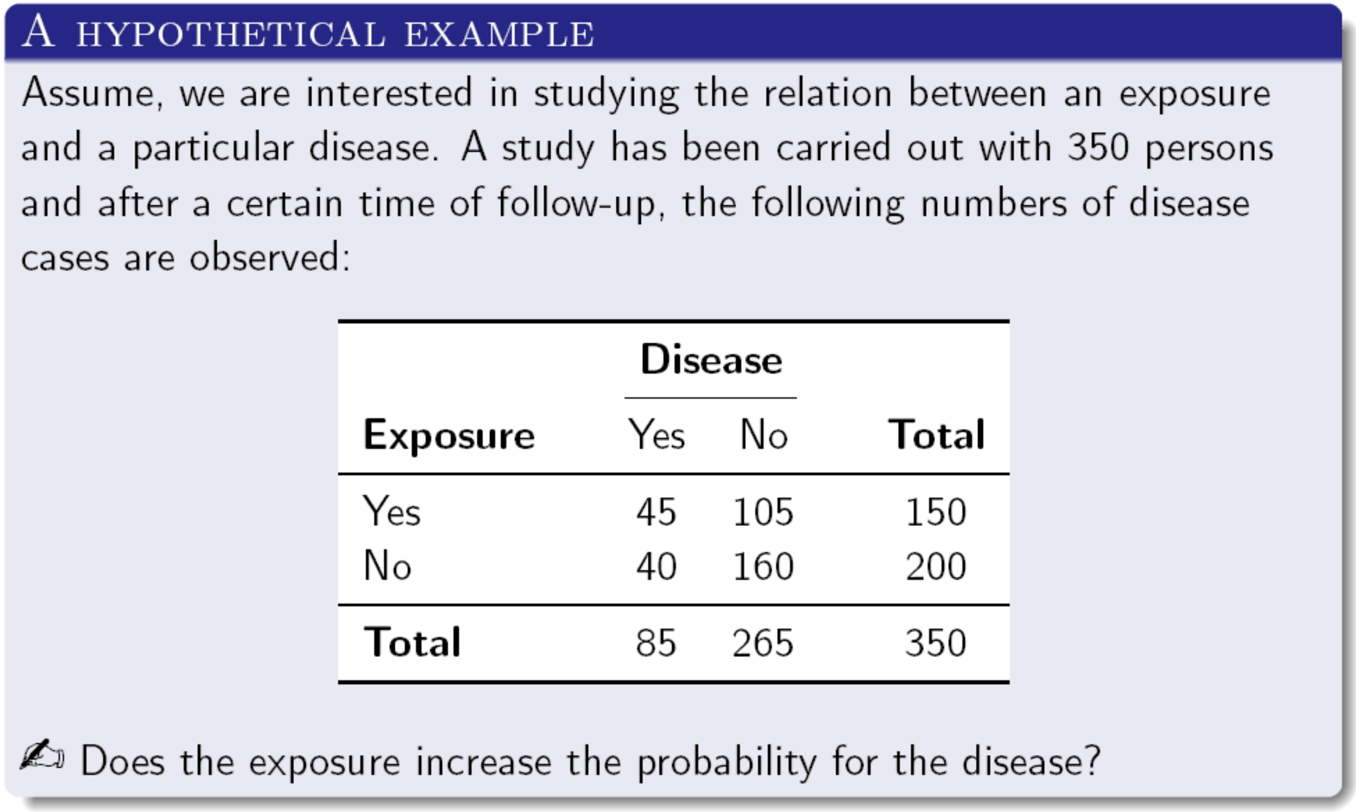
Biosciences: 3.- Health

3.4 Confounding

Index

- **Assessing the relationship between variables**
- **Confounding**
- **Interaction (effect modification)**
- **Making confounding**

How to assess relationship between variables.



A hypothetical example

Assume, we are interested in studying the relation between an exposure and a particular disease. A study has been carried out with 350 persons and after a certain time of follow-up, the following numbers of disease cases are observed:

	Disease		
Exposure	Yes	No	**Total**
Yes	45	105	150
No	40	160	200
Total	85	265	350

✍ Does the exposure increase the probability for the disease?

