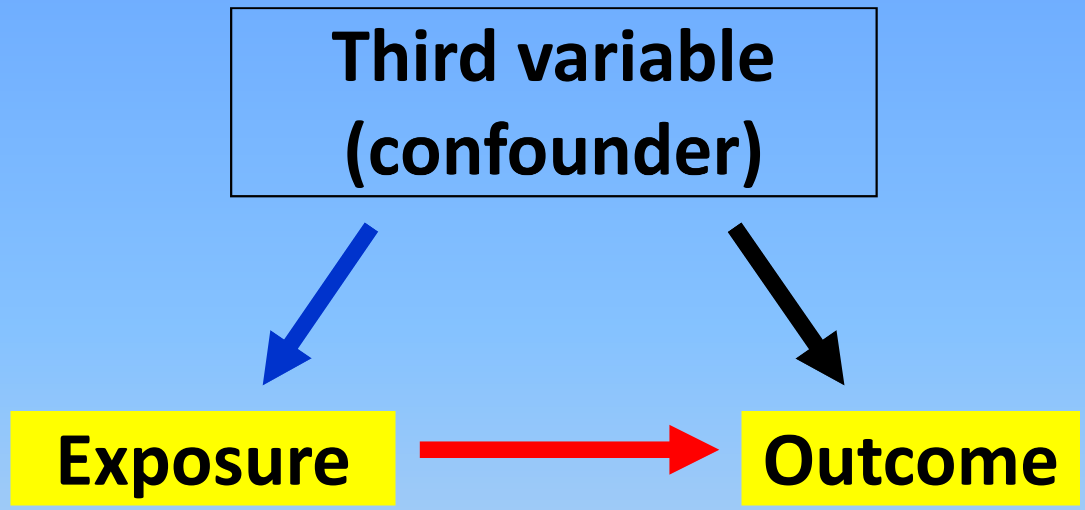
Matching and stratified analysis

Dr.Yongjua Laosiritaworn

Introductory on Field Epidemiology

1 July 2015, Thailand

Confounding



Must be a risk factor of outcome

Associated with exposure

Not an intermediate step between exposure – outcome

Control of Confounding

- **In study design**
  - Randomization
  - Restriction
- **In analysis**
  - Stratification
  - Adjustment/Standardization
  - Multivariable analysis
- **In study design and analysis**
  - Matching

Randomization

- Every individual has the same chance of being classified in either of the two groups.
- If sample size is big enough, two groups are comparable in terms of measured and unmeasured confounders.
- **Strength:**
    - Controls confounders even those unsuspected
    - Study groups are comparable
    - Permits evaluation of association between exposure and outcome for varying levels of the factor
- **Limitation:**
    - Not easy to perform
    - Ethical problems
    - Expensive

Restriction

- Putting admissibility criteria for subjects and limiting enrollment into the study to individuals who fall within a specified category or categories of the confounder.
- **Strength:**
  - Straightforward
  - Convenient if criteria are narrow
  - Inexpensive
- **Limitation:**
  - Reduces the number of subjects eligible to participate
  - Difficult if criteria are not narrow
  - Does not permit evaluation of association between exposure and outcome for varying levels of factor

Multivariable Analysis

- Analysis of data through construction of mathematical model that takes into account number of variables at the same time
- **Strength:**
  - Describes efficiently the association between exposure and outcome taking in consideration the impact of several other variables simultaneously.
- **Limitation:**
  - Many assumptions required for modeling
  - The choice of the appropriate model is complex and requires training and experience

Stratification

- Stratification is a technique used to control confounding in the analysis stage that involves the evaluation of the association within homogeneous categories or strata of the confounding factor
- Involves separating a sample into two or more subgroups according to specified levels of a third variable

Stratification

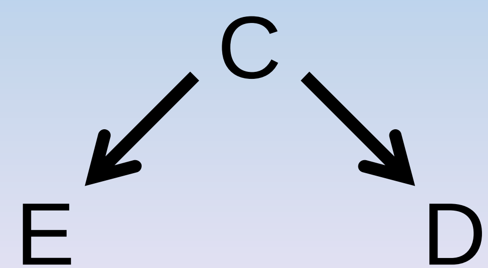
Example: A Case-control Study

Crude 2x2 table

	D+	D-	Total
E+	1000	838	1838
E-	100	262	362
Total	1100	1100	2200

OR = (1000 x 262)/(838 x 100)

= 3.13



Question: Is the OR distorted due to confounding?

