Controlling Bias & Confounding

Chihaya Koriyama

August 18th, 2016

ERROR VS. BIAS

Two types of errors:
---Error or bias?

- Random error
  - ⇨ is the nature of quantitative data.

- Systematic error (=**bias**)
  - ⇨ should be minimized at the designing stage.

Random error	Systematic error
Measured value (mm)	Measured value (mm)
53	48
47	48
48	48
49	48
51	48
52	48
50	48
Mean=50	Mean=48

God knows that the true value is **50**mm.

Which is a proper comparison?

- Using accurate data
- Using inaccurate data

↓

Can' t we use our data when it is NOT accurately measured?

Is the following study acceptable?

- We want to compare the mean of blood pressure levels between two groups.
- The blood pressure checker has a problem and always gives 5mmHg-higher than true values.
- All subjects were examined by the same blood pressure checker.

Proper comparison between groups :

1） Comparison using accurate data

2） Comparison using (in)accurate data

As long as the magnitude of random error and bias occur in a same manner among comparison groups.

Q1. What would be the problem in this study?

- Although the blood pressure checker has a problem, giving always 5mmHg-higher than true values, all subjects were examined by the same blood pressure checker.
- We reported the results of this study.

FOR DISCRETE VARIABLES, MEASUREMENTS ERROR IS CALLED CLASSIFICATION ERROR OR MISCLASSIFICATION

Two types of misclassification

- Non-differential misclassification
  - Misclassification of a study variable that is independent of other study variables
  - Systematic error may not be a critical issue as long as it occurs in all comparison groups.
- Differential misclassification
  - If the error occurs only in one specific group due to bias, the risk estimate deviate from null.

Non-differential Misclassification with Two Exposure Categories

Study setting:
The proportion of subjects with serum antibody against *helicobacter pylori* is high among gastric cancer patients.

Correct Data	H.P-positive	H.P-negative
GC Cases	240	200
Controls	240	600

OR = 3.0

If the kit to detect H.P antibody has 80% sensitivity…

Correct Data	H.P-positive	H.P-negative
GC Cases	240	200
Controls	240	600

OR = 3.0

20% of exposed subjects were misclassified

Sensitivity = 0.8

Specificity = 1.0	H.P-positive	H.P-negative
GC Cases	192	248
Controls	192	648

OR = 2.61

Q2. What is the number of each cell? Please calculate OR.

Sensitivity = 0.8

Specificity = 0.8

	H.P-positive	H.P-negative
GC Cases		
Controls		

OR = □

Sensitivity = 0.4

Specificity = 0.6

	H.P-positive	H.P-negative
GC Cases		
Controls		

OR = □

Q3. We learned that misclassification gives us wrong results. Is this bias?

Q4. How do you solve the problem of non-differential misclassification?

BIAS IN EPIDEMIOLOGIC STUDY

Different types of bias

- **Selection bias:**

  It occurs at sampling

- **Detection bias:**

  It occurs at diagnosis (outcome)

- **Information (measurement) bias:**

  It occurs at data collection

  □Recall bias

  □Family information bias

You need to avoid these biases as much as possible in your study design.

SELECTION BIAS

Study setting:

You suspect that exposure to electromagnetic field (EMF) increases the risk of childhood leukemia.
And, you conducted a case-control study.

- If parents of cases with leukemia, living in the neighborhood of power lines, suspect the association and tend to agree on participation to the study,

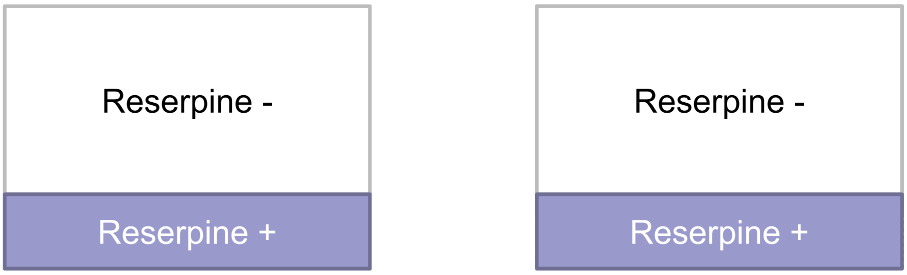
Q5. the association between EMF exposure and leukemia risk may become (stronger / weaker) than true association.

