Chapter 10. Controlling Confounding by Stratifying Data

January 11, 2024
Epidemiology (12)

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

Examples of confounding

- Fig 7-3 (Chapter 7) showed that the apparent effect of birth order on the prevalence at birth of Down syndrome was attributable to confounding. As Fig 7-4 showed, maternal age had stronger effect, and as Fig 7-5 showed, maternal age had strong effect in any birth order but birth order showed no clear effect in any maternal age-class. The birth order effect was confounded by maternal age.

- Fig 7-5 is graphic demonstration of **stratification**.

  - A kind of cross-tabulation of data

  - Similar example is showed in Table 1-2 (Chapter 1).

- Stratification is an effective and straightforward means to control confounding.
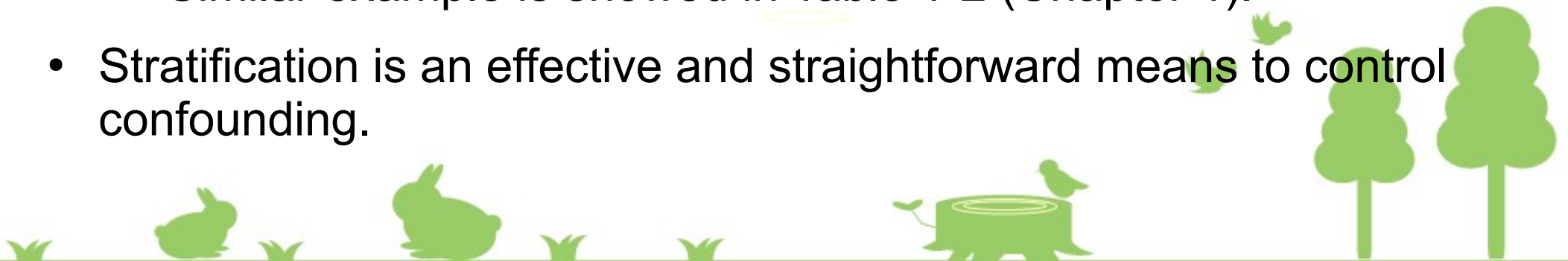
AN EXAMPLE OF CONFOUNDING

Table 10-1. Mortality rates among patients with trigeminal neuralgia categorized by sex

	Males	Females
Deaths	90	131
Person-years (pyr)	2465	3946
Mortality rate	36.5/1000 pyr	33.2/1000 pyr
Rate ratio	1.10	
90%CI	0.88-1.38	
Data from Rothman and Monson		

- Mortality rates in patients of trigeminal neuralgia （三叉神経痛 =recurrent proximal pain of the face ） were compared between males (exposed) and females (unexposed)

- Rate ratio 1.10 indicates slightly greater mortality rate for males than females

- The result is confounded by age (Females are more aged and mortality is higher in aged)

Table 10-2. Mortality rates among patients with trigeminal neuralgia categorized by sex and age category

	<65 Years		65+ years	
	Males	Females	Males	Females
Deaths	14	10	76	121
Person-years (pyr)	1516	1701	949	2245
Mortality rate (cases/1000 pyr)	9.2	5.9	80.1	53.9
Rate ratio	1.57		1.49	
90%CI	0.79-3.10		1.17-1.89	
Data from Rothman and Monson				

- Stratified comparison by two age categories showed higher rate ratios in both age categories.