Determine the OR of the exposure (E) separately for C+ and C

Crude 2x2 table

Crude OR = 3.13

	D+	D-	Total
E+	1000	838	1838
E-	100	262	362
Total	1100	1100	2200

In C+

	D+	D-	Total
E+	900	819	1719
E-	10	91	101
Total	910	910	1820

Stratum-specific OR = 10

In C-

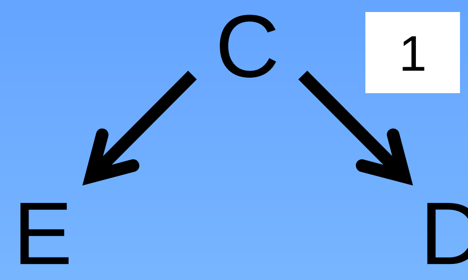
	D+	D-	Total
E+	100	19	119
E-	90	171	261
Total	190	190	368

Stratum-specific OR = 10

Adjusted OR = 10

Crude OR = 3.13

	D+	D-	Total
E+	1000	838	1838
E-	100	262	362
Total	1100	1100	2200



1. Determine, separately for E+ and E- , whether the confounder (C) and the outcome (D) are associated.

In E+

	D+	D-	Total
C+	900	819	1719
C-	100	19	119
Total	1000	838	1838

OR = 0.2

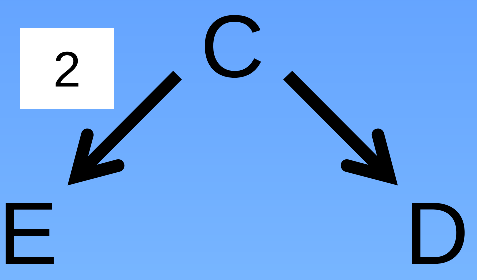
In E-

	D+	D-	Total
C+	10	91	101
C-	90	171	261
Total	100	262	362

OR = 0.2

Crude OR = 3.13

	D+	D-	Total
E+	1000	838	1838
E-	100	262	362
Total	1100	1100	2200



2. Determine, separately for D+ and D-, whether the confounder (C) and the Exposure (E) are associated.

In D+

	C+	C-	Total
E+	900	100	1000
E-	10	90	100
Total	910	190	1100

OR = 81

In D-

	C+	C-	Total
E+	819	19	838
E-	91	171	262
Total	910	190	1100

OR = 81

3. We must determine whether it is safe to assume C is not a link in the causal chain between RF and D.

Depends on existing content knowledges or theories

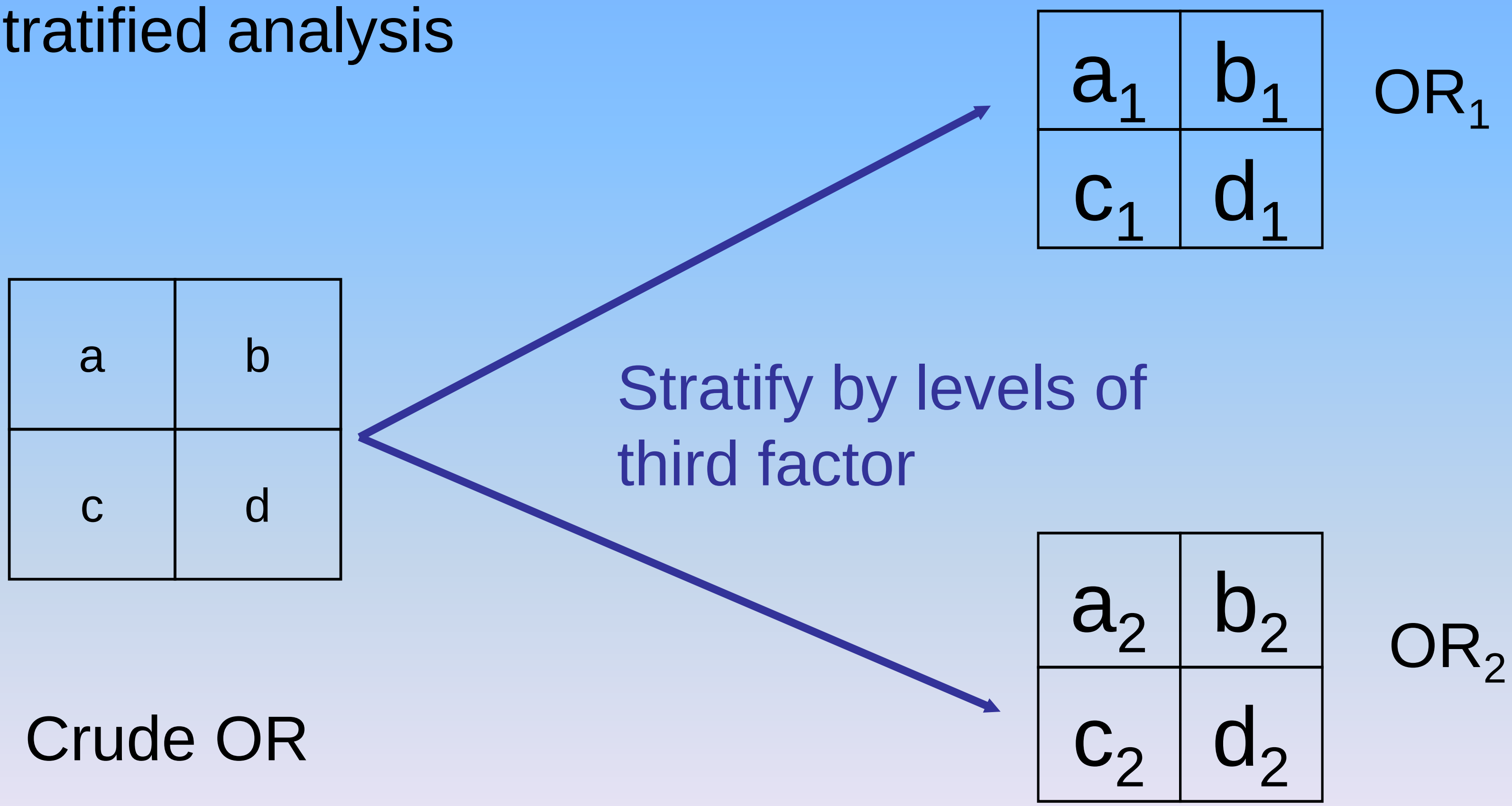
e.g. patho-physiology of diseases

If this assumption can be made we can conclude that C is a confounder of D.

Strategy to take into account a third factor in data analysis

