Confounding, mediation and colliding

What types of shared covariates does the sibling comparison design control for?

Arvid Sjölander and Johan Zetterqvist

Causal effects and confounding

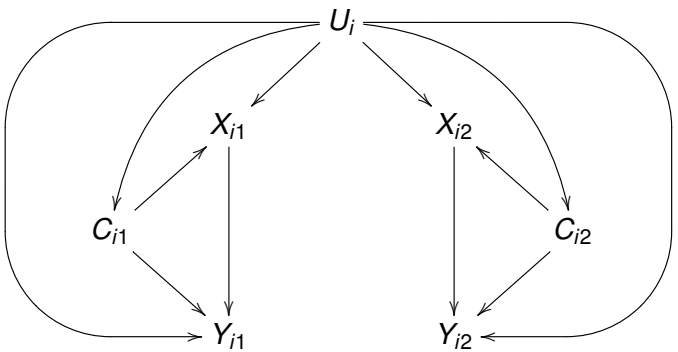
- A common aim of epidemiologic research is to estimate the causal effect of an exposure on an outcome
- In observational studies, confounding is always a concern

Sibling comparison designs

- A popular way to reduce confounding is to use a sibling comparison design
  - the exposure-outcome association is studied within families instead of between unrelated individuals
- It is routinly argued that the within-sibling association is controlled for all measured **and unmeasured** covariates that are shared (constant) within families
  - e.g. socioeconomic status and parental genetic makeup
- **But the sibling comparison design has subtle issues, that are not present in studies of unrelated subjects**

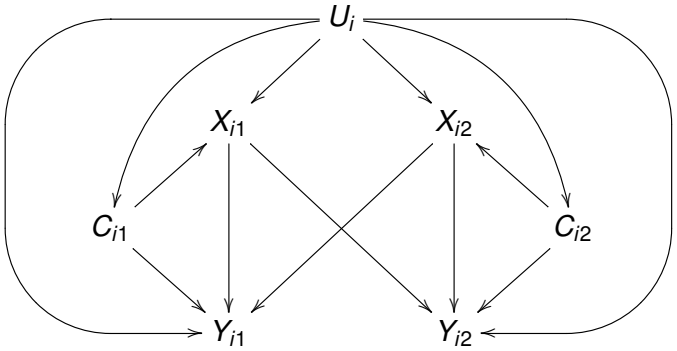
Amplification of bias due to unmeasured non-shared confounders

Frisell, T., Öberg, S., Kuja-Halkola, R., Sjölander, A. (2012). Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology*, **23**(5), 713-720.



Bias due to carryover effects

Sjölander, A., Frisell, T., Kuja-Halkola, R., Öberg, S., Zetterqvist, J. (2016). Carryover effects in sibling comparison designs. *Epidemiology*, **27**(6), 852-858.



Outline

Outline

Association between mother's age and ADHD in offspring

- Attention deficit/hyperactivity disorder (ADHD) is the most common neurodevelopmental disorder in childhood
  - prevalence $\sim$ 5%
- Studies have shown that young mothers are more likely to get children with ADHD than old mothers
- But not clear if the association is causal
  - possibly confounded by familial factors, e.g. socioeconomic status and genetics

A sibling comparison study

- To examine the role of familial confounding, Chang et al. (2014) carried out a sibling comparison study
- A cohort of 1495543 children born to 896389 mothers, between 1988 and 2003
  - 1.7 children per mother, on average
- Each child was followed from birth to ADHD diagnosis or 2009.12.31, whichever came first

Analyses

- Standard analysis: an ordinary Cox proportional hazards model
- Sibling comparison: a stratified Cox model, with one stratum per family
  - automatically controls for covariates that are shared (constant) in the family
  - model details later

Results

Table 2. Association between maternal age at each birth (MAEB) and offspring ADHD (hazard ratios with 95% confidence intervals)

MAEB	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
Binary[e]	2.24 (2.12–2.36)	1.57 (1.48–1.67)	0.90 (0.84–0.96)	0.81 (0.71–0.94)
Continuous	1.06 (1.05–1.06)	1.05 (1.04–1.05)	0.99 (0.99–1.00)	0.98 (0.97–0.99)

[a]Population-wide association, adjusted for offspring's sex, birth order and birth year in categories.
[b]In addition to Model 1, adjusted for paternal age at childbirth in categories.
[c]In addition to Model 2, adjusted for MAFB.
[d]Sibling-comparison, adjusted for unmeasured genetic and environmental factors shared by siblings and measured covariates.
[e]MAEB < 20 y.

