Observational Study

Yuan Meng

Meng-y16@mails.tsinghua.edu.cn

Story: London Taxi Drivers

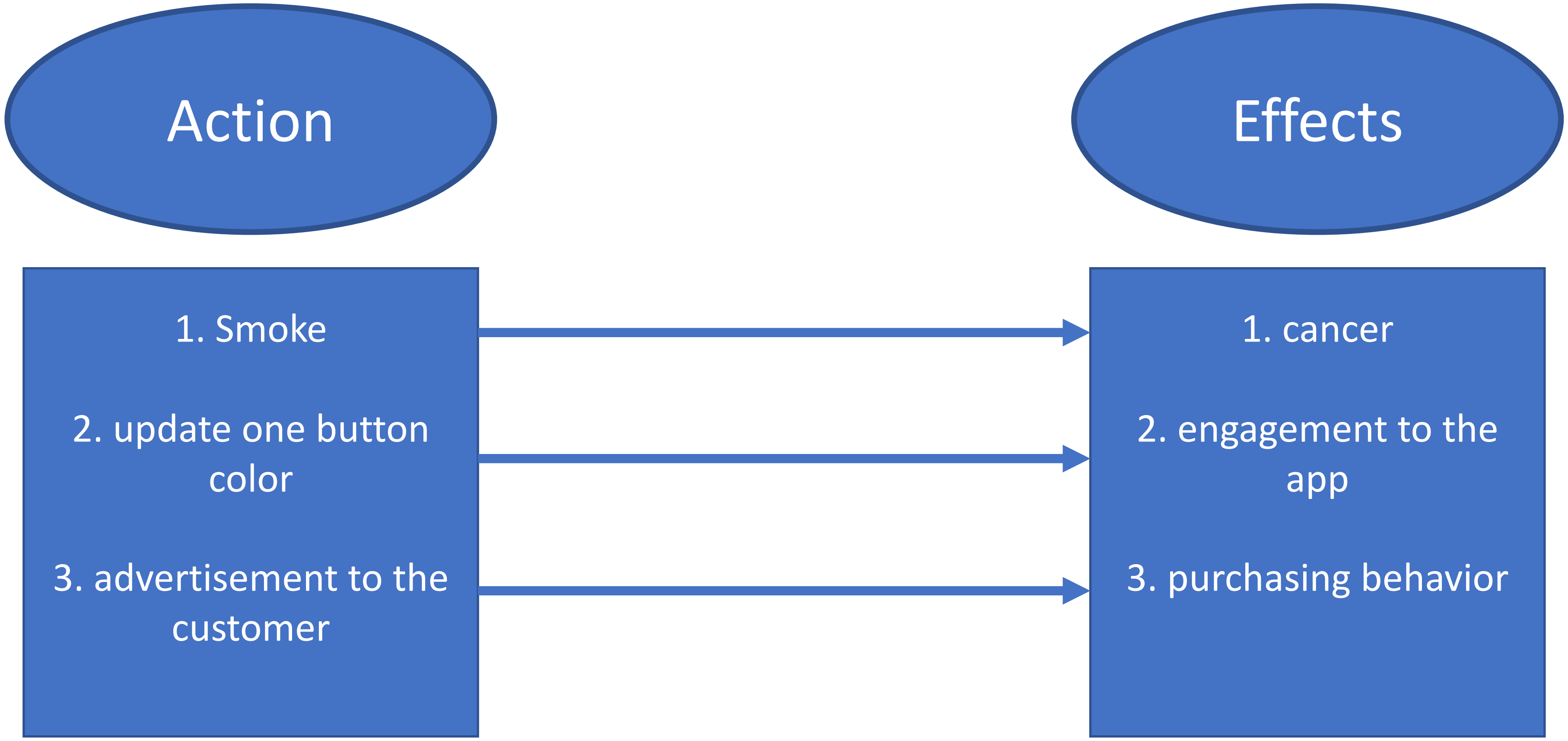
- Examples:

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Decision based on the causality ?

Causality examples (A causes B)

- Exposure/Action/Decision Effects



Causality——Rubin Causal Model(RCM)

average causal effect(ACE)

the treatment result for sample i, the control result for sample i,

$$ACE(Z \rightarrow Y) = E(Y_i(1) - Y_i(0)).$$

Treatment/control(untreatment) Potential outcome

However, it is often hard to obtain both Yi(1) and Yi(0) at the same time

We need to design Random Experiments (A/B tests) such that the distributions all variables (e.g. age, weight, height, gender, etc., excluding the treatment, e.g. smoking) have the same distribution in the treatment samples and control samples

$$\begin{aligned} ACE(Z \rightarrow Y) &= E(Y_i(1)) - E(Y_i(0)) \\ &= E(Y_i(1)|Z_i = 1) - E(Y_i(0)|Z_i = 0) \\ &= E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0), \end{aligned}$$

What if random experiments cannot be conducted?
e.g.:

- Too expensive
- Legally prohibited
- not ethnical
- There are large amount of existing and potentially useful data which were not generated as the result of a carefully designed random experiment.