1) Crude analysis

2) Stratified analysis



Crude OR

Strategy to take into account a third factor in data analysis

3) Compare stratified ORs : Woolf test for heterogeneity

4) Where is the crude OR?

Strategy to take into account a third factor in data analysis

5a)

Woolf test: $OR_1 \neq OR_2$

Crude OR → OR_1, OR_2

Third factor = Effect modifier

No computation of adjusted OR

Stratum-specific results of the association between exposure and outcome

Strategy to take into account a third factor in data analysis

5b)

Woolf test: OR1 ≈ OR2

OR_1

OR_2

Crude OR

Computation of Mantel-Haenszel adjusted OR

Strategy to take into account a third factor in data analysis

5b)

	D+	D-	Total
E+	a	b	N_1
E-	c	d	N_0
Total	M_1	M_0	N

Computation of Mantel-Haenszel Adjusted Odds Ratio
($OR_{M\text{-}H}$ or Adjusted OR)

$$OR_{M\text{-}H} = \frac{\Sigma\,[\,(a_i d_i) / N_i]}{\Sigma\,[\,(b_i c_i) / N_i]}$$

Strategy to take into account a third factor in data analysis

if $OR_{M-H} \neq OR_{Crude}$ (no statistical test; somebody suggest differ more than 10-15%)

and

if Third factor complies the conditions

then:

Third factor = Confounder

Crude OR is wrong

Proper measure of association between exposure and outcome given by adjusted OR_{M-H}

Strategy to take into account a third factor in data analysis

5c)
Woolf test: OR1 ≈ OR2

Crude OR → OR_1, OR_2

$OR_{M\text{-}H} \approx OR_{Crude}$

Third factor = no role

Use crude OR to measure the association between exposure and outcome

For Cohort Study (Count Data)