- Crude rate ratio was biased downwarded

- The effect of confounding can be removed by stratification

- Next, how to control the effect of confounding is explained

UNCONFOUNDED EFFECT ESTIMATES AND CONFIDENCE INTERVALS FROM STRATIFIED DATA

- The methods to estimate unconfounded effect
  - Stratified analysis
  - To aggregate the information over strata
    - Pooling
    - Standardization

- Pooling
  - Assume that the effect being estimated is constant across the strata
  - Each stratum provides a stratum-specific estimate
  - Overall effect (pooled estimate) is obtained as weighed average of stratum-specific estimate
  - The weights are assigned so that the strata with the most data get the most weight
  - If the effect differs by strata (heterogeneous), pooling is not applicable, but standardization is still applicable
    - See, pp.100 of https://cran.r-project.org/doc/contrib/Epicalc_Book.pdf
    - https://www.cogentoa.com/article/10.1080/25742558.2018.1478698.pdf
  - Mantel-Haenszel's method is commonly used to get pooled estimate https://doi.org/10.1093/jnci/22.4.719

- Cohort studies with risk data or prevalence data
  - Consider risk data. For analytic purposes, prevalence data may be treated the same as risk data
  - For the ith stratum, the table below is obtained.
  - Mantel-Haenszel's pooled risk difference (RD_{MH}) is obtained by [10-1] and Mantel-Haenszel's pooled risk ratio (RR_{MH}) is obtained by [10-2]

	Exposed	Unexposed	Total
Cases	a_i	b_i	M_{1i}
Noncases	c_i	d_i	M_{0i}
Total at risk	N_{1i}	N_{0i}	T_i

$$\mathrm{RD}_{MH} = \frac{\sum_i \frac{a_i N_{0i} - b_i N_{1i}}{T_i}}{\sum_i \frac{N_{1i} N_{0i}}{T_i}} \quad [10-1]$$

$$\mathrm{RR}_{MH} = \frac{\sum_i \frac{a_i N_{0i}}{T_i}}{\sum_i \frac{b_i N_{1i}}{T_i}} \quad [10-2]$$

Example: Stratification of risk data (Table 7-7 and 7-8 → Table 10-3)

Table 10-3. Risk of death for groups receiving tolbutamide or placebo in the university group diabetes program in 1970.					
	<55 Years		55+ Years		Total	
	Tolbutamide	Placebo	Tolbutamide	Placebo	Tolbutamide	Placebo
Deaths	8	5	22	16	30	21
Total at risk	106	120	98	85	204	205
Risk of death	0.076	0.042	0.224	0.188	0.147	0.102
Risk difference	0.034		0.036		0.045	
Risk ratio	1.81		1.19		1.44	

* Data from University Group Diabetes Program (Meinert, C. L., Knatterud, G. L., Prout, T. E., & Klimt, C. R. (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. Mortality results. Diabetes, 19 Suppl. 2: 789-830.)

- Age structure is different between tolbutamide group and placebo group: 48% (98/204) of tolbutamide group is at least 55 years old, but 41% (85/205) of placebo group is so.
- Within-stratum confounding is referred as residual confounding. The best number of strata depends on tha balance the needs to control confounding and the needs to avoid random error (see, BOX).
- Using R code (http://minato.sip21c.org/epispecial/codes-for-Chap10.R), it's easy to calculate. The code calculate the point estimate and confidence intervals.
    - If not stratified, RD is 0.045, but RD_{MH} is 0.035. Pooled estimate is a weighted average of the stratum-specific estimates of the effect and thus within this range. Crude estimate is not within this range, which includes the effect of tolbutamide and confounding effect of age. Tolbutamide group is older than placebo group, so that crude RD is larger than pooled RD.
    - If not stratified, RR is 1.44, but RR_{MH} is 1.33. Some age confounding is removed by stratification, but the RR in each stratum (1.81 and 1.19) is quite difference each other. The assumption of "constant effect across strata" may be violated. However, this null-hypothesis is not rejected by statistical significance (test of heterogeneity/homogeneity), so that calculating pooled estimate is still reasonable and convenient in this case.

Confidence intervals for pooled estimates

- Variance formulas for pooled analysis (RD$_{MH}$, RR$_{MH}$, IRD$_{MH}$, IRR$_{MH}$, OR$_{MH}$) are given right.