In **epidemiology** studies we use the "Exposure" variable (ex. smoking), whereas in **clinical trials** we use the "Treatment" variable (ex. Anti-retrovial tratment).

The Relative Risk

Definition

The **relative risk** (or **risk ratio**) is the ratio of the risk of disease (D) among exposed people (E) as compared to the risk among unexposed people ($\bar{E}$):

$$\mathrm{RR} = \frac{P(D|E)}{P(D|\bar{E})}. \tag{1}$$

The RR may also be defined as the ratio of two incidence rates.

A hypothetical example (cont.)

In the previous example, the relative risk amounts to

$$\mathrm{RR} = \frac{45/150}{40/200} = \frac{0.3}{0.2} = 1.5.$$

That is, the risk of disease is 1.5 times higher among exposed people.

	Disease		
Exposure	Yes	No	**Total**
Yes	45	105	150
No	40	160	200
Total	85	265	350

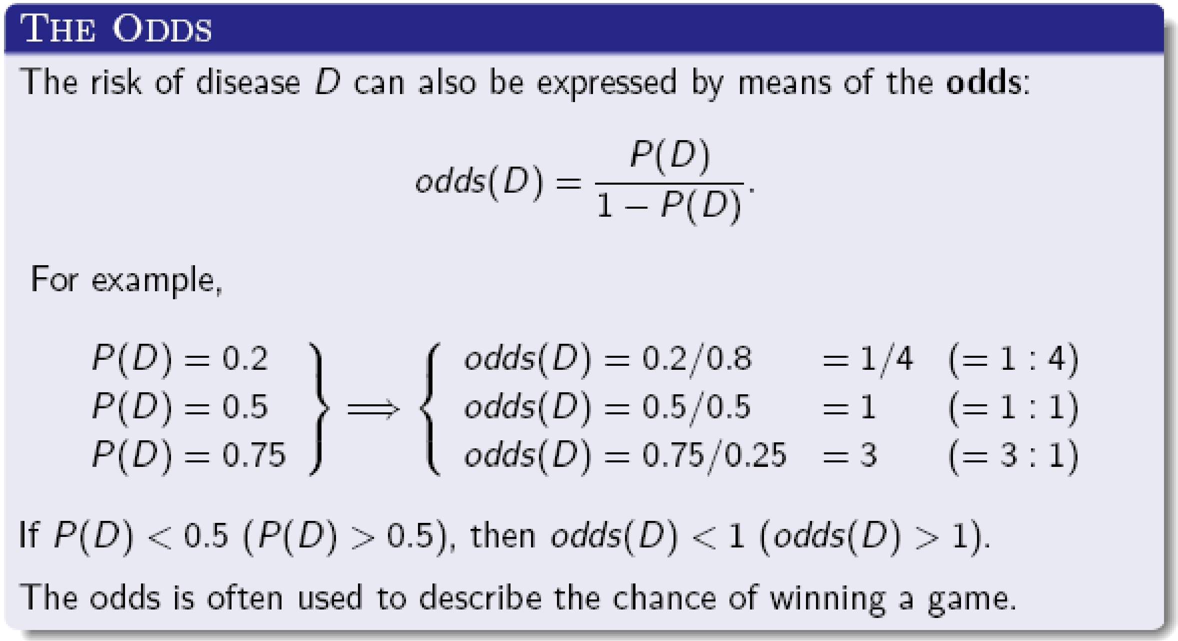
The Relative Risk (cont.)

Comments

- If RR > 1, there is a greater probability of D among exposed people. Hence, E is a (possible) **risk factor** for D.

  If RR < 1, E is a (possible) **protective factor** for D.

  RR $= 1$ indicates that there is no difference between both groups with respect to risk of disease. That is, D and E are independent.



The Odds

The risk of disease D can also be expressed by means of the **odds**:

$$odds(D) = \frac{P(D)}{1 - P(D)}.$$

For example,

$$\left.\begin{array}{l} P(D) = 0.2 \\ P(D) = 0.5 \\ P(D) = 0.75 \end{array}\right\} \Longrightarrow \left\{\begin{array}{lll} odds(D) = 0.2/0.8 & = 1/4 & (= 1:4) \\ odds(D) = 0.5/0.5 & = 1 & (= 1:1) \\ odds(D) = 0.75/0.25 & = 3 & (= 3:1) \end{array}\right.$$

If $P(D) < 0.5$ ($P(D) > 0.5$), then $odds(D) < 1$ ($odds(D) > 1$).

