Analysis of data from case-control studies

Isaac M. Malonza, MD, MPH

Department of Reproductive Health and Research
World Health Organization

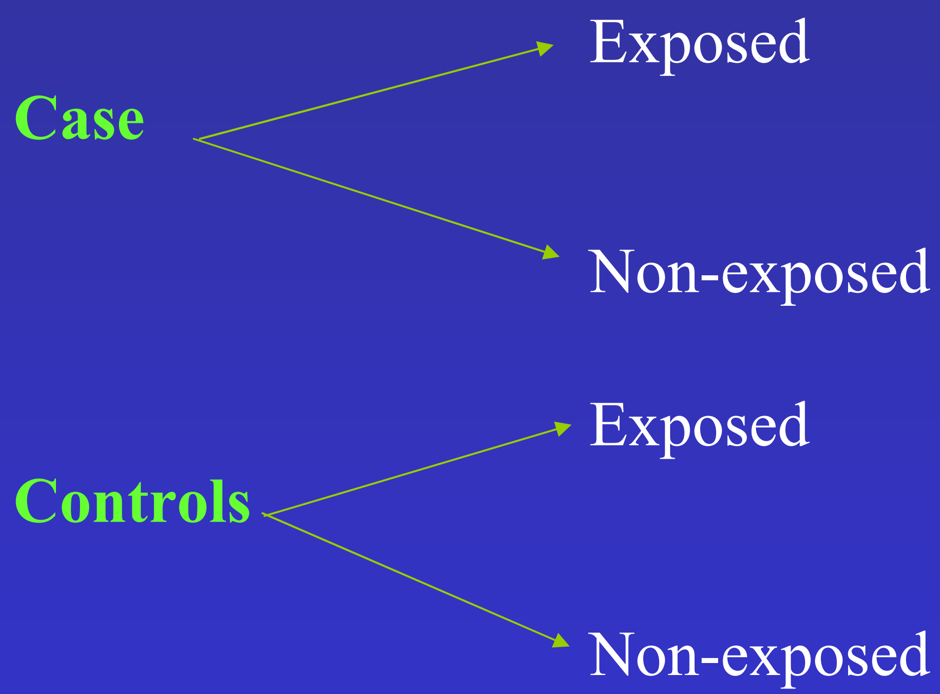
Objectives of this lecture

- Quick review of the design of case –control studies
- Calculating Odds ratios
- 95% confidence interval for Odds ratios
- Relationship between odds ratio and relative risk
- Interpretation of the odds ratio
- Analysis of data from matched case-control studies

Design of case-control studies

- Identify a group of individuals with the disease (cases)
- Select a group of individuals without the disease (controls)
- Determine the proportion of cases who were exposed and those that were not exposed
- Then do the same for control (exposed versus non-exposed)

Diagrammatic representation of a case-control study



Summarising data from case-control studies using a 2 by 2 table

	Cases	Controls	Total
Exposed	A	B	**(A+B) M_1=**
Non-exposed	C	D	**(C+D) M_2=**
Total	**A+C=N_1**	**B+D=N_2**	**M_{1+} M_2=T**

Proportion of cases exposed = A/(A+C)

Proportion of controls exposed = B/(B+D)

If disease is associated with exposure, we expect the proportion of cases who are exposed to be higher than the proportion of controls who are exposed, i.e

A/(A+C) greater than B/(B+D)

Hypothetical example: coronary heart disease (CHD) versus history of smoking

	CHD	Controls
Smoking	56	88
No smoking	44	112
Total	**100**	**200**
Proportions (exposed)	**56%**	**44%**

This implies that history of smoking may be associated with development of CHD.

