Biostatistics (MATH11230)
Confounding

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Confounding

Introduction to the concept

- Up to now, and as already alluded briefly before, there is a complication that we have been ignoring: the role of other factors when estimating the association between an exposure and a disease outcome.

- **Confounding** is a distortion in the estimated measure of association between an exposure and disease/health outcome that occurs when the exposuree that occurs when the exposure groups differ with respect to other factors that influence the outcome.

- Since the primary exposure of interest is rarely the only factor that differs between exposed and unexposed groups and that also affects the disease outcome, confounding is a common occurrence in epidemiology.

- If present, confounding may lead to an under or over estimate of the association measure of interest. It can even chenge the apparent direction of an effect.

- A research question in which we would want to consider confounding is, for instance, the following: does being overweight increases the risk of coronary heart disease, independently of cholesterol, hypertension, and diabetes?

Confounding

Confounding needs to be addressed

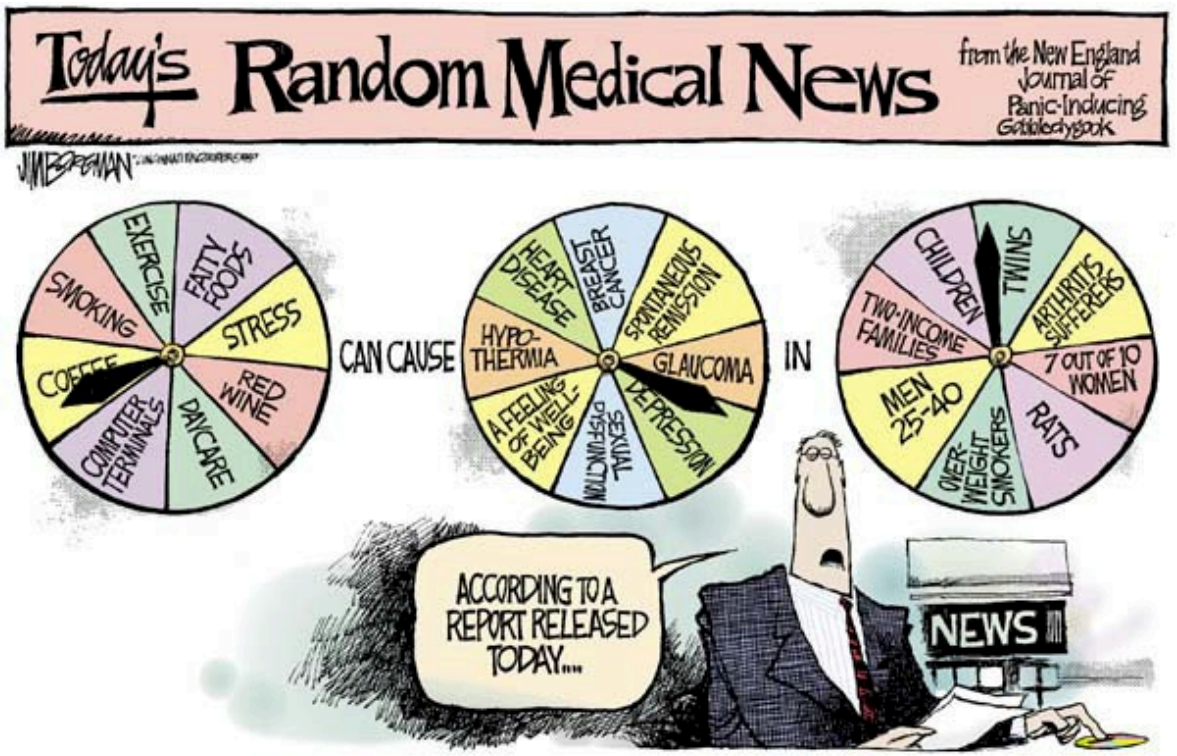
↪ From the BMJ editorial: "*The scandal of poor epidemiological research*", available here:

`https://www.bmj.com/content/329/7471/868.full`

"Confounding, the situation in which an apparent effect of an exposure on risk is explained by its association with other factors, is probably the most important cause of spurious associations in observational epidemiology."