The odds is often used to describe the chance of winning a game.

The Odds Ratio

Definition

The **odds ratio (OR)** is the ratio of the odds of disease among exposed people as compared to the odds among unexposed people:

$$\mathrm{OR} = \frac{odds(D|E)}{odds(D|\bar{E})} = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))}.$$

A hypothetical example (cont.)

In the previous example, the odds ratio amounts to:

$$\mathrm{OR} = \frac{45/150/105/150}{40/200/160/200} = \frac{45 \cdot 160}{105 \cdot 40} \approx 1.71.$$

That is, the odds of the disease is 1.71 times higher among exposed people as compared with unexposed people.

	Disease		
Exposure	Yes	No	**Total**
Yes	45	105	150
No	40	160	200
Total	85	265	350

Confounding

Consider the following example. Suppose we have the treatment variable (X=X1,X2), the outcome variable (Y=Y1,Y2) and a third variable (Z=Z1,Z2) which indicates two different centers:

Center 1 (Z1)

	Y+	Y-	
X+	100	50	**150**
X-	20	10	**30**
	120	**60**	**180**

OR = 1

Center 2 (Z2)

	Y+	Y-	
X+	10	20	**30**
X-	50	100	**150**
	60	**120**	**180**

OR = 1

Overall

	Y+	Y-	
X+	110	70	**180**
X-	70	110	**180**
	180	**180**	**360**

OR = 2,47

Note that we observe non treatment effect (X) on Y in both centers, but the overall result shows treatment effect (OR>1). Why is so?

This phenomenon is known as **Simpson's paradox** and it is a well-known problem in clinical trials or epidemiology called **confounding**.

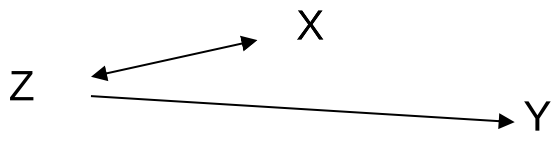
Block 3.4 – Confounding

Z is called a **confounder** (of the association between X and Y), if it is both related with X (collinearity) and a risk factor for Y (prediction). It causes a biased estimation of the association of interest.

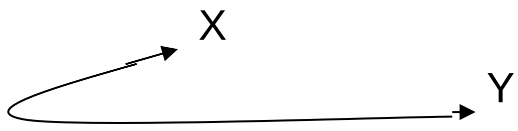
That is, in order for Z to become a confounder …

ZX: Z and X have their effects confounded.

ZY: Z is related to the effect Y (Z predicts Y).



If we "forget" Z, a spurious (false) relationship "comes up."



ZY Relationship is done: *Life Facts.*

	experimental		design
ZX relationship can be avoided in an	observational	study with a nice	analysis

***Confounding** literally means confusion of effects. A study might seem to show either an association or no association between an exposure and the risk of a disease. In reality, the seeming association or lack of association is due to another factor that determines the occurrence of the disease but that is also associated with the exposure.*

STROBE, Plos Medicine 2007

In order to avoid confounding, one needs to

a) be aware of all possible risk factors for the outcome of interest, and

b) control for Z by means of a **balanced design**, that is, an equal distribution of Z among treatment groups.

Center 1 (Z1)

	Y+	Y-	
X+	60	30	**90**
X-	60	30	**90**
	120	**60**	**180**

OR = **1**

Center 2 (Z2)

	Y+	Y-	
X+	30	60	**90**
X-	30	60	**90**
	60	**120**	**180**

OR = **1**

Overall

	Y+	Y-	
X+	90	90	**180**
X-	90	90	**180**
	180	**180**	**360**

OR = **1**

Interaction (effect modification)

Consider the following example, where we have a balanced design of exposure in both groups.

