# *Adjusting for selection bias in case control studies*

S.Geneletti, S.Richardson, N.Best

Department of Epidemiology and Public Health, Imperial College

01/07/2008

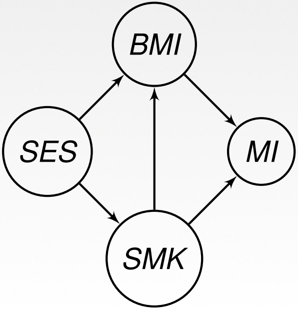
# OUTLINE

1. DAGs and conditional independence
2. Examples of selection bias in case-controls studies
3. DAG expression
4. Odds ratios
5. Bias breaking model
6. Application
7. Simulation
8. Further work

## DAGs

DAGs are *directed acyclic graphs*

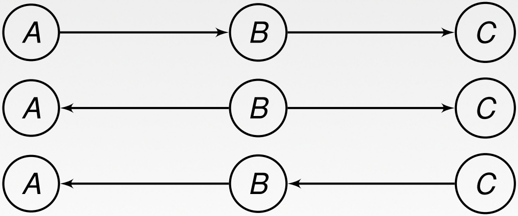
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



# CONDITIONAL INDEPENDENCE

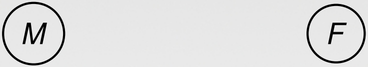
DAGs are used to encode *conditional independence statements*

- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C|B) = p(A|B)p(C|B)$
- In words if we know about $C$, knowing about $A$ gives us no extra clues about $B$ (and vice-versa)



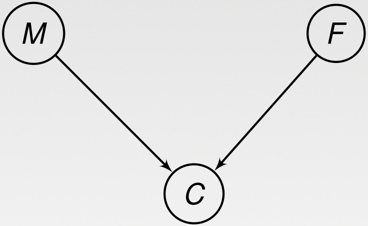
- *Causal interpretation* from *observational data is difficult*
- Need to make additional explicit assumptions
- Not all DAGs have others that are Markov Equivalent

# SIMPLE EXAMPLE - INHERITANCE



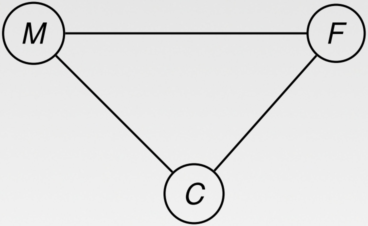
1. Male and female are independent

# SIMPLE EXAMPLE - INHERITANCE



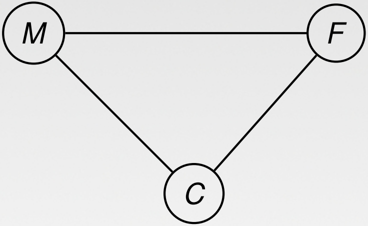
1. Male and female are independent
2. Then they meet and have a child

# SIMPLE EXAMPLE - INHERITANCE



1. Male and female are independent
2. Then they meet and have a child
3. Now they are dependent through child

# SIMPLE EXAMPLE - INHERITANCE



- In terms of conditional independence we have that
- Initially $M \perp\!\!\!\perp F$
- Later $M \not\!\perp\!\!\!\perp F | S$

## EXAMPLES of SELECTION BIAS

### *Case selection bias*

- Study in 70's found oestrogen use associated with endometrial cancer [2]
  - Selecting cases mainly amongst women with vaginal bleeding (associated to oestrogen use)
  - induces a false association between endometrial cancer and oestrogen use.

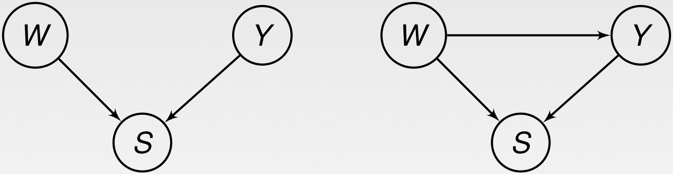
### *Control selection bias*

- Recent studies find a weak association between exposure to magnetic fields (EMF) and childhood leukaemia [3]
  - Eligible controls with lower SES are less likely to allow EMF measurements in their homes,
  - this induces a false association between leukaemia and EMF when only "full" controls included.

