

# Confounding II

Mabel Carabali

EBOH, McGill University

01-08-2022 Updated: ( 2024-09-25)

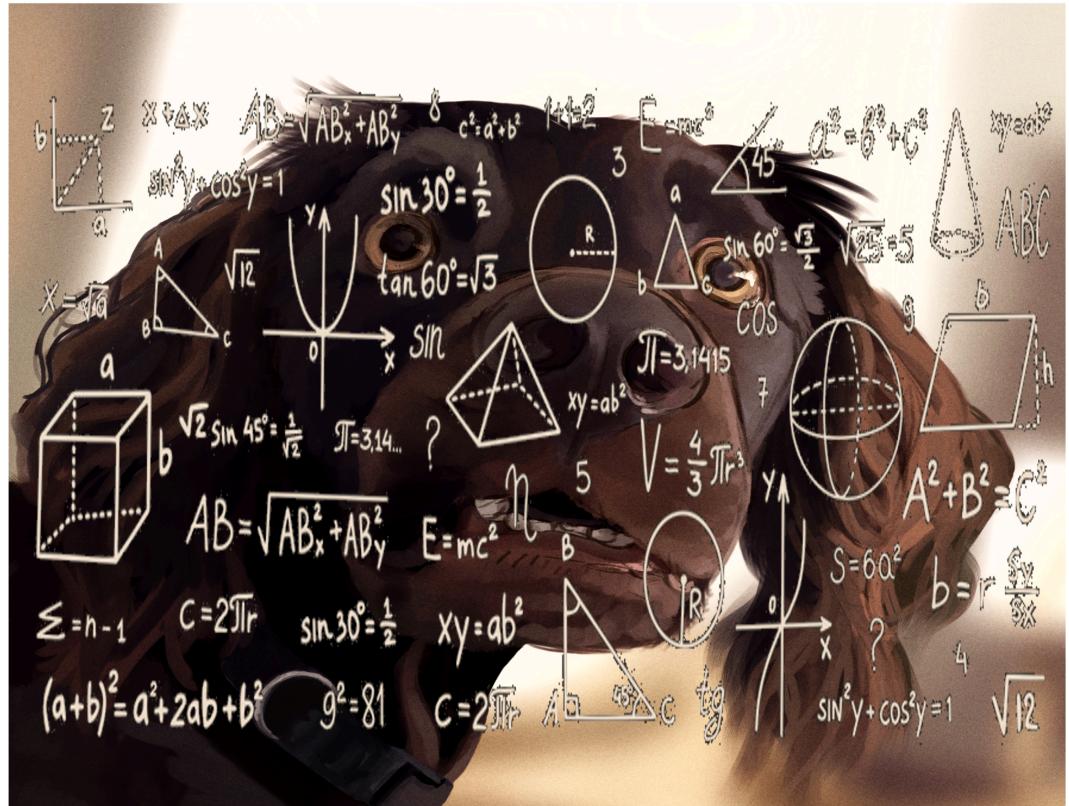
# The Structure of Confounding and worked examples !

## Conditions that allow a variable to be a confounder:

► Modern Epidemiology 4th, page 268

*The developments in causal inference over the past decades, summarized in Chapter 3, have made clear that this definition [...] the traditional criteria described from ME3... ] of a "confounder" is inadequate. It is inadequate because there can be a pre-exposure variable associated with the exposure and the outcome, the control of which introduces, rather than eliminates, bias [ME4;p268]*

# The Structure of Confounding??



# The Structure of Confounding

$$A \leftarrow L \rightarrow Y$$

This diagram shows two sources of association between treatment and outcome:

1. The path  $A \rightarrow Y$  that represents the causal effect of A on Y , and
2. The path  $A \leftarrow L \rightarrow Y$  between A and Y that includes the common cause  $L$ 
  - The path  $A \leftarrow L \rightarrow Y$  links A and Y through the common cause  $L$ , is the "**backdoor path**"

# The structure of Confounding

- In a causal DAG, a backdoor path is a non-causal path between treatment and outcome that remains even if all arrows pointing from treatment to other variables (i.e., the descendants of treatment) are removed.
- That is, the path has an arrow pointing into treatment.

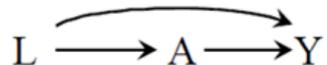


Figure 7.1

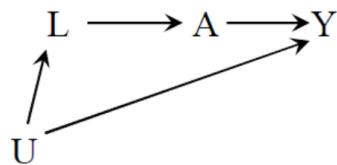


Figure 7.2

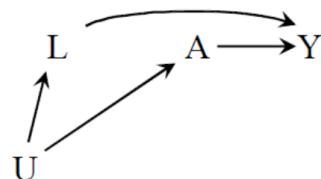


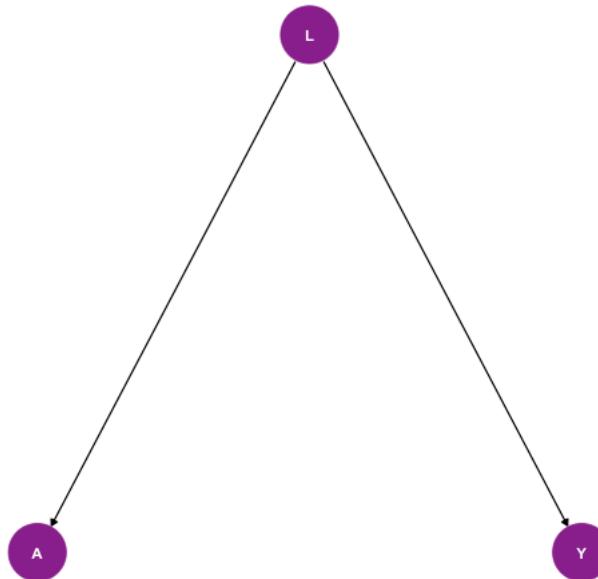
Figure 7.3

## Confounding and exchangeability

- The backdoor criterion, **does not** answer questions regarding the magnitude or direction of confounding.
- It is possible that some unblocked backdoor paths are weak and thus induce little bias, or that several strong backdoor paths induce bias in opposite directions and thus result in a weak net bias.
- Because unmeasured confounding is not an “all or nothing” issue, in practice, it is important to consider the expected direction and magnitude of the bias.

# Confounders ( $Y \leftarrow L \rightarrow A$ )

DAG Simple Confounding



## Simulated Example

```
set.seed(704); N <- 100;
L <- rbinom(N,1,0.5)
A <- ifelse(L==0,rbinom(N,1,0.25),
            rbinom(N,1,0.75))
Y <- ifelse(L==0,rbinom(N,1,0.20),
            rbinom(N,1,0.8))
#summary(L)
data <- data.frame(N, A, L, Y)
tab <- table(data$A, data$Y)
#tab; tab/margin.table(tab)
l6conf1<-epi.2by2(tab,
                     method = "cohort.count")
```

	Outcome +	Outcome -	Total	Inc risk *
## Exposed +	39	13	52	75.00 (61.05 to 85.97)
## Exposed -	14	34	48	29.17 (16.95 to 44.06)
## Total	53	47	100	53.00 (42.76 to 63.06)

\*Outcomes per 100 population units

# Confounders

## Crude

```
tbl6conf1 <- data.table::as.data.table(l6c  
kable(tbl6conf1, digits = 2) %>%  
  kable_paper()
```

var	est	lower	upper
Inc risk ratio	2.57	1.61	4.11
Inc odds ratio	7.29	3.01	17.63
Attrib inc risk *	45.83	28.40	63.26
Attrib fraction in exposed (%)	61.11	37.90	75.64
Attrib inc risk in population *	23.83	7.68	39.99
Attrib fraction in population (%)	44.97	21.55	61.40

```
l6strtab1<- data %>%  
 tbl_summary(by= L,  
  label=list(Y ="Outcome", A ="Exposure"),  
  #type = all_continuous() ~ "continuous1",  
  statistic = all_categorical() ~ c( "{n}"  
  missing = "no") %>%  
  modify_spanning_header(c("stat_1", "stat_2")) %>%  
  modify_caption("**Summary of covars distribution")  
l6strtab1
```

