Epidemiology (9)
Chapter 7. Dealing with Biases

Minato NAKAZAWA, Ph.D.

<minato-nakazawa@people.kobe-u.ac.jp>

Two types of errors

- Epidemiologists try to reduce both types of the following 2 errors
- Random error
  - Discussed in (Chapter 8)
  - Statistical
  - If the sample size is infinitely large, reduced to zero
- Systematic error
  - Focused in this chapter
  - a.k.a. Bias
  - Bias can also refer the attitude on the part of the investigator, but here the author refers any systematic error in a study
    - How the subjects have been selected → **Selection bias**
    - How the study variables are measured → **Information bias**
    - Some confounding factor that is not completely controlled → **Confounding**
  - Not affected by sample size (Fig 7-1)
- To estimate the average height of women in the city of Centerville (population of 500,000), using an official measuring tape, 100 randomly sampled women are measured
  - The source of errors includes how the measuring tape is held, which gives sometimes higher than, sometimes lower than true values, thus random error
  - When the sample size (n) increases from 100 to 1000 or to 10000, the effect of random error becomes less (Standard error of mean is SD/$\sqrt{n}$)
  - However, other errors are not affected by sample size. If official tape is made of cloth and shrank after wash, the measured values are systematically higher than true values → This is bias

Selection Bias

- It "stems from the procedures used to select subjects and from factors that influence study participation" → Association between exposure and disease differs by study participation → Since the association is unknown for non-participants, the existence of selection bias is inferred
- Screening test to detect colon cancer → If participation is voluntary, person who (1) is more health conscious or (2) especially worries one's cancer is more likely getting tested → In case of (1), screened subjects may show lower incidence of cancer, but in case of (2), screened subjects may show higher incidence → Both selection bias (so-called **self-selection bias**), difficult to quantify → To avoid such issue, randomized trial is necessary
- Biased choice of participants by investigator: (eg.) Comparing death rates between workers in a specific job and general population → Since general population includes many people who cannot work due to ill health, workers' death rates apparently show much lower than those of general population (so-called **healthy worker effect**)

Table 7-1. Healthy worker effect

	Exposed workers	General population		
		Workers	Nonworkers	Total
Deaths	50	4500	2500	7000
Person-time	1000	90000	10000	100000
Mortality rate (cases/yr)	0.05	0.05	0.25	0.07

- Table 7-1: Exposed workers' mortality rate 0.05 is 5/7 (71%) of 0.07 in general population. This is caused by selection bias, because
  - General population includes healthy majority and unworkable minority, among whom the former has the same mortality rate with exposed workers but the latter has much higher mortality rate

Actual example of selection bias

- Actual example of the effect of selection bias is seen in the efficacy of influenza vaccination in the elderly (https://www.nejm.org/doi/full/10.1056/NEJMoa070844).

- Among 713872 person-seasons, vaccinated had a 48% decrease in overall mortality during the season.

- However, 5-10% of deaths among the elderly during the season are attributable to influenza. Higher mortality in unvaccinated is probably because they had no priority in vaccination. It's selection bias.

- Jackson et al. (2006) showed the effect of selection bias in this issue (https://academic.oup.com/ije/article/35/2/337/694702). As shown in Figure 7-2, during the season, almost 50% of all causes deaths were prevented, but the preventive efficacy was greater before the season due to selection bias (Only those who are apparently healthy can get vaccinated). After the season, when bias is smaller, vaccine efficacy is theoretically zero.

- A similar trend was evident for the outcome of hospitalization for pneumonia or influenza, indicating strong selection bias.

RR adjusted for age and sex	Time period as to influenza season		
	Before	During	After
All cause	0.39 [0.33,0.47]	0.56 [0.52, 0.61]	0.74 [0.67, 0.80]
Pneumonia or influenza	0.72 [0.59, 0.89]	0.82 [0.75, 0.89]	0.95 [0.85, 1.07]

Matching in case-control studies (Another form of selection bias)

- **Matching** is often used to select comparative exposed and unexposed group in cohort study, but <u>matching to select controls in case-control study paradoxically results in selection bias</u>.
- <u>Controls must be sampled independently of the exposure</u>, but matching in case-control study typically violates this assumption.
- Commonly used matching factors are age, sex, geographic location, but some other specific factors may be included.
- The aim of matching is to prevent confounding, and thus matching factors are usually potential confounding factors. Confounding factors are associated with both exposure and disease, so that matching to select controls in case-control studies means that the sampling controls is not independent from exposure.
  → "If the exposure were perfectly correlated with one of the matching factors, <u>controls would then have exactly the same exposure distribution as the cases</u>, which would appear to indicate no effect of exposure, regardless of the actual effect that the exposure has."