# SELECTION BIAS DAG

*Basic premise*

Selection bias comes about by conditioning on a common child where we don’t know distribution of child given parents



- $Y$ is the outcome of interest, $W$ the exposure, $S$ the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

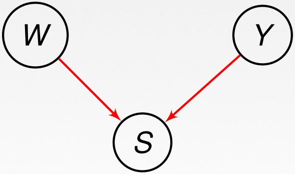
Problem - we don’t know $p(S|Y)$

# CONDITIONAL INDEPENDENCE in SB

DAGs in previous slide represent the following conditional (in)dependences :

- Left: $Y \perp\!\!\!\perp W$
- Right: None (and ME to $Y \rightarrow W$)

However, both share the same v-structure



which “charcterises” the selection bias problem.

## ODDS RATIO

### *True Odds ratio*

$$\begin{aligned}\psi &= \frac{p(Y=1|W=1)p(Y=0|W=0)}{p(Y=0|W=1)p(Y=1|W=0)}\\ &= \frac{p(W=1|Y=1)p(W=0|Y=0)}{p(W=0|Y=1)p(W=1|Y=0)}\end{aligned} \qquad (1)$$

### *Observed Odds ratio*

$$\psi^o = \frac{p(Y=1,W=1|S=1)p(Y=0,W=0|S=1)}{p(Y=0,W=1|S=1)p(Y=1,W=0|S=1)} \qquad (2)$$

## BIAS BREAKING MODEL

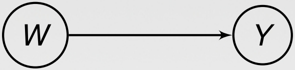
1. The problem can be addressed if we can find a bias breaking variable $B$ s.t. we can somehow separate exposure $W$ from selection $S$ for example we can assume **A1** that

$$W \perp\!\!\!\perp S | (Y, B) \tag{3}$$

2. It also necessary to find additional data
3. s.t. we can obtain an unbiased estimate of the distribution of $p(B|Y)$ - see why below.

## IDEA OF "SEPARATION"

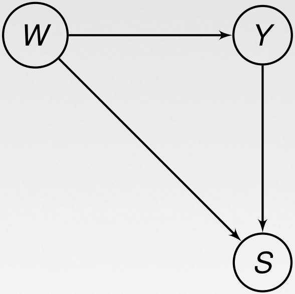
The conditional independence **A1** $W \perp\!\!\!\perp S | (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest

## IDEA OF “SEPARATION”

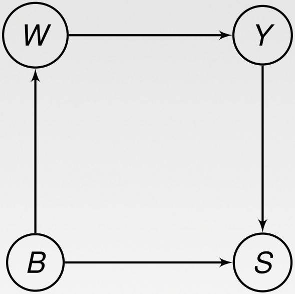
The conditional independence **A1** $W \perp\!\!\!\perp S | (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest
2. from the niusance selection bias mechanism

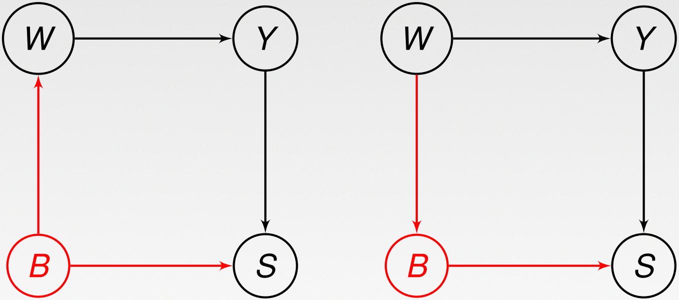
# IDEA OF “SEPARATION”

The conditional independence **A1** $W \perp\!\!\!\perp S | (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest
2. from the niusance selection bias mechanism
3. by using *B* to separate these mechanisms

# EXAMPLE DAGs



**BB MODEL**

Essential assumptions:

***A1*** *Have B such that* $W \perp\!\!\!\perp S | (Y, B)$ *holds*

***A2*** *Case and control selection are independent*
This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

***S1*** There is no selection bias in the cases i.e.
$p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1)$.

***S2*** Stratify $B$ if it is not discrete

## BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$\begin{aligned} p(W|Y = 0, S = 1, B) &= p(W|Y = 0, B) \text{ by } \textbf{A1} \text{ and} \\ \sum_B p(W|Y = 0, B)p(B|Y = 0) &= p(W|Y = 0) \end{aligned}$$

- Focus is on finding estimates of $p(B|Y)$ as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

## ESTIMATES OF $p(B|Y)$

There are various options depending on the source of additional data

*Data sources*

1. Partial study data OR
2. External (eg census) data.

... and also on the type of estimate:

*Type of estimate*

1. Conditional estimate - based on $p(B|Y)$ OR
2. Marginal estimate - based on $p(B)$ -
3. Marginal estimate valid to adjust for control selection bias when $p(B|Y = 0) \approx p(B)$.