Simpson's Paradox in naturally generated data

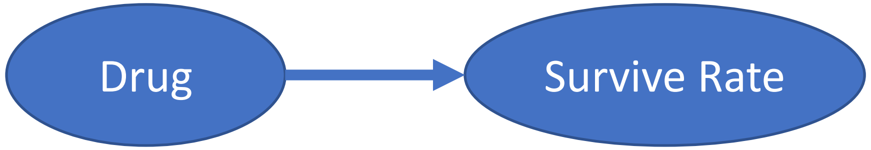


Table 1: Yule-Simpson's Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	**Survive**	**Die**	**Survive Rate**
Treatment	18	12	60%
Control	7	3	70%
Female	**Survive**	**Die**	**Survive Rate**
Treatment	2	8	20%
Control	9	21	30%

Treatment is better

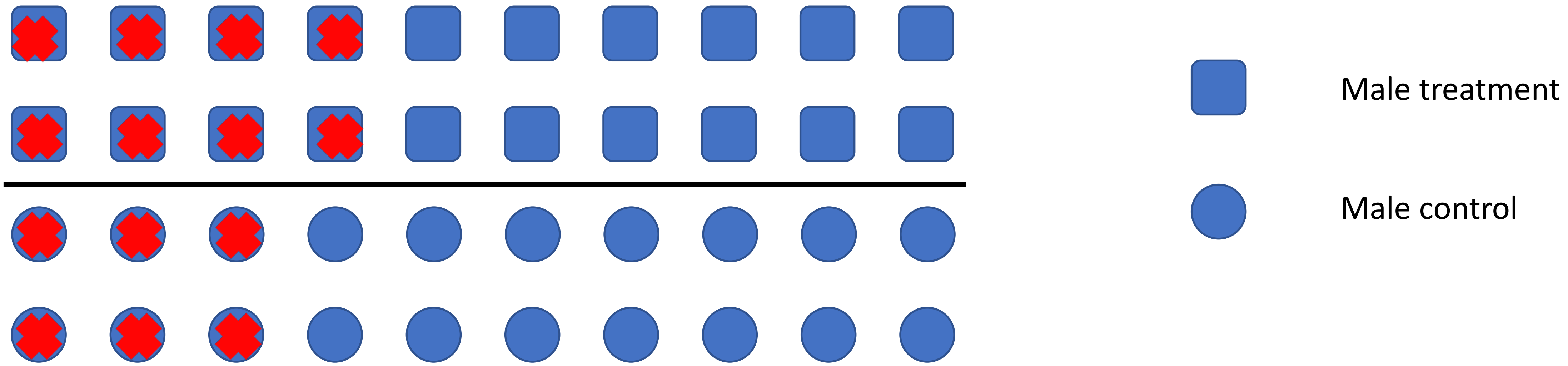
Control is better

Control is better

Simpson’s Paradox

Table 1: Yule-Simpson’s Paradox

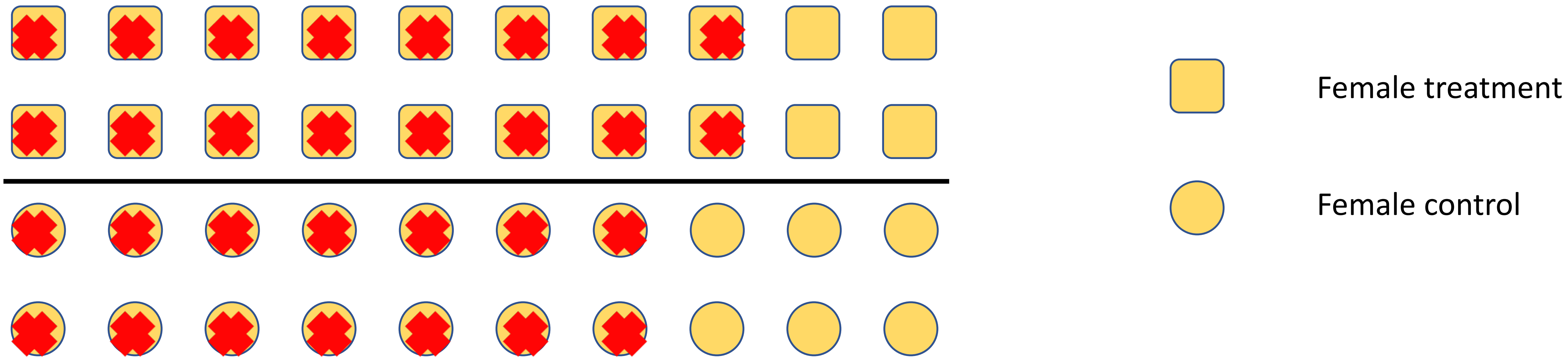
	Survive	Die	Survive Rate
Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%



