Observational studies and covariate adjustments

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Introduction to Statistics

So far

- In previous lectures we considered a randomized trial of epinephrine and cardiac arrest

- By randomizing the treatment, we guarantee that there are no systematic differences between treated and untreated
  - any difference in the outcome between treated and untreated must be due to the treatment itself

Problems with randomized trials

- Randomization is ideal from a statistical perspective, but often not feasible in practice
- This is particularly the case when the factor that we wish to study - the **exposure** - is not a medical treatment
  - unethical if the exposure of interest is known or suspected to be harmful (e.g. alcohol, nicotine)
  - impractical if the exposure of interest is difficult to control by intervention (e.g. BMI)
  - expensive
- Even if we manage to randomize the exposure, the study participant may not comply with the assignment
- **Most of our medical/scientific knowledge comes from observational (non-randomized) studies**

Problems with observational studies

- In observational studies, there are often systematic differences between exposed and unexposed in important predictors for the outcome - **covariates**
  - e.g. exposed may be older/younger/healthier/unhealthier etc than unexposed
- As a consequence, we may observe a difference in the outcome between exposed and unexposed
  - even if the exposure itself has no effect on the outcome

Adjustment for covariates

- To avoid misleading results, one may attempt to 'adjust/control' for measured covariates in the analysis
  - stratification, regression modeling, matching, propensity scores etc
- But what covariates should we adjust for?
  - some adjustments may change the results in unexpected directions
  - some adjustments may be unnecessary or even harmful, depending on what research question we have in mind
- **To determine what to adjust for we need to know something about the underlying mechanisms**

Statistical uncertainty

- In previous lectures we have argued that all estimates are associated with statistical uncertainty
  - due to the fact that we only have a limited random sample, and not the whole population
- We have used standard errors, confidence intervals and p-values to determine the reliability in the estimates
- The problems that we will discuss in this lecture are of a different kind
  - they would persist even with data on the whole population
- For pedagogical purposes we ignore statistical uncertainty altogether
  - no standard errors, confidence intervals or p-values

Statistical software

- In previous lectures we have done all calculations by hand
  - not feasible for more complex analysis
- There are several statistical computer programs
  - R, Python, SAS, Stata, SPSS...
- We will illustrate all analyses with R
  - free at `https://cran.r-project.org/`
  - relatively standard programming syntax, e.g. similar to C and Java
  - extremely comprehensive
- LOTS of online help/tutorials: check the Documentation page at `https://cran.r-project.org/`

Outline

Outline

Study design

- Suppose we wish to estimate the effect of alcohol intake on the risk of breast cancer
- A randomized trial is not feasible!
- We take a random sample of $n = 5000$ women from the target population of interest
- At enrollment, we measure several potential risk factors for breast cancer, including alcohol intake
- We follow the women for 10 years, and measure all breast cancer diagnoses during follow-up
- This is an observational study

Data (fictitious!)

```
> head(bc)
    age    edu   alc  dens bc  time hosp
1 33.28   High 10.59 53.30  0 10.00    0
2 38.61   High 12.66 54.59  0 10.00    0
3 48.64   High 10.43 52.41  0  7.27    0
4 65.41 Medium  7.95 53.70  1  0.53    0
5 30.08    Low 14.99 54.60  0 10.00    1
6 64.92 Medium  8.31 51.68  0  4.11    0
```

- `age`: age at enrollment
- `edu`: education level (Low, Medium, High)
- `alc`: alcohol intake (g/day) - **the exposure of interest**
- `dens`: breast dense volume (cm^3)
- `bc`: breast cancer, yes (=1) or no (=0) during follow-up - **the outcome of interest**
- `time`: time to breast cancer or end of follow-up, whichever came first
- `hosp`: hospitalized during follow-up, yes (=1) or no (=0)

Dichotomization of alcohol intake

- For the purpose of this lecture we dichotomize alcohol intake as

```
> bc$alc.binary <- as.numeric(bc$alc>mean(bc$alc))
```

- `alc.binary` is equal to 1 if for women with high alcohol intake, and equal to 0 for women with low alcohol intake
  - high/low defined as above/below the mean (=12.51 g/day)

