Cluster designs to adjust for unmeasured confounding

Arvid Sjölander

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

SIMSAM

Outline

Outline

The causal research question

- Most epidemiological/sociological research questions are centered around an **exposure** and an **outcome**
- Typically, we would like to know if the exposure has a causal effect on the outcome
  - e.g does smoking cause lung cancer?

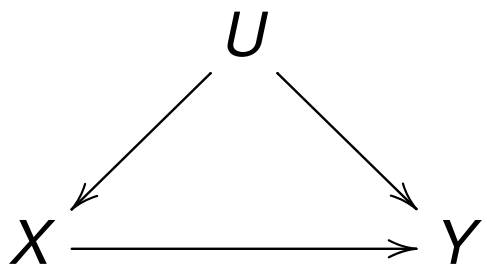
$$X \xrightarrow[?]{} Y$$

Randomized trials

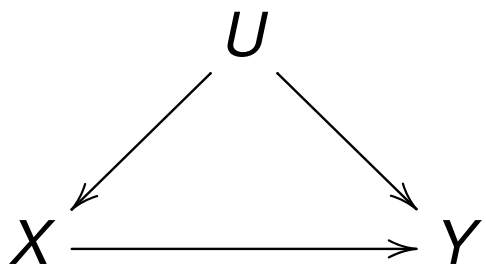
- Ideally, we can **randomize** the exposure
- Randomization eliminates systematic differences between exposed and unexposed
- The exposure-outcome association can be given a causal interpretation
- However, randomization is often infeasible due to ethical, practical and financial constraints

Observational studies

- In observational studies, **confounding** typically leads to systematic differences between exposed and unexposed
  - for instance, suppose there is a carcinogenic genotype that also leads to a craving for nicotine
  - then, smokers would have this genotype more often than non-smokers and for this reason get lung cancer more often