	D+	D-	Total
E+	a	b	N_1
E-	c	d	N_0
Total	M_1	M_0	N

Computation of Mantel-Haenszel Adjusted Risk Ratio
($RR_{M\text{-}H}$ or Adjusted RR)

$$RR_{M\text{-}H} = \frac{\Sigma\, [\, (a_i N_{0i}) \,/\, N_i]}{\Sigma\, [\, (c_i N_{1i}) \,/\, N_i]}$$

For Cohort Study (Person-Time Data)

	No. of Case	Person-Time
E+	a	T_1
E-	b	T_0
Total	M	T

Computation of Mantel-Haenszel adjusted Rate Ratio (IRR_{M-H} or Adjusted IRR)

$$IRR_{M-H} = \frac{\Sigma [(a_i T_{0i}) / T_i]}{\Sigma [(b_i T_{1i}) / T_i]}$$

Stratification

- **Strength:**
  - Easy for limited variables with limited number of categories
  - Permits evaluation of confounding and interaction
  - Permits evaluation of association between exposure and outcome for varying levels of the factor
- **Limitation:**
  - Difficult if many variables with varying number of categories are required

Matching

Matching

- **Ensures that confounding factor is equally distributed among both study groups**
  - Case – control studies: controls selected to match specific characteristics of cases
  - Cohort studies: unexposed selected to match specific characteristics of exposed
- **Balanced data set achieved**
  - Prevents confounding
  - Increase study precision / efficiency

Focus on case-control studies

Types of matching

1. Individual matching

- Controls selected for each individual case by matching variable / variables
- 1 case : 1 control - pairs of individuals
- 1 case : n controls – triplets, quadruplets, ...
- Continuous variable
  - Exact matching: e.g. age 42 yr vs 42 yr
  - Caliper matching: e.g. age 42 yr vs 42±5yr
- Categorical variable:
  - Stratum matching: e.g. male vs male

Types of matching

2. Frequency matching

- Controls selected in categories of matching variable according to the distribution of matching variable among cases
- Start recruit controls after we get all cases.

In both types, in analysis we must take matching design into account

- Stratified analysis

Individual matching (1:1)

- Echovirus meningitis outbreak, Germany, 2001
- Was swimming in pond "A" risk factor?
- Case control study with each case matched to one control

		Controls		
		Exposed	Unexposed	Total
Cases	Exposed	194	46	240
	Unexposed	6	29	35
Total		200	75	275

Concordant pairs

Discordant pairs

Source: A Hauri, RKI Berlin

Individual matching (1:1)

Matched 2x2 table

		Controls		
		Exposed	Unexposed	Total
Cases	Exposed	194	46	240
	Unexposed	6	29	35
Total		200	75	275

~~**OR = 2.6**~~

Unmatched 2x2 table

	Cases	Controls	Total
Exposed	**240**	**200**	**440**
Unexposed	**35**	**75**	**110**
	275	**275**	**550**

Individual matching: Analysis

- **Stratified analysis**
  - **Each pair, triplet, quadruplet, ... a stratum**
  - **Calculate Mantel-Haenszel odds ratio**

$$OR_{M\text{-}H} = \frac{\Sigma\,[\,(a_i\, d_i) / N_i]}{\Sigma\,[\,(b_i\; c_i) / N_i]}$$

	D+	D-	Total
E+	a	b	N_1
E-	c	d	N_0
Total	M_1	M_0	N

Individual matching 1:1 – 1 pair a stratum

Matched 2x2 table

		Controls	
		Exposed	Unexposed
Cases	Exposed	e	f
	Unexposed	g	h

Individual Matching (1:1): Analysis

		Controls	
		Exposed	Unexposed
Cases	Exposed	e	f
	Unexposed	g	h

$$OR_{M\text{-}H} = \frac{\Sigma [(a_i d_i) / N_i]}{\Sigma [(b_i c_i) / N_i]}$$

Situation e

	Case	Control	Total	ad/N	bc/N
Exposed	1	1	2	0/2	0/2
Unexposed	0	0	0		
Total	1	1	2		

Individual Matching (1:1): Analysis

		Controls	
		Exposed	Unexposed
Cases	Exposed	e	f
	Unexposed	g	h

$$OR_{M\text{-}H} = \frac{\Sigma\,[\,(a_i\, d_i) / N_i]}{\Sigma\,[\,(b_i\; c_i) / N_i]}$$