Center 1 (Z1)

	Y+	Y-	
X+	20	40	**60**
X-	40	20	**60**
	60	**60**	**120**

OR = 0,25

Center 2 (Z2)

	Y+	Y-	
X+	40	20	**60**
X-	20	40	**60**
	60	**60**	**120**

OR = 4

Overall

	Y+	Y-	
X+	60	60	**120**
X-	60	60	**120**
	120	**120**	**240**

OR = 1

If the association between X and Y differs across levels of a third variable Z, this is said to modify the effect of X on Y. That is, there is an interaction of X and Z.

In presence of interaction, an **overall measure is not meaningful**. Results should be presented for each category of the third variable.

Interaction does not depend on the distribution of Z among treatment groups (X).

Interaction

Effect of X on Y conditioning or adjusted by Z1.

The effect of X on Y changes under Z1 and Z2

Effect of X on Y conditioned or adjusted by Z1

Z1	Y+	Y-	Z2	Y+	Y-		Y+	Y-
X+	20	40	X+	60	30	X+	80	70
X-	40	20	X-	30	60	X-	70	80
OR=1/4			OR=4			OR≈1’31		

Effect of X on Y without conditioning by Z

Effect : ‘global’, ‘raw’, ‘aggregate’, ‘unadjusted’

Y+	Z1	Z2	Y-	Z1	Z2		Z1	Z2
X+	20	60	X+	40	30	X+	60	60
X-	40	30	X-	20	60	X-	90	90
OR=1/4			OR=4			OR=1		

X+	Z1	Z2	X-	Z1	Z2		Z1	Z2
Y+	20	60	Y+	40	30	Y+	51	330
Y-	40	30	Y-	20	60	Y-	155	26
OR=1/4			OR=4			OR=1		

Remark:

· An overall estimation of the effect of X on Y does not make sense.

· It appears although the design is balanced: X and Z are independent.

Confunding and effect modification

a) Confunding: different effect depending on adjustment or not.

"Argot" <u>Statistics' definition</u>: Z and X have their effects confounded by collinearity.

<u>Epidemiology's definition</u>: Z confounds the effect of X on Y. That is, Z and X are related (collinearity) and Z predicts the outcome Y.

b) Modification: different effect (adjusted) according to the attribute levels

"Argot":

Epidemiology: effect modification

Pharmacology: synergism and antagonism

Statistics: interaction

Biochemistry: catalyst, enzyme

Making confounding

1) Overall independence

Z1

	Y+	Y-	
X+	10	10	20
X-	10	10	20
	20	20	40

OR_{XY} = 1

OR_{ZX} = 1

OR_{ZY} = 1

Z2

	Y+	Y-	
X+	10	10	20
X-	10	10	20
	20	20	40

OR_{XY} = 1

Overall

	Y+	Y-	
X+	20	20	40
X-	20	20	40
	40	40	80

OR_{XY} = 1

2) Adding ZX relationship or *'collinearity'* (OR_{ZX} = **10^2**)
(We multiply x 10)

Z1

	Y+	Y-	
X+	*100*	*100*	**200**
X-	10	10	**20**
	110	110	220

OR_{XY} = 1

OR_{ZX} = 100 *(collinearity)*

OR_{ZY} = 1

Z2

	Y+	Y-	
X+	10	10	**20**
X-	*100*	*100*	**200**
	110	110	220

OR_{XY} = 1

Overall

	Y+	Y-	
X+	110	110	220
X-	110	110	220
	220	220	440

OR_{XY} = 1

3) Adding ZY relationship or *'Z predicts Y'* (OR_{ZY} = **5^2**)
(We multiply x 5)

Z1

	Y+	Y-	
X+	*500*	100	600
X-	*50*	10	60
	550	**110**	660

OR_{XY} = 1

OR_{ZX} = 100 *(collinearity)*

OR_{ZY} = 25 *(Z predicts Y)*

Z2

	Y+	Y-	
X+	10	*50*	60
X-	100	*500*	600
	110	**550**	660

OR_{XY} = 1

Overall

	Y+	Y-	
X+	510	150	660
X-	150	510	660
	660	660	1320

OR_{XY} = 11,6

Appendix

Another example of confounding

In the table is an example of Bishop et al (1975)-also analyzed by Freeman (1987) - about the evolution of a newborn (living, dying) depending on the length of maternal childbirth preparation care (> < 1) and the clinic (A,B).