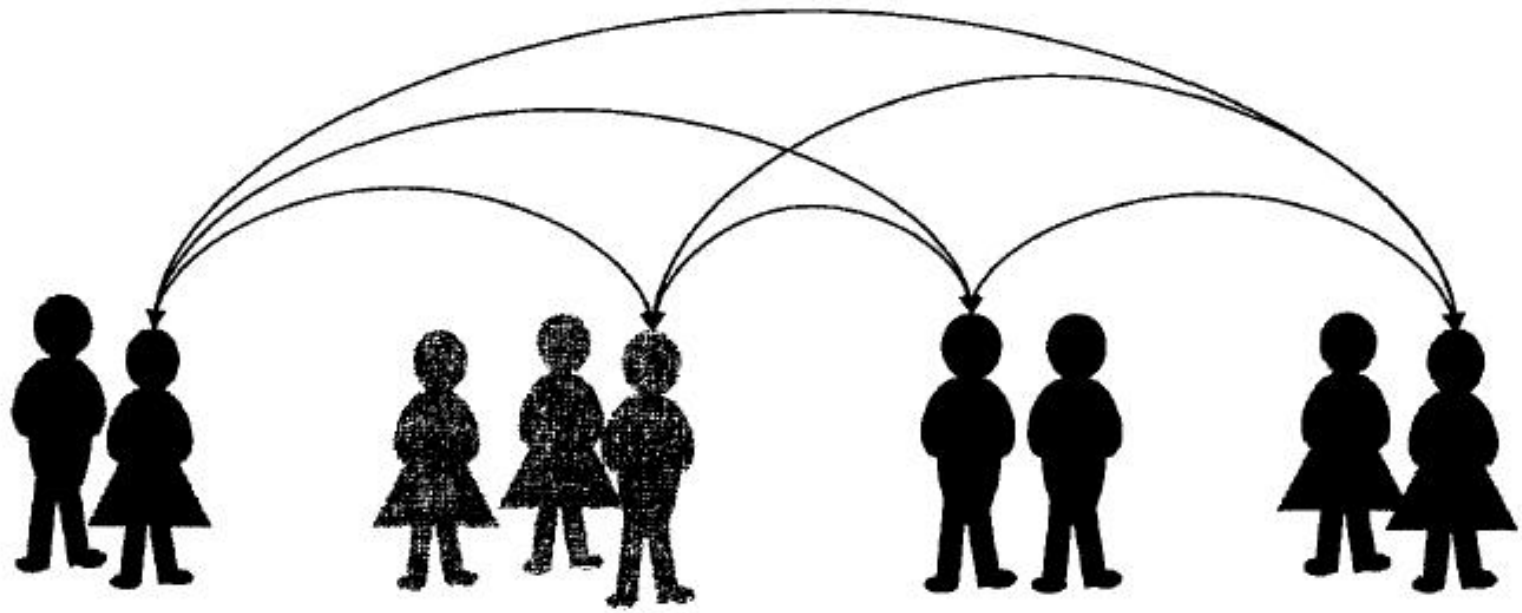
Confounder adjustment



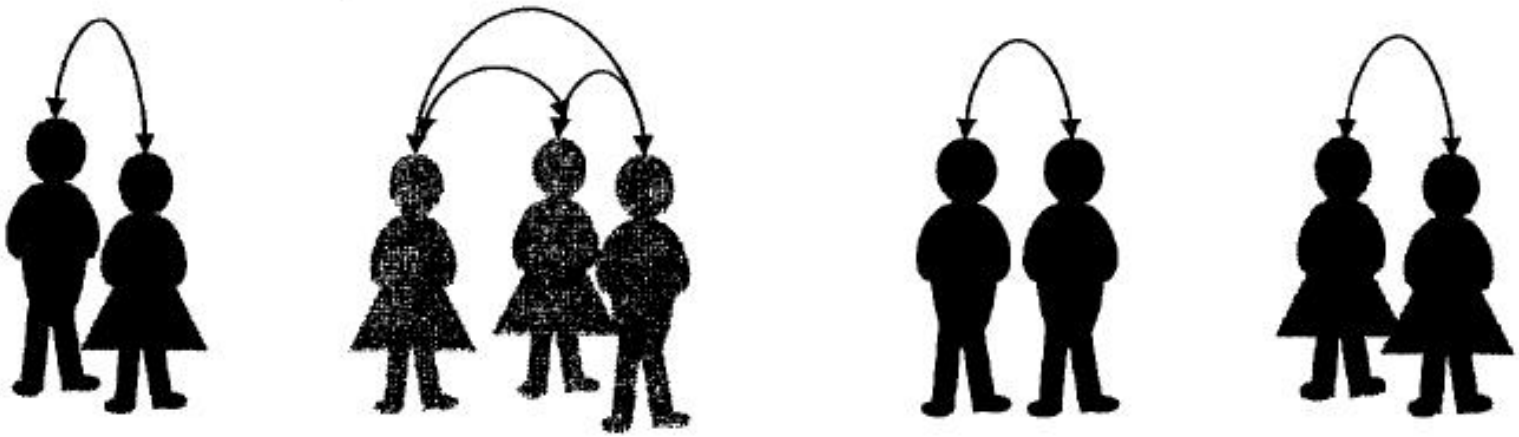
- Observed confounders can be adjusted for, e.g. by stratification or through regression modelling
- But many confounders may be difficult to measure, or unknown to the investigator
  - e.g. genetic factors and life-style factors

Cluster design

- Suppose that we can identify clusters, in which some confounders are (approximately) constant, e.g
  - neighbourhoods; socioeconomic status
  - school classes; teacher
  - MZ twin pairs; the whole genome
- By studying the exposure-outcome association **within** clusters rather than **between** unrelated subjects, we eliminate confounding by cluster-constant confounders