Simpson's Paradox

Table 1: Yule-Simpson's Paradox

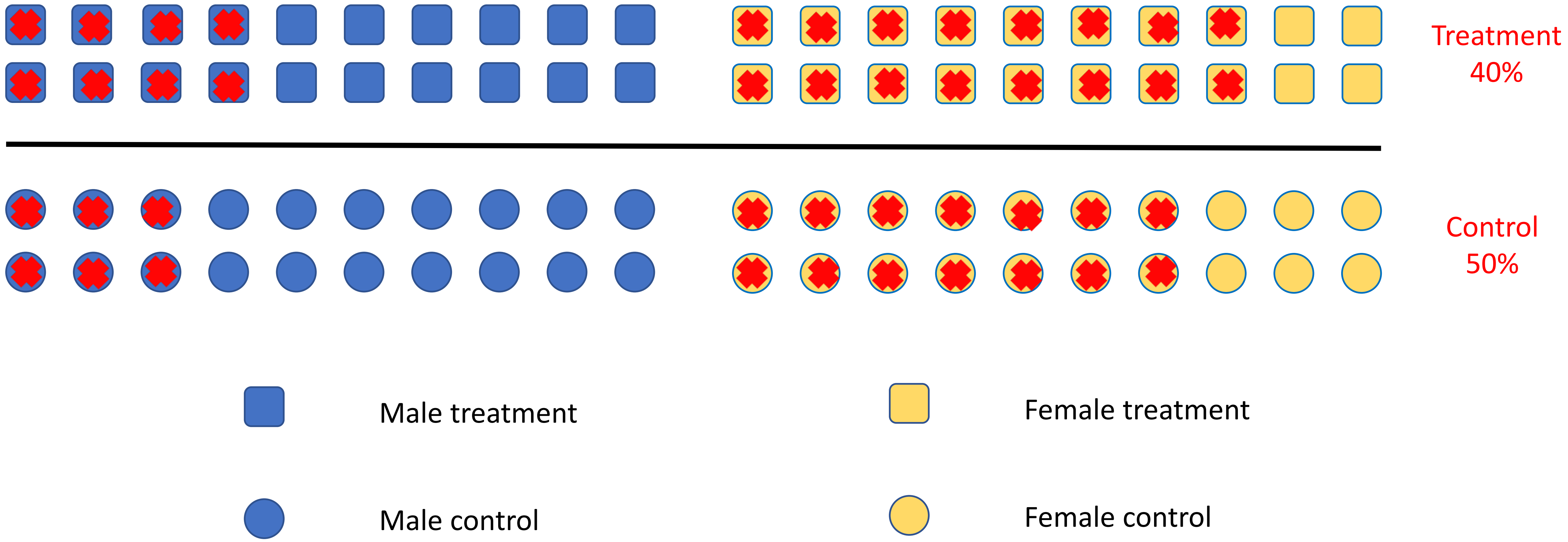
	Survive	Die	Survive Rate
Population			
Treatment	20	20	50%
Control	16	24	40%
Male			
Treatment	18	12	60%
Control	7	3	70%
Female			
Treatment	2	8	20%
Control	9	21	30%