1) Sort variables in groups X, Y, Z

2) Get OR values for care*dying giving clinic and overall

3) What do you think about the influence of care in the evolution?

	Clinic A		Clinic B		All	
	Dying	Living	Dying	Living	Dying	Living
Care <1	3	176	17	197	20	373
Care >1	4	293	2	23	6	316
OR	1’25		0’99		2’88	
$CI_{95\%}$OR*	0’28, 5’64		0’22, 4’57		1’12, 7’12	

** See next slide.*

Approximative confidence interval for OR

	Y+	Y-	
X+	**a**	**b**	a+b
X-	**c**	**d**	c+d
	a+c	b+d	

The (1-α)·100% confidence interval for OR is:

$\widehat{OR} \cdot \exp(\pm z_{1-\alpha/2} \cdot (\mathrm{Var}(\log(\widehat{OR})),$

where $\widehat{OR} = (a \cdot d) / (b \cdot c)$ *and* $\mathrm{Var}(\log(\widehat{OR})) = 1/a+1/b+1/c+1/d.$

Adjusted and unadjusted effects in General linear models (I)

a) Z and X independence: OR_{ZX} = 10·10 / 10·10 = 1

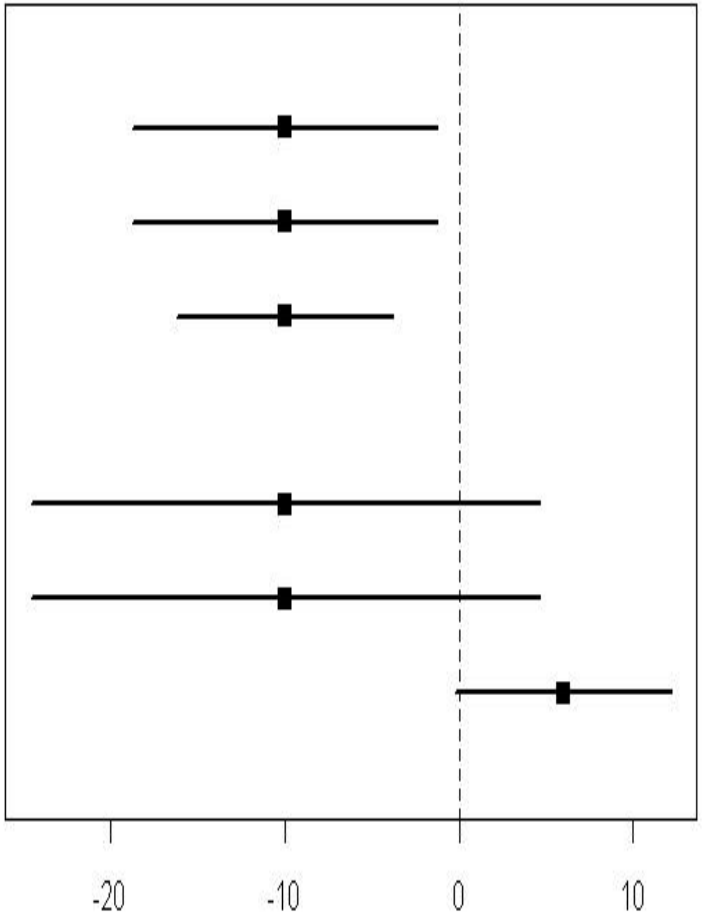
Mean (n)	Treated	Control	Mean (n)
Z1	110 (10)	120 (10)	115 (20)
Z2	130 (10)	140 (10)	135 (20)
Total	120 (20)	130 (20)	125 (40), SD=10

Equality of adjusted (-10) and raw effects (-10)

b) But Z and X collinearity: OR_{ZX} = 2·2 / 18·18 = 1/81

Mean (n)	Treated	Control	Mean (n)
Z1	110 (2)	120 (18)	119 (20)
Z2	130 (18)	140 (2)	131 (20)
Total	128 (20)	122 (20)	125 (40), SD=10

Different adjusted (-10) and raw effects (+6)



Adjusted and unadjusted effects in General**ized** linear models

In generalized linear models such propriety (additive SS decomposition) doesn’t exist, a challenge called ‘non collapsibility’ that requires the definition of 2 different concepts: ‘adjusted’ and ‘unadjusted’ effects.