What is this bias? How do you solve it?

- If parents of controls, living in the neighborhood of power lines, tend to agree on participation to the study,

Q6. the association between EMF exposure and leukemia risk may become (stronger / weaker) than true association.

Is Reserpine a cause of breast cancer?



Reserpine -

Reserpine +

Reserpine -

Reserpine +

Cases: Breast cancer patients

Controls: Patients at the same hospital

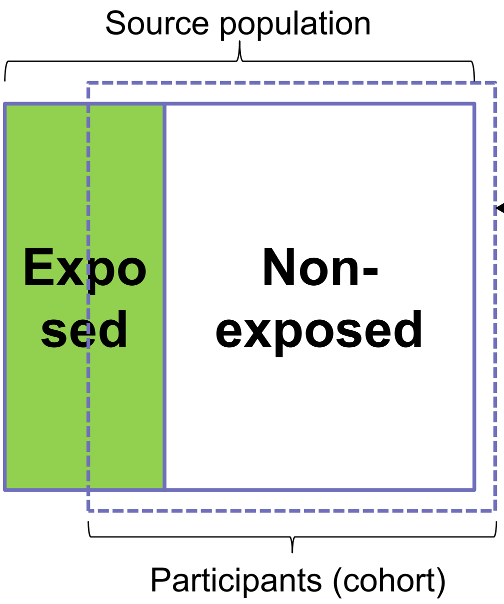
(Except who have cardiovascular diseases to which Reserpine is likely to be prescribed.)

Horwitz RI, Feinstein AR. Exclusion bias and the false relationship of reserpine and breast cancer. Arch Intern Med. 1985;145(10):1873-5.

Selection bias influences internal validity of the obtained results.

Q7. Is selection bias a matter in (prospective) cohort studies?

Selection bias: a cohort study



As a results, the proportion of exposed group may be different from that in the source population. However, it is not a problem as long as the incidence rates between participants and non-participants are the same.

HEALTHY WORKER EFFECT

Study setting:

You suspect that working at construction site is in danger, and thus, their mortality rate must be worse than general population.

Comparison mortality rate between labors at construction site and general population

	Labor at construction site	General population
Number of death	50	7,000
Person-year	1,000	100,000
Mortality rate	0.05	0.07

I am disappointed in my expectations…

Q8. Can you conclude that the mortality rate among labors working at construction site is lower than that of general population?

Q9. If you say "no", how do you solve this?

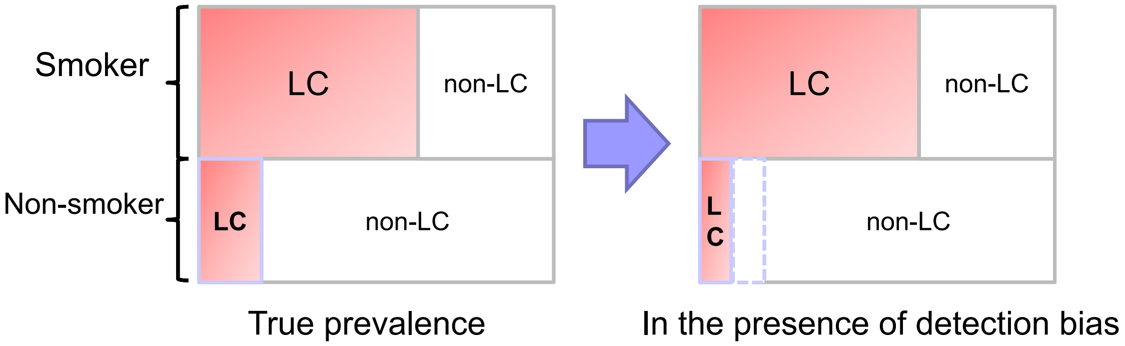
DETECTION BIAS

Study setting:

A doctor may examine the patient' s chest X-ray more carefully if he knew the patient is a heavy smoker but not for non-smoking patients.