Simpson's Paradox

Table 1: Yule-Simpson's Paradox

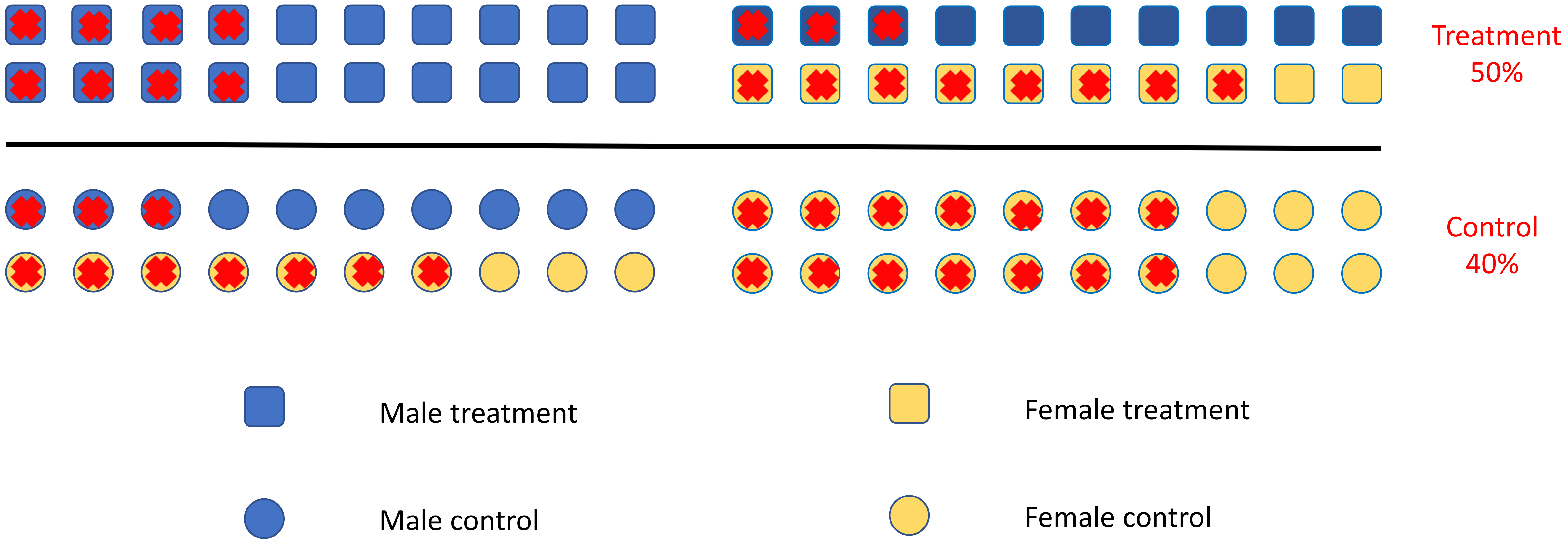
Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%



Simpson's Paradox

Table 1: Yule-Simpson's Paradox

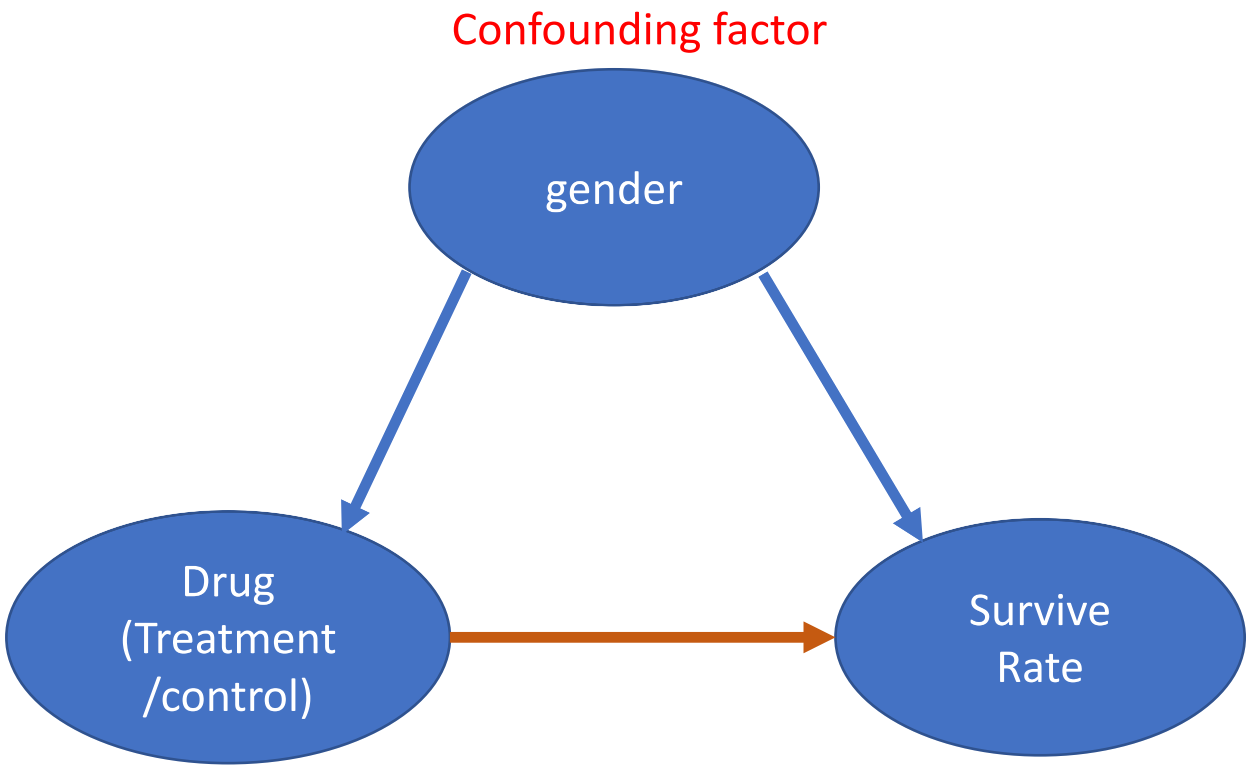
Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

