

# **Week 3: Matched Study Designs & Analysis**

## **Video 6: Matched Cohort Studies**

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH

Department of Epidemiology, Harvard TH Chan School of Public Health



**HARVARD T.H. CHAN**  
**SCHOOL OF PUBLIC HEALTH**

# Key Concepts

- Matched cohort studies
- Matched case-control studies
  - Impact of matching on validity and precision
  - Appropriate, unnecessary and over-matching
- Stratification in matched case-control studies
- Relative efficiency of matching ratios
- Notation
- Analysis of pair-matched case-control studies
  - Hypothesis tests
  - Point and interval estimation
- Relationship between McNemar estimator and  $OR_{MH}$

# Utility of Matching

- Matching can be done in the design phase of a study
  - In matched cohort studies, the exposed are matched to the unexposed group on one or more matching factors
  - In case-control studies, the cases are matched to the controls on one or more matching factors
- Matching is most useful in small studies with several confounding variables of nominal nature
- Individual matching and frequency matching are conceptually similar

# Matched Cohort Studies

## Motivation

- Matching can be used in cohort studies to reduce confounding and remove bias
- We can match on known confounders that are widely distributed in the population, such as gender
- We can also match on factors that are (relatively) unique to a given study participant such as:
  - sibship
  - genotype
  - multiple factors simultaneously (age, gender, ethnicity, neighborhood).

# Matched Cohort Studies

## Design

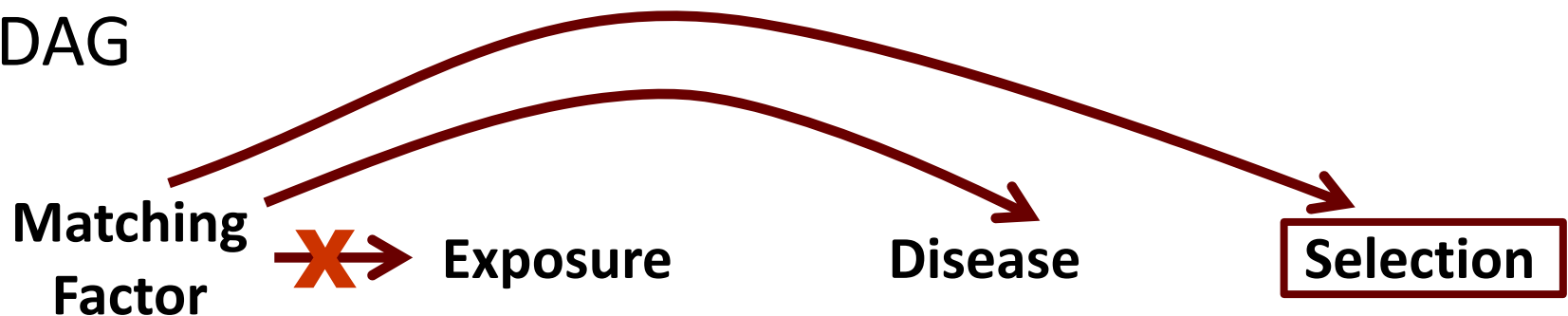
- In a cohort study, the exposed group is matched to the unexposed group on the matching factor(s)
- DAG
- Matching in a cohort study prevents an association between the exposure and the matching factor(s)

# Matched Cohort Studies

## Design

- In a cohort study, the exposed group is matched to the unexposed group on the matching factor(s)

- DAG



- Matching in a cohort study prevents an association between the exposure and the matching factor(s)

# Matched Cohort Studies

## Validity and Precision

- Matching prevents confounding even in crude (marginal) analysis
- If the planned analysis will focus on rates, optimal matching should focus on person-time (rather than persons), but this is unattainable

# Matched Cohort Studies

## Evaluable Associations

- The association between the matching factor(s) and the outcome of interest can be evaluated in cohort studies
- In matched cohort studies, one can evaluate whether the matching factor(s) modify the association between the primary exposure of interest and the outcome



**BREAK**

# **Week 3: Matched Study Designs & Analysis**

## **Video 7: Matched Case-control Studies**

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH

Department of Epidemiology, Harvard TH Chan School of Public Health



**HARVARD T.H. CHAN**  
**SCHOOL OF PUBLIC HEALTH**

# Key Concepts

- Matched cohort studies
- Matched case-control studies
  - Impact of matching on validity and precision
  - Appropriate, unnecessary and over-matching
- Stratification in matched case-control studies
- Relative efficiency of matching ratios
- Notation
- Analysis of pair-matched case-control studies
  - Hypothesis tests
  - Point and interval estimation
- Relationship between McNemar estimator and  $OR_{MH}$

# Matched Case-control Studies

## Motivation

- Matching can be used in case-control studies to improve precision, **but does not by itself remove bias.**
- Like in cohort studies, we can match on known confounders that are widely distributed in the population, such as gender.
- We can also match on factors that are (relatively) unique to case and control sets such as:
  - sibship
  - genotype
  - multiple factors simultaneously (age, gender, ethnicity, neighborhood).