Q10. The association between smoking and lung cancer risk may become (stronger / weaker) than what it should be.

The association between smoking and lung cancer risk becomes stronger.



Q11. How do you avoid detection bias?

INFORMATION BIAS

Study setting:

Suppose, you conducted a case-control study on relationship of prenatal infections and congenital malformations.

You asked mothers regarding *prenatal episode of infections by interview / questionnaire*.

Cases
(mothers of babies
with defect)

Controls
(mothers of
healthy babies)

Q12. What is a possible answers by control mothers?

Q13. How do you avoid /minimize the bias?

Controlling for misclassification

- - Blinding
  - □ prevents investigators and interviewers from knowing case/control or exposed/non-exposed status of a given participant
- - Form of survey
  - □ mail may impose less "white coat tension" than a phone or face-to-face interview
- - Questionnaire
  - □ use multiple questions that ask same information
- - Accuracy
  - □ Multiple checks in medical records & gathering diagnosis data from multiple sources

Lecture note of Dr. Dorak (http://www.dorak.info/epi)

Key concepts

- Bias

→ Should be minimized at the designing stage.

- Random errors

→ Is the nature of quantitative data.

- Non-differential misclassification

→ Is the nature of (inaccurate) measurement.

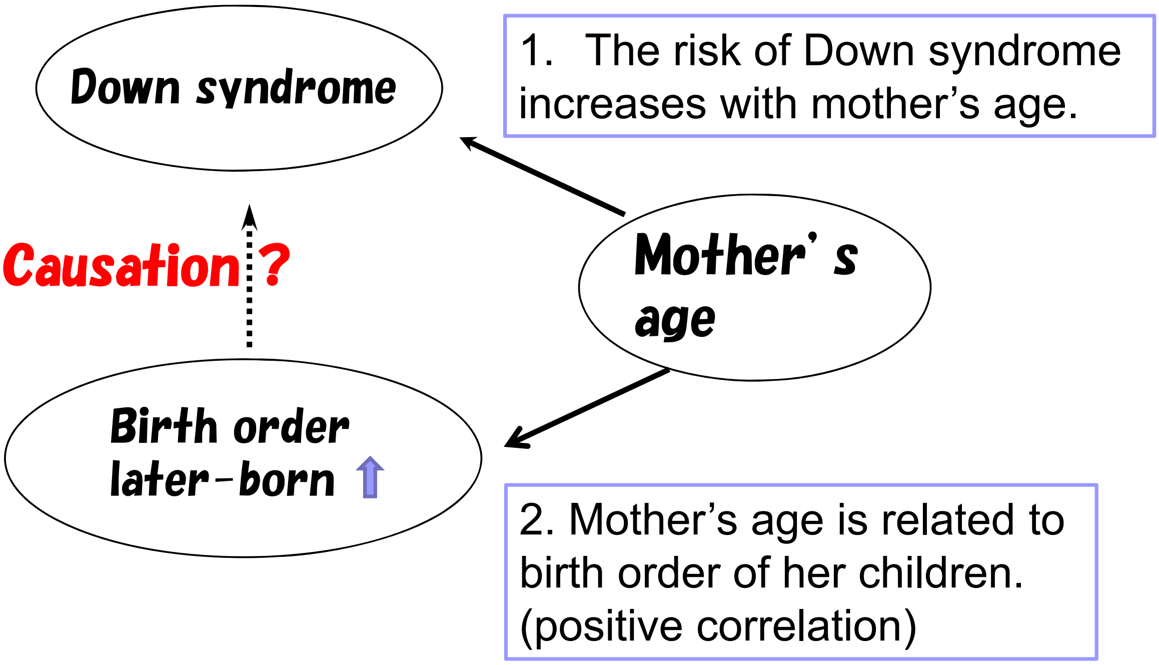
CONFOUNDING

3 conditions of Confounding

1. Confounders are **risk factors** for the outcome.
2. Confounders are **related to exposure** of your interest.
3. Confounders are **NOT on the causal pathway** between the exposure and the outcome of your interest.

Example of confounder

– mother's age is a confounder –



Down syndrome

Causation?