## Summary of covars distribution

L=0/L=1

Characteristic	1 > 0	1 > 1
	N = 51 <sup>1</sup>	N = 49 <sup>1</sup>

N

100	51 / 51 (100%)	49 / 49 (100%)
-----	----------------	----------------

Exposure	10 / 51 (20%)	38 / 49 (78%)
----------	---------------	---------------

Outcome	9 / 51 (18%)	38 / 49 (78%)
---------	--------------	---------------

<sup>1</sup> n / N (%)

# Confounders

L=0

Exposure	Outcome; L=0	
	0	1
0	36	5
1	6	4

```
tab1 <- table(data$A, data$Y, data$L)
#tab1
l6conf2<-epi.2by2(tab1, method = "cohort.count")
tabl6conf2 <- data.table::as.data.table(l6conf2$massoc.summary)
```

L=1

Exposure	Outcome; L=1	
	0	1
0	3	8
1	8	30

# Confounders

## Adjusted

Measure	Estimate 95% CIs		
	Est.	LB	UB
Measure	Est.	LB	UB
Inc risk ratio (crude)	2.57	1.61	4.11
Inc risk ratio (M-H)	1.42	0.87	2.30
Inc risk ratio (crude:M-H)	1.81		
Inc odds ratio (crude)	7.29	3.01	17.63
Inc odds ratio (M-H)	2.46	0.84	7.21
Inc odds ratio (crude:M-H)	2.96		
Attrib inc risk (crude) *	45.83	28.40	63.26
Attrib inc risk (M-H) *	16.69	-16.33	49.71
Attrib inc risk (crude:M-H)	2.75		

\*Outcomes per 100 population units

# Confounders?

Consider this DAG:

$$C \rightarrow E \rightarrow Y$$

- In this case, C is not a confounder because it does not have an independent effect on Y.
  - But there will be an observed association between C and Y, by virtue of their common association with E.
  - But it is not an independent association.

**That's why we should assess this criterion within levels of exposure.**

- Stratified by E, the association between C and Y is null if there is no direct effect (as shown in the DAG).

# Confounders ?

$$C \rightarrow E \rightarrow Y$$

```
set.seed(704)
N <- 100
C <- rbinom(N, 1, 0.5)
E <- ifelse(C==0, rbinom(N, 1, 0.8),
            rbinom(N, 1, 0.5))
Y <- ifelse(E==0, rbinom(N, 1, 0.2),
            rbinom(N, 1, 0.5))
#summary(C)
data1 <- data.frame(N, C, E, Y)
tab1C <- table(data1$E, data1$Y, data1$C)
```

Measure	Estimate 95% CIs		
	Est.	LB	UB
Inc risk ratio (crude)	1.91	1.36	2.70
Inc risk ratio (M-H)	1.75	1.14	2.69
Inc risk ratio (crude:M-H)	1.09		
Inc odds ratio (crude)	5.12	2.02	12.97
Inc odds ratio (M-H)	3.83	1.45	10.09
Inc odds ratio (crude:M-H)	1.34		
Attrib inc risk (crude) *	37.15	19.01	55.30
Attrib inc risk (M-H) *	31.20	-6.71	69.12
Attrib inc risk (crude:M-H)	1.19		

# Confounders ?

Figure 7.4 A version of the famous M-diagram again. No confounding, despite backdoor paths.

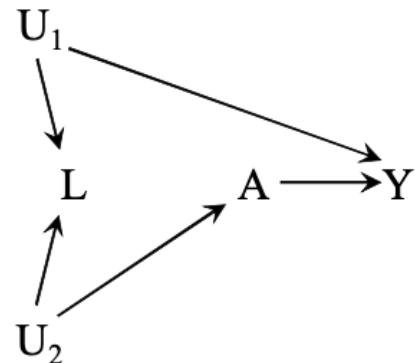


Figure 7.4

Here there are no common causes of treatment A and outcome Y, and therefore there is no confounding.

The back door path between  $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$  is locked because  $L$  is a collider on that path.

# Confounders

No common causes but L is a collider  
 $U_1 \rightarrow L \leftarrow U_2$

```
set.seed(704)
N <- 100
U1 <- rbinom(N, 1, 0.5)
U2 <- rbinom(N, 1, 0.5)
L <- ifelse(U1==1, rbinom(N, 1, 0.6),
            ifelse(U2==1, rbinom(N, 1, 0.6),
                   rbinom(N, 1, 0.5))) #L is affected by U1
A <- ifelse(U2==1, rbinom(N, 1, 0.5),
            rbinom(N, 1, 0.5)) #A is affected by U2
Y <- ifelse(A==1, rbinom(N, 1, 0.6),
            ifelse( U1==1, rbinom(N, 1, 0.6),
                   rbinom(N, 1, 0.5))) # Y is affected by A

#summary(C)
datanoconf2 <- data.frame(N, U1, U2, L, A, Y
tab.noconf2<- table(datanoconf2$A,
                     datanoconf2$Y,
                     datanoconf2$L)
```

Measure	Estimate 95% CIs		
	Est.	LB	UB
Measure	Est.	LB	UB
Inc risk ratio (crude)	1.73	1.05	2.86
Inc risk ratio (M-H)	1.70	1.04	2.78
Inc risk ratio (crude:M-H)	1.02		
Inc odds ratio (crude)	2.53	1.11	5.74
Inc odds ratio (M-H)	2.46	1.08	5.61
Inc odds ratio (crude:M-H)	1.03		
Attrib inc risk (crude) *	22.00	3.21	40.79
Attrib inc risk (M-H) *	21.33	0.74	41.92
Attrib inc risk (crude:M-H)	1.03		

# Confounders

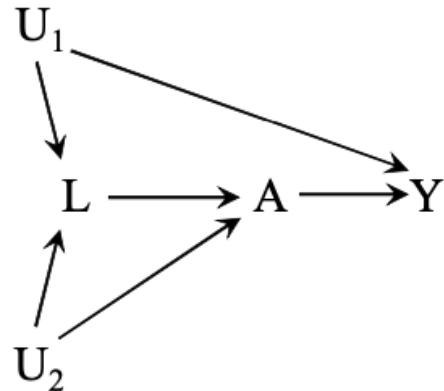


Figure 7.5

There is an arrow  $L \rightarrow A$ . The presence of this arrow creates an open backdoor path:

- $A \leftarrow L \leftarrow U_1 \rightarrow Y$ , because  $U_1$  is a common cause of  $A$  and  $Y$ , and so **confounding exists**.
- Conditioning on  $L$  would block that backdoor path but would simultaneously open a backdoor path on which  $L$  is a collider ( $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$ )

The bias is **intractable**: attempting to block the confounding path opens a selection bias path.

# Confounders

```
set.seed(704)
N <- 100
U1 <- rbinom(N, 1, 0.5)
U2 <- rbinom(N, 1, 0.5)
L <- ifelse(U1==1, rbinom(N, 1, 0.65),
            ifelse(U2==1, rbinom(N, 1, 0.65),
                   rbinom(N, 1, 0.15))) #L is affected by U
A <- ifelse(L==1, rbinom(N, 1, 0.65),
            ifelse(U2==1, rbinom(N, 1, 0.65),
                   rbinom(N, 1, 0.45))) #A is affected by L
Y <- ifelse(A==1, rbinom(N, 1, 0.65),
            ifelse(U1==1, rbinom(N, 1, 0.6),
                   rbinom(N, 1, 0.3))) # Y is affected by A

#summary(C)
data2 <- data.frame(N, U1, U2, L, A, Y)
tabL.intract <- table(data2$A, data2$Y,
                      data2$L)
```

Measure	Estimate 95%CIs		
	Est.	LB	UB
Inc risk ratio (crude)	1.91	1.16	3.13
Inc risk ratio (M-H)	2.12	1.18	3.80
Inc risk ratio (crude:M-H)	0.90		
Inc odds ratio (crude)	3.00	1.31	6.88
Inc odds ratio (M-H)	3.13	1.35	7.28
Inc odds ratio (crude:M-H)	0.96		
Attrib inc risk (crude) *	25.97	7.09	44.85
Attrib inc risk (M-H) *	28.01	5.35	50.67
Attrib inc risk (crude:M-H)	0.93		

# Confounders

Figure 7.7 is another non confounding example in which the traditional criteria lead to selection bias due to adjustment for L.