- Statistically significant association between mother's age and offspring ADHD in the standard analysis
- No (or inverse) association in the sibling comparison
- Authors concluded:
  'The association between early maternal age and offspring ADHD is mainly explained by genetic confounding'

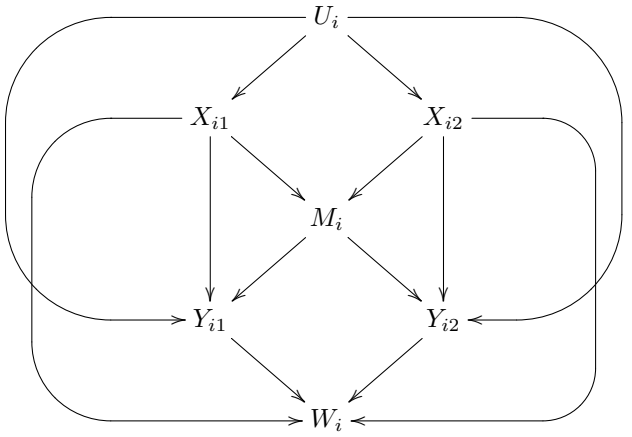
How about shared mediators?

- Familial environment (shared) is affected by the mother's age at all her childbirths
  - e.g. having multiple children at an early age may lead to a financially difficult family situation.
- If the familial environment in turn affects the risk of ADHD, then this covariate is a shared mediator
- Controlling for this covariate may therefore remove part of the true causal effect and attenuate the association

How about shared colliders?

- Familial environment is a dynamic condition which may change according to the development of the children
  - e.g. an ADHD diagnosis in one of the siblings may lead to a stressful situation for the whole family
- If so, then familial environment could also play the role of a shared collider
- Controlling for this covariate may induce 'collider-stratification bias', which may also attenuate the association

Confounding, mediation and colliding



- **What types of shared covariates does the sibling comparison design really control for?**

Outline

Outline

Statistical models that condition on the family

- Linear model for continuous outcome:

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

- Logistic model for binary outcome:

$$\text{logit}\{p(Y_{ij} = 1|\text{family } i, X_{ij})\} = \alpha_i + \beta X_{ij}$$

- Cox proportional hazards model for time-to-event outcome:

$$\log\{\lambda(y_{ij}|\text{family } i, X_{ij})\} = \alpha_i(y_{ij}) + \beta X_{ij}$$

- We will focus on the linear model, but all arguments and results hold for the logistic model and the Cox proportional hazards model as well

The role of the intercept

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

- The family-specific intercept α_i is intended to absorb those covariates that are shared in the family
- The exposure 'effect', β, is controlled for the covariates absorbed into α_i
- **Does α_i only absorb shared confounders, or shared mediators and colliders as well?**

In other words

- When fitting the conditional (on family) model

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij},$$

which of the following β's is really estimated?

$$\begin{aligned}
E(Y_{ij}|U_i, X_{ij}) &= \alpha_i^1 + \beta^1 X_{ij} \\
E(Y_{ij}|M_i, X_{ij}) &= \alpha_i^2 + \beta^2 X_{ij} \\
E(Y_{ij}|W_i, X_{ij}) &= \alpha_i^3 + \beta^3 X_{ij} \\
E(Y_{ij}|U_i, M_i, X_{ij}) &= \alpha_i^4 + \beta^4 X_{ij} \\
E(Y_{ij}|U_i, W_i, X_{ij}) &= \alpha_i^5 + \beta^5 X_{ij} \\
E(Y_{ij}|M_i, W_i, X_{ij}) &= \alpha_i^6 + \beta^6 X_{ij} \\
E(Y_{ij}|U_i, M_i, W_i, X_{ij}) &= \alpha_i^7 + \beta^7 X_{ij}
\end{aligned}$$