# Matched Case-control Studies

## Introduction of Selection Bias By Design

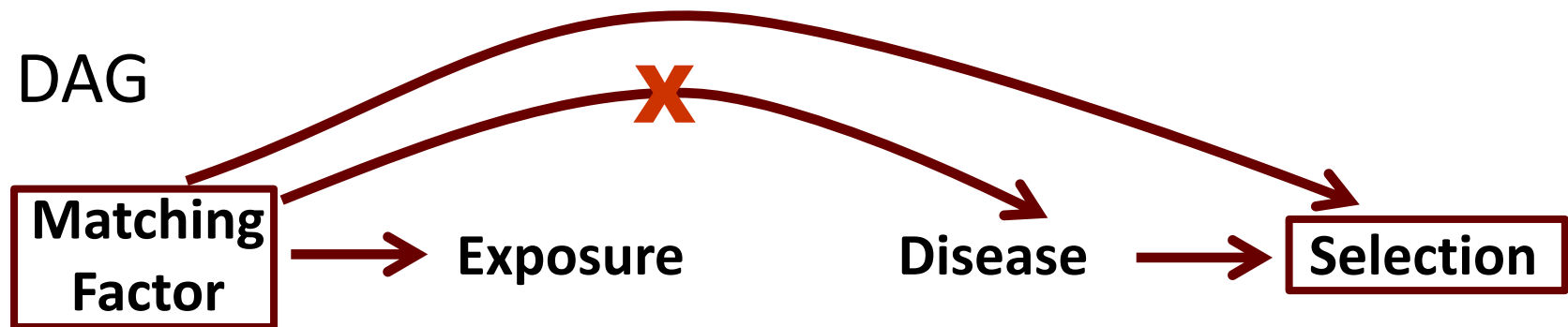
- In a case-control study, the case group is matched to the control (referent) group on the matching factor(s)
- DAG
- Matching in a case-control study abolishes any association between the matching factor(s) and the outcome of interest

# Matched Case-control Studies

## Introduction of Selection Bias By Design

- In a case-control study, the case group is matched to the control (referent) group on the matching factor(s)

- DAG



- Matching in a case-control study abolishes any association between the matching factor(s) and the outcome of interest

# Matched Case-control Studies

## Young Age at First Term Pregnancy and Cervical Cancer (1)

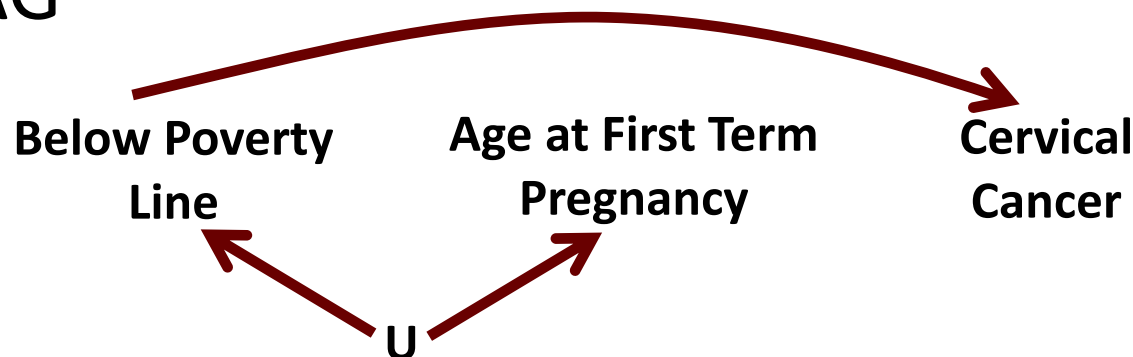
- Case-control study of the association between young age at first term pregnancy (<16 versus nulliparous)
  - Living below the poverty line during childhood is associated with full-term pregnancies at an early age and is also associated with a higher incidence of cervical cancer
  - Are you concerned about confounding?
  
- DAG

# Matched Case-control Studies

## Young Age at First Term Pregnancy and Cervical Cancer (1)

- Case-control study of the association between young age at first term pregnancy (<16 versus nulliparous)
  - Living below the poverty line during childhood is associated with full-term pregnancies at an early age and is also associated with a higher incidence of cervical cancer
  - Are you concerned about confounding?