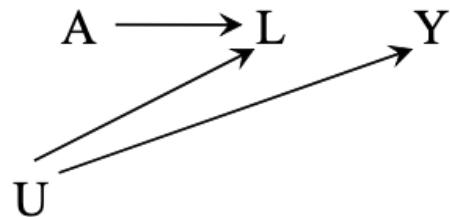


Figure 7.7

- The traditional criteria would not have resulted in bias had condition (3) been replaced by the condition that L is not caused by treatment.
  - *(3) it does not lie on a causal pathway between treatment and outcome.*

Replace condition (3) by the condition that “there exist variables A and Y such that there is conditional exchangeability within their joint levels  $Y^a \perp A|L, U$ ”. H&R, Technical Point 7.2

# Confounders

L is not on the "pathway"  $A \rightarrow Y$

```
set.seed(704)
N <- 100
U <- rbinom(N, 1, 0.5)
A <- rbinom(N, 1, 0.55) #A affects L
L <- ifelse(U==1, rbinom(N, 1, 0.65),
             ifelse(A==1, rbinom(N, 1, 0.65),
                    rbinom(N, 1, 0.25))) #L is affected by U
Y <- ifelse(U==1, rbinom(N, 1, 0.6),
             rbinom(N, 1, 0.25)) # Y is affected by U

datanoconf3 <- data.frame(N, U, L, A, Y)
tabL.noconf3 <- table(datanoconf3$A,
                      datanoconf3$Y,
                      datanoconf3$L)
```

Measure	Estimate 95% CIs		
	Est.	LB	UB
Inc risk ratio (crude)	0.98	0.67	1.42
Inc risk ratio (M-H)	0.98	0.61	1.57
Inc risk ratio (crude:M-H)	1.00		
Inc odds ratio (crude)	0.96	0.42	2.18
Inc odds ratio (M-H)	0.97	0.42	2.23
Inc odds ratio (crude:M-H)	0.99		
Attrib inc risk (crude) *	-1.10	-21.55	19.36
Attrib inc risk (M-H) *	-0.85	-25.69	23.99
Attrib inc risk (crude:M-H)	1.30		

## Surrogate confounders (Is L a confounder?)

In Figure 7.8, confounding of A on Y via unmeasured common cause U .

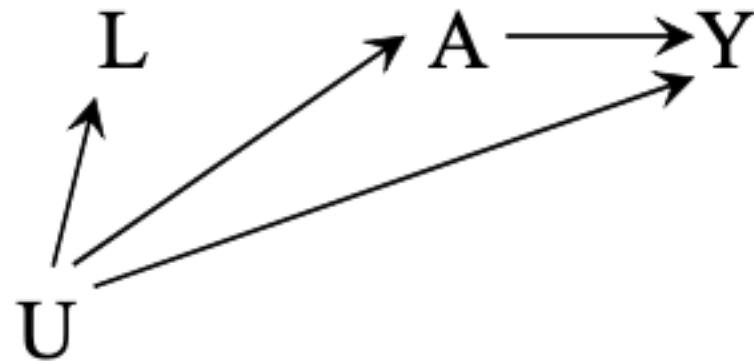


Figure 7.8

- Measured variable L is a proxy or surrogate for U . Adjust for the variable L?
- On the one hand, L is not a confounder because it does not lie on a backdoor path between A and Y .

# Confounders

## Surrogates when L is not highly correlated with U

```
set.seed(704)
N <- 100
U <- rbinom(N, 1, 0.8)
A <- ifelse(U==1, rbinom(N, 1, 0.65),
             rbinom(N, 1, 0.5)) #A is affected by U
L <- ifelse(U==1, rbinom(N, 1, 0.65),
             rbinom(N, 1, 0.5)) #L is affected by U
Y <- ifelse(A==1, rbinom(N, 1, 0.65),
            ifelse(U==1, rbinom(N, 1, 0.65),
                   rbinom(N, 1, 0.15))) # Y is affected by
dataconf4 <- data.frame(N, U, L, A, Y)
tabL.conf4 <- table(dataconf4$A,
                      dataconf4$Y,
                      dataconf4$L)
```

Measure	Estimate 95% CIs		
	Est.	LB	UB
Inc risk ratio (crude)	1.94	1.22	3.08
Inc risk ratio (M-H)	2.01	1.23	3.30
Inc risk ratio (crude:M-H)	0.97		
Inc odds ratio (crude)	3.29	1.39	7.78
Inc odds ratio (M-H)	3.46	1.41	8.50
Inc odds ratio (crude:M-H)	0.95		
Attrib inc risk (crude) *	28.52	8.61	48.44
Attrib inc risk (M-H) *	29.58	-5.76	64.92
Attrib inc risk (crude:M-H)	0.96		

## Surrogate confounders (Is L a confounder?)

- On the other hand, adjusting for L, which is associated with U , will indirectly adjust for some of the confounding caused by U .

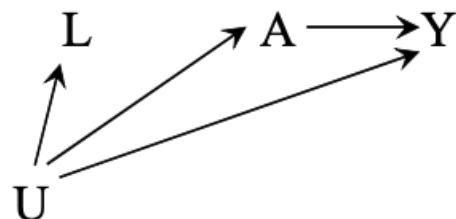


Figure 7.8

- In the extreme case that L were perfectly correlated with U then adjusting for L = adjusting for U.
- Therefore we will typically prefer to adjust, rather than not to adjust, for L.

# Confounders

## Surrogates when L and U correlated

```
set.seed(704)
N <- 100
U <- rbinom(N,1,0.8)
A <- ifelse(U==1, rbinom(N,1,0.65),
rbinom(N,1,0.5)) #A is affected by U
L <- ifelse(U==1, rbinom(N,1,0.95),
rbinom(N,1,0.5)) #L is affected by U
Y <- ifelse(A==1, rbinom(N,1, 0.65),
ifelse(U==1, rbinom(N,1,0.65),
rbinom(N,1,0.15))) # Y is affected by U and
dataconf5 <- data.frame(N, U, L, A,Y)
tabL.conf5 <- table(dataconf5$A,
dataconf5$Y,
dataconf5$L)
```

Measure	Estimate 95%CIs		
Measure	Est.	LB	UB
Inc risk ratio (crude)	1.94	1.22	3.08
Inc risk ratio (M-H)	1.74	1.06	2.88
Inc risk ratio (crude:M-H)	1.11		
Inc odds ratio (crude)	3.29	1.39	7.78
Inc odds ratio (M-H)	2.69	1.11	6.53
Inc odds ratio (crude:M-H)	1.22		
Attrib inc risk (crude) *	28.52	8.61	48.44
Attrib inc risk (M-H) *	23.16	-8.90	55.21
Attrib inc risk (crude:M-H)	1.23		

# Confounders cannot be descendants of treatment, but can be in the future of treatment

In Figure 7.11. L is a descendant of treatment A that blocks all backdoor paths from A to Y.

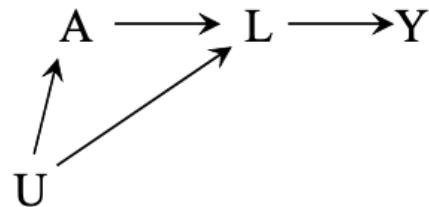


Figure 7.11

- Conditioning on L does not cause selection bias because no collider path is opened.
- Since the causal effect of A on Y is only through L, conditioning on L completely blocks this pathway.
- This shows that adjusting for a variable L that blocks all backdoor paths does not eliminate bias when L is a descendant of A.
- Since  $Y^a \perp\!\!\!\perp A|L$  implies adjustment for L eliminates all bias, there must not be conditional exchangeability,
- **and thus  $E[Ya = 1] - E[Ya = 0]$  is not identified.**