Confounding

On the less serious side...



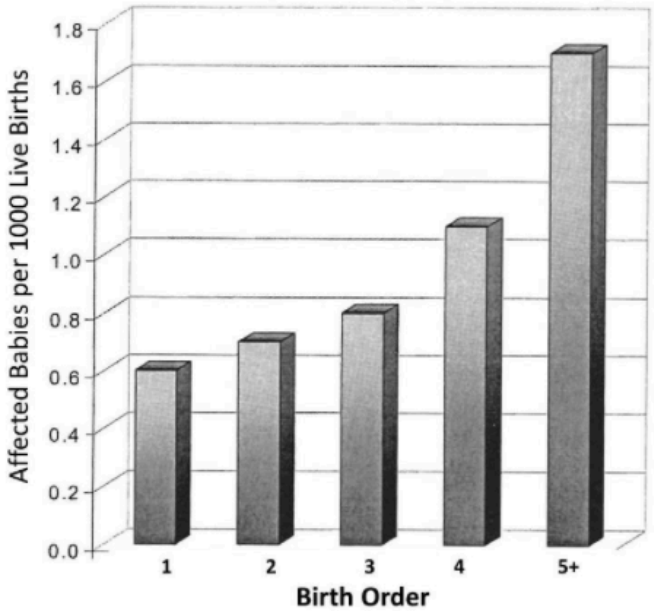
Cartoon by Jim Borgman, first published by the Cincinnati Inquirer and King Features Syndicate 1997 Apr 27; Forum section: 1 and reprinted in the New York Times, 27 April 1997, E4.

Confounding

Example

- The following example is a classic one and it is taken from Rothman (*Epidemiology: An Introduction*, 2002, p 101–105). The figures used in this example are also taken from this book.

- The figure below, based on data from the work of Stark and Mantel (1966), shows an increasing trend in the prevalence of Down syndrome with increasing birth order or, otherwise stated, an association between increasing birth order and risk of Down syndrome.

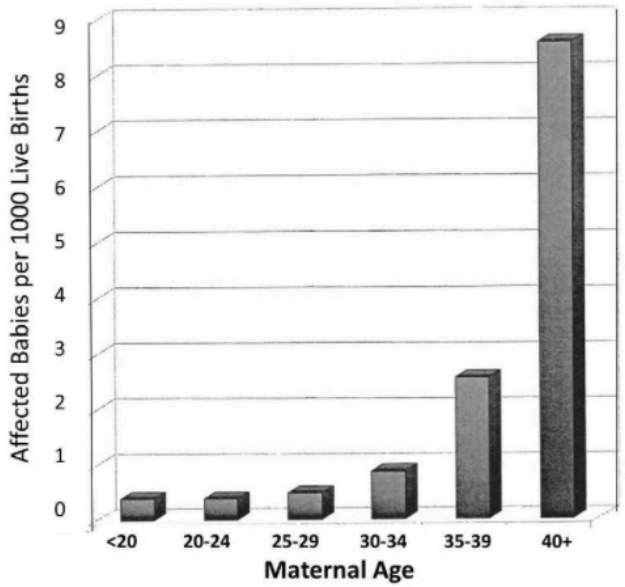


Confounding

Example

- The effect of birth order on the risk of Down syndrome, however, is a blend of whatever effect birth order has by itself and the effect of another variable that is closely correlated with birth order: the age of the mother.

- The figure below shows a striking relationship between maternal age at birth and the child's risk of being born with Down syndrome.