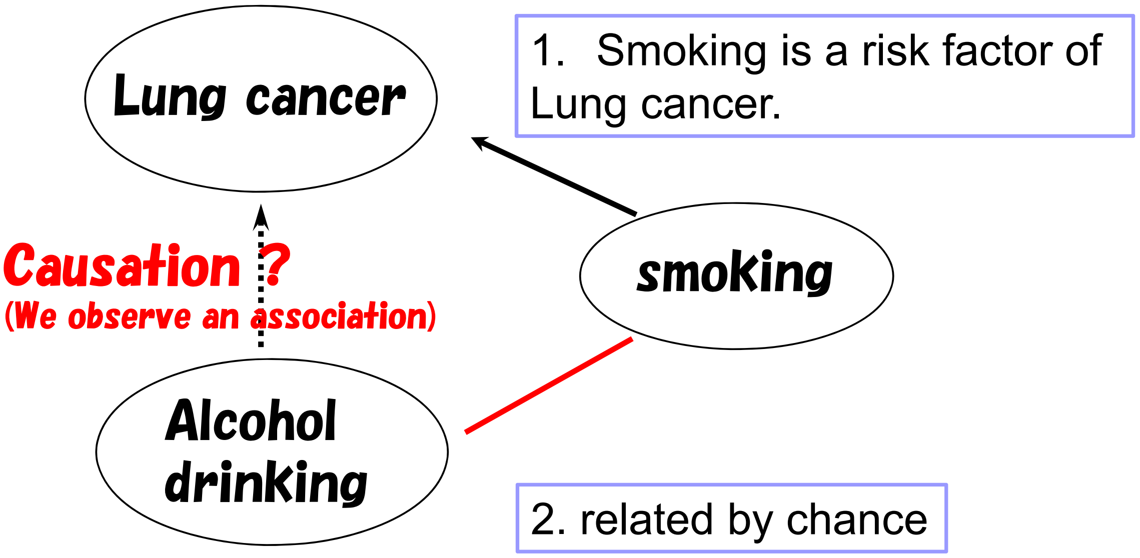
Birth order later-born ⇧

Mother's age

1. The risk of Down syndrome increases with mother's age.

2. Mother's age is related to birth order of her children. (positive correlation)

Example of confounder
− smoking is a confounder −



1. Smoking is a risk factor of Lung cancer.

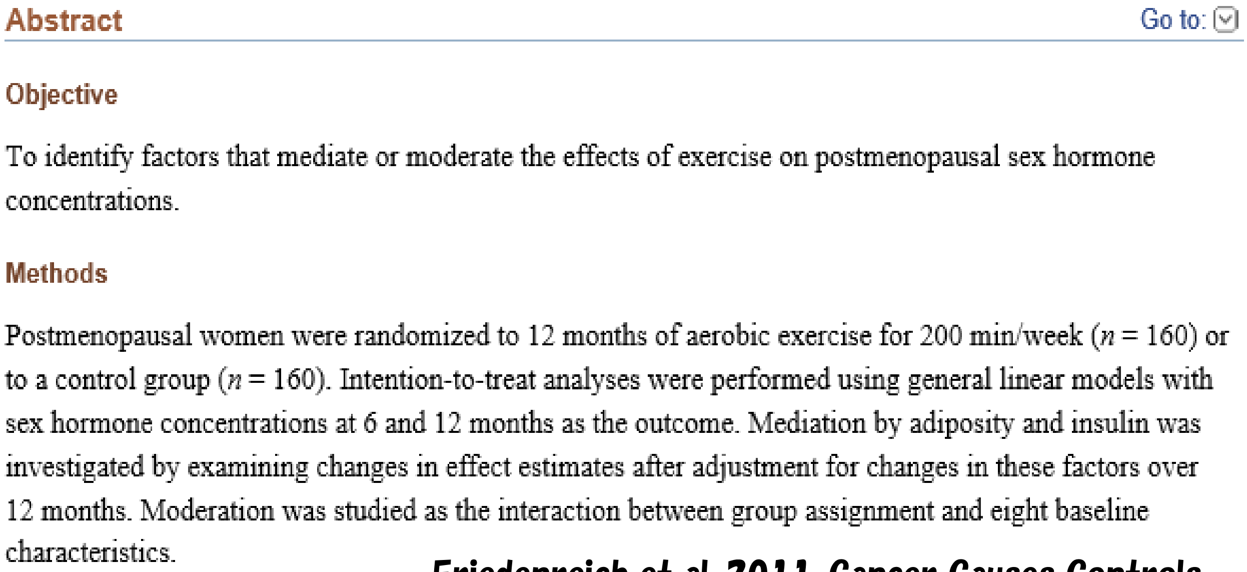
2. related by chance

How can we solve the problem of confounding?