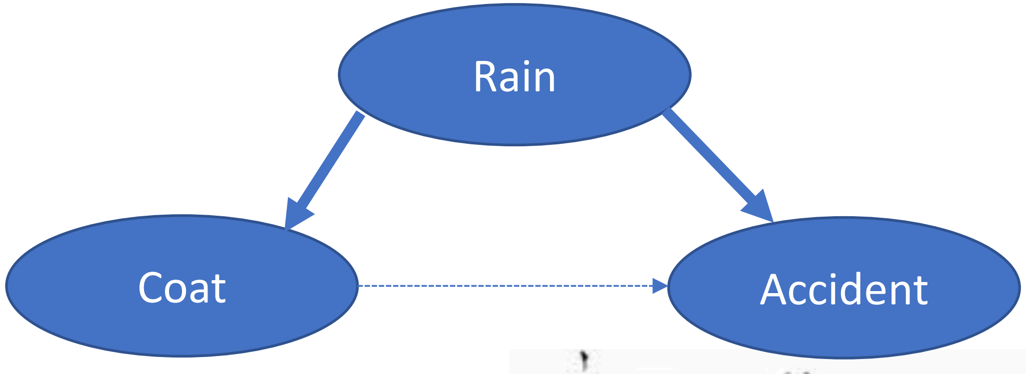


Observational Study

Table 1: Yule-Simpson's Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	**Survive**	**Die**	**Survive Rate**
Treatment	18	12	60%
Control	7	3	70%
Female	**Survive**	**Die**	**Survive Rate**
Treatment	2	8	20%
Control	9	21	30%





◆ Examples:

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Correlation is not causality
Causality really matters

How to deal with confounding factor?

- Fix the confounding factor, then conduct the analysis, then average the treatment effect based on the distribution of the confounding factor. E.g:
  - gender=female, analyze the treatment effect; gender=male, analyze the treatment effect; then analyze the overall effect based on the distribution of the gender.

$$
\begin{aligned}
ACE &= E(Y(1)) - E(Y(0)) \\
&= E[E(Y(1) \mid X)] - E[E(Y(0) \mid X)] \\
&= E[E(Y(1) \mid X, Z = 1)] - E[E(Y(0) \mid X, Z = 0)] \\
&= E[E(Y \mid X, Z = 1)] - E[E(Y \mid X, Z = 0)] \,.
\end{aligned}
$$

fix confounding factor X

average over confounding factor X

Myopia

- A study published in Nature [11] made the causal conclusion that children who sleep with the light on are more likely to develop myopia later in life.

  G. E. Quinn, C. H. Shin, M. G. Maguire, and R. A. Stone, “Myopia and ambient lighting at night,” Nature, vol. 399, no. 6732, pp. 113–113, 1999

- However, as it turns out, myopic parents tend to leave the light on more often, as well as pass their genetic predisposition to myopia to their children. Accounting for the confounding variable of parent’s myopia, the causal results were subsequently invalidated or substantially weakened.

  Gwiazda J, Ong E, Held R*, et al*. Myopia and ambient night-time lighting. *Nature* 2000;**404**:144.
  Zadnik K, Jones LA, Irvin BC*, et al*. Myopia and ambient night-time lighting. *Nature* 2000;**404**:143–4.