The risk ratio

- We will use risk ratios to measure the statistical association between alcohol intake and breast cancer
- $p_{\text{bc|high alc}}$ = risk of breast cancer among those with high alcohol intake
- $p_{\text{bc|low alc}}$ = risk of breast cancer among those with low alcohol intake

$$RR = \frac{p_{\text{bc|high alc}}}{p_{\text{bc|low alc}}}$$

- $RR > 1$: association
- $RR = 1$: no association
- $RR < 1$: inverse association

Data

```
> xtabs(formula=~alc.binary+bc, data=bc)
           bc
alc.binary    0    1
         0 2078  538
         1 2243  141
```

- *Calculate the risk ratio of breast cancer, comparing low and high alcohol intake*

Solution

```
           bc
alc.binary    0    1
         0 2078  538
         1 2243  141
```

$$p_{\text{bc|high alc}} = \frac{141}{2243 + 141} = 0.059$$

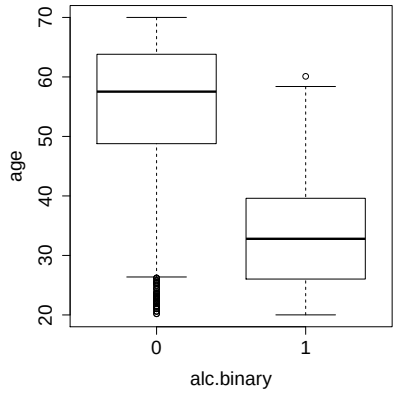
$$p_{\text{bc|low alc}} = \frac{538}{2078 + 538} = 0.201$$

$$RR = \frac{p_{\text{bc|high alc}}}{p_{\text{bc|low alc}}} = 0.28$$

- Breast cancer more common among women with low alcohol intake - an inverse association
- *What could possibly explain this?*

Outline

Difference in age



- The boxplots show a systematic difference in age between exposed and unexposed
  - women with low alcohol intake are generally older than women with high alcohol intake
- Perhaps this explains the inverse association between alcohol intake and breast cancer - let's try to adjust for age

Stratification

- The simplest way to adjust for age (or any other covariate) is by stratification
- We divide the sample into groups - strata - such that subjects have similar age within strata
  - within strata, the differences in age are reduced
- We analyze the strata separately

Dichotomization of age

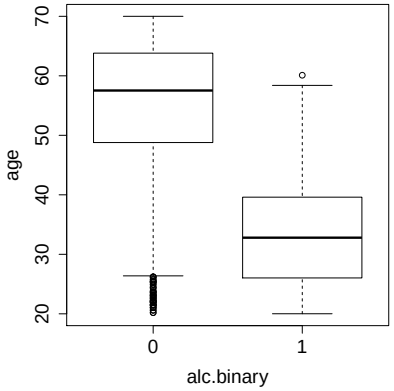
- For the purpose of stratification we dichotomize age as

```
> bc$age.binary <- as.numeric(bc$age>mean(bc$age))
```

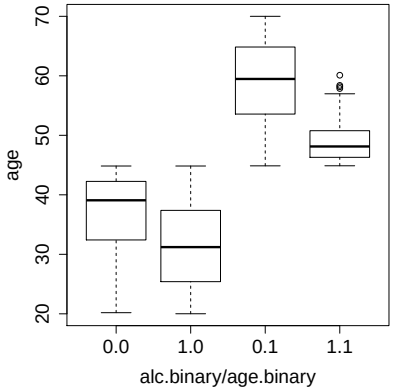
- `age.binary` is equal to 1 if for women with high age, and equal to 0 for women with low age
  - high/low defined as above/below the mean (=44.85 years)

Reduced difference in age by stratification

- Before stratification:



- After stratification:



Data stratified by age

```
> xtabs(formula=~alc.binary+bc+age.binary, data=bc)
, , age.binary = 0

           bc
alc.binary    0    1
         0  386   37
         1 2016  112

, , age.binary = 1

           bc
alc.binary    0    1
         0 1692  501
         1  227   29
```