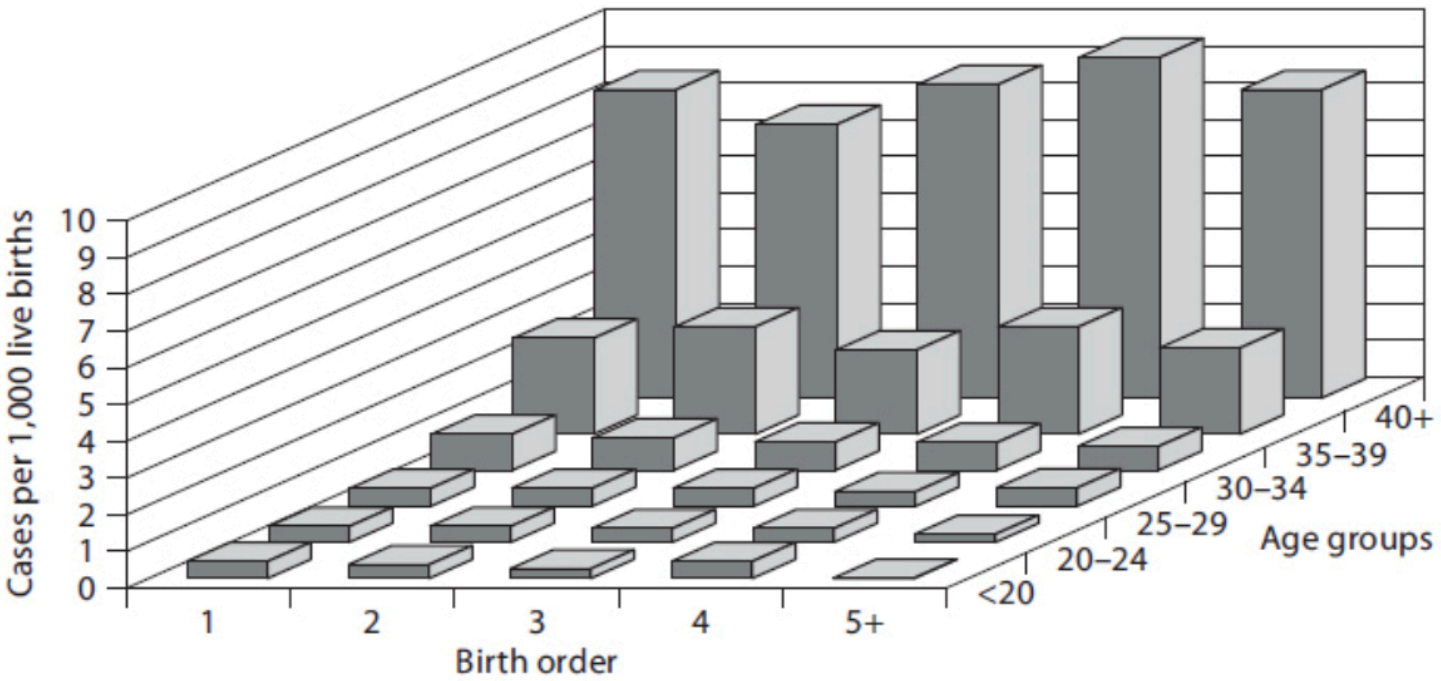
Confounding

Example

- ↪ It is reasonable to expect that mother's giving birth to their fifth baby are, on average, older than mother's giving birth to their first child.
- ↪ Therefore, the comparison of high-birth order babies with lower-birth order babies is to some extent a comparison of babies born to older mothers with babies born to younger mothers.
- ↪ That is, the relationship between birth order and of Down syndrome is confounded by age.
- ↪ But is the converse also true? That is, is the effect of maternal age on Down syndrome also confounded by birth order?
- ↪ It is possible, but only if birth order has really some independent effect on the likelihood of Down syndrome. i.e., an effect that birth order is linked to maternal age.
- ↪ To sort this out, Rothman (2002, p 104) suggests to look at both effects simultaneously as in the plot in the next slide.

Confounding

Example



Confounding

Example

- From the figure in the previous slide, we can observe that within each category of maternal age, looking from left to right, there is no discernible trend with birth order.
- Therefore, the apparent trend in the first plot in slide 5, was due entirely to confounding by maternal age.
- In other words, if one controls for maternal age, there is no evidence that birth order has any impact on the prevalence of Down syndrome.
- On the other hand, within each category of birth order there is clearly a marked increase in prevalence as maternal age increases within all five levels of birth order.
- That is, even after taking birth order into account the association with maternal age persists.