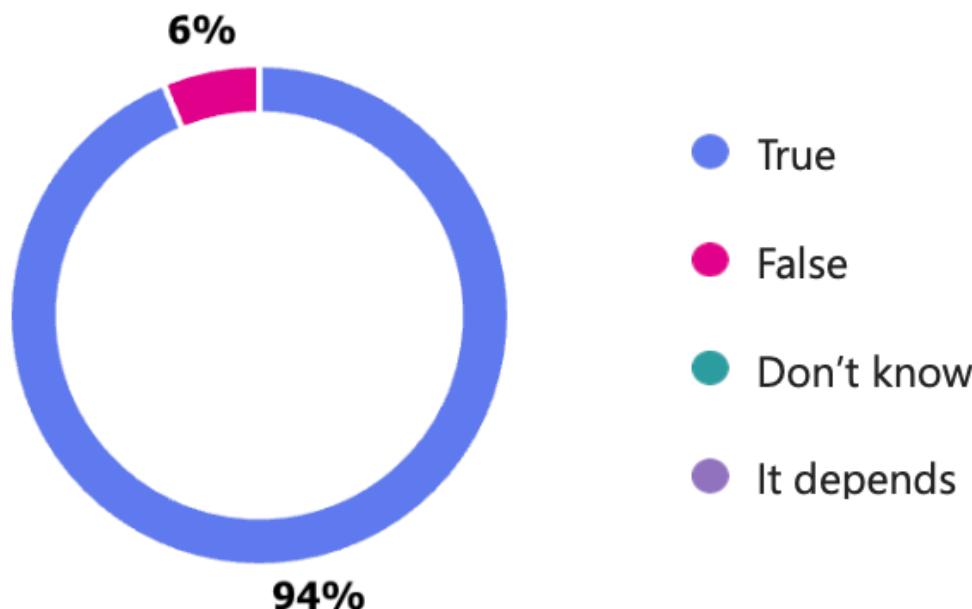
## Do we know what a confounder is?



Confounding Variable Joke

# How to adjust for confounding

7. Strategies to control for confounding include restriction, stratification plus adjustment, matching, and regression.  
(0 point)



# How to adjust for confounding

- **Randomization is the best method.**
  - In conditionally randomized experiments given covariates  $L$ , the common causes (i.e., the covariates L) are measured and thus the adjusted (standardization or IP weighting) association measure is expected to equal the effect measure.
- Subject-matter knowledge to identify adjustment variables is ***discretionary in "ideal" randomized experiments.***
- On the other hand, **subject-matter knowledge is key (a must!) in observational studies** in order to identify and measure adjustment variables (e.g., for regression adjustment).

# How to adjust for confounding

- Causal inference from observational data relies on the **uncheckable assumption** that we have used our knowledge to identify and measure a set of variables  $L$  that is a sufficient set for confounding adjustment:
  - The set of non-descendants of treatment that includes enough variables to block all backdoor paths.
- Under this assumption of no unmeasured confounding or of conditional exchangeability given  $L$ , standardization and Inverse Probability (IP) weighting can be used to compute the average causal effect in the population.

# Standardization

## Why standardize?

- To control for confounding
- To summarize many estimates into one
- Is a weighted average of measures of occurrence across a distribution (say, age).
- Can be applied to any measure of occurrence or measure of effect
- Weights are chosen based on the population of interest

(ME3, pg. 49)

# Standardized measures of association and effect

- Let  $I_k$  represent strata specific incidence rates and
- let  $I_k^*$  represent another schedule of such rates (perhaps based on a different exposure distribution)
- Let  $T_k$  represent person-time at risk in each strata

$$I_s = \left( \frac{\sum_{k=1}^K T_k I_k}{\sum_{k=1}^K T_k} \right)$$

$$I_s^* = \left( \frac{\sum_{k=1}^K T_k I_k^*}{\sum_{k=1}^K T_k} \right)$$

- Then the standardized rate ratio is:  $IR_s = I_s / I_s^*$
- The standardized rate difference is:  $IR_s = I_s - I_s^* = \sum T_k (I_k - I_k^*)$

(ME3, pg. 67)

# Standardized measures of association and effect

- Note that the standardized rate difference is a weighted average of stratum-specific rate differences

## Interpretation of both measures:

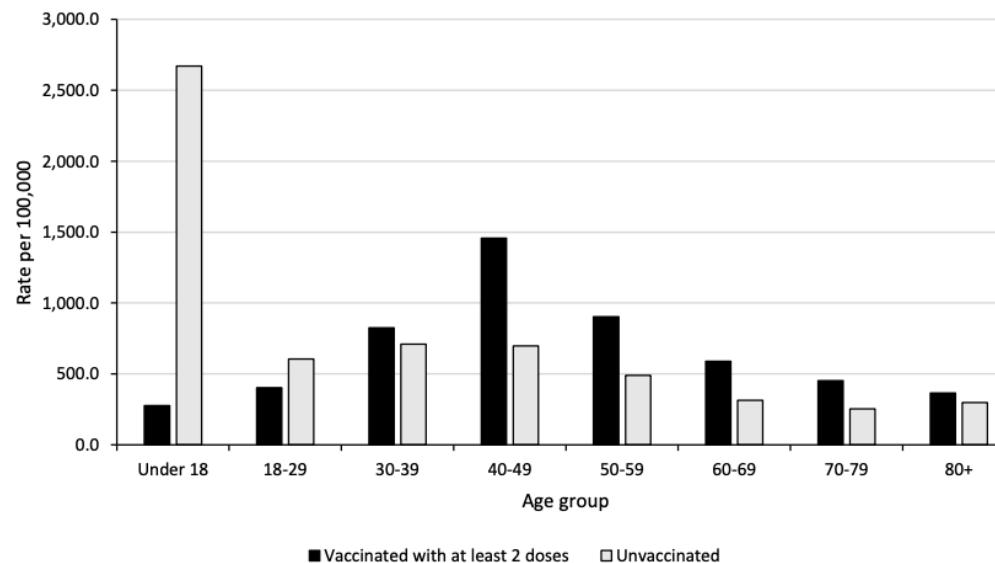
- Effects of exposure on this population.
  - For the standardized rate ratio we need to assume that the relative distribution of person-time would be unaffected by exposure.
  - Standardized **risk ratios** do not require this assumption because the denominators do not use person-time.

# Example: COVID-19 vaccine effectiveness in the UK

UK Health Security Agency "COVID-19 vaccine surveillance report", Week 41

**Figure 2. Rates (per 100,000) by vaccination status from week 37 to week 40 2021**

(a) COVID-19 cases



**Rates (per 100,000) by vaccination status from week 37 to week 40 2021**

## Example: COVID-19 vaccine effectiveness in the UK (2)

Numbers by variant are reported by Public Health England.

**Table 5. Attendance to emergency care and deaths of sequenced and genotyped Delta cases in England by vaccination status (1 February 2021 to 12 September 2021)**

Variant	Age group (years)**	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	≥14 days post dose 2	Un-vaccinated
Delta cases	<50	497,105	119,611	49,527	30,359	83,009	85,407	248,803
	≥50	95,587	35,596	7,602	314	7,129	71,991	8,551
	All cases	593,572	155,252	58,003	30,674	90,138	157,400	257,357
Cases with an emergency care visit§ (exclusion†)	<50	16,709	N/A	167	1,051	2,494	2,518	10,479
	≥50	5,445	N/A	21	30	448	3,747	1,199
	All cases	22,162	N/A	196	1,081	2,942	6,265	11,678
Cases with an emergency care visit§ (inclusion#)	<50	22,719	N/A	273	1,364	3,060	3,162	14,860
	≥50	10,102	N/A	50	64	755	6,532	2,701
	All cases	32,834	N/A	336	1,428	3,815	9,694	17,561
Cases where presentation to emergency care resulted in overnight inpatient admission§ ((exclusion†))	<50	3,490	N/A	95	174	352	453	2,416
	≥50	2,784	N/A	10	18	184	1,908	664
	All cases	6,280	N/A	111	192	536	2,361	3,080
Cases where presentation to emergency care resulted in overnight inpatient admission§ (inclusion#)	<50	6,230	N/A	144	283	565	721	4,517
	≥50	6,167	N/A	33	42	393	3,913	1,786
	All cases	12,407	N/A	187	325	958	4,634	6,303

From: Table 5. Attendance to emergency care and deaths of sequenced and genotyped Delta cases in England by vaccination status (1 February 2021 to 12 September 2021) [here](#).)