- *Calculate the risk ratio for young and old separately*

Solution

```
, , age.binary = 0

           bc
alc.binary    0    1
         0  386   37
         1 2016  112
```

$$p_{\text{bc|high alc, young}} = \frac{112}{2016+112} = 0.053$$
$$p_{\text{bc|low alc, young}} = \frac{37}{386+37} = 0.087$$

$$RR_{\text{young}} = \frac{p_{\text{bc|high alc, young}}}{p_{\text{bc|low alc, young}}} = 0.60$$

Solution, cont'd

```
, , age.binary = 1

           bc
alc.binary    0    1
         0 1692  501
         1  227   29
```

$$p_{\text{bc|high alc, old}} = \frac{29}{227+29} = 0.113$$
$$p_{\text{bc|low alc, old}} = \frac{501}{1692+501} = 0.186$$

$$RR_{\text{old}} = \frac{p_{\text{bc|high alc, old}}}{p_{\text{bc|low alc, old}}} = 0.61$$

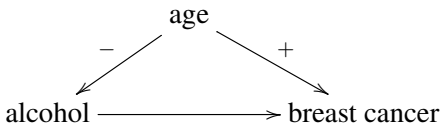
Unadjusted RR vs adjusted RR's

$$RR = 0.28$$
$$RR_{\text{young}} = 0.60$$
$$RR_{\text{old}} = 0.61$$

- When adjusting for age, the inverse association between alcohol intake and breast cancer becomes weaker
- What mechanism makes this happen? Should we adjust for age?

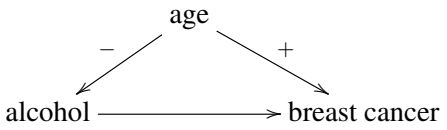
One possible explanation

- Suppose that
  - old women drink less alcohol than young women
  - old women have higher risk of breast cancer than young women



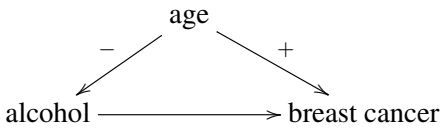
- If so, then those with a high alcohol intake tend to be young, and may therefore have a low risk of breast cancer
  - an inverse non-causal association between alcohol intake and breast cancer

The unadjusted RR



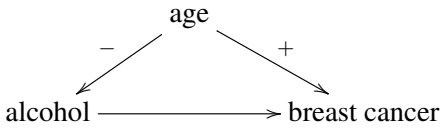
- The unadjusted RR combines
  - the causal effect of alcohol
  - the inverse non-causal association due to the influence of age
- If the influence of age is strong enough, then this could explain why the unadjusted RR is below 1

The adjusted RR's



- Within strata, subjects have similar age
- Thus, when stratifying on age the inverse non-causal association due to the age is partly removed
- This could explain why the adjusted RR's are bigger than the unadjusted RR

Confounding



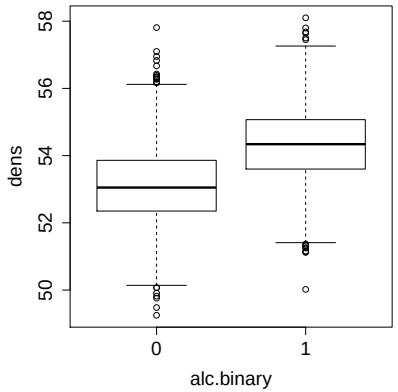
- We say that there is **confounding** when the exposure and the outcome have common causes
- The common causes are called **confounders**

Confounder adjustment

- Failing to adjust for a confounder induces non-causal associations between the exposure and the outcome
  - the causal effect is overestimated or underestimated
- **We should always adjust for potential confounders**

Outline

Difference in breast density



- The boxplot shows a systematic difference in breast density between exposed and unexposed
  - women with low alcohol intake generally have a lower breast density than women with high alcohol intake
- Perhaps we should adjust for breast density as well?

