Strategies for Data Analysis: Cohort and Case-control Studies

Post-Graduate Course, Training in Research in Sexual Health, 24 Feb 05

Isaac M. Malonza, MD, MPH
Department of Reproductive Health and Research
World Health Organization

Objectives of the lecture

- Analyses tables of basic characteristics
- Review the design of Cohort studies
- Review the design of Case –control studies
- Calculating Absolute Risk, Relative Risks, Risk Difference, and Odds Ratios (ORs)
- 95% confidence interval for Relative Risk and Odds Ratio
- Relationship between Odds Ratio and Relative Risk
- Interpretation of Relative Risk and Odds Ratio
- Data analysis from Matched Case-control Studies
- Confounding and Effect modification

Analysis Table of Basic Characteristics

Characteristics of participating women

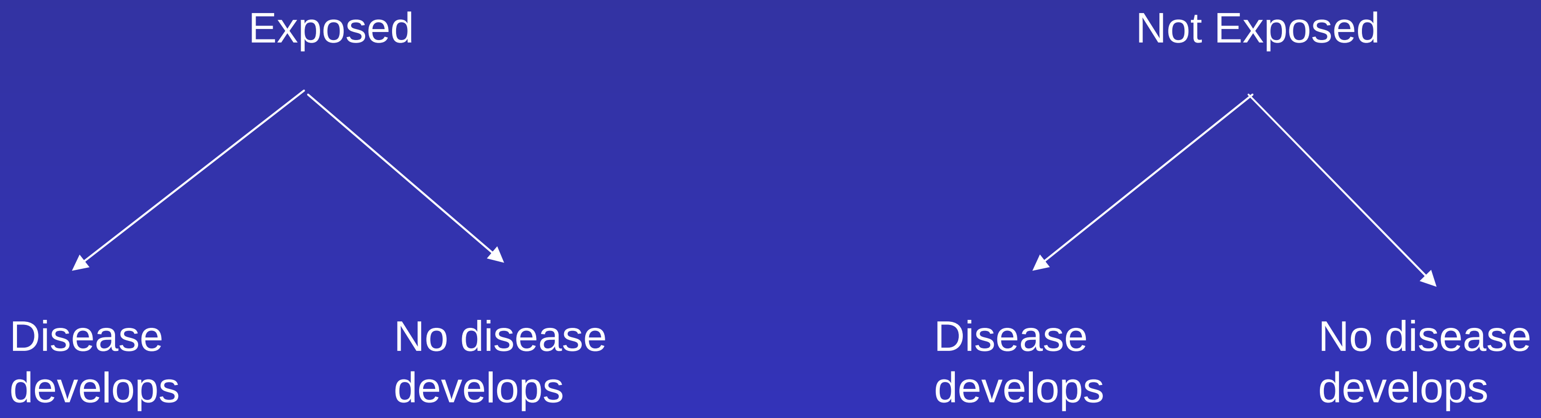
Characteristic	Rapid HIV test	ELISA test
Age in years(range)	23 (18-43)	23(18-44)
Marital status:		
Single	67(11%)	62(10%)
Married	548(88%)	554(89%)
Other	10(2%)	4(1%)
Occupation		
Housewife	368(59%)	382(62%)
Unemployed	64(10%)	46(8%)
Parity(range)	1(0-9)	1(0-8)

(Malonza et al, 2003)

Design of Cohort Studies

- Investigator selects a group of individuals :
  - – exposed to the factor of interest (Exposed)
  - – not exposed to the factor of interest (Not exposed)
- Follows both groups to determine the incidence of disease (case) in the in two groups
- if exposure is associated with disease, we would expect that the incidence of disease among the exposed is greater than the incidence of disease among the non-exposed group
- since we identify new cases of disease as they occur in both groups, we can determine a temporal relationship between exposure and the development of disease
- Definition: A cohort is a group of individuals who share a common experience or condition

Diagrammatic Representation of a Cohort Study



Exposed

Not Exposed

Disease develops

No disease develops

Disease develops

No disease develops

	Disease (yes)	Disease (no)	Totals	Incidence
Exposed	a	b	a+b	a/(a+b)
Not exposed	c	d	c+d	c/(c+d)

- a/(a+b) equals the incidence of disease among the exposed
- c/(c+d) equals the incidence of disease in the non-exposed

Objectives of Cohort Studies

- To estimate incidence, rate of occurrence and risk of disease
- To measure and compare the incidence of disease in one or more study cohorts
- To determine the aetiology of disease