- DAG





# Matched Case-control Studies

## Young Age at First Term Pregnancy and Cervical Cancer (2)

- Stratification (No matching)

Above Poverty Line			
	Pregnancy <16	Nulliparous	Total
Case	a <sub>1</sub>	b <sub>1</sub>	40
Control	c <sub>1</sub>	d <sub>1</sub>	90

Below Poverty Line			
	Pregnancy <16	Nulliparous	Total
Case	a <sub>2</sub>	b <sub>2</sub>	60
Control	c <sub>2</sub>	d <sub>2</sub>	10

- Is the analysis valid?
- Notice the imbalance in the number of cases and controls in each stratum. This will lead to a decrease in *precision*.
- Could we design a study with the same total number of participants, but a better balance in the number of cases and controls in each stratum?

# Matched Case-control Studies

## Young Age at First Term Pregnancy and Cervical Cancer (3)

- Stratification (Matching)

Above Poverty Line			
	Pregnancy <16	Nulliparous	Total
Case	$a_1$	$b_1$	40
Control	$c_1$	$d_1$	40

Below Poverty Line			
	Pregnancy <16	Nulliparous	Total
Case	$a_2$	$b_2$	60
Control	$c_2$	$d_2$	60

- Is the analysis valid?
- Is the analysis more precise than the unmatched design?

# Matched Case-control Studies

## Validity and Precision

- Matching can improve precision (efficiency) but it is not, in itself, a measure to achieve (or improve) validity
- If the matching factor(s) are associated with the exposure of interest, matching will cause the exposure distribution among the control group to be more similar to the cases than the distribution of exposure in the study base
- This is a form of selection bias (collider-stratification bias)
- The bias which is introduced by matching has a tendency towards the null value
- This bias can be eliminated by accounting for the matching factor(s) in the analysis e.g. by conditioning

# Matched Case-control Studies

## Evaluable Associations

- The association between the matching factor(s) and the outcome of interest cannot be evaluated in case-control studies
- In matched case-control studies, one can evaluate whether the matching factor(s) modify the association between the primary exposure of interest and the outcome

# Matched Case-control Studies

## Analysis

- Confounding and matching should be looked upon conditionally on other confounding factors, which are already accounted for
- If you have matched (for better or worse), take matching into account in the analysis (except if matching factor and exposure are clearly and definitely not associated)

**BREAK**

## **Week 3: Matched Study Designs & Analysis**

### **Video 8: Appropriate Matching, Validity, and Precision in Case-control Studies**

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH

Department of Epidemiology, Harvard TH Chan School of Public Health



**HARVARD T.H. CHAN**  
SCHOOL OF PUBLIC HEALTH

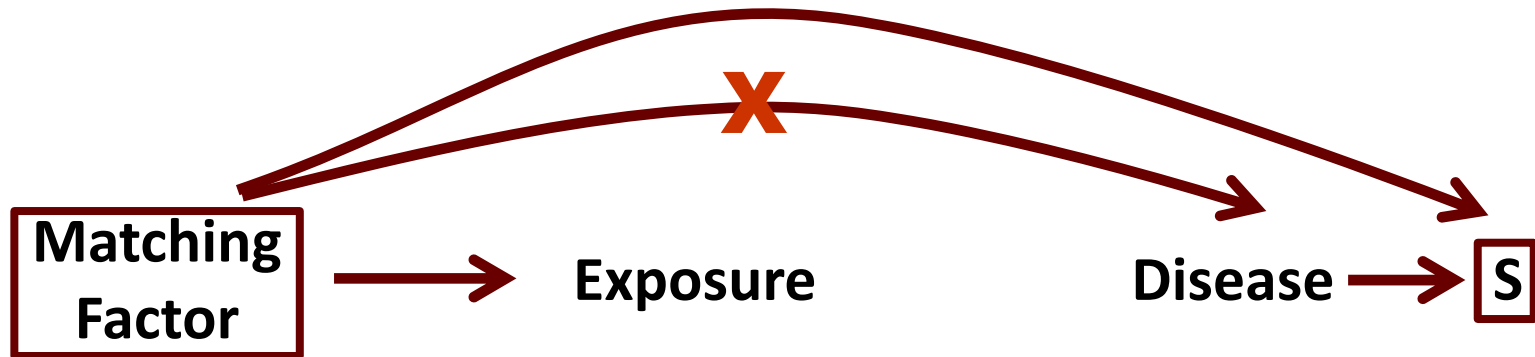
# Key Concepts

- Matched cohort studies
- Matched case-control studies
  - Impact of matching on validity and precision
  - Appropriate, unnecessary and over-matching
- Stratification in matched case-control studies
- Relative efficiency of matching ratios
- Notation
- Analysis of pair-matched case-control studies
  - Hypothesis tests
  - Point and interval estimation
- Relationship between McNemar estimator and  $OR_{MH}$



