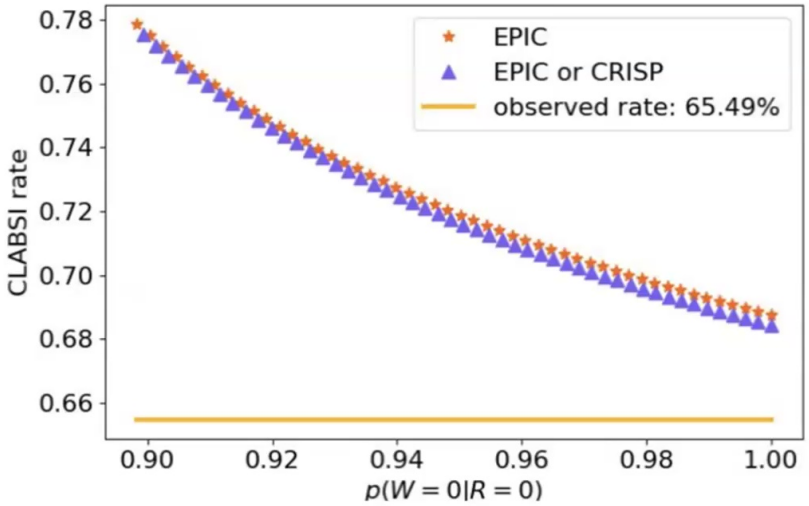
$$q_{w_0|r_0} \in \begin{cases} [\max_c p_{w_0|x_0,c}, 1] \text{ if } \exists \tilde{c}, p_{w_0|x_0,\tilde{c}} > p_{w_0|x_1} \\ [0, \min_c p_{w_0|x_0,c}] \text{ if } \exists \tilde{c}, p_{w_0|x_0,\tilde{c}} < p_{w_0|x_1} \\ (0,1) \setminus \{p_{w_0|x_1}\} \text{ if } \forall c, p_{w_0|x_0,c} = p_{w_0|x_1} \end{cases}$$



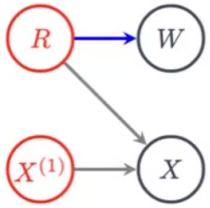
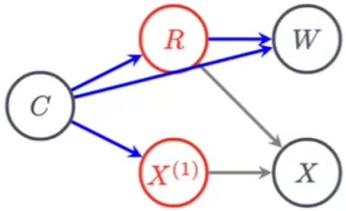
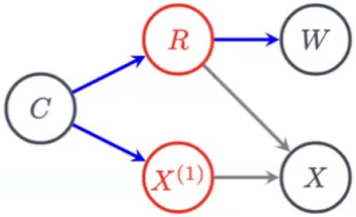
Proxy-augmented ZI MAR



Proxy-augmented ZI MAR *



CLABSI rate in [0.68, 1.0]

Proxy-augmented ZI MCAR	Proxy-augmented ZI MAR *	Proxy-augmented ZI MAR
		
(A1) $W \perp\!\!\!\perp X^{(1)} \mid R$, **(A2)** The matrix $\mathbf{p}_{W\mid R}$ is invertible.	**(A1†)** $\forall i, W_i \perp\!\!\!\perp X^{(1)}, R_{-i} \mid C, R_i$. **(A2†)** The matrix $\mathbf{p}_{W\mid R,c}$ is invertible for every value c.	**(A1*)** $\forall i, W_i \perp\!\!\!\perp X^{(1)}, C, R_{-i} \mid R_i$. **(A2*)** The matrix $\mathbf{p}_{W\mid R}$ is invertible.
$\forall x \neq 0, p_{w_0\mid x} = p_{w_0\mid x_1}$	$\forall c, \forall x \neq 0, p_{w_0\mid x,c} = p_{w_0\mid x_1,c}$.	$\forall c, \forall x \neq 0, p_{w_0\mid x,c} = p_{w_0\mid x_1}$, *either* $\forall c \left(p_{w_0\mid x_0,c} \leq p_{w_0\mid x_1}\right)$ *or* $\forall c \left(p_{w_0\mid x_0,c} \geq p_{w_0\mid x_1}\right)$
$q_{w_0\mid r_1} = p_{w_0\mid x_1}$ $q_{w_0\mid r_0} \in \begin{cases} [p_{w_0\mid x_0}, 1] \text{ if } p_{w_0\mid x_0} > p_{w_0\mid x_1} \\ [0, p_{w_0\mid x_0}] \text{ if } p_{w_0\mid x_0} < p_{w_0\mid x_1} \\ (0,1) \setminus \{p_{w_0\mid x_0}\} \text{ if } p_{w_0\mid x_0} = p_{w_0\mid x_1} \end{cases}$	$q_{w_0\mid r_1,c} = p_{w_0\mid x_1,c}$ $q_{w_0\mid r_0,c} \in \begin{cases} [p_{w_0\mid x_0,c}, 1] \text{ if } p_{w_0\mid x_0,c} > p_{w_0\mid x_1,c} \\ [0, p_{w_0\mid x_0,c}] \text{ if } p_{w_0\mid x_0,c} < p_{w_0\mid x_1,c} \\ (0,1) \setminus \{p_{w_0,\mid x_0,c}\} \text{ if } p_{w_0\mid x_0,c} = p_{w_0\mid x_1,c} \end{cases}$	$q_{w_0\mid r_1} = p_{w_0\mid x_1}$ $q_{w_0\mid r_0} \in \begin{cases} [\max_c p_{w_0\mid x_0,c}, 1] \text{ if } \exists \tilde{c}, p_{w_0\mid x_0,\tilde{c}} > p_{w_0\mid x_1}, \\ [0, \min_c p_{w_0\mid x_0,c}] \text{ if } \exists \tilde{c}, p_{w_0\mid x_0,\tilde{c}} < p_{w_0\mid x_1}, \\ (0,1) \setminus \{p_{w_0\mid x_1}\} \text{ if } \forall c, p_{w_0\mid x_0,c} = p_{w_0\mid x_1}. \end{cases}$