## Example: COVID-19 vaccine effectiveness in the UK

### Let's play with the numbers (1): check the risk difference (RD)

```
#157400 - 2361 #exposed without outcome
#257357 - 30801 #unexposed without outcome
l6UKdata<-c(2361,155039, 3080, 254277)
l6UKest<- epi.2by2(l6UKdata, method = "cohort.count")
l6UKest

##          Outcome +    Outcome -    Total      Inc risk *
## Exposed +        2361      155039    157400    1.50 (1.44 to 1.56)
## Exposed -        3080      254277    257357    1.20 (1.16 to 1.24)
## Total           5441      409316    414757    1.31 (1.28 to 1.35)
##
## Point estimates and 95% CIs:
## -----
## Inc risk ratio                  1.25 (1.19, 1.32)
## Inc odds ratio                 1.26 (1.19, 1.33)
## Attrib risk in the exposed *   0.30 (0.23, 0.38)
## Attrib fraction in the exposed (%) 20.21 (15.85, 24.35)
## Attrib risk in the population *  0.12 (0.06, 0.17)
## Attrib fraction in the population (%) 8.77 (6.64, 10.86)
##
## -----
## Uncorrected chi2 test that OR = 1: chi2(1) = 69.360 Pr>chi2 = <0.001
## Fisher exact test that OR = 1: Pr>chi2 = <0.001
## Wald confidence limits
## CI: confidence interval
## * Outcomes per 100 population units
```

# Example: COVID-19 vaccine effectiveness in the UK

- Missing something?

Variant	Age group (years)**	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	≥14 days post dose 2	Un-vaccinated
Delta cases	<50	497,105	119,611	49,527	30,359	83,009	85,407	248,803
	≥50	95,587	35,596	7,602	314	7,129	71,991	8,551
	All cases	593,572	155,252	58,003	30,674	90,138	157,400	257,357
Cases with an emergency care visit§ (exclusion‡)	<50	16,709	N/A	167	1,051	2,494	2,518	10,479
	≥50	5,445	N/A	21	30	448	3,747	1,199
	All cases	22,162	N/A	196	1,081	2,942	6,265	11,678
Cases with an emergency care visit§ (inclusion#)	<50	22,719	N/A	273	1,364	3,060	3,162	14,860
	≥50	10,102	N/A	50	64	755	6,532	2,701
	All cases	32,834	N/A	336	1,428	3,815	9,694	17,561
Cases where presentation to emergency care resulted in overnight inpatient admission§ ((exclusion‡))	<50	3,490	N/A	95	174	352	453	2,416
	≥50	2,784	N/A	10	18	184	1,908	664
	All cases	6,280	N/A	111	192	536	2,361	3,080
Cases where presentation to emergency care resulted in overnight inpatient admission§ (inclusion#)	<50	6,230	N/A	144	283	565	721	4,517
	≥50	6,167	N/A	33	42	393	3,913	1,786
	All cases	12,407	N/A	187	325	958	4,634	6,303

From: Table 5. Attendance to emergency care and deaths of sequenced and genotyped Delta cases in England by vaccination status (1 February 2021 to 12 September 2021) [here](#).) Note: The totals do not exactly sum up to the previous table, as age was missing in a few cases.

## Example: COVID-19 vaccine effectiveness in the UK

### Let's play with the numbers (2) - Standardization

#### Outcomes among people under 50 years

```
l6UKdatu50<-c(453,84954, 2416, 246387)
l6UKt1u50<- epi.2by2(l6UKdatu50, method = "cohort.count")
l6UKt1u50$tab
```

	Outcome +	Outcome -	Total	Inc risk *
## Exposed +	453	84954	85407	0.53 (0.48 to 0.58)
## Exposed -	2416	246387	248803	0.97 (0.93 to 1.01)
## Total	2869	331341	334210	0.86 (0.83 to 0.89)

#### Outcomes among people $\geq 50$ years

```
l6UKdatm50<-c(1908 , 70083, 664, 7887)
l6UKt1m50<- epi.2by2(l6UKdatm50, method = "cohort.count")
l6UKt1m50$tab
```

	Outcome +	Outcome -	Total	Inc risk *
## Exposed +	1908	70083	71991	2.65 (2.53 to 2.77)
## Exposed -	664	7887	8551	7.77 (7.21 to 8.35)
## Total	2572	77970	80542	3.19 (3.07 to 3.32)

## Example: COVID-19 vaccine effectiveness in the UK

Let's play with the numbers (3) check the risk differences (RD)

### Outcomes among people under 50 years

Measure	Estimate 95%CIs		
Measure	Est.	LB	UB
Inc risk ratio	0.55	0.49	0.60
Inc odds ratio	0.54	0.49	0.60
Attrib inc risk *	-0.44	-0.50	-0.38
Attrib fraction in exposed (%)	-83.08	-102.34	-65.65
Attrib inc risk in population *	-0.11	-0.16	-0.06
Attrib fraction in population (%)	-13.12	-14.92	-11.34

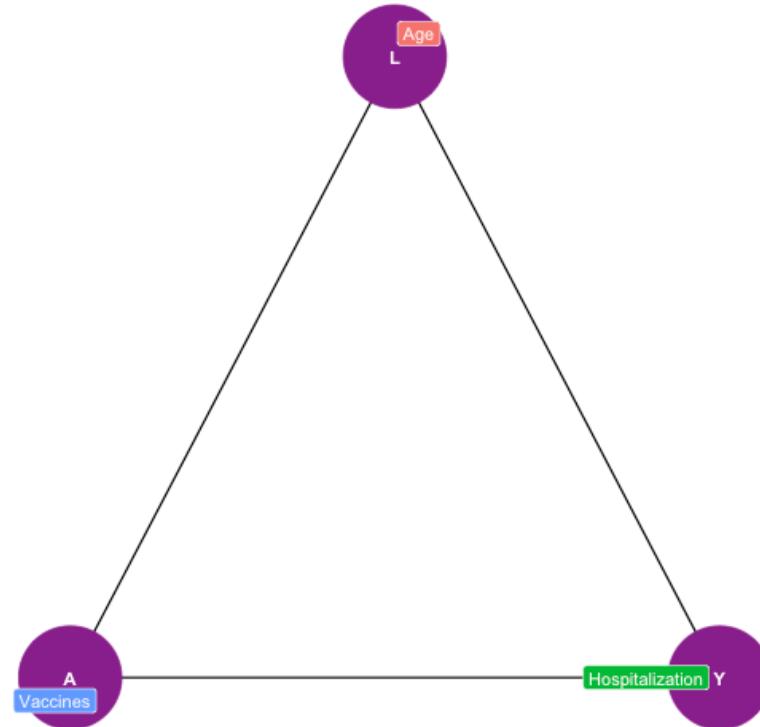
### Outcomes among people $\geq 50$ years

Measure	Estimate 95%CIs		
Measure	Est.	LB	UB
Inc risk ratio	0.34	0.31	0.37
Inc odds ratio	0.32	0.30	0.35
Attrib inc risk *	-5.11	-5.69	-4.54
Attrib fraction in exposed (%)	-192.99	-219.11	-169.00
Attrib inc risk in population *	-4.57	-5.15	-3.99
Attrib fraction in population (%)	-143.17	-159.10	-128.21

## Example: COVID-19 vaccine effectiveness in the UK

### Confounding?

DAG of Age, Vaccines and COVID-19 Hospitalization Confounding



We know that IRL the "L" includes a vector / set of potential covariates that could be considered as Confounders... this is an illustration only!

## Direct standardization

Suppose we want to estimate  $E[Y^a = 1] - E[Y^a = 0] = RD$ .