# Appropriate Matching

## Matching Factor is a Confounder

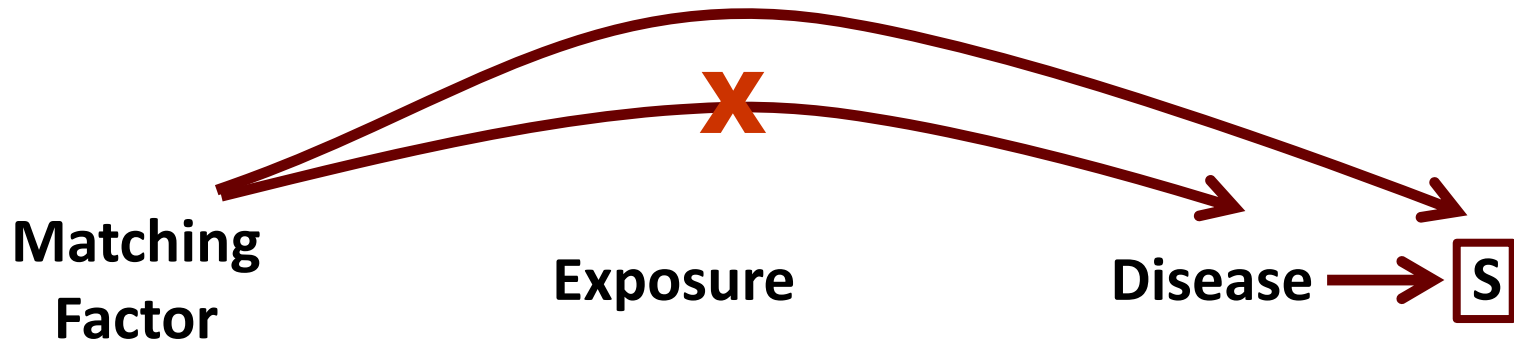


Design	Analysis	
	Stratified	Not Stratified
Match	V PPP	BIAS
Do not match	V P	BIAS

V=Valid; PPP=maximum precision; PP=Slightly reduced precision; P=reduced precision

# Unnecessary Matching

## Matching Factor Unrelated to Exposure

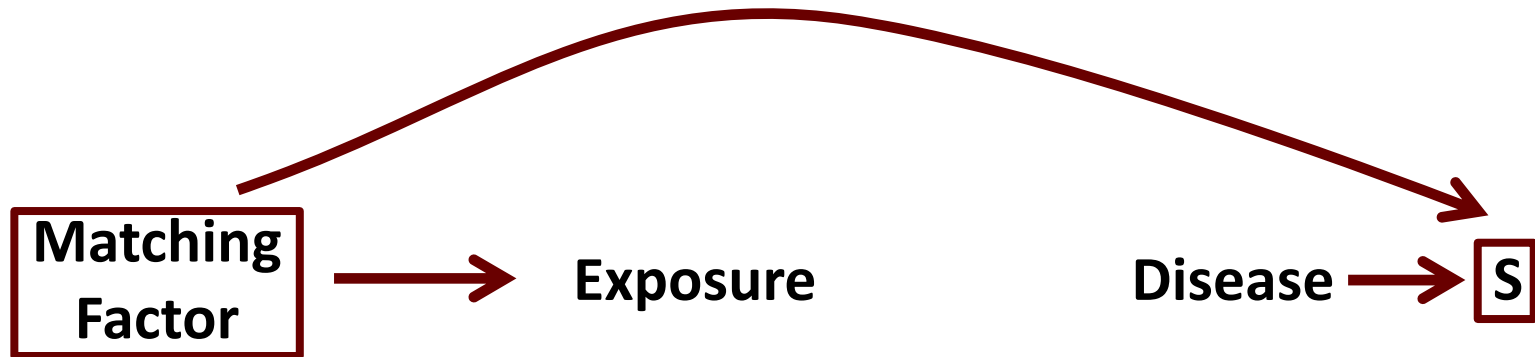


	Analysis	
Design	Stratified	Not Stratified
Match	<b>V PP</b>	<b>V PPP</b>
Do not match	<b>V PP</b>	<b>V PPP</b>

V=Valid; PPP=maximum precision; PP=Slightly reduced precision; P=reduced precision