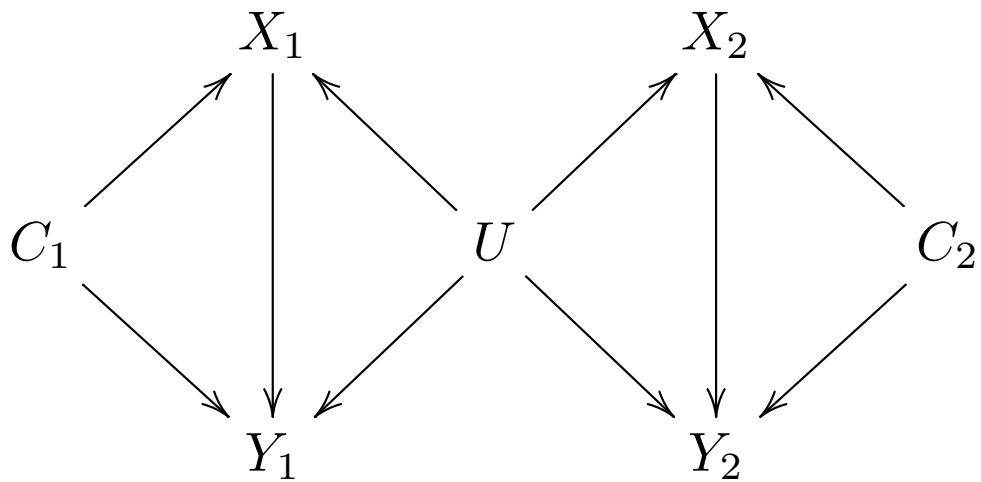
Between-family analyses



Sibling-comparison analyses

The aim of within-cluster analysis

- X = exposure, Y = outcome, C = cluster-varying confounders, U = cluster-constant confounders



- The aim of within-cluster analysis is to adjust for U

Generality of the method

- The cluster design is a very general method that is used in various fields...
  - epidemiology
  - biostatistics
  - sociology
  - econometrics

  ... and for various types of data
  - continuous
  - binary
  - survival
  - repeated measures

Family design

- The family design is a special case of the cluster design, where the cluster is a set a genetically related individuals
  - e.g. cousins, siblings or twins
- In epidemiology, families is the most common cluster type
- In sociology and econometrics, other cluster types are commonly used as well
  - e.g. school classes or neighbourhoods

Outline

Motivating example

- Maternal smoking during pregnancy (SDP) is associated with several psychiatric outcomes in the offspring
  - low academic achievement, low cognitive ability (CA), criminal behavior, drug abuse etc
- However, it can been questioned whether these associations can be given causal interpretations

Data

- To study the association between SDP and CA we use a dataset borrowed from Kuja-Halkola et al. (2014)
- This dataset comprises 207175 males from 185336 families, born in Sweden between 1983 and 1992
- The exposure, SDP, is coded as 1 (= yes) or 0 (= no)
- The outcome, cognitive ability, is measured on a 9-grade scale and obtained from the Swedish Conscription Registry
  - higher values are better
- Additional measured covariates are: maternal age at childbirth and birthyear
  - these are family-varying

First rows in data frame

```
> CA[1:10,]

   famid cascore sdp matage byear
1     64       5   1     30     0
2    157       6   0     34     1
3    239       5   1     25     5
4    254       4   1     21     9
5    259       5   0     30     7
6    266       4   0     29     5
7    385       5   1     23     3
8    385       6   1     25     5
9    440       9   1     39     3
10   520       4   0     23     0
```

- `byear` is coded in years since 1983

Unadjusted linear regression

$$E(cascore|sdp) = \alpha + \beta sdp$$

```
>fit <- lm(formula=cascore~sdp,data=CA)
>summary(fit)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.287678   0.004751 1113.01   <2e-16 ***
sdp         -0.633506   0.009217  -68.73   <2e-16 ***
```

- The mean cascore is 0.63 units lower for children to smokers, than for children to non-smokers
  - $p < 2 \times 10^{-16}$; the difference is highly significant

Linear regression vs linear GEE

- The linear regression assumes
  - that individuals are independent
  - normality and homoscedasticity (constant variance) of the cascore
- Easy to fix by using a linear generalized estimating equation (GEE) model instead
- The mean model is the same as in the linear regression, but
  - the standard errors are corrected for familial clustering
  - all distributional assumptions are avoided, apart from the mean model itself

The `drgee` package

- There are several packages in `R` for fitting GEE models
- We will use the package `drgee` (Zetterqvist and Sjölander, 2015a,b)
- Several advantages compared to other packages
  - more standard user interface
  - at least as fast (often faster)
  - several facilities for 'doubly robust estimation' (beyond the scope of this seminar)
  - **performs within-cluster analysis with conditional GEEs**

Unadjusted linear GEE

$$E(cascore|sdp) = \alpha + \beta sdp$$

```
>library(drgee)
>fit <- gee(formula=cascore~sdp,data=CA,
   clusterid="famid")
>summary(fit)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.287678   0.004979 1061.93   <2e-16 ***
sdp         -0.633506   0.009317  -67.99   <2e-16 ***
```

- Same estimates as in the linear regression, but slightly larger standard errors