Observational Study

Table 1: Yule-Simpson's Paradox

	Survive	Die	Survive Rate
Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

$$\widehat{ACE}_{unadj} = \widehat{P}(Y=1 \mid Z=1) - \widehat{P}(Y=1 \mid Z=0)$$
$$= 0.50 - 0.40 = 0.10 > 0.$$

(0.6*(30/40)+0.2*(10/40))-(0.7*(10/40)+0.3*(30/40))=0.10

male

female

$$\widehat{ACE}_{adj}$$
$$= \{\widehat{P}(Y=1 \mid Z=1, X=1) - \widehat{P}(Y=1 \mid Z=0, X=1)\}\widehat{P}(X=1)$$
$$+ \{\widehat{P}(Y=1 \mid Z=1, X=0) - \widehat{P}(Y=1 \mid Z=0, X=0)\}\widehat{P}(X=0)$$
$$= (0.60 - 0.70) \times 0.5 + (0.20 - 0.30) \times 0.5$$
$$= -0.10 < 0.$$

(0.6*(20/40)+0.2*(20/40))-(0.7*(20/40)+0.3*(20/40))=-0.10

Methods for Observational study

- Propensity score for complex confounding factor

multi-dimensional confounding factor

e.g.:gender,weight,height

Propensity score: $e(X) = P(Z = 1 \mid X)$

$$\widehat{ACE} = \frac{1}{N} \sum_{i=1}^{n} \left[\frac{Y_i Z_i}{\widehat{e}(X_i)} - \frac{Y_i(1 - Z_i)}{1 - \widehat{e}(X_i)} \right].$$

The higher the propensity score of confounding factor X_i, the lower the weight for X_i.

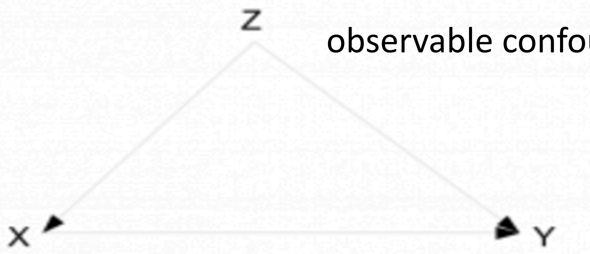
  - Ya Xu and Nanyu Chen. 2016. Evaluating Mobile Apps with A/B and Quasi A/B Tests. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, New York, NY, USA, 313-322. DOI: https://doi.org/10.1145/2939672.2939703

- Matched Design: Matching samples from treatment group and control group with the similar confounding factor
  - VIDEO QUALITY IMPACTS VIEWER BEHAVIOR

Observational study based on Causal Diagram ——Judea Pearl[1995]

Give the causal diagram from the domain knowledge, where the arrows represent the causal relationship between two variables, and the data from the real world (partial variables in the diagram are observable, then estimate the ACE based on the diagram).
Two typical structure to identify the ACE as follows

- Backdoor criterion

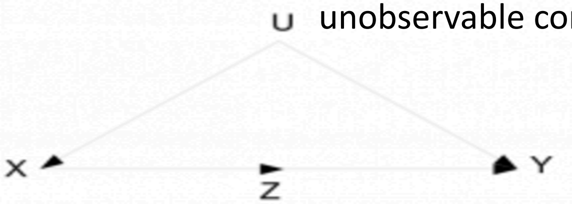


observable confounding factor

Such confounding factor can be dealt with using previously mentioned methods (adjustment, propensity score, matched design)

(a) The Backdoor Criterion for identifying ACE from X to Y is satisfied, because Z blocks all the backdoor paths.

- Front-door criterion



unobservable confounding factor

U: user i might want to buy the product anyway.

X: ads targeted at user i

Y: whether user i purchases the product

variable Z which is not influenced by the confounding factor U and on the path from X to Y

Regroup with Z, where Z indicates whether user i actually saw the ads

(b) The Frontdoor Criterion for identifying ACE from X to Y is satisfied, because Z blocks all the frontdoor paths.

Daniel N. Hill, Robert Moakler, Alan E. Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. 2015. Measuring Causal Impact of Online Actions via Natural Experiments: Application to Display Advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15). ACM, New York, NY, USA, 1839-1847. DOI: http://dx.doi.org/10.1145/2783258.2788622

Negative control——Detecting Confounding and Bias in Observational Studies

- negative controls—is designed to detect both suspected and unsuspected sources of spurious causal inference.