Risk and Absolute Risk

Risk

- Definition: The proportion of individuals who develop a disease over a specified period of time

$$\text{Risk} = \frac{\text{Number of people who develop disease}}{\text{Total population followed up}}$$

e.g. 1000 people were observed for 3 years

950 did not develop disease

50 developed disease, Risk =50/1000=0.05

Absolute risk

- Definition: The incidence of a disease in a population
- Does not consider the incidence of disease in the unexposed, therefore cannot decide whether exposure is associated with disease

Association between Exposure and Disease

- How do we determine that a certain exposure is associated with a disease of interest?
- Use data from a cohort or case-control study
- determine whether there is excess risk of the disease in persons who have been exposed
- Let us use a hypothetical investigation of a disease outbreak
- the suspect foods were identified and for each food, the incidence of disease was calculated for those who ate (exposed) and those who did not eat (not exposed) the type of food

Food borne Disease Outbreak: Calculating Excess Risk

	A (%sick)	B (%sick)	risk	risk
Food	Ate	Not eaten	A/B	A-B
Fish	60	30	2.00	30
Rice	78	67	1.16	11
Meat	72	50	1.44	22

methods of calculating excess risk:

1)calculate the *ratio* of attack rate in those who ate to those who did not eat (A/B)-*risk ratio*

2)subtract the risk in those who did not eat from those who ate (A-B)-*risk difference*

Excess Risk

- To determine whether a certain exposure (specific food) is associated with a certain disease (diarrhea), we need to determine whether there is excess risk
- Excess risk=comparison of risk of disease in exposed population to risk of disease in non-exposed population:
  - Ratio of the risks: $\frac{\text{Disease risk in exposed}}{\text{Disease risk in non-exposed}}$
  - Difference in the risks (or of the incidence rates):

    (disease risk in exposed-disease risk in non-exposed)

Risk Ratio and Risk Difference

- Question: does the method we use to calculate Excess Risk make any difference?
- Consider a hypothetical example of two communities X and Y:

	X	Y
Incidence (%)		
in exposed	40	90
in non-exposed	10	60
Difference in risks	30	30
Risk ratio	4.0	1.5

Relative Risk

- Cohort and case-control studies are designed to determine whether there is an association between exposure and disease
- if an association exists, we would like to know how strong it is
- For cohort studies, the question to ask is:

  What is the ratio of risks of disease in exposed persons to the risk of disease in non-exposed individuals?
- This ratio is called the Relative Risk:

$$\text{Relative risk} = \frac{\text{Risk in exposed}}{\text{Risk in non-exposed}}$$

Calculating the Relative Risk

$$\text{Relative risk}=\frac{\text{incidence in exposed}}{\text{incidence in non-exposed}}=\frac{a/a+b}{c/c+d}$$

- Example: Smoking versus CHD

	Developed CHD	Did not develop CHD	Totals	incidence per 1,000 per year
Smokers	82	2,918	3,000	27.3
Non smokers	86	4,914	5,000	17.2

Incidence among exposed= 82/3000=27.3 per 1000

Incidence among non-exposed= 86/5000=17.2 per 1000

Relative risk= 27.3/17.2= 1.58

Incidence Density Relative Risk

- In most studies, not all enrolled persons are followed up for the entire duration of the study
- the time each person (person-time) contributes to the study is therefore taken into consideration
- person-time is therefore used as the denominator instead of number of persons enrolled
- this type of relative risk is called the incidence density relative risk (IDR)

Analysis Table for IDR

	Exposed	Unexposed	Total
Cases	a	b	m_1
Person-time	n_1	n_0	t

$$\text{IDR} = \frac{a/n_1}{b/\ n_0}$$

a=number of cases among the exposed,
b=the number among the unexposed,
n_1=person-time among the exposed, and
n_0=person-time among the unexposed,

Confidence Interval for Relative Risk

- Confidence interval = RR $^{(1 \pm z/x)}$

  where z is the normal variate (1.96),

  and $$x^2 = \frac{(t-1)*[(a*d)-(b*c)]^2}{n_1*n_2*m_1*m_0}$$

- Confidence interval that include 1 implies no association between exposure and disease