Table 7-2. Hypothetical data showing risk for disease during 1 year by exposure status and sex

	Sex	Population	Risk	No. Cases
Exposed	Male	90000	5.00%	4500
	Female	10000	1.00%	100
Unexposed	Male	10000	0.50%	50
	Female	90000	0.10%	90

* Being male is associated with exposure and is a risk factor for disease

- Assumptions
  - Exposure to an agent (10% of a community) multiplies the risk tenfold
  - Male has 5 times greater risk than female
  - Consider 100000 males and 100000 females in the population
- The imbalance of males between exposed and unexposed will confound the effect of exposure.
  - Though the effect of exposure increases the risk of disease tenfold, risk among all exposed (4600/100000) and among all unexposed (140/100000) results in RR 32.9 (much larger than 10).

Matching in case-control studies (cont'd)

Table 7-3. Hypothetical cohort study

	Sex	Population	Risk	No. Cases
Exposed	Male	9000	5.00%	450
	Female	1000	1.00%	10
Unexposed	Male	9000	0.50%	45
	Female	1000	0.10%	1

* Based on 10% sample of exposed from the population and 10000 unexposed matched by sex

- Take 10% sample of exposed and sex-matched unexposed
- No imbalance of males between exposed and unexposed
- RR is (460/10000)/(46/10000)=10
- Matching prevented the confounding by male sex

Table 7-4. Hypothetical case-control study

	Exposed	Unexposed	Total
Cases	4600	140	4740
Controls	4114	626	4740

- In case-control study, total cases 4740 is included. Sex-matched controls include 4500+50=4550 males and 100+90=190 females.
- 90% of males were exposed, 10 % of females were exposed, and thus 4550x0.9+190x0.1=4095+19=4114 were exposed in total. 4550x0.1+190x0.9=455+171= 626 were unexposed (Table 7-4).
- As the estimate of RR,
  OR = (4600/140)/(4114/626)=4.99.. ≈ 5.0
  (Underestimated!!)

Table 7-5. Case-control data from Table 7-4, stratified by sex

	Males		Females	
	Exposed	Unexposed	Exposed	Unexposed
Cases	4500	50	100	90
Controls	4095	455	19	171

- Stratified by sex (Table 7-5), case-control data gives the same estimates of RR,
  OR=(4500/50)/(4095/455)=10 for males and
  OR=(100/90)/(19/171)=10 for females.
- In case-control studies, the selection bias by matching can be removed by appropriate analytic methods (in this case, stratified analysis, but regression models is also applicable).

Matching in case-control studies (cont'd)

- Without matching (randomly sampled), half of controls (expected numbers are 4740/2=2370) would be males and females, but 190 cases are females and 4550 cases are males
- Matching in case-control study cannot improve validity (no improvement in confounding), but can improve efficacy of a stratified analysis
- However, improvement of efficiency in matched case-control study is unclear
- If the matched variable is related to exposure, matching on it will introduce selection bias.
- But if it's not related to disease, it can be ignored (not a confounding). In this case, stratified analysis by matched variable is needed, but the efficiency is not improved

- Another issue: Small numbers within strata, if there are more confounding variables to be matched → the case and all matched controls within a set will have the same value for exposure (all exposed or all unexposed) → Such stratum (**concordant set**) cannot contribute to analysis → loss of efficiency
- Matching in case-control studies can be expensive and cannot improve validity. Efficiency might be improved, but not guaranteed (due to concordant set, sometimes efficiency is lost).
- Therefore, matching should be avoided in case-control studies except for some special cases.

Special settings for appropriate matching in case-control studies

- **Convenience matching**
  - Some types of matching may simply be convenient way to identify controls
  - Risk-set sampling (see, Chapter 5) is done for convenience: Matching on time as a means of selecting controls proportional to their person-time contribution to source population of cases
    - If matching factor may not be related to exposure and thus matching may not introduce selection bias, it can be ignored
    - If matching factor (time variable in risk-set) is related to exposure, it has to be controlled in the analysis
    - (eg.) Mobile phone use and brain cancer that matched risk set on time of occurrence of brain cancer. If mobile phone use changes over time, matched sets by time have to be retained.