Dichotomization of breast dense volume

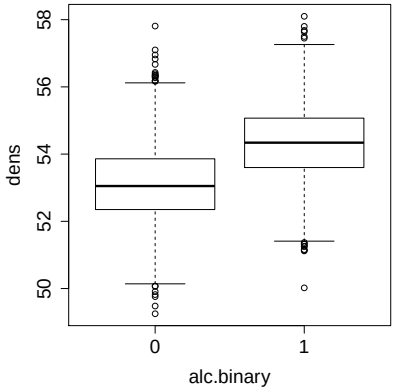
- For the purpose of stratification we dichotomize breast dense volume as

```
> bc$dens.binary <-
  as.numeric(bc$dens>mean(bc$dens))
```

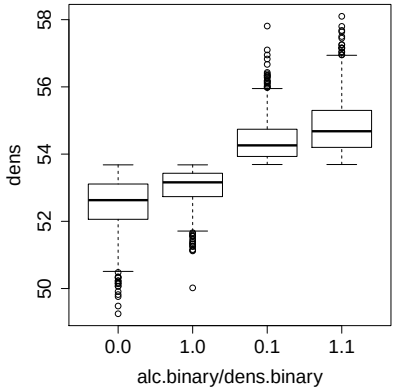
- `dens.binary` is equal to 1 if for women with high breast dense volume, and equal to 0 for women with low breast dense volume
  - high/low defined as above/below the mean (=53.69 cm^3)

Reduced difference in breast density by stratification

- Before stratification:



- After stratification:



Data stratified by breast density

```
> xtabs(formula=~alc.binary+bc+dens.binary,
  data=bc)
, , dens.binary = 0

           bc
alc.binary    0    1
         0 1461  352
         1  620   27

, , dens.binary = 1

           bc
alc.binary    0    1
         0  617  186
         1 1623  114
```

- *Calculate the risk ratio for those with high and low breast density separately*

Solution

```
, , dens.binary = 0

           bc
alc.binary    0    1
         0 1461  352
         1  620   27
```

$$p_{\text{bc|high alc, low dens}} = \frac{27}{620+27} = 0.042$$

$$p_{\text{bc|low alc, low dens}} = \frac{352}{1461+352} = 0.194$$

$$RR_{\text{low dens}} = \frac{p_{\text{bc|high alc, low dens}}}{p_{\text{bc|low alc, low dens}}} = 0.21$$

Solution, cont'd

```
, , dens.binary = 1

           bc
alc.binary    0    1
         0  617  186
         1 1623  114
```

$$p_{\text{bc|high alc, high dens}} = \frac{114}{1623+114} = 0.066$$

$$p_{\text{bc|low alc, high dens}} = \frac{186}{617+186} = 0.232$$

$$RR_{\text{high dens}} = \frac{p_{\text{bc|high alc, high dens}}}{p_{\text{bc|low alc, high dens}}} = 0.28$$

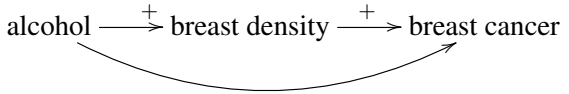
Unadjusted RR vs adjusted RR's

$$\begin{aligned} RR &= 0.28 \\ RR_{\text{low dens}} &= 0.21 \\ RR_{\text{high dens}} &= 0.28 \end{aligned}$$

- When adjusting for breast density, the inverse association between alcohol intake and breast cancer becomes stronger (at least in one stratum)
- What mechanism makes this happen? Should we adjust for breast density?

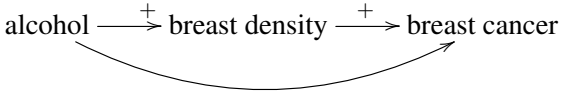
One possible explanation

- Suppose that
  - alcohol increases breast density
  - high breast density increases breast cancer risk



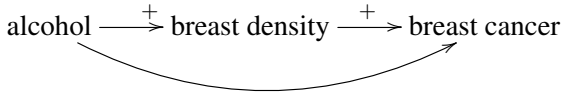
- If so, then those with a high alcohol intake tend to get high breast density, and may therefore have a high risk of breast cancer
  - a positive causal effect of alcohol intake on breast cancer