## HYPOSPADIAS CASE CONTROL STUDY

*Story*

- Hypospadias is a congenital malformation of newborn boys
- Is it associated to gestational age or smoking? [4, 5]
- Concern that controls have a higher SES than cases-selection bias?
- SES measured using the Carstairs score - an area (ward) level index of deprivation ([6])

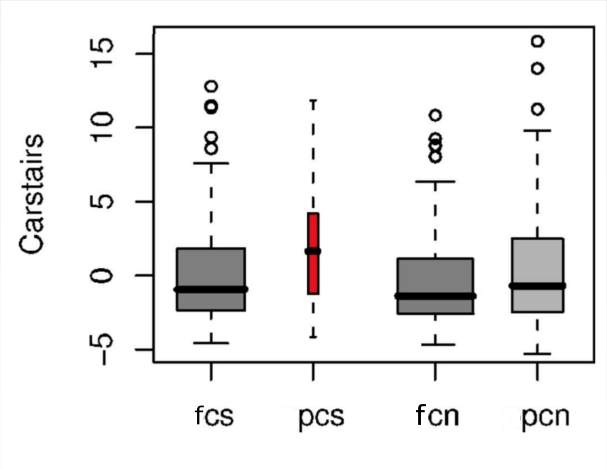
# HYPOSPADIAS CASE CONTROL STUDY

*Data*

Due to data collection process we had

- Carstairs score of people who participated - full participants (indexed by f)
- Carstairs score of many people who were asked to participate but declined as their ward was known - partial participants (indexed by p)
- Finally, Carstairs score of people who lived in the region the study was conducted from census

# Boxplot



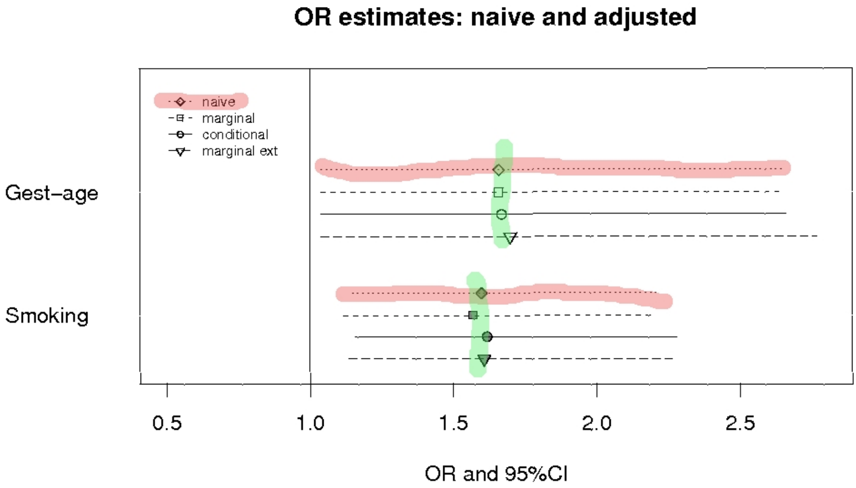
Is there also case selection bias? partial participant cases (pcs) have low SES (high Carstairs)

## HYPOSPADIAS CASE CONTROL STUDY

To adjust for selection bias we need additional data to get an “unbiased” estimate of $p(B|Y)$

- Pooling partial and full participant data and assuming this is a representative sample of the target population gives us an *internal adjustment* can estimate:
  - $p(B|Y)$ conditional
  - as well as $p(B)$ marginal - this is assuming that $p(B|Y = 0)$ can be approximated by $p(B)$
- Using data from the census means that we can do an external adjustment based on just $p(B)$ marginal ext, again assuming $p(B|Y = 0) \approx p(B)$

# RESULTS


OR estimates: naive and adjusted
naive
marginal
conditional
marginal ext
Gest–age
Smoking
0.5
1.0
1.5
2.0
2.5
OR and 95%CI