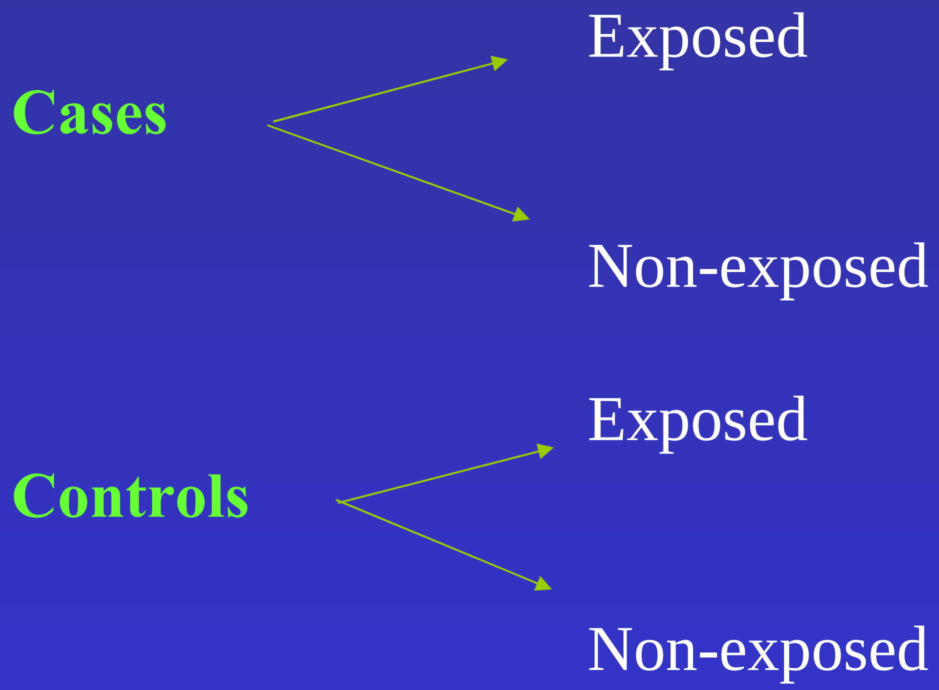
Interpreting the Relative Risk

- If RR=1 risk in exposed equals risk in non-exposed (no effect/association)
- If RR>1 risk in exposed greater than risk in non-exposed (positive association, possibly causal)
- If RR<1 risk in exposed less than risk in non-exposed (negative association, possibly protective)
- If confidence interval includes 1, then the RR is not significant (no association)

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



Cases

Exposed

Non-exposed

Controls

Exposed

Non-exposed

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

Summarize 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**

- OR $= \dfrac{\#\text{ Pairs (case exposed and control not exposed)}}{\#\text{ Pairs (control exposed and case not exposed)}}$

Hypothetical Example: Matched Case-control Study

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 Number of Controls per Case

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 ($0 \leq m \leq R+1$)

OR_{MH} =
$\{R\ F_{1,0} + (R-1)F_{1,1} + (R-2)\ F_{1,2} +….+ F_{1,R-1}\}/ \{ F_{0,1} + 2F_{0,2} + 3F_{0,,3} +….+ RF_{0,R}\}$

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 each 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 multiple outcomes, e.g. different types of cancer, then you have to calculate OR for each type of cancer.

Calculating the 95% Confidence Interval (CI) for Odds Ratios

- 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, 1959, Miettinen, 1976): $\mathbf{CI=OR^{(1\pm Z/x)}}$

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

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

Confounding

- Definition:
  - factor is associated with both exposure and outcome
  - but factor not a result of exposure
  - hides the true relationship between exposure and outcome
- Controlling for confounding
  - Study design stage: *Matching* (individual or group)
  - Data analysis stage: *Stratification* and *Adjustment*

Controlling for Confounding (1)

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

Controlling for Confounding (2)

High education

OC	cases	controls	OR
+	5	42	
-	3	33	1.31

Low education

OC	cases	controls	OR
+	45	9	
-	47	16	1.70

All volunteers

OC	cases	controls	OR
+	50	51	
-	50	49	0.96

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

Effect Modification

- Definition:
  - present when the relationship between exposure and outcome is different for various subgroups in the study population
  - detected by stratifying the analysis by each stratum and comparing the RRs for the strata

Maternal age (<25yrs)	Anemia	No Anemia	
Low birthweight	13	4	
Normal birthweight	18	17	IR=2.2

Maternal age (>25yrs)	Anemia	No Anemia	
Low birthweight	5	4	
Normal birthweight	37	31	IR=1.0

(To rule out of confounding: Overall IR=1.7, Adjusted =1.6)

References

- Malonza IM, Richardson BA, Kreiss JK, Bwayo JJ, Stewart GC. The effect of rapid HIV-1 testing on uptake of perinatal HIV-1 interventions: a randomized clinical trial. *AIDS* 2003; 17:113-118.

- Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst* 1959; **22**:719-748.

- Gordis L. *Epidemiology*. Philadelphia, Pennsylvania: W.B. Saunders Company, 1996.

- CDC, FHI and WHO. *An epidemiologic approach to reproductive health*. Geneva, Switzerland: World Health Organization, 1994.