- If controlling the variables in the analysis is impossible, matching in case-control studies may be allowed.
  - (eg.) An investigator wishes to control for early-childhood environmental/genetic effects by controlling for family (specifically by using **sibling controls**). Selecting sibling controls by matching on sibship during subject ascertainment is possible.

- Except for such special settings, the drawbacks of matching in case-control studies outweighs any advantage.

Information Bias

- Information bias = Systematic error because of the erroneous information, often referred as **misclassification** (when the variable is measured on a categorical scale).
  - **Differential** misclassification
    - Misclassification differs according to the value of other study variables (especially exposure and disease).
  - **Nondifferential** misclassification
    - Unrelated to other study variables

- Common type of information bias
  - **Recall bias**: Occurring in case-control studies, when subjects are interviewed after the disease occurrence
    - Patients can recall more exposure information than healthy controls → differential
    - (eg.) **Maternal recall bias**: mothers who born babies with birth defects can more accurately recall exposure during early pregnancy because adverse pregnancy outcome serves as a stimulus to consider potential causes, but mothers of normal babies have no comparable stimulus → differential
    - General problems for remembering and reporting exposures are nondifferential
  - Prevention of recall bias: To frame the questions to aid accurate recall, to select controls with accurate recall, to get information from medical record (recorded before the outcome is known) rather than interview, ...

Information Bias (cont'd)

- Similar to recall bias may occur in cohort study: Unexposed people are underdiagnosed for disease more than exposed.
    - (eg.) To assess the effect of tobacco smoking on the occurrence of emphysema（気腫）by cohort study, if no examination to check the diagnosis is conducted, diagnosis of emphysema is often missed, and thus it may more likely to be diagnosed in smokers than in nonsmokers → differential misclassification of disease (**biased follow-up**)
    - This bias can be avoided by conducting examination for emphysema as part of study itself
- Differential misclassification can exaggerate or underestimate an effect.

- **Nondifferential misclassification** (more common): exposure or disease (or both) is misclassified but the misclassification does not depend on a person's status for the other variables
    - (eg.) To examine the relationship between consumption of red wine and emphysema: assuming that consumption of red wine is not related to smoking, diagnosis of emphysema is not affected by whether the subject consume more red wine or less → somebody who develops emphysema may not be diagnosed, but such misclassification may occur in the same probability regardless the consumption of red wine
- Nondifferential misclassification can only "dilute" (closer to the null or no-effect value than actual effect).

Example of **dilution** in nondifferential misclassification in dichotomous exposure

	Correct classification		Nondifferential misclassification			
	High-fat diet		20% of No → Yes		20% N → Y	
					20% Y → N	
	No	Yes	No	Yes	No	Yes
MI cases	450	250	360	340	410	290
Controls	900	100	720	280	740	260
RR	OR=5.0		OR=2.4		OR=2.0	

- Case-control study to assess the relation between eating high-fat diet and subsequent heart attack (myocardial infarction: MI)
- According to arbitrary cutoff of eating high-fat diet or not and measurement error (very common in dietary assessment study), everybody may be misclassified regardless with heart attack
- In the table 7-6 (shown left), true OR is (250/450)/(100/900)=5.0
- If 20% of not eating high-fat diet people are misclassified as eating high-fat diet, 450 → 450x0.8=360, 250 → 250+450x0.2 = 340, 900 → 900x0.8=720, and 100 → 100+900x0.2=280. OR is (340/360)/(280/720) = 2.4. Less than ½ of the effect is seen. Excess RR (5 −1=4) is reduced to 1.4 (=2.4 − 1).
- If, 20% misclassification occurred in both direction, 450 → 450x0.8+250x0.2 = 410, 250 → 250x0.8+450x0.2 = 290, 900 → 900x0.8+100x0.2 = 740, 100 → 100x0.8+900x0.2 = 260. OR is (290/410)/(260/740) = 2.0. ¾ of the effect is nullified.