Z and X independence OR_{ZX} = 18·18 / 18·18 = 1

N : Healthy / Dead	Treated	Control	OR
Z1	18: 15 / 3	18: 9 / 9	(15·9)/(3·9) = 5
Z2	18: 9 / 9	18: 3 / 15	(9·15)/(9·3) = 5
Total	36: 24 /12	36: 12 /24	(24·24)/(12·12)=4

different adjusted (5) and raw effects (4)

Unadjusted effects depend upon case-mix composition in the sample

Adjusted effects are always higher

Generalized linear models: counts, logistic, survival

Adjusted and unadjusted effects in General linear models (II)

Orthogonal columns in the design matrix X implies additive SS decomposition

Let the model be $\mathbf{Y} = \mathbf{X}\ \beta + \varepsilon$ subdividing in T components

matrix **X** $\mathbf{X} = \{ \mathbf{X}_1 , \mathbf{X}_2 , \ldots , \mathbf{X}_T \}$

and the vector β: $\beta' = \{ \beta_1' , \beta_2' , \ldots , \beta_T' \}$

so that: $E(\mathbf{Y}) = \mathbf{X}\ \beta = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \ldots + \mathbf{X}_T \beta_T$

If the columns of $\mathbf{X}_t$ are orthogonal ($\mathbf{X}_t' \mathbf{X}_{t'} = \mathbf{0}$) among them, then:

(1) $SS(\mathbf{b}) = SS(\mathbf{b}_1) + SS(\mathbf{b}_2) + \ldots + SS(\mathbf{b}_T)$

$= \mathbf{b}_1' \mathbf{X}_1' \mathbf{Y} + \mathbf{b}_2' \mathbf{X}_2' \mathbf{Y} + \ldots + \mathbf{b}_T' \mathbf{X}_T' \mathbf{Y}$

(2) $\mathbf{b}_t$ is the estimator of β_t

(3) $SC(\mathbf{b}_i) = \mathbf{b}_i' \mathbf{X}_i' \mathbf{Y}$

Independent of the inclusion of the remaining part of the orthogonal matrix in the model.

Exercise: $\mathbf{X} = \{ \mathbf{X}_1 , \mathbf{X}_2 \}$ with $\mathbf{X}_1' \mathbf{X}_2 = \mathbf{X}_2' \mathbf{X}_1 = \mathbf{0}$ (Orthogonal)

Check: (1) $\mathbf{b}_1 = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{Y}$ independent of the inclusion of $\mathbf{b}_2$ in the model

(2) $SS(\mathbf{b}_1, \mathbf{b}_2) = SS(\mathbf{b}_1) + SS(\mathbf{b}_2)$

$$SS(b_1)^{\text{in model } Y=f(X1)} = SS(b_1)^{\text{in model } Y=f(X1,X2)}$$

Meaning of “independent” variables

To study:
- Bivariant correlation matrix over the X.
- Multiple correlation coefficient between one predictor and the remaining ones: $R_{i,rest}^2$
- VIF (*Variance Inflation Factor*).
- P.C.A. among the predictors in order to identify the dimension of **X**.

  are the last eigenvalues close to 0?

  Condition index: $\sqrt{(\lambda_1/\lambda_k)}$ If >15 → Beware
  If >30 → Danger

To think:

The statistical modeling reports the effect of an "independent" variable.

But, to what extent can act on one variable without changing the other?

It makes no sense to introduce variables whose great "collinearity" raises the suspicion that, actually, you can not act on one "independently" of each other.

Select variables that provide independent information.

If it’s necessary, the predictors can be transformed:
- a priori, according to some criterion (sum, subtraction,...)
- a posteriori, through PCA (in order to identify components which influence in several variables)

Adjustments according to types of variables

Post variables: do not ever adjust

Pre variables: they should be included in adjustment.

If collinearity, they eliminate confounding effects.

If a predictor of response, they lower the residual variance and improve estimations.

Simultaneous ("concomitants"): is there an independent meaning?

Should they be combined in single measure?

Intermediate: Adjustment estimates the direct effect.

Do not adjust estimates overall effect.