# Overmatching

Matching Factor Related to Exposure **ONLY**

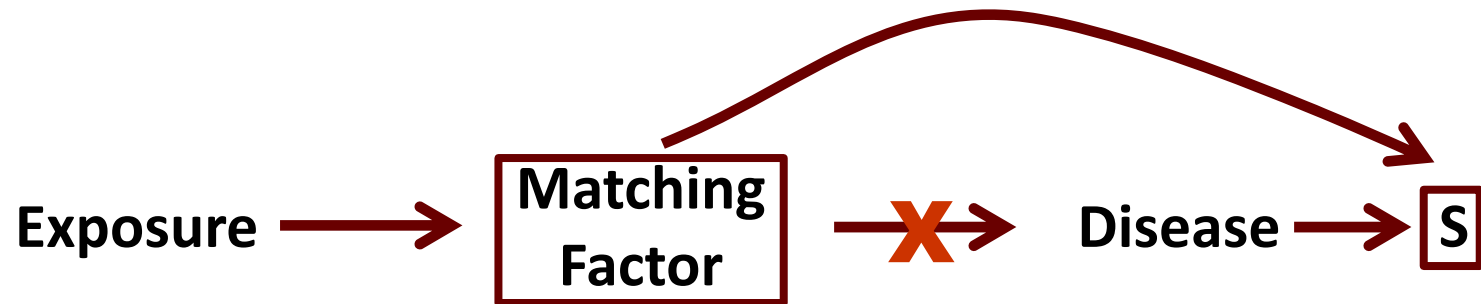


Design	Analysis	
	Stratified	Not Stratified
Match	V P	BIAS
Do not match	V P	V PPP

V=Valid; PPP=maximum precision; PP=Slightly reduced precision; P=reduced precision

# Matching on an Intermediate

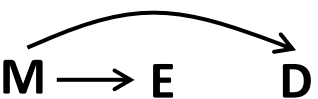



## Another Form of Overmatching



Design	Analysis	
	Stratified	Not Stratified
Match	<b>BIAS</b>	<b>BIAS</b>
Do not match	<b>BIAS</b>	<b>V PPP</b>

V=Valid; PPP=maximum precision; PP=Slightly reduced precision; P=reduced precision

# Effect of Matching on Validity and Precision in Case-control Studies

	Design	Analysis	
		Stratified	Not Stratified
 <b>Appropriate Matching</b>	Match Do not match	V PPP V P	BIAS BIAS
 <b>Unnecessary Matching</b>	Match Do not match	V PP V PP	V PPP V PPP
 <b>Over-matching</b>	Match Do not match	V P V P	BIAS V PPP
 <b>Match on Intermediate</b>	Match Do not match	BIAS BIAS	BIAS V PPP

V=Valid; PPP=maximum precision; PP=Slightly reduced precision; P=reduced precision

**BREAK**

# **Week 3: Matched Study Designs & Analysis**

## **Video 9: Matched Analysis and Matching Ratios**

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH

Department of Epidemiology, Harvard TH Chan School of Public Health



**HARVARD T.H. CHAN**  
**SCHOOL OF PUBLIC HEALTH**

# Key Concepts

- Matched cohort studies
- Matched case-control studies
  - Impact of matching on validity and precision
  - Appropriate, unnecessary and over-matching
- Stratification in matched case-control studies
- Relative efficiency of matching ratios
- Notation
- Analysis of pair-matched case-control studies
  - Hypothesis tests
  - Point and interval estimation
- Relationship between McNemar estimator and  $OR_{MH}$



# Analysis of Matched Case-Control Studies

## Stratification

- In case-control studies, matching is used to improve efficiency but it does not remove confounding unless the analysis accounts for the matching factors(s) with stratification/adjustment.
- In general, matching should be followed by stratification
- In the stratified analysis of a matched case-control study, a matched set is exactly equivalent to a stratum in an unmatched stratified analysis.
- In a matched case-control study, cases are matched to controls with respect to important potential confounders.
- Just as in an unmatched study, the data are stratified in the analysis with respect to all important confounders.

# Analysis of Matched Case-Control Studies

## Unique Combination of Confounders

- When each matched set represents a unique level of potential confounders, matched analysis strategies are required
- Examples:
  - Each matched pair represents a set of twins
  - Each case is age and gender matched to a neighborhood control
  - Each case is age and gender matched to a friend control
- When the "level" of potential confounders is unique to each case in the study or nearly so, one must match in order to obtain appropriate controls.
- Without matching it is unlikely that there would be control data in any strata for which there were cases.
- It is most convenient in this case to use methods for matched data in the analysis.