The conditional exchangeability allows us to say  $Y^a \perp\!\!\!\perp A|L$

According to the law of total expectation:

$$E[Y^a = 1] = \sum_x E[Y^a = 1|X = x]Pr(x) ;$$

$$E[Y^a = 0] = \sum_x E[Y^a = 0|X = x]Pr(x)$$

- $\sum_x$  means sum over all values x that occur in the study population.
- $Pr(x)$  refers to the distribution of x in that population.

$$RD = E[Y^a = 1] - E[Y^a = 0] =$$

$$\sum_x E[Y^a = 1|X = x]P(x) - \sum_x E[Y^a = 0|X = x]P(x)$$

## Example: COVID-19 vaccine effectiveness in the UK

### Let's play with the numbers - Standardization

#### Outcomes among people under 50 years

	Outcome +	Outcome -	Total	Inc risk *
## Exposed +	453	84954	85407	0.53 (0.48 to 0.58)
## Exposed -	2416	246387	248803	0.97 (0.93 to 1.01)
## Total	2869	331341	334210	0.86 (0.83 to 0.89)

#### Outcomes among people $\geq 50$ years

	Outcome +	Outcome -	Total	Inc risk *
## Exposed +	1908	70083	71991	2.65 (2.53 to 2.77)
## Exposed -	664	7887	8551	7.77 (7.21 to 8.35)
## Total	2572	77970	80542	3.19 (3.07 to 3.32)

## Example: COVID-19 vaccine effectiveness in the UK

### Let's play with the numbers (4) - Standardization

To compute the PO using observed data, we need the consistency assumption

$$RD = \sum_x E[Y|A=1, X=x]P(x) - \sum_x E[Y|A=0, X=x]Pr(x)$$

Standardized risk in the vaccinated :

$$(453/85,407 \times 334,210/414,752 + 1,908/71,991 \times 80,542/414,752) \approx 0.94\%$$

$$R_{vax} = 0.94$$

Standardized risk in the unvaccinated :

$$(2,416/248,803 \times 334,210/414,752 + 664/8,551 \times 80,542/414,752) \approx 2.29\%$$

$$R_{unvax} = 2.29$$

**Standardized RD = -1.35** from  $(0.94\% - 2.29\% = -1.35\%) \neq 0.3$  in the crude estimates.

**Standardized RR = 0.41** from  $(0.0094 / 0.0229) \neq 1.25$  in the crude estimates.

# Example: COVID-19 vaccine effectiveness in the UK

UK Health Security Agency "COVID-19 vaccine surveillance report", Week 41

(b) Cases presenting to emergency care (within 28 days of a positive test) resulting in overnight inpatient admission

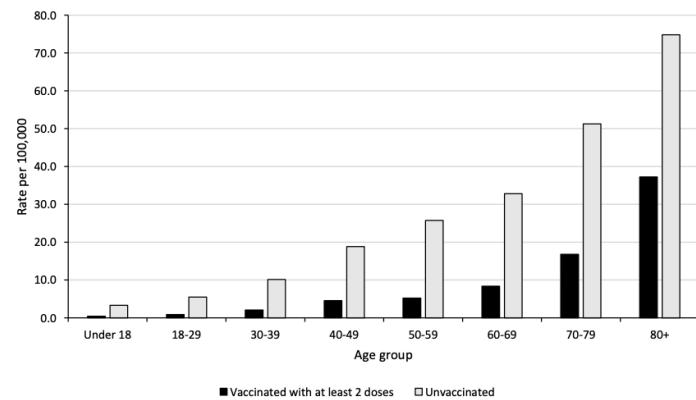
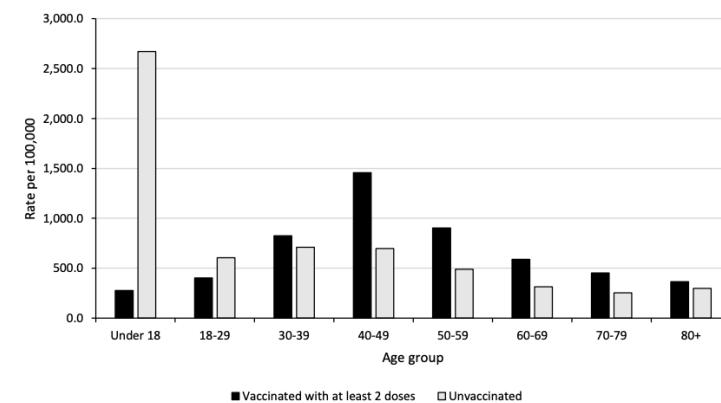


Figure 2. Rates (per 100,000) by vaccination status from week 37 to week 40 2021

(a) COVID-19 cases



Cases presenting to emergency care (within 28 days of a positive test) resulting in overnight inpatient admission.[here](#).

# What about the Mantel-Haenzel Methods?

- *Cochran-Mantel-Haenzel* methods are useful for associations, when only few covariates are involved in the calculation.
- Takes the effect in each strata of  $L$  or  $Z$  (our third variable),
- Combines these measures across  $L$  using calculated weights <sup>1</sup>, for example example:

$$RD_{M-H} = \left( \frac{\sum_l (RD_l w_l)}{\sum_l w_l} \right) = \left( \frac{RD_0 w_0 + RD_1 w_1}{w_0 + w_1} \right)$$

- Are expected to work in closed cohorts and **assumes homogeneity across strata!!**
  - Limited use in a set of covariates  $L$  and in presence of Effect measure modification and or interaction.

<sup>1</sup> There are specific formulas for RD, RR and ORs as well

## Standardized measures of association and effect

- No assumption of homogeneity, "agnostic of the distribution", Model-based direct standardization <sup>1</sup> are used when  $L(X, E, A)$  consists of a large vectors of covariates.

Involves two steps:

- Fitting a regression model for the outcome given exposure and covariates
- Averaging the exposure effect over the covariate distribution of the standard population.

<sup>1</sup> More on "advanced" techniques to address confounding empirically after we deal with regressions.

# Standardized Morbidity Ratio (SMR)

- A generalization to standardization when the standard population is the exposed sub-population.
- In this case, the standardized rate ratio becomes:

$$I_s = \left( \frac{\sum_{k=1}^K T_k I_k}{\sum_{k=1}^K T_k I_k^*} \right) = \left( \frac{\sum_{k=1}^K A_k}{\sum_{k=1}^K T_k I_k^*} \right)$$

[Numerator] cases occurring in exposed (**Observed**)

[Denominator] cases **expected** to occur in absence of exposure if exposure doesn't affect person time at risk

(ME3, pg. 68-69)

# How to adjust for confounding

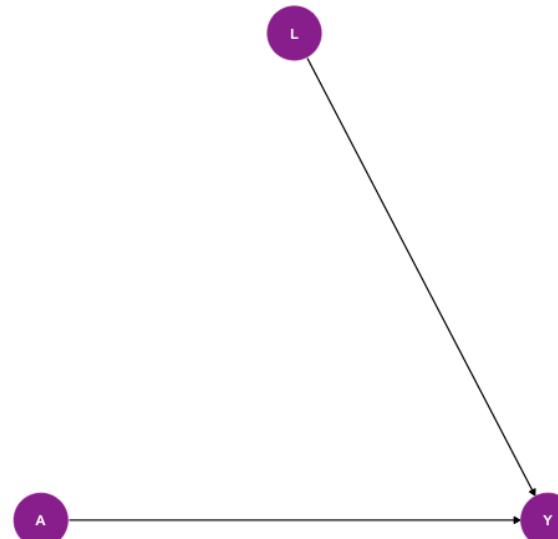
Standardization and Inverse Probability (IP) weighting are not the only methods.

$$IPW_z = \left( \frac{1}{Pr(A = a | L = z)} \right)$$

Often using regression models, **assuming the model specification is correct!** 😬

**IPW removes the arrow from  $L \rightarrow A$ :**

DAG - effect of IPW on Confounding



# How to adjust for confounding

Two categories of methods for confounding adjustment:

**1) G-methods (including G-formula, IP weighting, and G-estimation).** These exploit conditional exchangeability in subsets defined by  $L$  to estimate the causal effect of  $A$  on  $Y$  in the entire population or in any subset of the population.

- Under the assumption of conditional exchangeability given  $L$ , g-methods simulate  $A - Y$  associations in the population if backdoor paths involving variables  $L$  did not exist; simulated  $A - Y$  associations can then be attributed to the effect of  $A$  on  $Y$ .
- IP weighting achieves this by creating a pseudo-population in which  $A$  is independent of measured confounders  $L$ , by “deleting” the arrow from  $L \rightarrow A$ .

# How to adjust for confounding

## 2) Stratification-based methods (including Stratification, Restriction, Matching).

Methods that exploit conditional exchangeability in subsets defined by  $L$  to estimate the association between  $A$  and  $Y$  in those subsets only.

Stratification-based methods estimate the association between  $A$  and  $Y$  in one or more subsets of the population in which the treated and the untreated are assumed to be exchangeable.