## HYPOSPADIAS CASE CONTROL STUDY

*Conclusions*

- There appears to be no selection bias mediated by SES
- Naive and adjusted are all very similar
- Do not read too much into small differences
- Validates the study results

## SIMULATIONS

*Set-up*

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, $B$ is also a confounder
- $B$ has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

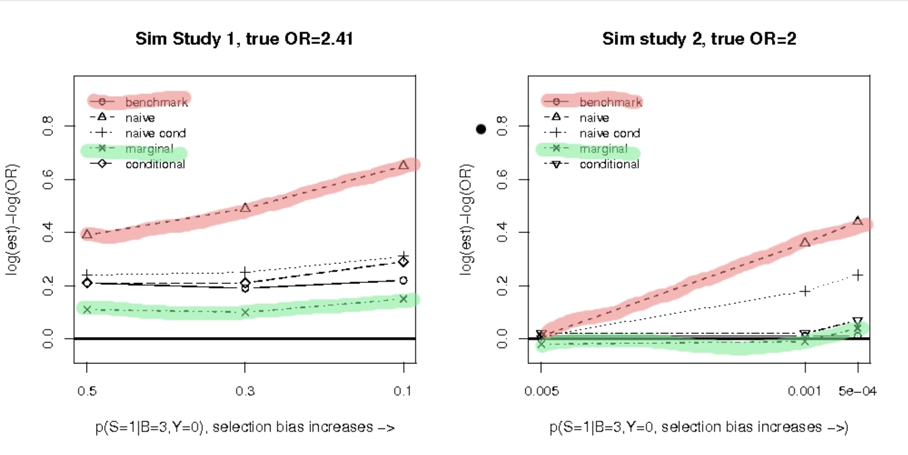
## SIMULATIONS

*Monitor*

1. No bias estimate (Logistic regression coefficient with *B* as covariate in data that is not biased)
2. Naive estimate
3. Logistic regression coefficient with *B* as covariate
4. Marginal estimator based on all data on *B*
5. Marginal estimator based only on external data on *B*

we compare our estimators to logistic regression coefficients as these are standard approaches in Epidemiology

# RESULTS


Sim Study 1, true OR=2.41
benchmark
naive
naive cond
marginal
conditional
log(est)−log(OR)
0.0
0.2
0.4
0.6
0.8
0.5
0.3
0.1
p(S=1|B=3,Y=0), selection bias increases −>
Sim study 2, true OR=2
benchmark
naive
naive cond
marginal
conditional
log(est)−log(OR)
0.0
0.2
0.4
0.6
0.8
0.005
0.001
5e−04
p(S=1|B=3,Y=0, selection bias increases −>)

# FINAL COMMENTS

*Conclusions*

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Similar to post-stratification [7]
6. Comes out in next issue of Biostatistics

*Further work*

1. Have developed Baysian version
2. Are applying it to EMF data from the US [8]

## BIBLIOGRAPY

[1] A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B (Statisical Methodology)*, 41(1):1–31, 1979.

[2] RI. Horwitz and AR. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, 299(20):1089–1094, 1978.

[3] G. Mezei and L. Kheifets. Selection bias and its implications for case-control studies: a case study of magnetic field exposure and childhood leukaemia. *International Journal of Epidemiology*, 35:397–406, 2006.

[4] G. Ormond, M.J. Nieuwenhuijsen, P. Nelson, N. Izatt, S. Geneletti, M. Toledano, and P. Elliott. Folate supplementation, endocrine disruptors and hypospadias: case-control study. under review in BMJ, 2008.

[5] M. Nieuwenhuijsen, P. Nelson, and P. Elliott. Occupational exposure of pregnant women in the south east of England. *Epidemiology*, 15(4):S165, 2004.

[6] V. Carstairs and R. Morris. *Deprivation and Health in Scotland*. Aberdeen University Press, Aberdeen, 1991.

[7] A. Gelman. Struggles with survey weighting and regression modelling. *Statistical Science*, 22:153–164, 2007.

[8] E.E. Hatch, R.A. Kleinerman, M.S. Linet, R.E. Tarone, W.T. Kaune, A. Anssi, B. Dasul, L.L. Robison, and S. Wacholder. Do confounding or selection factors of residential wire codings and magnetic fields distort findings of electromagnetic field studies? *Epidemiology*, (11):189–198, 2000.