Technical note

- If $(U_i, X_{i1}, X_{i2}, M_i, Y_{i1}, Y_{i2}, W_i)$ have a multivariate normal distribution, then all the models below are correct

$$\begin{aligned}
E(Y_{ij}|U_i, X_{ij}) &= \alpha_i^1 + \beta^1 X_{ij} \\
E(Y_{ij}|M_i, X_{ij}) &= \alpha_i^2 + \beta^2 X_{ij} \\
E(Y_{ij}|W_i, X_{ij}) &= \alpha_i^3 + \beta^3 X_{ij} \\
E(Y_{ij}|U_i, M_i, X_{ij}) &= \alpha_i^4 + \beta^4 X_{ij} \\
E(Y_{ij}|U_i, W_i, X_{ij}) &= \alpha_i^5 + \beta^5 X_{ij} \\
E(Y_{ij}|M_i, W_i, X_{ij}) &= \alpha_i^6 + \beta^6 X_{ij} \\
E(Y_{ij}|U_i, M_i, W_i, X_{ij}) &= \alpha_i^7 + \beta^7 X_{ij}
\end{aligned}$$

- So not a problem of model misspecification per se

Conditional maximum likelihood

- The answer lies in the way the model is fitted, rather than in the model itself
- The standard way to fit conditional (on family) models is to use conditional maximum likelihood (CML)
  - leads to conditional logistic regression (for the logistic model) and stratified Cox regression (for the Cox proportional hazards model)
- Eliminates the family-specific intercept α_i by conditioning on a sufficient statistic

Underlying assumptions in CML

- Let $\mathbf{Z}_i$ be the set of shared covariates that we control for by conditioning on 'family i'
- We may thus reformulate the model

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

as

$$E(Y_{ij}|\mathbf{Z}_i, X_{ij}) = \alpha_i + \beta X_{ij}$$

- The CML estimator of β in this model is consistent if

$$\text{A: } Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, \mathbf{Z}_i$$
$$\text{B: } Y_{ij} \perp X_{ij'} | X_{ij}, \mathbf{Z}_i$$

Comparison with conditional logistic regression

- Suppose that X_{ij} and Y_{ij} are binary, and that we use the model

$$\text{logit}\{p(Y_{ij} = 1|\mathbf{Z}_i, X_{ij})\} = \alpha_i + \beta X_{ij}$$

- Let D be the exposure-discordant pairs
- Arbitrary order siblings so that $D \Leftrightarrow X_{i1} = 1$ and $X_{i2} = 0$

D	$Y_{i2} = 0$	$Y_{i2} = 1$
$Y_{i1} = 0$	n_{00}	n_{01}
$Y_{i1} = 1$	n_{10}	n_{11}

$$\exp(\hat{\beta}) = \frac{n_{10}}{n_{01}} \to \frac{p(Y_{i1} = 1, Y_{i2} = 0|D)}{p(Y_{i1} = 0, Y_{i2} = 1|D)},$$

- Not generally equal to $\exp(\beta)$

Comparison with conditional logistic regression, cont'd

- However, under

$$\text{A: } Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, \mathbf{Z}_i$$
$$\text{B: } Y_{ij} \perp X_{ij'} | X_{ij}, \mathbf{Z}_i$$

we have:

$$\begin{aligned}
\frac{p(Y_{i1} = 1, Y_{i2} = 0|D)}{p(Y_{i1} = 0, Y_{i2} = 1|D)} &= \frac{E\{p(Y_{i1} = 1, Y_{i2} = 0|D, \mathbf{Z}_i)|D\}}{E\{p(Y_{i1} = 0, Y_{i2} = 1|D, \mathbf{Z}_i)|D\}} \\
&\overset{\text{A}}{=} \frac{E\{p(Y_{i1} = 1|D, \mathbf{Z}_i)p(Y_{i2} = 0|D, \mathbf{Z}_i)|D\}}{E\{p(Y_{i1} = 0|D, \mathbf{Z}_i)p(Y_{i2} = 1|D, \mathbf{Z}_i)|D\}} \\
&\overset{\text{B}}{=} \frac{E\{p(Y_{i1} = 1|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 0|X_{i2} = 0, \mathbf{Z}_i)|D\}}{E\{p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}} \\
&= \frac{E\left\{\frac{p(Y_{i1}=1|X_{i1}=1,\mathbf{Z}_i)p(Y_{i2}=0|X_{i2}=0,\mathbf{Z}_i)}{p(Y_{i1}=0|X_{i1}=1,\mathbf{Z}_i)p(Y_{i2}=1|X_{i2}=0,\mathbf{Z}_i)} p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\right\}}{E\{p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}} \\
&= \frac{E\{\exp(\beta)p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}}{E\{p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}} \\
&= \exp(\beta)\frac{E\{p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}}{E\{p(Y_{i1} = 0|X_{i1} = 1, \mathbf{Z}_i)p(Y_{i2} = 1|X_{i2} = 0, \mathbf{Z}_i)|D\}} = \exp(\beta)
\end{aligned}$$