Confounding

A never ending journey?

- ↪ Because most health problems have many risk factors, there is a lot of room for potential confounding.

- ↪ It is interesting to read Rothman's (2002, p 105) take on this:

> *"The research process of learning about and controlling for confounding can be thought of as a walk through a maze toward a central goal. The path through the maze eventually permits the scientist to penetrate into levels that successively get closer to the goal: in [the example of maternal age and Down syndrome] the apparent relations between Down syndrome and birth order can be explained entirely by the effect of mother's age, but that effect in turn will ultimately be explained by other factors that have not yet been identified. As the layers of confounding are left behind, we gradually approach a deeper causal understanding of the underlying biology. Unlike a maze, however, this journey toward biologic understanding does not have a clear endpoint, in the sense that there is always room to understand the biology in a deeper way."*

Confounding

Key criteria

↪ The three conditions for considering a factor C to be a potential confounder of the relationship between the exposure E and the diasease D are:

1. C is a risk factor for D, independent of E, but not a direct consequence of D.
2. C is associated with E, but not a direct consequence of E.
3. C is not an intermediate step in the causal path between E and D. That is, if E affects C, which in turn affects D, then we should not adjust for the effect of C in our analysis of the $E - D$ association.

↪ Note that we have used the word 'potential' because not all variables that meet these criteria will actually turn out to confound the data – we figure this out during the analysis.

Confounding

Key criteria

- ↪ Criterion 1 for confounding means that among the unexposed there should be an association between the confounder and the disease outcome. One could calculate, e.g., the relative risk/odds ratio between C and D in the unexposed group.
- ↪ On the other hand, criterion 2 means that the confounder variable should have a different distribution in the in the exposed and unexposed groups.
- ↪ With respect to criterion 3: suppose that one is interested in the association between saturated fat and coronary heart disease (CHD).
- ↪ Scientific studies have shown that saturated fat increases LDL (low-density lipoprotein) levels and, in turn, elevated LDL levels are a major cause of CHD disease.
- ↪ Thus, LDL levels are an intermediate step in the causal pathway between saturated fat and CHD.

Confounding

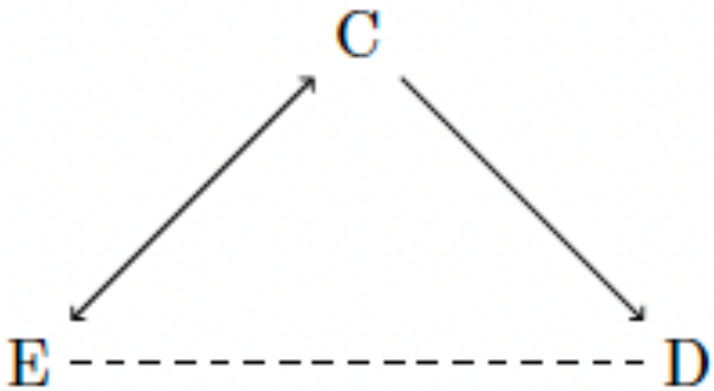
Key criteria

- **Question**: would we want to control for LDL levels?
- If we control for LDL levels, we essentially control for the saturated fat level, and we would likely find an adjusted estimate of the relative risk or odds ratio relating saturated fat to CHD status to be close to the null value of one (or very much attenuated).
- This is the reason why variables that reside in the causal pathway between E and D are not confounders, i.e., we do not control for them in the design/analysis stages. These variables are known in the literature as **mediators**.

Confounding

Causal diagram

↪ We can describe the possible relationships between the disease, exposure, and confounder using a so-called causal diagram



↪ Here one sided arrows ($\rightarrow$) indicate causality, two sided arrows ($\leftrightarrow$) indicate association, and dashed lines ($- - -$) represent the relationship being studied.

Confounding

Implications

- ↪ The implications of confounding is that the crude/unadjusted measure of association reflects the effects of both the exposure of interest and confounders.