Odds ratio (1)

	Cases	Controls
Exposed	A	B
Non-exposed	C	D
	A+C	**B+D**

- **A** divided by **(A+C)** is the **probability** that a **case** was **exposed**
- **C** divided by **(A+C)** is the **probability** that a **case** was **not exposed**
- **A/(A+C)** divided by **C/(A+C)** is a **ratio of two probabilities** which is called **odds**
- **Odds** of a **case** being **exposed** = **A/(A+C)** divided by **C/(A+C) = A/C**

Odds ratio (2)

- the **odds** of an event is defined as the ratio of the number of ways the event can occur to the number of ways the event cannot occur, i.e.

$$\textbf{Odds} = \frac{\text{No. of ways event can occur}}{\text{No. of ways event cannot occur}}$$

- **A/C** is the **odds** that a **case** was **exposed**
- **B/D** is the **odds** that a **control** was **exposed**

Odds ratio (OR) = **A/C** divided by **B/D** =**AD/BC**

Definition: **OR** in **case-control** studies is defined as the ratio of the **odds that the cases were exposed to the odds that the controls were exposed.**

Odds ratio from cohort studies

- **A** divided by **B** is the **odds** that the **exposed** will develop **disease**
- **C** divided by **D** is the **odds** that the non- **exposed** will develop **disease**
- **OR=A/B** divided by **C/D=AD/BC**
- Therefore, **AD/BC** represents the odds ratio in both case-control and cohort studies,
- **OR** in a **cohort studies** is defined as **the ratio of the odds that the exposed persons will develop disease to the odds that the non-exposed will develop the disease.**

Recapitulate

- Note that **AD/BC** has a different meaning depending on whether its from a case-control or cohort study

- **OR** in **case-control** studies is defined as the ratio of the **odds that the cases were exposed to the odds that the controls were exposed**

OR in a **cohort studies** is defined as **the ratio of the odds that the exposed persons will develop disease to the odds that the non-exposed will develop the disease**

Interpreting the odds ratio

- If OR=1, the exposure is not related to the disease (no association)
- If OR>1, the exposure is positively related to the disease (possible causal)

If OR<1, the exposure is negatively related to the disease (possible protective)

Calculating OR from case-control studies

	CHD	Controls
Smoking	56	88
No smoking	44	112

OR= (56 X 112) / (88 X 44) = 6272 / 3872 = 1.6

Indicating that smoking increases the odds of developing CHD

Suppose we rearrange the order of columns

	CHD	Controls
No Smoking	**44**	112
Smoking	56	88

OR= (44 X 88) / (112 X 56) = 3872 / 6272 = 0.6

Indicating that non-smoking reduces the odds of developing CHD

	CHD	Controls
Smoking	112	44
No smoking	88	56

OR=1.6, indicating the odds of not developing CHD are increased for non-smokers

Odds ratio from matched pairs case - control study

- Controls may be matched to each case according to a certain factor, e.g. age, sex, race
- Analysis is done for case-controls pairs, not by individual subjects
- What types of combinations are possible?
- Assume that exposure is **dichotomous** (either exposed or not exposed)
- Possibilities:

1. Both cases and controls exposed
2. Neither case nor control was exposed
3. Case exposed, but control not exposed
4. Control exposed, but case not exposed

- 1 and 2 are called **concordant** pairs
- 3 and 4 are **discordant** pairs

- we can summarise the data into a 2 X 2 table:

		Controls	
		Exposed	Not exposed
Cases	Exposed	a	b
	Not exposed	c	d

Note: a, b, c, d, represent pairs

- concordant pairs (**a** and **d**) had the same exposure experience, therefore they cannot tell anything about the relationship between **exposure** and **outcome**
- calculation of OR is based on the discordant pairs, **b** and **c**
- **OR=b/c**
- Definition: **OR** in a **matched case-control study** is defined as the **ratio of the number of pairs a case was exposed and the control was not to the number of ways the control was exposed and the case was not**

Hypothetical example: matched case/control

Cases	Controls
E	N
E	E
N	N
E	N
N	E
N	N

		Controls	
		Exposed	Not exposed
Cases	Exposed	1	2
	Not exposed	1	2

OR=2/1=2.0

Matched case/control study with R controls per case controls

cases	0	1	2	…	R
exposed	F_{10}	F_{11}	F_{12}	…	F_{1R}
Not exposed	F_{00}	F_{01}	F_{02}	…	F_{0R}

F_{10}=no. of times the case is exposed and none of the controls are exposed

F_{11}=no. of times the case is exposed and one of the controls are exposed

M =total no. of exposed subjects in a matched set (0 = m = R+1)