Confounding

- **Simple definition**: confusion of effects
- The effect of the exposure is mixed with the effect of another variable, leading to bias
- (eg.) Stark CR, Mantel N (1966) Effects of Maternal Age and Birth Order on the Risk of Mongolism and Leukemia. *Journal of the National Cancer Institute*, 37(5): 687–698.
  - (Fig 7-3) The prevalences of Down syndrome by birth order increased from about 0.6/1000 at first birth to 1.7/1000 at fifth or greater order births
  - Higher order births occur in elder mothers, so that fig 7-3 mixes the effects of birth order and mother's age → <u>The effect of birth order on the prevalence of Down syndrome is confounded by the effect of mother's age</u>
  - (Fig 7-4) The prevalences of Down syndrome by mother's age increased from 0.4/1000 in younger than 20 years to 8.5/1000 in mothers with age 40 or elder.

- From Fig 7-4, whether the effect of mother's age is confounded by birth order is unknown
- (Fig 7-5) The prevalences of Down syndrome at birth by both birth order and mother's age simultaneously.
  - Within each category of birth order, looking from the front to the back, the same striking trend in prevalence of Down syndrome with increasing maternal age
  - Within each category of mother's age, looking from left to right, no discernible trend with birth order
- The maternal age effect is not confounded by birth order, but by other factors, <u>because age is just a marker of time</u>. Biologic events occurring during a woman's aging process may truly affect the increase of Down syndrome. Age is a proxy for unidentified events. If we identify such events, we may find that maternal age has no effect after controlling the biologic changes correlated with age.

Confounding (cont'd)

- Strict definition: The confounding variable must have an effect on the outcome to be confounding.
  - Theoretically a confounding variable should be a cause of the disease, but in practice, it may be only a proxy or a marker for a cause.
  - Anyway, a confounder is a predictor of disease occurrence, whereas not all predictor is confounder.
  - (eg.) Age would not be confounding unless the age distributions of people in the various exposure categories differed. If the age distribution in different exposure category is same, the comparison between different exposure categories is not distorted by age.

- The results of a <u>randomized</u> trial designed to assess how well three treatments for diabetes prevented fatal complications. (Even randomized, confounding may occur)
- Table 7-7 shows crude data comparing mortality between exposure groups. The excess mortality in tolbutamide group (0.147 – 0.102 =0.045) means the additional risk of 4.5% of dying over 7 years compared to placebo group. (But confounded by age)
- Table 7-8 shows excess mortality due to tolbutamide (approx. 3.5%) in either age group, though it's still somewhat confounded by age (stratification is done only for 2 age categories)
- 4.5% is almost 30% overestimate of the adverse effect of tolbutamide than 3.5%

Table 7-7. Mortality during 7 yr follow-up

	Tolbutamide	Placebo
Deaths	30	21
Surviving	174	184
Total	204	205
Mortality risk	0.147	0.102

Table 7-8. Stratified by 2 age groups

	Age < 55		Age > 55	
	T	P	T	P
Dead	8	5	22	16
Surviving	98	115	76	69
Total	106	120	98	85
Mortality	0.076	0.042	0.224	0.188
Difference	0.034		0.036	

Confounding (cont'd)

- Confounding can cause a bias in either direction
- Properties of a confounding factor (Note: it's not definition)
  - A confounder must be associated with the disease
  - A confounder must be associated with exposure
  - A confounder must not be an effect of exposure
- Is confounding in a randomized experiment a bias?
  - It's an example of random error rather than systematic error
  - Confounding in an experiment can be controlled using the same methods to control confounding in nonexperimental studies

- Control of confounding: 3 methods
  - Randomization
    - Can be used only in experiments
  - Restriction
    - Selecting subjects for a study who all have the same value or almost the same value for a variable that would otherwise be a confounding variable
    - Can be used in any epidemiologic study
    - May work contrary to generalizability of the result (representativeness issue), but it's the nature of science
  - Matching in cohort studies
    - The index series is exposed cohort, unexposed cohort can be matched to have same age distribution (if age is potential confounding) → "**frequency matching**"
    - To take the exposed subjects one by one and to find for each of them an unexposed subject that has a matching age → "**individual matching**"
    - Expensive, except for the cases if all potential subjects and their data are already stored in a data warehouse or database.
  - Since no method prevents confounding completely, these are best viewed as methods to limit confounding.