"Prevention" at study design

- ✓Limitation
- ✓Randomization in an intervention study
- ✓Matching in a cohort study

Notice: Matching does not always prevent the confounding effect in a case-control study.



Abstract

Go to: ⌄

Objective

To identify factors that mediate or moderate the effects of exercise on postmenopausal sex hormone concentrations.

Methods

Postmenopausal women were randomized to 12 months of aerobic exercise for 200 min/week (*n* = 160) or to a control group (*n* = 160). Intention-to-treat analyses were performed using general linear models with sex hormone concentrations at 6 and 12 months as the outcome. Mediation by adiposity and insulin was investigated by examining changes in effect estimates after adjustment for changes in these factors over 12 months. Moderation was studied as the interaction between group assignment and eight baseline characteristics.

Friedenreich et al. 2011, Cancer Causes Controls

Q14. What is the main factor (exposure) in this study?

Q15. Did randomization work well to prevent confounding imbalances?

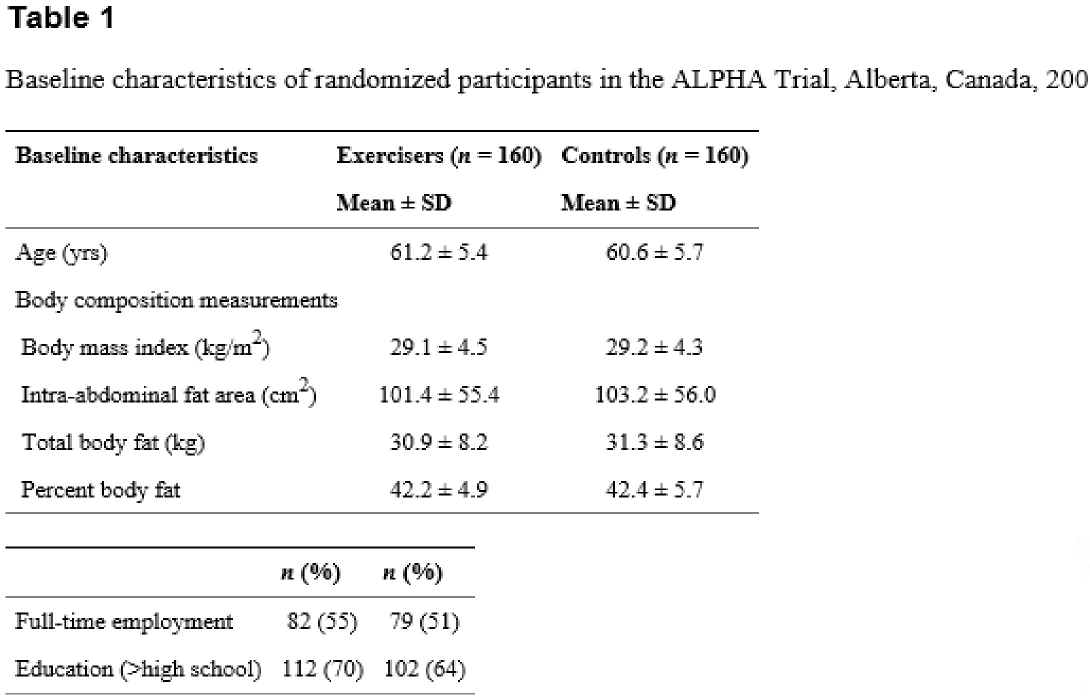


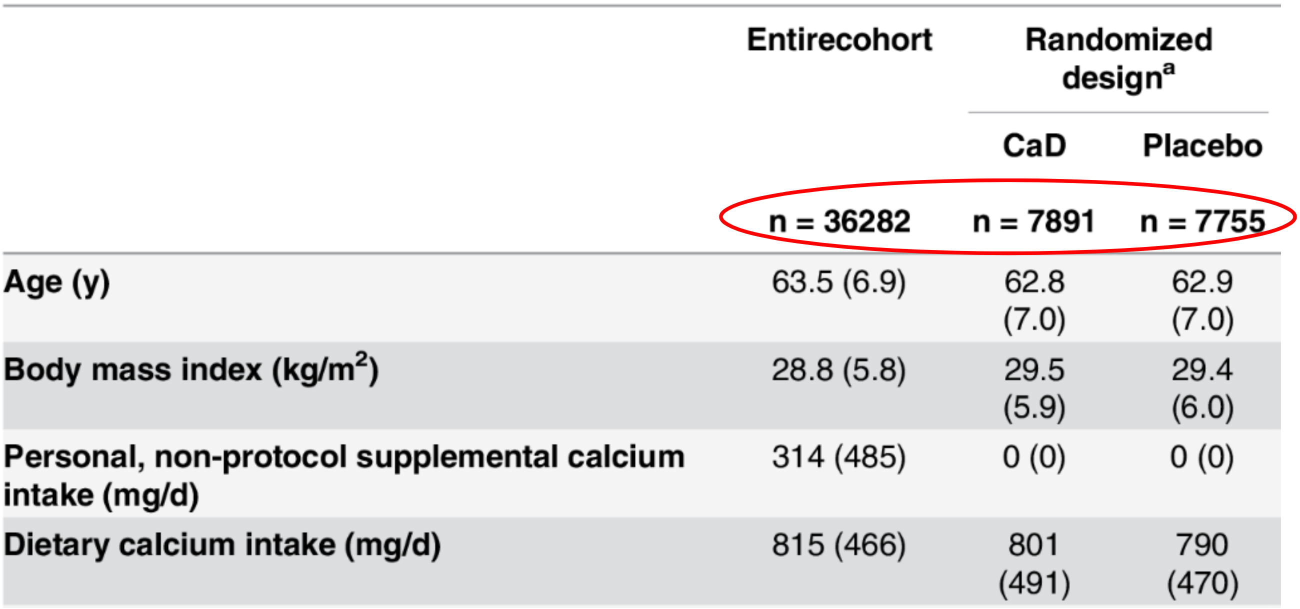
Table 1

Baseline characteristics of randomized participants in the ALPHA Trial, Alberta, Canada, 2003–2007, *n* = 320