- Confidence intervals (for 90%CI, Z=1.645, for 95%CI, Z=1.96) are given as follows
  - For difference measures, point estimate ± Z x square root of Var(*D$_{MH}$)
  - For ratio measures, exponential of the natural log of point estimate ± Z x square root of Var(ln(*R$_{MH}$))

- Calculating 90%CI for Table 10-3, RD$_{MH}$ ranges from -0.018 to 0.088 and RR$_{MH}$ ranges from 0.87 to 2.0 (By R codes, RD$_{MH}$ ranges from -0.018 to 0.088 and RR$_{MH}$ ranges from 0.87 to 2.03)

- Point estimate and 90%CI for Table 1-2, RR$_{MH}$ is 1.21, ranging from 1.06 to 1.38, whereas crude RR is 0.76, ranging from 0.65 to 0.88. Strongly confounded by age

Table 10-4. Variance formulas for pooled analysis

$$\text{Var}(\text{RD}_{MH}) = \frac{\sum_i (\frac{N_{1i}N_{0i}}{T_i})^2 \left[\frac{a_i c_i}{N_{1i}^2(N_{1i}-1)} + \frac{b_i d_i}{N_{0i}^2(N_{0i}-1)} \right]}{(\sum_i \frac{N_{1i}N_{0i}}{T_i})^2}$$

$$\text{Var}[\ln(\text{RR}_{MH})] = \frac{\sum_i (\frac{M_{1i}N_{1i}N_{0i}}{T_i^2} - \frac{a_i b_i}{T_i})}{\left(\sum_i \frac{a_i N_{0i}}{T_i}\right)\left(\sum_i \frac{b_i N_{1i}}{T_i}\right)}$$

$$\text{Var}(\text{IRD}_{MH}) = \frac{\sum_i (\frac{PT_{1i}PT_{0i}}{T_i})^2 (\frac{a_i}{PT_{1i}^2} + \frac{b_i}{PT_{0i}^2})}{\left(\sum_i (\frac{PT_{1i}PT_{0i}}{T_i})\right)^2}$$

$$\text{Var}[\ln(\text{IRR}_{MH})] = \frac{\sum_i \left(\frac{M_{1i}PT_{1i}PT_{0i}}{T_i^2}\right)}{\left(\sum_i \frac{a_i PT_{0i}}{T_i}\right)\left(\sum_i \frac{b_i PT_{1i}}{T_i}\right)}$$

$$\text{Var}[\ln(\text{OR}_{MH})] = \frac{\sum_i G_i P_i}{2\left(\sum_i G_i\right)^2} + \frac{\sum_i (G_i Q_i + H_i P_i)}{2\left(\sum_i G_i \sum_i H_i\right)} + \frac{\sum_i H_i Q_i}{2\left(\sum_i H_i\right)^2}$$

$$\text{where} \quad G_i = \frac{a_i d_i}{T_i}, H_i = \frac{b_i c_i}{T_i}, P_i = \frac{(a_i + d_i)}{T_i}, Q_i = \frac{(b_i + c_i)}{T_i}$$

Cohort studies with incidence rate data

	Exposed	Unexposed	Total
Cases	a_i	b_i	M_i
Person-time at risk	PT_{1i}	PT_{0i}	T_i

- Pooled estimates of IRD and IRR using Mantel-Haenszel's method are given in [10-4] and [10-5].

$$\text{IRD}_{MH} = \frac{\sum_i \frac{a_i PT_{0i} - b_i PT_{1i}}{T_i}}{\sum_i \frac{PT_{1i} PT_{0i}}{T_i}} \quad [10-4]$$

$$\text{IRR}_{MH} = \frac{\sum_i \frac{a_i PT_{0i}}{T_i}}{\sum_i \frac{b_i PT_{1i}}{T_i}} \quad [10-5]$$

Table 10-5. Mortality rates for current and past clozapine users, overall and by age category

	10-54 years		55-94 years		Total	
	Curr	Past	Curr	Past	Curr	Past
Deaths	196	111	167	157	363	268
Person-Years	62119	15763	6085	2780	68204	18543
Rate ($\times 10^5$ yr)	315.5	704.2	2744	5647	532.2	1445
IRD ($\times 10^5$ yr)	-388.7		-2903		-912.8	
IRR	0.45		0.49		0.37	

Walker et al. (https://www.ncbi.nlm.nih.gov/pubmed/9345668)

- Clozapine (to treat schizophrenia) is thought to affect mortality primarily for current users. Past users are the reference for judging the effect of current use.