- ↪ Given that we know confounding has occurred, we seek an adjusted association measure that reflects only the direct effect of the exposure of interest.

Controlling for confounding...

...at the study design

- One can control for confounding either in the study design stage or in the analysis stage, the latter provided one has information on the status of the confounding variable in the study subjects.

- In terms of study design, three approaches are possible:

  - Restriction.

  - Matching.

  - Randomisation.

Controlling for confounding...

...at the study design: restriction

- We have just seen that one necessary condition for confounding to occur is that the confounding variable must be distributed unequally between the exposed and unexposed groups.

- A strategy that therefore is logic in order to avoid confounding is **to restrict** the admission into the study to subjects who share the same value (or levels) of the confounding variable(s).

- For example, in the study looking at the association between birth order and risk of Down syndrome we found that age was confounding (in that specific case, entirely driving) the mentioned association. Let us suppose that the only confounding variable of concern was, in fact, age.

- Under such an assumption, restricting the study to babies born from mothers in the same age band (e.g., 20 to 24 years old) would have avoided confounding as the mothers' age distributions were similar in the different groups of babies being compared.

- A possible more mundane example: restricting the study population to nonsmokers when studying the association of air pollution with lung cancer.

Controlling for confounding...

...at the study design: restriction

- The strategy of restriction is simple and works beautifully in terms of controlling confounding but often it is not a realistic approach because it limits the study too much.
- In the example in the previous slide, by restricting the babies participating in the study to those born from mothers in a given (and narrow) age band, we would loose the ability to generalise the findings.
- Further, **residual confounding** can occur if one does not restrict narrowly enough.
- For instance, suppose that maternal's age would have been restricted to $30 - 39$. However, from the figure in slide 6, we could observe that the risk of Down syndrome is quite different from ages $30 - 34$ and $35 - 39$.

Controlling for confounding...

...at the study design: matching

- ↪ Instead of restriction, one could also ensure that the study groups do not differ with respect to possible confounders by **matching** the comparison groups.
- ↪ Suppose we were to compare how high birth order babies (4, 5+) vs low birth order babies (1, 2, 3) affects the risk of Down syndrome. Age is again the confounder. Then, under matching, for every high birth order baby born from a mother in a given age band, we would enrol a low birth order baby from a mother in the same age band.
- ↪ Maternal age would still be a risk factor for Down syndrome, but by enforcing that the maternal distribution to be the same in the groups of high and low birth orders, we have negated an association between the exposure (birth order) and the confounding variable (mother's age).
- ↪ Obviously, the type of matching just described is only valid for exposure-based (cohort) sampling designs.
- ↪ A similar reasoning can be followed in case-control studies.
- ↪ Data arising from matched studies require special analytical techniques (because the unexposed or control group is not a random sample from the study population).

Controlling for confounding...

...at the study design: randomisation

- **Randomisation**, typically employed in experimental studies (clinical trials), is a powerful method to prevent confounding.

- If a large number of individuals are allocated to treatment groups in a random fashion, where each individual has equal probability of being in any treatment group, then it is reasonable to expect that the groups will have similar distributions of age, gender, behaviours, and virtually all other unknown possible confounder factors.

- Remember that randomisation is not achieved in observational studies as the investigator does not assign individuals to the exposure/unexposed groups!

Controlling for confounding...

...in the analysis stage: stratification

- One way to solve for the problem of confounding in the **analysis stage** is to restrict comparisons to subjects who have the same value (or values in a similar range) of the confounding variable C.
- The subsets defined by the levels of C are called **strata**, and so this process is known as **stratification**.
- It leads to separate estimates of the OR/RR for the $E - D$ association in each stratum.
- We shall note that C does not necessarily need to be a binary variable. For instance, we might allow for the confounding effect of age by splitting it into several discrete categories (e.g., 20 to 30 years old, 30 to 40 years old).
- We should nevertheless be aware that residual confounding can still occur when we discretise a continuous confounder variable.
- Of course, the best way to overcome this is to ensure that the groups are as narrow as possible, but this may lead to very few (or in the extreme case even none!) individuals in some stratum. This is especially true when we have more than one confounder variable (e.g., if age and gender are the confounders we are adjusting for and if we discretise age in six categories, we will have 12 strata).