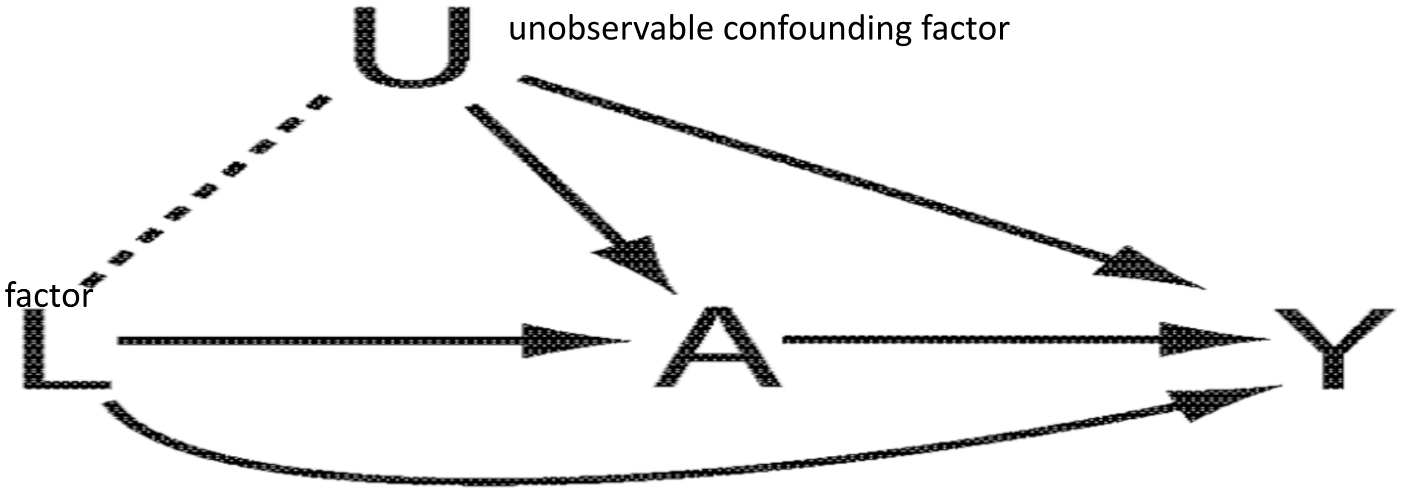


FIG 1.
Causal diagram for the effect of an exposure of interest (A) on an outcome of interest (Y), with confounders L (assumed measured) and U (assumed uncontrolled) that cause both A and Y. The dashed line between L and U indicates that either may cause the other, and they may share common causes.

Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass)*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb.

The negative control outcome.

- We want to know whether U affects A and A->Y, but U is unobservable.
- A negative control outcome (N) should be an outcome such that *the set of common causes of exposure A and outcome Y* should be as identical as possible to *the set of common causes of A and N*. Also N is not caused by A.

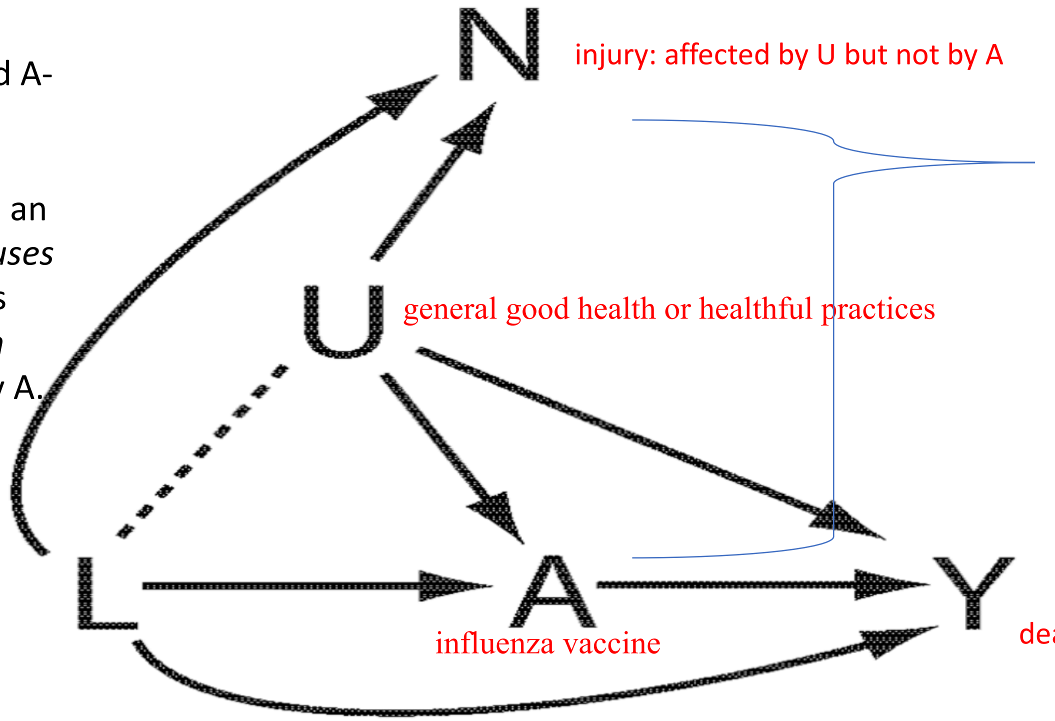


FIG 2.
Causal diagram showing an ideal negative control outcome N for use in evaluating studies of the causal relationship between exposure A and outcome Y. N should ideally have the same incoming arrows as Y, except that A does not cause N; to the extent this criterion is met, N is called U-comparable to Y.