OR_{MH} =

$\{R\ F_{1,0} + (R\text{-}1)F_{1,1} + (R\text{-}2)\ F_{1,2} + \ldots + F_{1,R-1}\} / \{\ F_{0,1} + 2F_{0,2} + 3F_{0,,3} + \ldots + RF_{0,R}$

Example:

Previous history of induced abortion among women with ectopic pregnancy and matched controls

	controls				
cases	0	1	2	3	4
Exposed	3	5	3	0	1
Not exposed	5	1	0	0	0

OR_{MH} ={4x3 + 3x5 +2x3 +1x0}/{1+2x0+3x0+4x0}=33/1=33

Calculating OR from data with continuos exposure

Daily cigarette consumption

	<5	5-14	15-24	25-49	50+
Lung cancer	26	208	196	174	45
Controls	65	242	201	118	23

smoking	Lung cancer	controls
5-14	208	242
<5	26	65

OR=2.1

- We can therefore calculate **OR** for other smoking categories compared to **<5** group
- We get a list of OR as shown in the next slide

Daily cigarette consumption

	<5	5-14	15-24	25-49	50+
Lung cancer	26	208	196	174	45
Controls	65	242	201	118	23
OR	1	2.1	2.4	3.7	4.9

Smoking more that 5 cigarettes per day increases the odds of developing lung cancer

Suppose we had a continuous outcome, e.g. causes of death, then you have to calculate OR for each cause of death.

Calculating the 95% confidence interval for ORs

- Epidemiologic studies usually involve only a sample of the entire population
- However, the main interest is to use the sample to make conclusions about the entire population
- Question: how does the OR from the sample differ from that for the entire population?
- We would like to be 95% confident that the population OR lies within a certain range
- This range is referred to as the **confidence interval** (CI)

CI for the OR (Mantel and Haenszel, 19959, Miettinen, 1976): **$CI = OR^{(1 \pm Z/x)}$**

Where **Z** is the normal variate and **x =square root of** $\frac{(T-1) \times (AD-BC)^2}{N_0 \times N_1 \times M_1 \times M_0}$

Estimating the CI from "The Cancer and Steroid hormone study, 1987"

	Ovarian cancer	Controls	Total
OC use	250	2,696	**2,946**
NO OC	242	1,532	**1,774**
Total	**492**	**4,228**	**4,720**

Step 1: calculate the $X^2 = \frac{4{,}719 \times (250 \times 1{,}532 - 242 \times 2{,}696)}{2{,}696 \times 1{,}532 \times 250 \times 242} = 31.51$, X=5.61

Step 2: Lower limit: $OR^{(1-Z/x)}$, where Z is 1.96, =0.5

Step 3: Upper limit, $OR^{(1+Z/x)}$, =0.7

Controlling for confounding

Example of **Education, cervical cancer** and **OC use:**

OC non users

Education	cancer	controls
High	3	33
Low	47	16
Total	50	49
%high	**6%**	**67%**

All women

Education	cancer	controls
High	8	75
Low	92	25
Total	100	100
%high	**8%**	**75%**

Conclusion: women with cervical cancer were more likely than controls to have ‘low’ level of education

Confounding (2)

High	OC	cases	controls	OR
	+	5	42	
	-	3	33	1.31

Low	OC	cases	controls	OR
	+	45	9	
	-	47	16	1.70

Total	OC	cases	controls	OR
	+	50	51	
	-	50	49	0.96

$$\text{Standardized OR} = \frac{(5 \times 33)/83 + (45 \times 16)/117}{(42 \times 3)/83 + (9 \times 47)/117} = 1.59$$

Relationship between OR and RR

- Relative risk = incidence in exposed/incidence in non-exposed
- cannot measure RR directly from a case-control study
- OR is a good estimate of RR when:

1)the disease or event is rare

2) cases are representative of the all people with the disease with regard to exposure

3) controls are representative of all people without disease in the population

- Example:

	cases	controls
exposed	200	9800
non exposed	100	9900

RR=(200/10,000)/(100/10,000) = 2.0

OR=2.02