Adjusted linear GEE

$$E(cascore|sdp, matage, byear) = \alpha + \beta sdp + \gamma_1 matage + \gamma_2 byear$$

```
>fit <- gee(formula=cascore~sdp+matage+byear,data=CA,
   clusterid="famid")
>summary(fit)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.017742	0.024808	161.96	<2e-16	***
sdp	-0.584265	0.009343	-62.54	<2e-16	***
matage	0.047065	0.000843	55.83	<2e-16	***
byear	-0.016562	0.001542	-10.74	<2e-16	***

- The mean cascore is 0.58 units lower for children to smokers, than for children to non-smokers
  - $p < 2 \times 10^{-16}$; the difference is highly significant

Limitations of the GEE model

- The standard GEE model can only adjust for measured confounders
- However, there is still a huge potential for unmeasured confounding
- In particular by family-constant confounders such as
  - parental socioeconomic status
  - parental education level
  - parental genetic factors
- *Is there anything clever we can do about the unmeasured family-constant confounders?*

A linear CGEE model

- Let *i* denote 'family'
- A linear CGEE model ('C' for 'conditional') is a GEE model that conditions on the family

$$E(cascore|i, sdp) = \alpha_i + \beta sdp$$

- The 'family-effect' is modeled by the family-specific intercept α_i

The GEE model vs the CGEE model

- The GEE model

$$E(cascore|sdp) = \alpha + \beta sdp$$

  assesses the exposure-outcome association **between** unrelated individuals

- The CGEE model

$$E(cascore|i, sdp) = \alpha_i + \beta sdp$$

  assesses the exposure-outcome association **within** families

The role of the family-specific intercept

$$E(cascore|i, sdp) = \alpha_i + \beta sdp$$

- By conditioning on the family, the CGEE model adjusts the exposure coefficient β for all family-constant confounders, e.g.
  - parental socioeconomic status
  - parental education level
  - parental genetic factors
- Conceptually same as including all the family-constant confounders in a vector X

$$E(cascore|X, sdp) = \underbrace{\gamma X}_{\alpha_i} + \beta sdp$$

- We can think of α_i as a term γX that represents the family-constant confounders

Unadjusted linear CGEE

$$E(cascore|i, sdp) = \alpha_i + \beta sdp$$

```
> fit <- gee(formula=cascore~sdp,data=CA,
   clusterid="famid",cond=TRUE)
> summary(fit)

        Estimate Std. Error z value Pr(>|z|)
sdp      0.11434    0.04595   2.488   0.0128 *
```

- The mean cascore is 0.11 units higher for children to smokers, than for children to non-smokers
  - $p = 0.012$; the difference is significant
- No estimates of the α_i's

Conclusion so far

- **Unadjusted linear GEE**: the mean cascore is 0.63 units lower for children to smokers, than for children to non-smokers
  - $p < 2 \times 10^{-16}$; the difference is highly significant
- **Adjusted linear GEE**: the mean cascore is 0.58 units lower for children to smokers, than for children to non-smokers
  - $p < 2 \times 10^{-16}$; the difference is highly significant
- **Unadjusted linear CGEE**: the mean cascore is 0.11 units higher for children to smokers, than for children to non-smokers
  - $p = 0.012$; the difference is significant
- *What if we adjust for both unmeasured family-constant confounders and measured family-varying confounders?*

Adjusted linear CGEE

$$E(cascore|i, sdp, matage, byear) = \alpha_i + \beta sdp + \gamma_1 matage + \gamma_2 byear$$

```
> fit <- gee(formula=cascore~sdp+matage+byear,
   data=CA,clusterid="famid",cond=TRUE)
> summary(fit)