Controlling for confounding...

...in the analysis stage: stratification

- ↪ Unless it appears that the association between the exposure and the outcome varies markedly between the strata (more on this later), we will usually wish to combine the evidence from the separate strata and summarise the association, controlling for the confounding effect of C.

- ↪ The general approach to pooling is to take weighted averages of the stratum-specific estimates (of the odds ratio/relative risk) using weights that reflect the reliability of the estimates. We particularly focus, mainly for historical reasons, on the Mantel-Haenszel methods.

- ↪ Regression techniques, in particular, logistic regression, which we will learn in the next lectures, provide another and more modern alternative for calculating summary estimates.

Controlling for confounding...

...in the analysis stage: stratification

- We start with the odds ratio.
- The table below shows the notation we will use for the 2×2 contingency table in stratum i.

	D	not D	Totals
E	a_i	b_i	$a_i + b_i$
not E	c_i	d_i	$c_i + d_i$
Totals	$a_i + c_i$	$b_i + d_i$	$n_i = a_i + b_i + c_i + d_i$

- It is exactly the same as the table we have used before but with the subscript i added, to refer to stratum i.
- The estimate of the odds ratio for stratum i is

$$\widehat{\text{OR}}_i = \frac{a_i d_i}{b_i c_i}.$$

Controlling for confounding...

...in the analysis stage: stratification

↪ According to Mantel and Haenszel (1959), the weights are given by

$$\omega_i = \frac{b_i c_i}{n_i}, \quad i = 1, \ldots, I.$$

↪ Therefore, the adjusted Mantel–Haenszel estimate of the odds ratio is given by

$$\begin{aligned}
\widehat{\mathrm{OR}}_{\mathrm{MH}} &= \frac{\sum_{i=1}^{I} \omega_i \widehat{\mathrm{OR}}_i}{\sum_{i=1}^{I} \omega_i} \\
&= \frac{\sum_{i=1}^{I} \frac{b_i c_i}{n_i} \frac{a_i d_i}{b_i c_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{n_i}} \\
&= \frac{\sum_{i=1}^{I} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{n_i}}.
\end{aligned}$$

Controlling for confounding...

...in the analysis stage: stratification

- As we did for the single odds ratio, a confidence interval for OR_{MH} is derived using the variance of $\log \widehat{\text{OR}}_{\text{MH}}$.

- As Jewell (2003, p 132) highlights:

  "It took almost 30 years from the introduction of the estimator to: (1) establish that $\log \widehat{OR}_{MH}$ *has an approximately Normal sampling distribution and (2) find a formula to estimate its sampling variance."*

- The estimator proposed by Robins et al. (1986) works well if there are either a few strata with substantial data in each or lots of strata with only a few observations in each.

- The formula of the estimate proposed by these authors is as follows:

$$\widehat{\text{var}}(\log \widehat{\text{OR}}_{\text{MH}}) = \frac{\sum_{i=1}^{I} \left(\frac{a_i+d_i}{n_i}\right)\left(\frac{a_i d_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{a_i d_i}{n_i}\right)^2} + \frac{\sum_{i=1}^{I} \left(\frac{a_i+d_i}{n_i}\frac{b_i c_i}{n_i} + \frac{b_i+c_i}{n_i}\frac{a_i d_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{a_i d_i}{n_i}\right)\left(\sum_{i=1}^{I} \frac{b_i c_i}{n_i}\right)} + \frac{\sum_{i=1}^{I} \left(\frac{b_i+c_i}{n_i}\right)\left(\frac{b_i c_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{b_i c_i}{n_i}\right)^2}.$$

Controlling for confounding...