Baseline characteristics	Exercisers (*n* = 160)	Controls (*n* = 160)
	Mean ± SD	Mean ± SD
Age (yrs)	61.2 ± 5.4	60.6 ± 5.7
Body composition measurements		
Body mass index (kg/m^2)	29.1 ± 4.5	29.2 ± 4.3
Intra-abdominal fat area (cm^2)	101.4 ± 55.4	103.2 ± 56.0
Total body fat (kg)	30.9 ± 8.2	31.3 ± 8.6
Percent body fat	42.2 ± 4.9	42.4 ± 5.7

	n (%)	*n* (%)
Full-time employment	82 (55)	79 (51)
Education (>high school)	112 (70)	102 (64)

Q16. Did randomization work well to prevent confounding imbalances?



	Entirecohort	Randomized design[a]	
		CaD	Placebo
	n = 36282	n = 7891	n = 7755
Age (y)	63.5 (6.9)	62.8 (7.0)	62.9 (7.0)
Body mass index (kg/m^2)	28.8 (5.8)	29.5 (5.9)	29.4 (6.0)
Personal, non-protocol supplemental calcium intake (mg/d)	314 (485)	0 (0)	0 (0)
Dietary calcium intake (mg/d)	815 (466)	801 (491)	790 (470)

Bolland et al. 2015 PLoS One

It is not desirable to use statistical significance testing (p value) to assess baseline differences in a trial.

- A large number of subjects improves confounding imbalances. However, it does not guarantee no confounding effect.
- Randomization is intended to prevent confounding. The outcome of a random process, however, is predictable only if aggregated over many repetitions.