- Hence the  $A \rightarrow Y$  association in each subset is entirely attributed to the effect of  $A$  on  $Y$ .
- Stratification/restriction do not delete the arrow from  $L \rightarrow A$ , but instead calculate the association within strata of  $L$ , since within each level of  $L$ , there is no  $L \rightarrow A$  association to cause confounding.

# How to adjust for confounding

All these methods require conditional exchangeability given the measured covariates  $L$  to identify the effect of treatment  $A$  on outcome  $Y$ .

- When interested in the effect in the entire population, conditional exchangeability is required in all strata defined by  $L$ ;
- When interested in the effect in a subset of the population, conditional exchangeability is required in that subset only.
- Achieving conditional exchangeability may be an unrealistic goal in many observational studies but expert knowledge can be used to get as close as possible to that goal.
- At the very least, investigators should generally avoid adjustment for variables affected by either the treatment or the outcome.

# How to adjust for confounding

Thoughtful and knowledgeable investigators could believe that various causal structures, possibly leading to different conclusions regarding confounding, are equally plausible.

- DAGs simply allow us to have that discussion.
- Existence of common causes of treatment and outcome does not depend on the adjustment method (although it does depend on the target population).
- Adjustment for measured confounding will generally imply a change in the estimate, but not necessarily the other way around.
- Changes in estimates may occur for reasons other than confounding,
  - **including selection bias when adjusting for non-confounders and the use of non-collapsible effect measures.**

H & R write:

*"Attempts to define confounding based on change in estimates have been long abandoned because of these problems." This is overstated. When using a DAG and collapsible measures, the method is a reasonable and practical strategy."*

## A note on stratification and non-collapsibility

Comparing crude to adjusted estimates is reliable for RR and RD, but not for OR unless: a) rare outcome or b)  $OR \approx RR$  due to design (e.g. case-cohort).

Recall the case of  $C \rightarrow E \rightarrow Y$

Measure	Estimate 95% CIs		
Measure	Est.	LB	UB
Inc risk ratio (crude)	1.91	1.36	2.70
Inc risk ratio (M-H)	1.75	1.14	2.69
Inc risk ratio (crude:M-H)	1.09		
Inc odds ratio (crude)	5.12	2.02	12.97
Inc odds ratio (M-H)	3.83	1.45	10.09
Inc odds ratio (crude:M-H)	1.34		
Attrib inc risk (crude) *	37.15	19.01	55.30
Attrib inc risk (M-H) *	31.20	-6.71	69.12
Attrib inc risk (crude:M-H)	1.19		

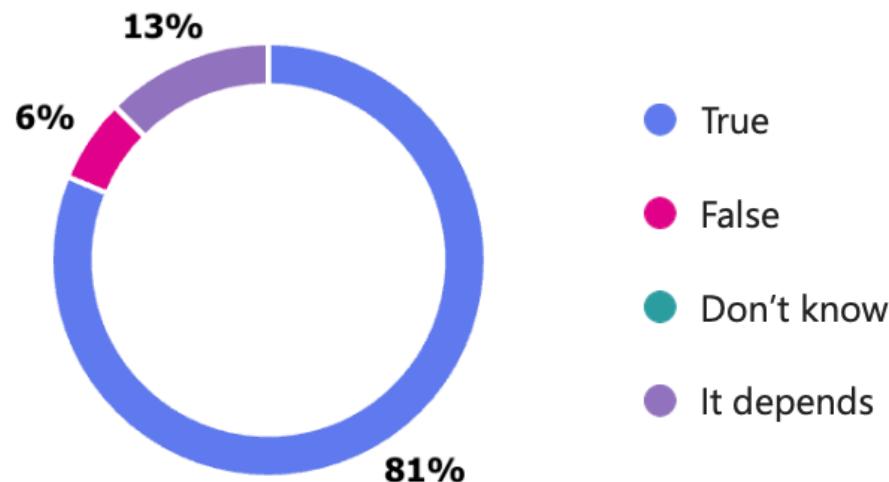
# A note on stratification and non-collapsibility

- We can say a measure of the association between A and Y is collapsible across L if the adjusted association,  $RR_{AY}|L$ , is equal to the crude association,  $RR_{AY}$ , where L is not a confounder — This means that a crude measure of association will not change if we adjust for a variable that is not a confounder ( $L$ )
- The odds (OR) and incidence density ratios (IDR) fail this property and are considered non collapsible effect measures
- For the OR, the crude measure may be closer to the null than the pooled/adjusted OR, particularly with a common outcome
- Therefore, for some measures, our simple crude vs. adjusted comparison **may suggest confounding when there really isn't!**

## Change in estimate??

**Not Really!!!**

8. The presence of confounding is suspected when the size of the association of interest changes meaningfully after adjustment by one of these methods. (0 point)



# Structural confounding, violation of Positivity

High correlations between confounder and exposure: violation of the “positivity assumption”. When this is “structural” (in the sense of a high correlation that exists because of causal relations in the source population), Oakes calls this “structural confounding”.

Table 3. Distribution of Racial Segregation (Number of Census Tracts per Cell<sup>a</sup>) According to Level of Neighborhood Deprivation in Wake and Durham Counties, North Carolina, 1999–2001<sup>b</sup>

County and Quartile of Percent Black	Quartile of NDI			
	NDI1 (Low)	NDI2	NDI3	NDI4 (High)
Durham County (n = 53 tracts)				
%BL1 (low)	10	2	1	1
%BL2	4	6	3	0
%BL3	0	5	4	4
%BL4 (high)	0	0	5	8
Wake County (n = 105 tracts)				
%BL1 (low)	23	4	0	0
%BL2	3	12	10	1
%BL3	1	8	12	5
%BL4 (high)	0	2	4	20

Abbreviations: %BL, percent black; NDI, neighborhood deprivation index.

<sup>a</sup> Cells are defined as the intersection between quartile of NDI and quartile of percent black.

<sup>b</sup> Cells with italicized numbers represent those with too few contexts ( $\leq 1$  tract per cell) for meaningful comparisons.

Oakes JM. Advancing neighbourhood-effects research selection, inferential support, and structural confounding. *Int J Epidemiol*. 2006 Jun;35(3):643–7.

Messer et al. Effects of Socioeconomic and Racial Residential Segregation on Preterm Birth: A Cautionary Tale of Structural Confounding *AJE* 2010; Mar 15;171(6):664–73.

# Structural confounding, violation of Positivity

## Data Generation Process

```
set.seed(704); n=500
ses1 <- sample(1:12, n, replace = TRUE);
ses1[ses1>=10]<-0
ses2 <- cut(ses1, breaks = c(0, 5, 10, 15),
            labels = c("0", "1", "2"))
ses2[is.na(ses2 )]<- "2"
exposure<- ifelse(ses2=="1",
                    rbinom(n,1,0.45),
                    ifelse(ses2=="0", rbinom(n,1,0.5),
                           ifelse(ses2=="2", rbinom(n,1,0.0001),
                                  rbinom(n,1,0.2))))
outcome<- ifelse(ses2=="0", rbinom(n,1,0.75),
                  ifelse(ses2=="1", rbinom(n,1,0.25),
                         rbinom(n,1,0.25)))
data.strconf <- data.frame(outcome, exposure,
                           ses2)
##          ses2
## exposure 0   1   2
##          0 101 107 123
##          1 104  65   0
strconf2 <- glm(outcome ~ exposure,
```

## Regression Results Crude/Unadjusted

	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	0.663	0.532	0.827	-3.657	0.000
exposure	1.677	1.154	2.437	2.713	0.007

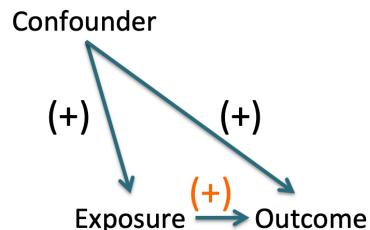
## Adjusted

	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	2.461	1.676	3.612	4.598	0.000
exposure	1.011	0.634	1.613	0.046	0.963
as.factor(ses2)1	0.119	0.074	0.190	-8.860	0.000
as.factor(ses2)2	0.168	0.097	0.290	-6.399	0.000

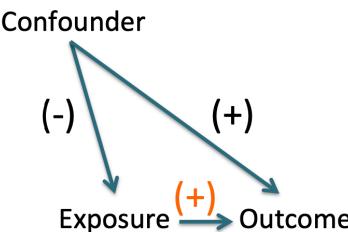
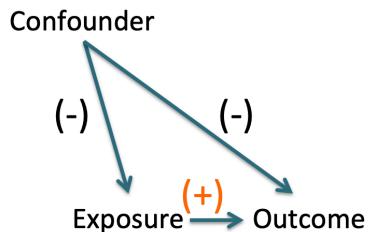
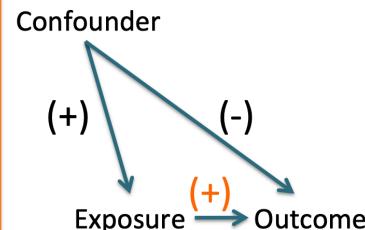
# Which way will the confounding go?