# Analysis of Matched Case-Control Studies

## Several Individuals with Same Combination of Confounders

- When each matched set represents a level of potential confounders that is observed repeatedly during the study, all study subjects with the same level of potential confounders can be analyzed within a single stratum.
- Matching is used in the design to improve efficiency, but without matching we would still be likely to find control data in the strata in which cases are observed.
- For example, imagine a study where cases are matched to several controls on gender, race and five-year age group. Except in those gender, race and age strata for which there is only one case, controls are not uniquely comparable to the case to whom they were matched during the data collection phase of the study. Rather, controls are comparable to all cases with the same age, gender and five-year age group.

# Matching Ratios

- Cases can be matched to controls in a 1:1 ratio, a 1:2 ratio or any arbitrary ratio, including variable matching ratios across strata
- Increasing the number of controls per case shrinks the variance of the adjusted OR, however, the incremental benefit of further increasing the ratio diminishes as the ratio goes up.
- The efficiency gain per number of controls in each matched set increases much more slowly as the underlying relative risk increases, especially when exposure prevalence in the controls is low
- For more, see Breslow and Day, The Design and Analysis of Cohort Studies, 1987, p. 303

# Matching Ratios

## Relative Efficiency of the Test of the Null Hypothesis

$H_A: RR =$	$Pr(A=1   Y=0)$	Number of Controls/Case						
		1	2	4	8	16	32	$\infty$
1	0-100%	50%	66%	80%	88%	95%	97%	100%
2	5%	30%	53%	71%	83%	92%	95%	100%
5	5%	20%	35%	54%	73%	84%	93%	100%

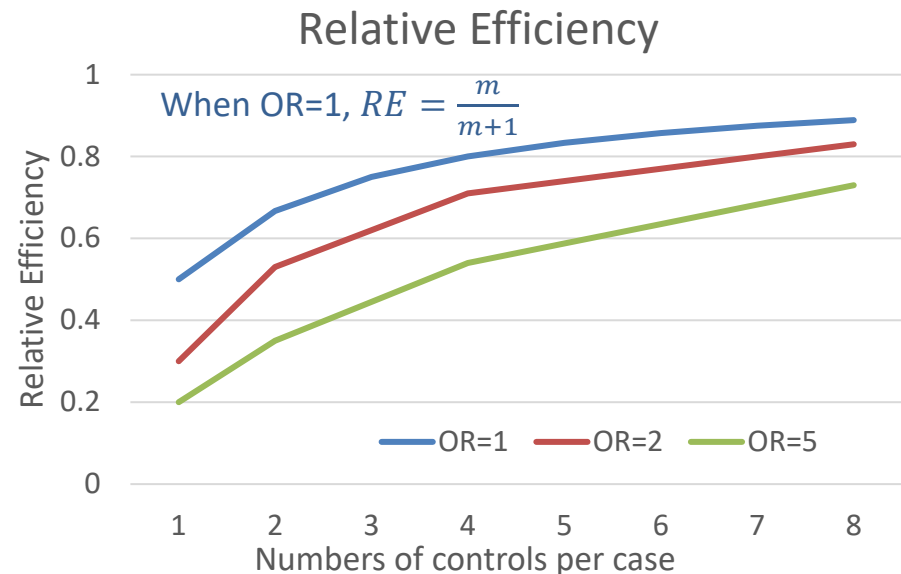
- $Pr(A=1 | Y=0)$  is the exposure prevalence in the controls
- Relative efficiency is the ratio of the variance of a particular design relative to having infinite controls (equivalent to conducting the full cohort analysis)
- Under the null, the efficiency of a particular matching ratio relative to the full cohort design is proportional to  $m/(m+1)$ , where  $m$  is the ratio of controls:cases.

# Matching Ratios

## Relative Efficiency of the Test of the Null Hypothesis

$H_A: RR =$	$Pr(A=1   Y=0)$	Number of Controls/Case						
		1	2	4	8	16	32	$\infty$
1	0-100%	50%	66%	80%	88%	95%	97%	100%
2	5%	30%	53%	71%	83%	92%	95%	100%
5	5%	20%	35%	54%	73%	84%	93%	100%

- Relative efficiency is the ratio of the variance of a particular design relative to having infinite controls (equivalent to conducting the full cohort analysis)
- Under the null, the efficiency of a particular matching ratio relative to the full cohort design is proportional to  $m/(m+1)$ , where  $m$  is the ratio of controls:cases.
- $Pr(A=1 | Y=0)$  is the exposure prevalence in the controls