- The death rates are much greater for older patients. Among current users, 9% (6085/68204) person-time is in older age category, but among past users, 15% (2780/18543) person-time is in older age category.

- IRD_{MH} is -720.0 x 10^{-5}(/yr) (90%CI, -867.7 to -572.3 x 10^{-5}(/yr)), smaller in absolute value than the crude IRD, -912.8 x 10^{-5}(/yr).

- IRR_{MH} is 0.47 (90%CI 0.41 to 0.54) is larger than crude IRR, 0.37.

Case-Control Studies

	Exposed	Unexposed	Total
Cases	a_i	b_i	M_{1i}
Controls	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

$$\text{OR}_{MH} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}} \quad [10-6]$$

- Pooled incidence rate ratio (or pooled risk ratio, depending on control sampling) is estimated as a pooled odds ratio from [10-6]

- Further references

  – Cornfield (1951)
  https://www.ncbi.nlm.nih.gov/pubmed/14861651
  explains how to estimate P(D|E) from case-control data, using Bayes theorem, but equals to OR as a value.

  – Breslow (1981)
  https://academic.oup.com/biomet/article/68/1/73/238096
  explains exact method to estimate confidence intervals

  – Miettinen (1974)
  https://www.jstor.org/stable/pdf/2285664.pdf

  – Miettinen and Nurminen (1985)
  https://doi.org/10.1002/sim.4780040211

Table 10-6. Infants with congenital heart disease and Down syndrome, and healthy controls, by maternal spermicide use before conception and maternal age at delivery

	<35 years			35+ years		
	Yes	No	Total	Yes	No	Total
Cases	3	9	12	1	3	4
Controls	104	1059	1163	5	86	91
Total	107	1068	1175	6	89	95
OR	3.39			5.73		

Data from Rothman
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1649884/)

- Crude OR from Table 10-6 is 3.50 (90%CI is 1.34 to 9.18)

- OR_{MH} is 3.78 (90%CI is 1.43 to 10.0)

- Original article wrote that prevalence ratio (PR) estimated from OR (Cornfield, 1951). Crude PR with exact 90%CI was 3.6 (1.2 to 9.0). Pooled PR with exact 90%CI was 3.9 (1.3 to 9.8) (Exact CI was obtained by Breslow, 1981 and Miettinen, 1974).

STANDARDIZATION

- A method of combining category-specific rates into a single summary value by taking a weighted average.

- The weights is taken from a standard population or distribution.

- For the comparison of mortality rates over long-term different years, the common age structure is used as weights (in Japan, 1985 population).

- Suppose

  - Male's mortality rate: 10/1000 yr^{-1}

  - Female's mortality rate: 5/1000 yr^{-1}

- Any standard population can be used

  - If males and females have equal weights, a weighted average is 7.5/1000 yr^{-1}

  - If males share 5% and females share 95% in a standard population, standardized rate is 5.25/1000 yr^{-1}

- If all strata had similar rates, the choice of weights doesn't matter.

- The main reason to standardize is to facilitate comparison.

- Both standardization and pooling are taking weighted average, but weights are different. In pooling using Mantel-Haenszel method, weights reflect the amount of information in strata. But in standardization, weights has no relation with the amount of information in each stratum.

- Pooling requires the assumption that the effect is same in all strata (assumption of *uniformity of effect*). Even when this assumption is violated, pooling may still be reasonable. The effect is not strictly expected as uniform across strata. When the effect is strikingly different for different strata, we can still use standardization to get a summary measure of the effect.