          Estimate Std. Error z value Pr(>|z|)
sdp       0.007273   0.045596   0.160  0.87327
matage -0.074017   0.023758  -3.115  0.00184 **
byear  -0.029434   0.023884  -1.232  0.21781
```

- The association almost completely disappears!

A cautionary note on the interpretation of the results

- It is tempting to interpret the null finding as 'no causal effect of SDP on CA'
- But some caution is warranted; other explanations are
  - unmeasured family-varying confounding
  - sampling variability
- However, these may not realistically 'explain away' a large causal effect
  - unmeasured family-varying confounding would have to almost perfectly balance with a causal effect
  - the sampling variability is small; the 95% CI for β is $(-0.08, 0.10)$, which is rather narrow

Outline

Fixed effects regression models

- In our case study we used a conditional (on the family) linear model, which is suitable for continuous outcomes
- Analog models exist for other types of outcomes
  - conditional logistic regression (binary outcomes)
  - stratified Cox regression (time-to-event outcomes)
- All these models are referred to as **fixed effects regression models**

Fixed effects vs random effects models

- The term ‘fixed effect’ means that the cluster-specific intercepts are treated as fixed constants
- In contrast, random/mixed/frailty models treat the intercept as a random variable
- **Beware! Random/mixed/frailty models do not adjust for cluster-constant confounding**
  - the intercept and the exposure are assumed to be independent, thus excluding cluster-constant confounding a priori

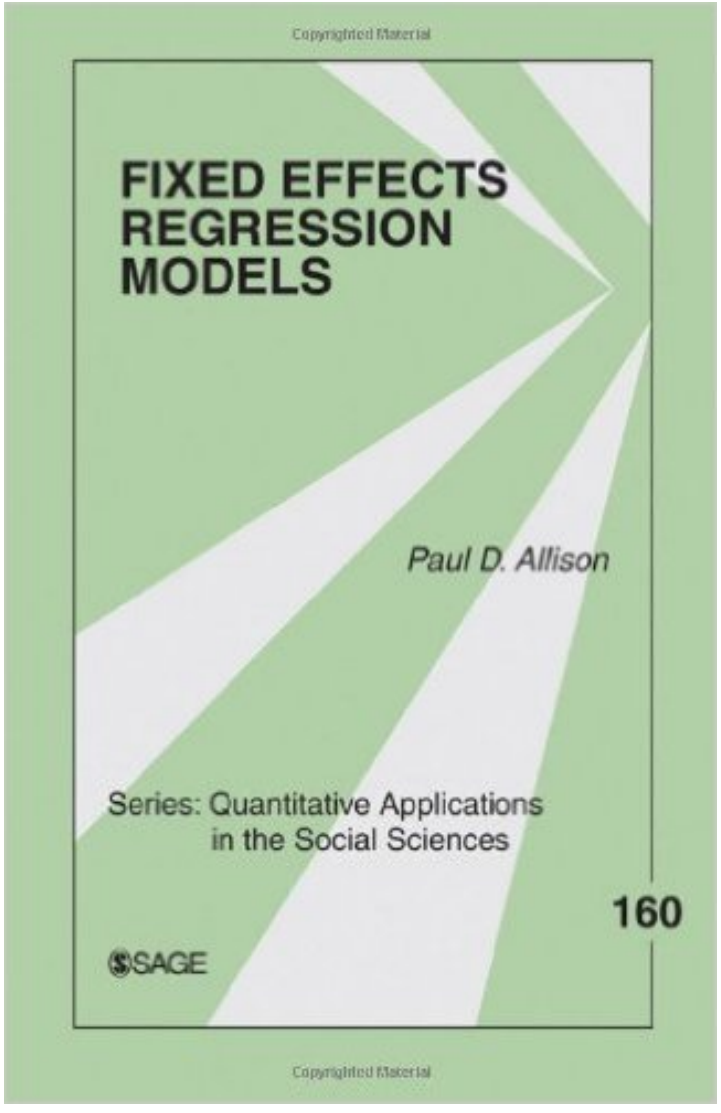
Using the individual as his/her own control

- In our case study we compared each individual to his/her own siblings
  - eliminates confounding by all factors that are constant in the family
- If both the exposure and the outcome are time-varying, then we may compare the individual to his/herself
  - i.e. compare exposed and unexposed time-periods within the same individual
  - eliminates confounding by all factors that are constant in the individual, e.g. genes
- Many variations on the theme
  - case-crossover design
  - self-controlled case series
  - case-time-control design

Conclusions

- The cluster design is a powerful tool to adjust for unmeasured confounding
  - implicit adjustment for all measured and unmeasured confounders that are constant within the cluster
- An important special case is the family design, which is made feasible by the large Swedish population registers
- Another is the case-crossover design, where is 'cluster' is a set of repeated measures on the same individual

Recommended reading



References

Kuja-Halkola, R., D'Onofrio, B., Larsson, H., and Lichtenstein, P. (2014). Maternal smoking during pregnancy and adverse outcomes in offspring: Genetic and environmental sources of covariance. *Behavior Genetics* **44,** 456–467.

Zetterqvist, J. and Sjölander, A. (2015a). Doubly robust estimation with the r package drgee. *Epidemiologic Methods* doi: 10.1515/em-2014-0021.

Zetterqvist, J. and Sjölander, A. (2015b). *drgee: doubly robust generalized estimating equations*. version 1.1.3.