DAGs if exposure & outcome are positively associated

**Positive confounding:**  
unadjusted > adjusted



**Negative confounding:**  
unadjusted < adjusted



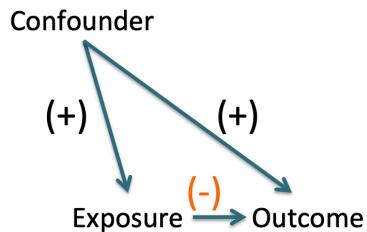
6

Vander Stoep A, et al. A didactic device for teaching epidemiology students how to anticipate the effect of a third factor on an exposure-outcome relation. AJE 1999; 15;150(2):221.

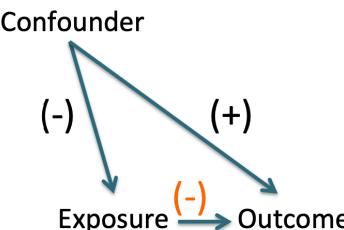
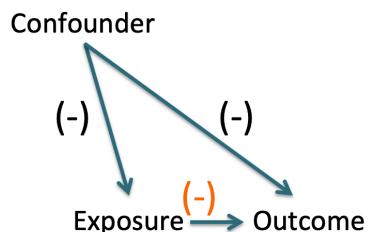
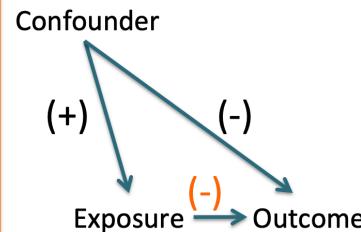
# Which way will the confounding go?

DAGs if exposure & outcome are negatively associated

**Negative confounding:**  
unadjusted > adjusted



**Positive confounding:**  
unadjusted < adjusted



7

These schematics are just illustrations, it depends on the strength (degree of correlation) of the covariates!!, simulations works better than "blanket" type of statements

# Positive, negative, and “qualitative” confounding

- Confounding may lead to an overestimation or an underestimation of the true magnitude of an effect.
- **Positive confounding:** the magnitude of the unadjusted vis-à- vis the adjusted association is exaggerated.
- **Negative confounding:** the magnitude of the unadjusted vis-à- vis the adjusted association is underestimated.
- **Qualitative confounding:** An extreme case when confounding results in an inversion of the direction of the association.

# Magnitude of confounding

- The magnitude of confounding will depend on the strength of the confounder-exposure AND confounder-outcome associations.
- Conversely, if there is no association between the confounder - exposure OR no association between the confounder-outcome then no confounding of the main effect could be present.
- The strength of the confounder-exposure and confounder- outcome associations bounds the confounding effect
  - e.g., if  $RR_{crude} = 2$  and the confounder-outcome relation is 2 (a doubling of risk), then the confounder would have to be perfectly correlated with the exposure in order to fully explain the main effect of  $RR=2$

## How strong the the *unmeasured confounding* should be to explain away my estimated association?

**E values:** respond to this question for ratio <sup>1</sup> measures, how?

$$E-value = RR + \sqrt{RR \times (RR - 1)}$$

- E-value is the minimum value of the association between  $U \rightarrow A$  and  $U \rightarrow Y$  that will be capable of attenuating the observed association towards the null.

- Example: RR=1.33;  $1.33 + \sqrt{1.33 \times (1.33 - 1)} = 1.99$  then, if there was an  $U$ , it should:
  - 1) double the risk among unexposed and/or exposed ( $RR_{UY} = 2$ ), AND
  - 2) be twice as prevalent among exposed than among unexposed ( $RR_{AU} = 2$ )

To completely explain away the observed association, but a weaker confounder (given the E-value), say 1.5 or 1.3, would not.

<sup>1</sup> E values are debatable for some but still a straightforward calculation and useful information to have. Versions of the E-value exists for ORs and HRs. E-value calculator.

# Statistical significance?

## In general, NO!

- But if you MUST use p-values, set the criteria on the high side (e.g.  $p < 0.30$ ). This way you adjust for some non-confounders, but you don't miss many true confounders.

Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. AJE 1989;129(1):125–37.

- Residual confounding (unmeasured L's ( $U_1, U_2$ , etc), categorization, measurement error, etc):

Kaufman JS, et al. Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race.

Epidemiology 1997; 8(6):621–8. Ogburn EL, Vanderweele TJ. Bias attenuation results for nondifferentiably mismeasured ordinal and coarsened confounders. Biometrika. 2013;100(1):241– 248. PMID: 24014285

# Residual confounding

Residual confounding occurs when adjustment does not completely remove the confounding effect of a given variable(s):

## 1) Misclassification of confounding variables

- (e.g., the variable is an imperfect proxy for the characteristic we want to adjust for)

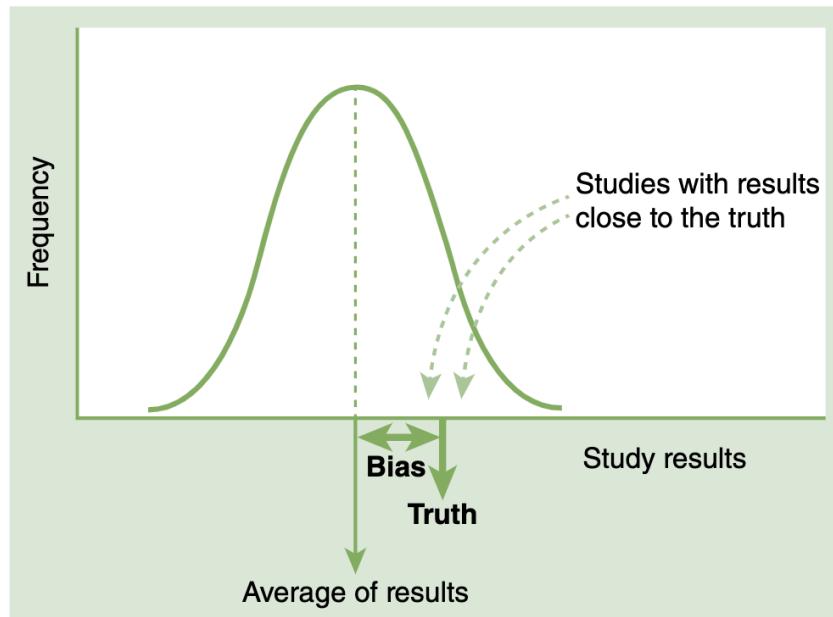
## 2) Improper modeling of the confounding variable

- (e.g., if we are studying air pollution and lung cancer and want to control for smoking, we should measure smoking in a way that best predicts lung cancer—i.e., pack-years not ever-never)

## 3) Other important confounders are not included (also known as unmeasured confounding or omitted variable bias)

# Validity and Bias:

- The epidemiologist's goal: the most **VALID and PRECISE** estimate possible of the causal effect of exposure on disease.
- Error comes from sampling variability (lack of precision) and bias (lack of validity).



# Confounded<sup>1</sup> ?

what are other  
words for  
confounded?

confused, bewildered, perplexed,  
baffled, befuddled,  
disconcerted, blasted,  
nonplussed, bemused, lost



Thesaurus.plus

<sup>1</sup> We all are!! We will have more on this and empirical examples after we deal with regressions.

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