Key concepts

- **Confounding**

→ Indicative of true association. Can be controlled at the designing or analysis stage.

We can do something even after conducting the survey.

Diagnosis of confounder

A case-control study for lung cancer

Is alcohol drinking a risk factor of LC?

		Lung cancer	Control
Alcohol	High	33	1,667
volume	Low	27	2,273

Odds ratio = (33*2273) / (1667*27) = 1.67

Diagnosis of confounder (contnd.)

Stratified by **smoking status (suspected confounder)**

		Smokers		Non-smokers	
		LC	Control	LC	Control
Alcohol volume					
	High	24	776	9	891
	Low	6	194	21	2,079
Odds ratio		24*194 / 776*6 = **1**		9*2079 / 891*21 = **1**	

An example of matching in a cohort study

	Exposed	Un-exposed
Lung cancer	1200	525
subjects	11000	11000

RR＝(1200/11000) / (525/11000) =2.3

Sex is a possible confounding factor.

Let’s see RR after stratification by sex

	Male		Female	
	Exp.	Un-exp.	Exp.	Un-exp.
Lung cancer	200	500	1000	25
subjects	1000	10000	10000	1000

Total: RR＝(1200／11000) / (525／11000) =2.3

Male: RR ＝(200／1000) / (500／10000) =4

Female: RR＝(1000／10000) / (25／1000)=4

Exposed and un-exposed group was matched by sex

	Male		Female	
	Exp.	Un-exp.	Exp.	Un-exp.
Lung cancer	2000	500	1000	250
subjects	10000	10000	10000	10000

Total: RR＝(3000／20000) / (750／20000) =4

Male: RR ＝(2000／10000) / (500／10000) =4

Female: RR＝(1000／10000) / (250／10000)=4

An example of matching in a case-control study

	case			control		
	male	female	total	male	female	total
Exposed	80	10	90	60	4	64
Non-exp.	20	90	110	40	96	136
Total	100	100	200	100	100	200

OR (total) ＝(90 x 136) / (110 x 64) =1.7

OR (male) ＝(80 x 40) / (20 x 60) =2.6

OR (female)＝(10 x 96) / (90 x 4)=2.6

How can we solve the problem of confounding?

“Treatment ” at statistical analysis

- ✓Stratification by a confounder
- ✓Multivariable / multiple analysis

Mantel-Haenszel odds ratio

- Stratification by confounding factor
  - □ After stratification by confounding factor, common OR, OR_{MH}, among all strata should be calculated.
  - □ Assumption: there is a common OR among all strata → there is no significant difference in ORs among all strata by homogeneity test.

An example of Mantel-Haenszel estimation 1

Calculate the common OR among all strata

smoking	Case	Control	
+	a_i	b_i	M_{1i}
-	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

$OR_c = \Sigma W_i OR_i / \Sigma w_i$

i :”i” th stratum、W_i :weight of “i” th stratum

How can we solve the problem of confounding?

“Treatment “ at statistical analysis

- ✓Stratification by a confounder
- ✓Multivariable / multiple analysis

Regression model

Paired?	Outcome variable	Proper model
No	Continuous	Liner regression model
	Binomial	Logistic regression model
	Categorical (≥3)	Multinomial (polytomous) logistic regression model
	Time length of the event including censoring	Cox proportional hazard model
Yes	Continuous	Mixed effect model, Generalized estimating equation
	Categorical (≥3)	Generalized estimating equation

How many explanatory variables can we use in a model?

Model	Number of explanatory variables	Example
Linear regression model	**Sample size / 15**	Up to around 6-7 variables in **100 subjects**
Logistic regression model	**Smaller sample size of outcome / 10**	Up to 10 variables if the numbers of cases and controls are **100** and 300, respectively.
Cox proportional hazard model	**The number of event / 10**	Up to 9 variables if you have **90 events** out of 150 subjects

ATTENTION!

- **When you include categorical variable in your model, you have to count that variable as (the number of categories – 1).**
  - □ For example, the variable of age group used in the previous practice, we have to count it as "two" (=3 categories – 1) variables.