Situation f

	Case	Control	Total	ad/N	bc/N
Exposed	1	0	1	1/2	0/2
Unexposed	0	1	1		
Total	1	1	2		

Individual Matching (1:1): Analysis

		Controls	
		Exposed	Unexposed
Cases	Exposed	e	f
	Unexposed	g	h

$$OR_{M\text{-}H} = \frac{\Sigma [(a_i d_i) / N_i]}{\Sigma [(b_i c_i) / N_i]}$$

Situation g

	Case	Control	Total	ad/N	bc/N
Exposed	0	1	1	0/2	1/2
Unexposed	1	0	1		
Total	1	1	2		

Individual Matching (1:1): Analysis

		Controls	
		Exposed	Unexposed
Cases	Exposed	e	f
	Unexposed	g	h

$$OR_{M\text{-}H} = \frac{\Sigma [(a_i d_i) / N_i]}{\Sigma [(b_i c_i) / N_i]}$$

Situation h

	Case	Control	Total	ad/N	bc/N
Exposed	0	0	0	0/2	0/2
Unexposed	1	1	2		
Total	1	1	2		

Individual Matching (1:1): Analysis

	ad/N	bc/N
Situation e	0	0
Situation f	1/2	0
Situation g	0	1/2
Situation h	0	0

$$OR_{M-H} = \frac{\sum [a_i d_i / N_i]}{\sum [b_i c_i / N_i]} = \frac{0\text{e} + 1/2\text{f} + 0\text{g} + 0\text{h}}{0\text{e} + 0\text{f} + 1/2\text{g} + 0\text{h}} = \frac{\text{f}}{\text{g}}$$

$$= \frac{\sum \text{discordant pairs where case exposed}}{\sum \text{discordant pairs where control exposed}}$$

Individual Matching (1:1): Analysis

Echovirus meningitis outbreak, Germany, 2001
Was swimming in pond “A” risk factor?
Case control study with each case matched to one control

		Controls		
		Exposed	Unexposed	Total
Cases	Exposed	194	46	240
	Unexposed	6	29	35
Total		200	75	275

$$OR_{M\text{-}H} = \frac{f}{g} = \frac{46}{6} = 7.67$$

Matching 1 case to n controls - analysis

- **Same principle as 1:1 matching (pair = stratum)**

- **Constitute**
  - Triplet (1 case, 2 controls) yields 2 pairs
  - Quadruplet (1 case, 3 controls) yields 3 pairs

- **Stratified analysis**
  - Each triplet, quadruplet, ... a stratum
  - Only discordant pairs (within triplets, quadruplets, ..) contribute to the OR_{M-H} estimate:

$$OR_{M-H} = \frac{\text{Sum of discordant pairs with exposed case} \quad (Ca+/Co-)}{\text{Sum of discordant pairs with exposed control} \quad (Ca-/Co+)}$$

Matching: 1 case to 2 controls (triplets)

Controls: exposed (+) unexposed (-)

Cases			
Exposed	a + / + + 0 DPs	b + / + - 1 DP	c + / - - 2 DPs
Unexposed	d - / + + 2 DPs	e - / + - 1 DPs	f - / - - 0 DPs

$$OR_{MH} = \frac{(a \times 0DPs_{Ca+/Co-}) + (b \times 1DP_{Ca+/Co-}) + (c \times 2DPs_{Ca+/Co-})}{(d \times 2DPs_{Ca-/Co+}) + (e \times 1DPs_{Ca-/Co+}) + (f \times 0DP_{Ca-/Co+})}$$

Matching: 1 case to 3 controls (quadruplets)

Controls: exposed (+) unexposed (-)

Cases				
Exposed	**a** + / + + + 0 DPs	**b** + / + + - 1 DP	**c** + / + - - 2 DPs	**d** + / - - - 3 DPs
Unexposed	**e** - / + + + 3 DPs	**f** - / + + - 2 DPs	**g** - / + - - 1 DPs	**h** - / - - - 0 DPs

$$OR_{MH} = \frac{(a \times 0DPs)+(b \times 1DP)+(c \times 2DPs)+(d \times 3DPs)}{(e \times 3DPs)+(f \times 2DPs)+(g \times 1DP)+(h \times 0DPs)}$$