The unadjusted RR



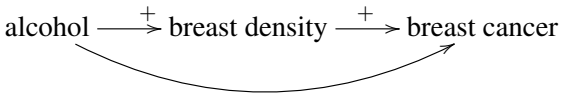
- The unadjusted RR combines
  - the direct effect of alcohol
  - the effect mediated through breast density

The adjusted RR's



- Within strata, subjects have similar breast density
- Thus, when stratifying on breast density the positive effect mediated through breast density is partly removed
- This could explain why one of the adjusted RR's is smaller than the unadjusted RR

Mediation



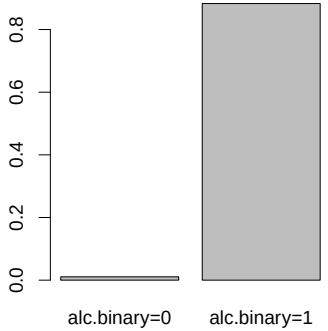
- A variable on the causal pathway between exposure and outcome is called a **mediator**
- Typically, an exposure effect is mediated through numerous mediators (biological, physiological, chemical ...)
- When we talk about the 'direct' exposure effect, it is always relative to a specific (set of) mediator(s)
  - e.g. 'direct' = 'not through breast density'

Mediator adjustment

- Adjusting for a mediator reduces/eliminates the mediated effect
- **Whether we should adjust for potential mediators or not depends on the research question**
  - if we are interested in the total (direct + mediated) effect, then we should not adjust for the mediator
  - if we are interested in the direct effect, then we should adjust for the mediator

Outline

Difference in hospitalization



- The barplot shows a systematic difference in hospitalization between exposed and unexposed
  - women with low alcohol intake generally have a lower risk of being hospitalized than women with high alcohol intake
- Perhaps we should adjust for hospitalization as well?

Data stratified by hospitalization

```
> xtabs(formula=~alc.binary+bc+hosp, data=bc)
, , hosp = 0

           bc
alc.binary    0    1
         0 2069  519
         1  274    5

, , hosp = 1

           bc
alc.binary    0    1
         0    9   19
         1 1969  136
```

- *Calculate the risk ratio for those who are hospitalized and those who are not hospitalized separately*

Solution

```
, , hosp = 0

           bc
alc.binary    0    1
         0 2069  519
         1  274    5
```

$$p_{\text{bc|high alc, not hosp}} = \frac{5}{274+5} = 0.018$$
$$p_{\text{bc|low alc, not hosp}} = \frac{519}{2069+519} = 0.200$$

$$RR_{\text{not hosp}} = \frac{p_{\text{bc|high alc, not hosp}}}{p_{\text{bc|low alc, not hosp}}} = 0.09$$

Solution, cont'd

```
, , hosp = 1

           bc
alc.binary    0    1
         0    9   19
         1 1969  136
```

$$p_{\text{bc|high alc, hosp}} = \frac{136}{1969+136} = 0.065$$
$$p_{\text{bc|low alc, hosp}} = \frac{19}{9+19} = 0.679$$

$$RR_{\text{hosp}} = \frac{p_{\text{bc|high alc, hosp}}}{p_{\text{bc|low alc, hosp}}} = 0.10$$

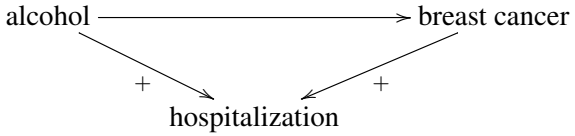
Unadjusted RR vs adjusted RR's

$$RR = 0.28$$
$$RR_{\text{not hosp}} = 0.09$$
$$RR_{\text{hosp}} = 0.10$$

- When stratifying on hospitalization, the inverse association between alcohol intake and breast cancer becomes stronger
- What mechanism makes this happen? Should we adjust for hospitalization?