**BREAK**

# Week 3: Matched Study Designs & Analysis

## Video 10: Notation and Analysis

EPI202 – Epidemiologic Methods II

Murray A. Mittleman, MD, DrPH

Department of Epidemiology, Harvard TH Chan School of Public Health



**HARVARD T.H. CHAN**  
SCHOOL OF PUBLIC HEALTH



# Key Concepts

- Matched cohort studies
- Matched case-control studies
  - Impact of matching on validity and precision
  - Appropriate, unnecessary and over-matching
- Stratification in matched case-control studies
- Relative efficiency of matching ratios
- Notation
- Analysis of pair-matched case-control studies
  - Hypothesis tests
  - Point and interval estimation
- Relationship between McNemar estimator and  $OR_{MH}$

# Abortion and Ectopic Pregnancy

## (Trichopoulos et al., 1969)

- Cases who had previously had at least one earlier pregnancy were matched to controls by order of pregnancy, maternal age and paternal age.
- The exposure of interest in this study was a history of induced abortions terminating any preceding pregnancy.
- The published study used a 1:4 matching ratio, but here we will look at the first control only.

# History of Induced Abortion and Ectopic Pregnancy Data

Pair	Case Exposure Status	Control Exposure Status
1	-	-
2	+	-
3	+	-
4	-	-
5	-	+
6	+	-
7	+	-
8	-	-
9	+	+
10	+	-
11	+	-
12	-	-
13	+	+
14	+	-
15	+	-
16	+	+
17	-	-
18	+	+

# Abortion and Ectopic Pregnancy

## Notation for Matched Case-Control Data

- 2x2 table for matched case-control data

		Controls	
		E	$\bar{E}$
Cases	E	$f_{11}$	$f_{10}$
	$\bar{E}$	$f_{01}$	$f_{00}$

- The data for this study can be displayed with this layout

		Controls	
		E	$\bar{E}$
Cases	E	4	8
	$\bar{E}$	1	5

# Abortion and Ectopic Pregnancy

## Discordant and Concordant Pairs

- The **discordant pairs** are the pairs for which a case is exposed and the control unexposed, or the case is unexposed and the control exposed.
- The **concordant pairs** are those pairs where both the case and the control are exposed, or both are unexposed.
- In these data,
  - $f_{10} = 8$ : discordant pairs with an exposed case and unexposed control
  - $f_{01} = 1$ : discordant pairs with an unexposed case and exposed control
  - $f_{11} = 4$ : concordant pairs where both the case and control are exposed
  - $f_{00} = 5$ : concordant pairs where neither the case nor the control are exposed

# Notation for Matched Case-Control Data

## MH, McNemar, and variance

- The Mantel-Haenszel  $\chi^2$  test statistic has the form:

$$Z^2 = \frac{[f_{10} - f_{01}]^2}{f_{10} + f_{01}} \sim \chi_1^2$$

- This test is also known as McNemar's test.

- The Mantel-Haenszel odds ratio is estimated as:

$$\hat{OR}_{MH} = \frac{f_{10}}{f_{01}}$$

- The variance for the  $\ln \hat{OR}_{MH}$  has the form:

$$\text{Var}[\ln(\hat{OR}_{MH})] = \frac{1}{f_{10}} + \frac{1}{f_{01}}$$

# Abortion and Ectopic Pregnancy

## Null and Alternative Hypotheses

- $H_0$ : There is no association between a history of induced abortion and subsequent risk of ectopic pregnancy after accounting for the matching factors.
- $H_A$ : There is an association between a history of induced abortion and subsequent risk of ectopic pregnancy after accounting for the matching factors.

# Abortion and Ectopic Pregnancy

## Test Statistic

$$Z^2 = \frac{[f_{10} - f_{01}]^2}{f_{10} + f_{01}} = \frac{[8 - 1]^2}{8 + 1} = 5.44$$
$$\Pr[\chi_1^2 > 5.44] = 0.02$$

- Assuming no residual confounding, no confounding by other risk factors, no information bias, no selection bias or any other source of bias, these data are not very consistent with the state of nature described by the null. If the null were true, we would expect to see associations this strong or stronger only 2% of the time. There is sufficient evidence in these data to reject the null hypothesis at the 2-sided  $\alpha = 0.05$  level and conclude that there is a statistically significant association between history of induced abortion and subsequent ectopic pregnancy.