Matching by 'intervention' vs 'nature'

A: $Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, \mathbf{Z}_i$
B: $Y_{ij} \perp X_{ij'} | X_{ij}, \mathbf{Z}_i$

- That the CML estimator requires assumptions A and B is rarely mentioned in standard textbooks
- Probably since A and B hold by design in studies that have been matched unrelated subjects on $\mathbf{Z}_i$ by 'intervention'
- Not guaranteed to hold in studies that have been matched 'nature', e.g. sibling comparison studies
  - whether A and B hold or not depends on what we consider $\mathbf{Z}_i$ to be

Implication

A: $Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, \mathbf{Z}_i$
B: $Y_{ij} \perp X_{ij'} | X_{ij}, \mathbf{Z}_i$

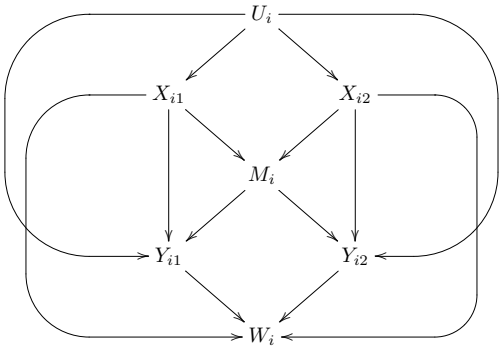
- Since A and B are sufficient for the CML estimator to be consistent we can turn the argument around and use A and B to identify $\mathbf{Z}_i$:
- **If there is a $\mathbf{Z}_i$ such that A and B hold, then the CML estimator of β in model**

$$E(Y_{ij} | \text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

consistently estimates β in model

$$E(Y_{ij} | \mathbf{Z}_i, X_{ij}) = \alpha_i + \beta X_{ij}$$

Causal diagram revisited



A: $Y_{i1} \perp Y_{i2} | X_{i1}, X_{i2}, \mathbf{Z}_i$
B: $Y_{ij} \perp X_{ij'} | X_{ij}, \mathbf{Z}_i$

- A and B hold if and only if $\mathbf{Z}_i = (U_i, M_i)$

Conclusion

$$\begin{aligned}
E(Y_{ij} | U_i, X_{ij}) &= \alpha_i^1 + \beta^1 X_{ij} \\
E(Y_{ij} | M_i, X_{ij}) &= \alpha_i^2 + \beta^2 X_{ij} \\
E(Y_{ij} | W_i, X_{ij}) &= \alpha_i^3 + \beta^3 X_{ij} \\
E(Y_{ij} | U_i, M_i, X_{ij}) &= \alpha_i^4 + \beta^4 X_{ij} \\
E(Y_{ij} | U_i, W_i, X_{ij}) &= \alpha_i^5 + \beta^5 X_{ij} \\
E(Y_{ij} | M_i, W_i, X_{ij}) &= \alpha_i^6 + \beta^6 X_{ij} \\
E(Y_{ij} | U_i, M_i, W_i, X_{ij}) &= \alpha_i^7 + \beta^7 X_{ij}
\end{aligned}$$

- **The CML estimator of β in model**

$$E(Y_{ij} | \text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

consistently estimates β^4 in model

$$E(Y_{ij} | U_i, M_i, X_{ij}) = \alpha_i^4 + \beta^4 X_{ij}$$

- Similar for the logistic model and the Cox model

Causal interpretation

- Let $Y_{ij}^{x,m}$ be the potential outcome for a subject with levels $X_{ij} = x$ and $M_i = m$
- In the absence of non-shared confounders, β is the controlled direct effect

$$E(Y_{ij}^{x+1,m}|U_i) - E(Y_{ij}^{x,m}|U_i)$$

The ADHD study revisited

Table 2. Association between maternal age at each birth (MAEB) and offspring ADHD (hazard ratios with 95% confidence intervals)

MAEB	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
Binary[e]	2.24 (2.12–2.36)	1.57 (1.48–1.67)	0.90 (0.84–0.96)	0.81 (0.71–0.94)
Continuous	1.06 (1.05–1.06)	1.05 (1.04–1.05)	0.99 (0.99–1.00)	0.98 (0.97–0.99)

[a]Population-wide association, adjusted for offspring's sex, birth order and birth year in categories.
[b]In addition to Model 1, adjusted for paternal age at childbirth in categories.
[c]In addition to Model 2, adjusted for MAFB.
[d]Sibling-comparison, adjusted for unmeasured genetic and environmental factors shared by siblings and measured covariates.
[e]MAEB < 20 y.