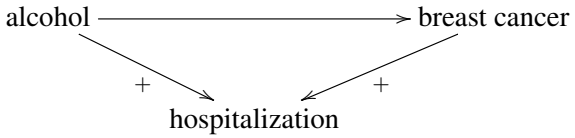
One possible explanation

- Suppose that
  - high alcohol intake increases the risk of hospitalization
  - breast cancer increases the risk of hospitalization



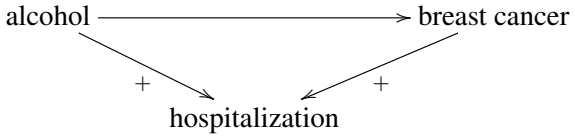
- Consider the stratum of hospitalized: why were these women hospitalized?
  - high alcohol intake?
  - breast cancer?

One possible explanation, cont'd



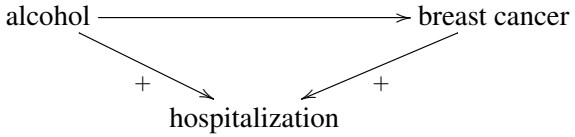
- For those with low alcohol intake, alcohol is not likely the cause
  - perhaps breast cancer then?
- Once alcohol has been ruled out as the cause of hospitalization, breast cancer becomes more likely
  - an inverse non-causal association between alcohol and breast cancer

The adjusted RR's and the unadjusted RR



- The adjusted RR's combine
  - the causal effect of alcohol
  - the inverse non-causal association due to stratification on hospitalization
- In contrast, the unadjusted RR only depends on the causal effect of alcohol
  - (and possibly confounding by age)
- This could explain why the adjusted RR's are smaller than the unadjusted RR

Colliding



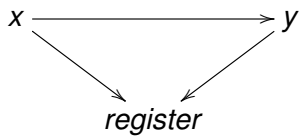
- We say that there is **colliding** when the exposure and the outcome have common effects
- The common effects are called **colliders**

Collider adjustment

- Adjusting for a collider induces non-causal associations between the exposure and the outcome
  - the causal effect is overestimated or underestimated
- **We should never adjust for potential colliders**

Selection bias

- Sometimes, collider adjustment is inevitable
- In register based research, both exposure and outcome may often affect whether the subject is in the register



- Restricting the analysis to those in the register (stratifying on the register) may induce non-causal associations
  - **selection bias**

Outline

Example

- Suppose we are given a data set containing
  - an outcome y
  - an exposure x
  - an additional covariate z
- Suppose we don't know anything about what these variables represent

Example, cont'd

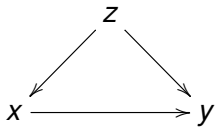
	$z=0$		$z=1$	
	$y=0$	$y=1$	$y=0$	$y=1$
$x=0$	320	80	180	120
$x=1$	60	40	40	160

$$
\begin{aligned}
RR &= 2.33 \\
RR_{z=0} &= 2 \\
RR_{z=1} &= 2
\end{aligned}
$$

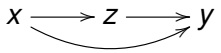
- When adjusting for z, the association between x and y becomes weaker
- *Should we adjust for z?*

Three possible explanations

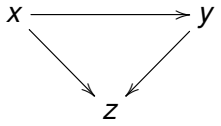
- z is a confounder, should adjust:



- z is a mediator, should maybe adjust, depending on the research question:



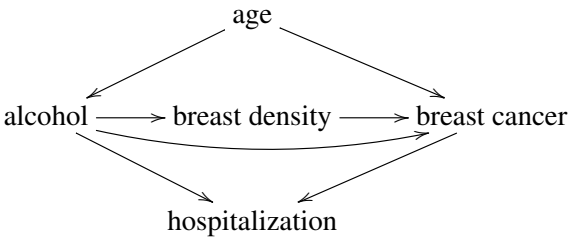
- z is a collider, should not adjust:



The need for subject matter knowledge

- There is no way that data can help us to distinguish between confounders, mediators and colliders
- We must rely on subject matter knowledge about the problem at hand
- For instance, it is unlikely that breast density affects alcohol intake, but the reverse may very well be true

Summary



- Always adjust for confounders
- Adjust for mediators if estimating direct effects, but not if estimating total effects
- Never adjust for colliders
- Use subject matter knowledge to distinguish between these types of variables