...in the analysis stage: stratification

↪ A $100(1-\alpha)\%$ confidence interval for $\log \mathrm{OR}$ based on the Mantel-Haenszel summary estimator, is given by

$$(\log \widehat{\mathrm{OR}}_{\mathrm{MH}} - z_\alpha \sqrt{\widehat{\mathrm{var}}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}})}, \log \widehat{\mathrm{OR}}_{\mathrm{MH}} + z_\alpha \sqrt{\widehat{\mathrm{var}}(\log \widehat{\mathrm{OR}}_{\mathrm{MH}})}),$$

where z_α is the $(1-\alpha/2)$th percentile of the standard normal distribution.

↪ The associated confidence interval for $\mathrm{OR}_{\mathrm{MH}}$ is obtained as usual by exponentiating the limits of the interval for $\log \mathrm{OR}_{\mathrm{MH}}$

Controlling for confounding...

...in the analysis stage: stratification

↪ The Mantel-Haenszel estimate of the relative risk takes the following form

$$\widehat{\text{RR}}_{\text{MH}} = \frac{\sum_{i=1}^{I} \omega_i \widehat{\text{RR}}_i}{\sum_{i=1}^{I} \omega_i},$$

where

$$\widehat{\text{RR}}_i = \frac{a_i/(a_i+b_i)}{c_i/(c_i+d_i)}, \quad \omega_i = \frac{c_i(a_i+b_i)}{n_i}.$$

↪ The adjusted Mantel-Haenszel estimate of the relative risk can then be written as

$$\widehat{\text{RR}}_{\text{MH}} = \frac{\sum_{i=1}^{I} \frac{a_i(c_i+d_i)}{n_i}}{\sum_{i=1}^{I} \frac{c_i(a_i+b_i)}{n_i}}.$$

Controlling for confounding...

...in the analysis stage: stratification

- To construct the associated confidence interval, we again work on the logarithmic scale.
- An estimate of the sampling variance of $\log \widehat{\mathrm{RR}}_{\mathrm{MH}}$ is given by

$$\widehat{\mathrm{var}}(\log \widehat{\mathrm{RR}}_{\mathrm{MH}}) = \frac{\sum_{i=1}^{I}[(a_i + b_i)(c_i + d_i)(a_i + c_i) - a_i c_i n_i]/n_i^2}{(\sum_{i=1}^{I} a_i(c_i + d_i)/n_i)(\sum_{i=1}^{I} c_i(a_i + b_i)/n_i)}.$$

- Confidence intervals for $\log \mathrm{RR}_{\mathrm{MH}}$ and, subsequently, for $\mathrm{RR}_{\mathrm{MH}}$ are then obtained as in the odds ratio case.

Controlling for confounding

↪ A simple and direct way to determine whether a risk factor caused confounding is to compare the estimated measure of association before and after adjusting for the confounding variable.

↪ If the difference between the two measures of association is 10% or more, in absolute value, most epidemiologists would consider that to be evidence of confounding. Specifically, if the following standardised difference is above the 10% threshold, then confounding is present

$$\frac{\widehat{\text{OR}}_{\text{unadjusted}} - \widehat{\text{OR}}_{\text{adjusted}}}{\widehat{\text{OR}}_{\text{unadjusted}}}, \qquad \frac{\widehat{\text{RR}}_{\text{unadjusted}} - \widehat{\text{RR}}_{\text{adjusted}}}{\widehat{\text{RR}}_{\text{unadjusted}}}.$$

↪ Other investigators will determine if the variable meets the three criteria listed before and if that is the case, the variable is regarded as a confounder.

Controlling for confounding

Example

- The following example is from Kirkwood and Sterne (2003, p 178). The tables in the next slides are also taken from these authors' book.
- The table below shows hypothetical results from a cohort study carried out to compare the prevalence of antibodies to leptospirosis in rural and urban areas of the West Indies. Rural residence is the exposure of interest.

	Leptospirosis antibodies			
Type of area	Yes	No	Total	Odds
Rural	60 (30%)	140 (70%)	200	0.429
Urban	60 (30%)	140 (70%)	200	0.429
Total	120	280	400	

- Since the number of individual with and without antibodies are identical in urban and rural areas, the odds ratio is exactly one, thus leading to the conclusion that there is no association between urban/rural residence and leptospirosis antibodies.