- The null finding by Chang et al. (2014) does not imply the absence of a causal effect
- It may also be explained by a causal effect that is mediated through familial environment
  - because the sibling comparison design implicitly controls for all shared mediators
- It may not be explained by 'collider-stratification' bias
  - because the sibling comparison design does not make implicit control for any shared colliders

Outline

Recap

- The CML estimator of β in model

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

consistently estimates β^4 in model

$$E(Y_{ij}|U_i, M_i, X_{ij}) = \alpha_i^4 + \beta^4 X_{ij}$$

...provided that the latter model is correctly specified

- What if the latter model is incorrect?

General result for linear model

- Suppose that X_{ij} is binary and define

$$\Delta(U_i, M_i) = E(Y_{ij}|U_i, M_i, X_{ij} = 1) - E(Y_{ij}|U_i, M_i, X_{ij} = 0)$$

- The CML estimator of β in model

$$E(Y_{ij}|\text{family } i, X_{ij}) = \alpha_i + \beta X_{ij}$$

generally converges to

$$E\{\Delta(U_i, M_i)|D\}$$

General result for logistic model

- Suppose that X_{ij} is binary and define

$$OR(U_i, M_i) = \frac{p(Y_{ij} = 1|U_i, M_i, X_{ij} = 1)p(Y_{ij} = 0|U_i, M_i, X_{ij} = 0)}{p(Y_{ij} = 0|U_i, M_i, X_{ij} = 1)p(Y_{ij} = 1|U_i, M_i, X_{ij} = 0)}$$

- The CML estimator of $\exp(\beta)$ in model

$$\text{logit}\{p(Y_{ij} = 1|\text{family } i, X_{ij})\} = \alpha_i + \beta X_{ij}$$

generally converges to

$$E\{d(U_i, M_i)OR(U_i, M_i)|D\},$$

where $d(U_i, M_i)$ is a weight that is proportional to

$$p(Y_{ij} = 0|U_i, M_i, X_{ij} = 1)p(Y_{ij} = 1|U_i, M_i, X_{ij} = 0)$$

General result for Cox proportional hazards model

- Suppose that X_{ij} is binary and define

$$HR(U_i, M_i) = \frac{\lambda(y_{ij}|U_i, M_i, X_{ij} = 1)\}}{\lambda(y_{ij}|U_i, M_i, X_{ij} = 0)\}},$$

where we have assumed that the hazard ratio is constant across time y_{ij}, but not across U_i and M_i

- The CML estimator of $\exp(\beta)$ in model

$$\log\{\lambda(y_{ij}|\text{family } i, X_{ij})\} = \alpha_i(y_{ij}) + \beta X_{ij}$$

generally converges to

$$E\{b(U_i, M_i)HR(U_i, M_i)|D\},$$

where $b(U_i, M_i)$ is a weight that is proportional to

$$\{1 + HR(U_i, M_i)\}^{-1}$$

Outline

No clear cut

- The sibling comparison design controls for all shared confounders and mediators, but no shared colliders
- In practice, the distinction between confounders, mediators and colliders may be somewhat artificial
  - e.g. familial environment may act as all three types of covariates
- Still a useful distinction, since helps to understand the theoretical properties of the design and models

Other estimation alternatives

- ML, GEE, random effects/frailty model, between-within (BW) model
- ML gives identical results as CML in linear models, but is inconsistent in non-linear models
  - the incidental parameter problem
- GEE does not control for any shared confounders
  - since estimates marginal (over families) association
- Random effects/frailty model is inconsistent if there is shared confounding
  - since the model assumes that the intercept is independent of exposure, thus ruling out shared confounding a priori
- BW model gives (nearly) identical results as CML in (non-)linear models

References

Chang, Z., Lichtenstein, P., D'Onofrio, B., Almqvist, C., Kuja-Halkola, R., Sjölander, A., and Larsson, H. (2014). Maternal age at childbirth and risk for adhd in offspring: a population-based cohort study. *International Journal of Epidemiology* **43,** 1815–1824.

Sjölander, A, and Zetterqvist, J. (In press). Confounders, mediators or colliders – What types of shared covariates does the sibling comparison design control for? *Epidemiology*.