# Relationship between McNemar and Mantel-Haenszel Estimators (1)

- Recall that in case-control studies:

$$\hat{OR}_{MH} = \frac{\sum \frac{a_i d_i}{T_i}}{\sum \frac{b_i c_i}{T_i}}$$

- In a matched-pair case control study with exactly 1 case and 1 control per stratum, there are only 4 possible layouts for the 2x2 tables that could result in any given stratum:

<u>Layout A</u>			<u>Layout B</u>			<u>Layout C</u>			<u>Layout D</u>		
E $\bar{E}$			E $\bar{E}$			E $\bar{E}$			E $\bar{E}$		
Case	1	0	Case	1	0	Case	0	1	Case	0	1
Control	1	0	Control	0	1	Control	1	0	Control	0	1

## Relationship between McNemar and Mantel-Haenszel Estimators (2)

- Assume that there are A matched pairs with layout A, B matched pairs with layout B, C matched pairs with layout C, and D matched pairs with layout D.

- The Mantel-Haenszel estimator in this case is given below:

$$\hat{OR}_{MH} = \frac{\sum \frac{a_i d_i}{T_i}}{\sum \frac{b_i c_i}{T_i}} = \frac{A * \left(\frac{1*0}{2}\right) + B * \left(\frac{1*1}{2}\right) + C * \left(\frac{0*0}{2}\right) + D * \left(\frac{0*1}{2}\right)}{A * \left(\frac{0*1}{2}\right) + B * \left(\frac{0*0}{2}\right) + C * \left(\frac{1*1}{2}\right) + D * \left(\frac{1*0}{2}\right)} = \frac{B * \left(\frac{1}{2}\right)}{C * \left(\frac{1}{2}\right)} = \frac{B}{C}$$

- Thus, in the setting of a matched pair case-control study, the Mantel-Haenszel estimator reduces to the ratio of the number of discordant pairs in which only the case is exposed (layout B) to the number of pairs in which only the control is exposed (layout C).
- The pairs in which the case and the control are either both exposed or both unexposed do not contribute any information to the estimate of the odds ratio.

# Abortion and Ectopic Pregnancy

$$\hat{OR}_{MH}$$

- Recall that  $f_{10} = 8$  and  $f_{01} = 1$

$$\hat{OR}_{MH} = f_{10}/f_{01} = 8$$

- Assuming no residual confounding, no confounding by other risk factors, no information bias, no selection bias or any other source of bias, there is an eight-fold higher rate of ectopic pregnancy among women that had a history of induced abortion in previous pregnancies compared to women with no history of induced abortions after accounting for the matching factors.

# Abortion and Ectopic Pregnancy

## 95% Confidence Interval for the $\text{Var}[\ln(\hat{\text{OR}}_{\text{MH}})]$

- $\text{Var} \ln(\hat{\text{OR}}_{\text{MH}}) = \frac{1}{f_{10}} + \frac{1}{f_{01}} = \frac{1}{8} + \frac{1}{1} = 1.125$
- $\ln(\hat{\text{OR}}_{\text{MH}}) \pm 1.96 \sqrt{\text{Var}[\ln(\hat{\text{OR}}_{\text{MH}})]}$   
 $= 2.079 \pm 1.96 \sqrt{1.125} = (0.00055, 4.158)$   
 $= e^{(0.00055, 4.158)} = (1.00, 63.96)$
- Although the observed eight-fold higher rate among the exposed compared to the unexposed is statistically significant at the 2-sided  $\alpha = 0.05$  level, the 95% confidence interval spans a wide range of values.
- This indicates that these data do not have a great deal of power to precisely estimate the odds ratio.

# Matched Case-Control Data Comparison with Crude Analysis

## Matched Data Layout

		Controls	
		E	$\bar{E}$
Cases	E	4	8
	$\bar{E}$	1	5

$$\hat{OR}_{MH} = 8/1 = 8$$

## Crude Data Layout

		E	$\bar{E}$	
Cases	12	6	18	
Controls	5	13	18	

$$\hat{OR}_{Crude} = [12*13]/[5*6] = 5.2$$

- In general, odds ratio from a crude analysis of matched case-control data is expected to be biased towards the null.

## Crude Analysis of Matched Case-Control Data Results in a Bias Towards the Null

- Why?
- Matching makes the cases and controls more similar to each other than they would be if simple random sampling were used.
- With simple random sampling, the prevalence of the confounder among the controls represents the confounder prevalence in the pool of person time from which the cases arose.
- With matching, the prevalence of the confounder among the controls is closer to that of the cases than the full population.
- Because the confounder is associated with exposure, the distribution of the exposure of interest more closely resembles the distribution in the cases than the distribution in the study base.

# 1:R Matched Data

- Formulas exist for hypothesis testing and point and interval estimation for matched case control analysis for any 1:R case-control matching ratio and for studies with a varying number of controls matched to each case.
- Formulas exist for hypothesis testing and point and interval estimation in 1:1 and 1:R matched cohort studies. These formulas are different from the formulas used in pair matched case-control studies.

**BREAK**