Numerator: (Ca+/Co-)

Denominator: (Ca-/Co+)

Frequency (group) matching

Controls selected in categories of matching variable according to the distribution of matching variable among cases; confounding factor is equally distributed

Age (yrs)	Cases	Controls, matched
0-14	10	10
15-29	15	15
30-44	35	35
>44	25	25
Total	85	85

Frequency matching: Analysis

Age (yrs)	Cases	Controls, matched
0-14	10	10
15-29	15	15
30-44	35	35
>44	25	25
Total	85	85

Stratum 1

0-14 yrs	Cases	Controls	Total
Exp	6	1	7
N_exp	4	9	13
Total	10	10	20

Stratum 2

15-29 yrs	Cases	Controls	Total
Exp	7	5	12
N_exp	8	10	18
Total	15	15	30

Stratum 3
Stratum 4

Strata according to categories / levels of confounding variable used for frequency matching.

Why stratified analysis when matching?

- **Matching eliminates confounding, however, introduces bias**
- **Controls not representative of source population as selected according to matching criteria (selection bias)**
- **Cases and controls more alike.**
  By breaking match, OR usually <u>underestimated</u>
- **Matched design => matched analysis**

Analysis of matched data

- **Frequency matching**
  - With many strata (matching for > 1 confounder, numerous nominal categories) - sparse data problem
  - Multivariate analysis
- **Individually matched data - conditional logistic regression**
  - Logistic regression for matched data
  - “Conditional“ on using discordant pairs only
  - Matching variable itself cannot be analysed
  - Testing for interaction of matching variable possible

Overmatching

- **Matching variable “too closely related” associated with with exposure (not disease)**
  (increase frequency of exposure-concordant pairs)

  => association obscured

- **Matching variable is not a confounder**
  (associated with disease, but not exposure)

  => statistical efficiency reduced

- **Matching process too complicated**

  => difficulty in finding controls

Example: Overmatching

- **20 cases of cryptosporidiosis**
- **? associated with attendance at local swimming pool**
- **Two matched case-control studies**
  - Controls from same general practice and nearest date of birth
  - Cases nominated controls (friend controls)

Overmatching

		Controls	
		Exposed	Unexposed
Cases	Exposed	1	15
	Unexposed	1	3

GP, age - matched

OR_{MH} = f/g = 15/1 = 15

		Controls	
		Exposed	Unexposed
Cases	Exposed	13	3
	Unexposed	1	3

Friend - matched

OR_{MH} = f/g = 3/1 = 3

Advantages of matching

- **Useful method in case-control studies to control confounding**
- **Can control for complex environmental, genetic, other factors**
  - Siblings, neighbourhood, social and economical status, utilization of health care
- **Can increase study efficiency, optimise resources in small case-contol studies**
  - Overcomes sparse-data problem by balancing the distribution of confounders in strata
  - Case-control study (1:1) is the most statistically efficient design
  - When number of cases is limited (fixed) statistical power can be increasesd by 1:n matching (< 1:4 power gain small)
- **Sometimes easier to identify controls**
  - Random sample may not be possible

Disadvantages of matching

- Cannot assess the main effect of matching variable on the disease
- Overmatching on exposure will bias OR towards 1
- Complicates statistical analysis (additional confounders?)
- Residual confounding by poor definition of strata

- Sometimes difficult to identify appropriate controls
- If no controls identified, lose case data

Final Messages

- **Do not match routinely**
  - "Unless one has very good reasons to match, one is undoubtedly better of avoiding the inclination."
- **Useful technique if employed wisely**
  - Prevents confounding (balanced data sets)
  - Can control for complex factors (difficult to measure)
  - Increase precision / efficiency
- **If you match**
  - **make sure you match on a confounder**
  - **do matched analysis**

Further Readings

- **Epidemiology Kept Simple, 2nd , B. Gertsman.**
- **Epidemiology: Concepts and Methods, 1st Ed., WA. Oleckno.**
- **Modern Epidemiology, 3rd Ed., KJ. Rothman et al.**

Thank you

Acknowledgements

Dr. Panithee Thammawijaya