- In the example of Table 10-5, rates of past clozapine users can be standardized using the person-years of current users as weights (1144/100000 yr^{-1}). When rates of current users are standardized using same standard, it becomes crude rate in current users (532.2/100000 yr^{-1}).

- The standardized IRD 532.2-1144=-612(/100000 yr^{-1}), slightly smaller than IRD$_{MH}$.

- The standardized IRR 532.2/1144 = 0.47 is essentially identical to the result obtained through pooling.

CALCULATION OF *P* VALUES FOR STRATIFIED DATA

- What is an SMR?

  - Standardized Mortality (Morbidity) Ratio

  - In Table 10-5, the ratio of actual number of death of current users to the expected number of death of current users if they experienced the same death rate as past users (past users are reference pop)

  - It's completely same as standardized RR. The SMR is the standardization of which standard is always exposed group.

  - Sometimes called as "indirect standardization", a misnomer.

- CI is better than p-value, but it's given for completeness.

- For risk, prevalence, or case-control data

  - For Table 10-6, $\chi = 2.41$

  - $(1 - pnorm(2.41)) \times 2 = 0.016$

$$\chi = \frac{\sum_i a_i - \sum_i \frac{N_{1i} M_{1i}}{T_i}}{\sqrt{\sum_i \frac{N_{1i} N_{0i} M_{1i} M_{0i}}{T_i^2 (T_i - 1)}}}$$

To evaluate the effect of unmeasured confounding, new methods are suggested. If we assume true RR (TRR) as some fixed value, the bias can be defined as
bias = 1/TRR x [RR+√(RR(RR-TRR))]
When TRR=1, this bias is called as E-value.
(See, https://annals.org/aim/article-abstract/2643434)

- For rate data

  - For Table 10-5, $\chi = -9.55$

  - $(1 - pnorm(9.55)) \times 2 < 10^{-21}$

$$\chi = \frac{\sum_i a_i - \sum_i \frac{PT_{1i} M_i}{T_i}}{\sqrt{\sum_i M_i \frac{PT_{1i} PT_{0i}}{T_i^2}}}$$

- MEASURING CONFOUNDING

  - It's reasonable to assess how much confounding a given variable produces in a body of data

  - It's possible to predict the amount of confounding from the general characteristics of confounding variables (the associations of a confounder with both exposure and disease)

  - By removing the effect of confounding, how much it's removed can be assessed.

  - Example: Comparison between Table 1-1 and 1-2. Crude RR (Table 1-1) was 0.76, which means the risk among smokers being 24% lower than nonsmokers. RR_{MH} (Table 1-2) was 1.21, which means risk among smokers being 21% higher than nonsmokers. The discrepancy between crude RR and RR_{MH} is a direct measure of the age confounding.

  - Statistical significance test (the association between confounder and exposure or disease) should not be used to evaluate the presence/absence of confounding.

- Stratification by two or more variables

  - Explanation and examples in this chapter used few strata with one stratification variable

  - To control the sex and age with 5 age categories, 10 strata are produced. Mantel-Haenszel method can be applied to such combination strata without any change

  - Difficulty is the trade-off of the number of strata and the amount of information included in each strata

- Stratification after matching

  - In cohort study, matching may avoid the potential confounding

  - But it's worthwhile to conduct stratified analysis by matched factors. It can lead to narrower CI than would be obtained if matching had been ignored.

  - In case-control study, matching may cause selection bias, which must be removed in the analysis and stratification by the matched sets can accomplish it.

- Importance of stratification

  - The methods to control confounding by stratification explained in this chapter is low-technology and easy to use

  - Sometimes sophisticated multivariate statistical models (explained in Chapter 12) are favored by some researchers.

  - Stratification has several advantages over multivariate analysis

    - With stratified analysis, researcher can visualize the distribution of subjects

    - Reader of the research can also visualize the distribution

    - Fewer assumptions are needed compared to multivariate models.