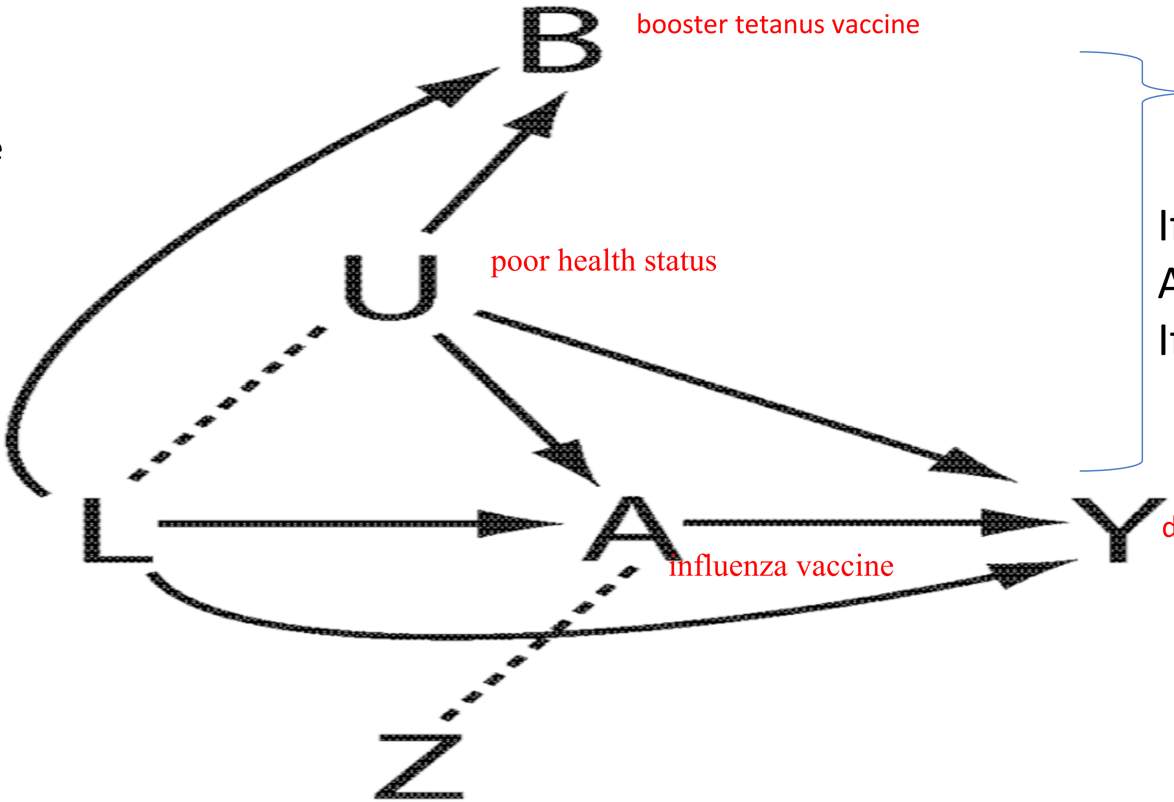
- Negative control solution: Test whether there is association between A & N

If yes, U is a confounding factor for A->Y;
If no, A->Y is not biased by U.

Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass)*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb.

The negative control exposure.

- A negative control exposure B should be an exposure such that *the common causes of A and Y* are as nearly identical as possible to the *common causes of B and Y .* And B does not cause cause Y



- Negative control solution: Test whether there is association between B&Y.

If yes, U is a confounding factor for A->Y;
If no, A->Y is not biased by U.

FIG 3.
Causal diagram showing an ideal negative control exposure B for use in evaluating studies of the causal relationship between exposure A and outcome Y. B should ideally have the same incoming arrows as A; to the extent this criterion is met, B is called U-comparable to A. Z is an instrumental variable of the A-Y relationship and is depicted to illustrate the difference between an instrumental variable and a negative control variable.

Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass)*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb.

Thanks