Controlling for confounding

Example

↪ However, when the same data is subdivided (*stratified*) according to sex, the risk of having antibodies is higher in rural areas for both males and females.

(a) Males.

	Antibodies			
Type of area	Yes	No	Total	Odds
Rural	36 (72%)	14 (28%)	50	2.57
Urban	50 (50%)	50 (50%)	100	1.00
Total	86	64	150	

$\text{OR} = 2.57/1 = 2.57$ (95% CI = 1.21 to 5.45), $P = 0.011$

(b) Females.

	Antibodies			
Type of area	Yes	No	Total	Odds
Rural	24 (16%)	126 (84%)	150	0.19
Urban	10 (10%)	90 (90%)	100	0.11
Total	34	216	250	

$\text{OR} = 0.19/0.11 = 1.71$ (95% CI = 0.778 to 3.78), $P = 0.176$

Controlling for confounding

Example

- The disappearance of the association when the data from males and females are combined is caused by a combination of two factors:
  - Females in both areas are much less likely than males to have antibodies (being male is indeed a known risk factor for leptospirosis).
  - The samples from the rural and urban areas have different gender compositions. The proportion of males is $100/200$ (50%) in the urban sample but only $50/200$ (25%) in the rural sample.
- Gender here is a confounding variable because it is related to both the outcome variable (presence of leptospirosis antibodies) and to the exposure groups being compared (rural and urban).

Controlling for confounding

Example

- ↪ Ignoring gender led to a bias in the results.
- ↪ Analysing males and females separately provided evidence of a difference between rural and urban areas for males but not for females.
- ↪ Note that the 95% CI for OR in males does not include the null value but the 95% CI for females does include the null value of no association. However, the two 95% CIs substantially overlap.

Controlling for confounding

Example

↪ Let us now calculate what is needed for the (sex-) adjusted Mantel-Haenszel estimate of the odds ratio.

Stratum i	$\widehat{\text{OR}}_i$	ω_i
Males ($i = 1$)	2.57	4.67
Females ($i = 2$)	1.71	5.04

↪ Therefore,

$$\widehat{\text{OR}}_{\text{MH}} = \frac{4.67 \times 2.57 + 5.04 \times 1.71}{4.67 + 5.04} = 2.13.$$

↪ Before controlling for gender, the odds ratio estimate was one. After controlling for the confounding effect of gender, the odds of leptospirosis antibodies are more than doubled in rural compared to urban areas.

Controlling for confounding...

...ain the analysis stage: stratification

- After doing the necessary calculations, we have that $\widehat{\text{var}}(\log \widehat{\text{OR}}_{\text{MH}}) = 0.0747284$.
- The corresponding 95% CI for OR_{MH} is

$$\left(\exp\{\log 2.13 - 1.96\sqrt{0.0747284}\}, \exp\{\log 2.13 + 1.96\sqrt{0.0747284}\}\right) = (1.246, 3.640).$$

- After adjusting for confounding by gender, the odds for leptospirosis antibodies is between 1.25 and 3.64 higher in people living in rural areas than in urban areas.
- Note that the crude odds ratio estimate we found earlier before adjusting for gender, is not even included in the 95% CI for the adjusted OR.
- The aforementioned rule of thumb $(1 - 2.13)/1$ is, in absolute value, way above the 10% threshold.

Effect modification

- One of the assumptions underpinning the validity of Mantel-Haenszel methods is that the exposure-disease association is actually consistent across strata and that the only reason for differences in the observed association measures between strata is sampling variation.

- If this is not true, then it makes little sense to combine the odds ratios/relative risks from the different strata. In this case one should report the $E - D$ association measure separately in each stratum.

- If the effect of E on D varies according to the level of C then we say that C modifies the effect of E on D. In other words, there is **effect modification**.

- Equivalently, it is also commonly said that there is **interaction** between the effects of E and C (on D).

- We will learn more about effect modification when learning about logistic regression.