

## *Adjusting for selection bias in case control studies*

S.Geneletti, S.Richardson, N.Best

Department of Epidemiology and Public Health, Imperial College

07/03/2008

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

## DAGs

DAGs are *directed acyclic graphs*

1. All arrows have direction
2. No cycles  $A \rightarrow B \rightarrow A$

Can be used to encode *conditional independence statements*

1.  $A \perp\!\!\!\perp C | B$  means  $p(A, C | B) = p(A | B)p(C | B)$
2. Arrows are *not* causal unless extra assumptions made -  
time ordering, intervention

## Examples of selection bias

### *Case selection bias*

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

### *Control selection bias*

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

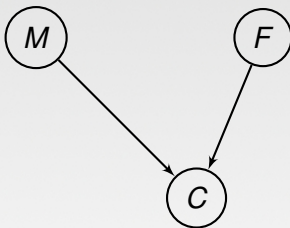
## Simple example - inheritance

M

F

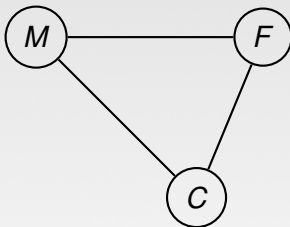
1. Male and female are independent

## Simple example - inheritance



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

## Simple example - inheritance

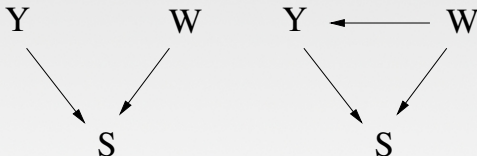


1. Male and female are independent
2. Then they meet and have a child
3. Now they are dependent through child

## 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



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

Problem - we don't know  $p(S|Y)$

## Conditional Independence

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

1. Left:  $Y \perp\!\!\!\perp W$
2. Right: None (and equivalent to  $Y \rightarrow W$ )

However, both share the same **v-structure**

$$Y \rightarrow S \leftarrow W$$

which “characterises” the selection bias problem.

## Odds ratios

### *True Odds ratio*

$$\begin{aligned}
 \psi &= \frac{p(Y = 1|W = 1)p(Y = 0|W = 0)}{p(Y = 0|W = 1)p(Y = 1|W = 0)} \\
 &= \frac{p(W = 1|Y = 1)p(W = 0|Y = 0)}{p(W = 0|Y = 1)p(W = 1|Y = 0)} \\
 &= \frac{\pi^1 \times (1 - \pi^0)}{(1 - \pi^1) \times \pi^0}
 \end{aligned} \tag{1}$$

### *Observed Odds ratio*

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

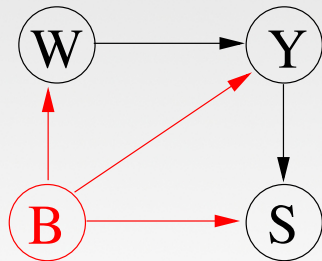
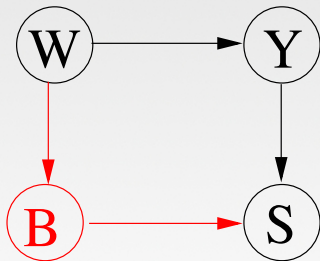
## Bias Breaking Model

1. The problem is impossible to overcome without **additional data**
2. It can be addressed if we find a **bias breaking** variable  $B$  s.t. we can somehow **separate** exposure  $W$  from selection  $S$  i.e we can assume **A1** that

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

3. **AND** s.t. we can obtain an **unbiased estimate** of the distribution of  $p(B|Y)$

## Example DAGs



**BB model cont.**

1. Find  $B$  such that **A1** holds
2. Assume **A2**: case and control selection are independent
3. Assume **A3**: there is no selection bias in the cases i.e.  
 $p(W = 1|Y = 1, S = 1) = p(W = 1|Y = 1)$ .
4. Stratify  $B$  if it is not discrete
5. Estimate  $p(W = 1|Y = y)$  - can do because by **A1**

$$p(W|Y, S, B) = p(W|Y, B) \text{ and } \sum_B p(W|Y, B)p(B|Y) = p(W|Y)$$

Focus is on finding estimates of  $p(B|Y)$  as  $p(W|Y, B)$  is estimated by stratum specific proportion of exposed cases/controls in the study

## Estimates of $p(B|Y)$

There are various options depending on the source of additional data

### *Data sources*

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

... and also on the type of estimate:

### *Type of estimate*

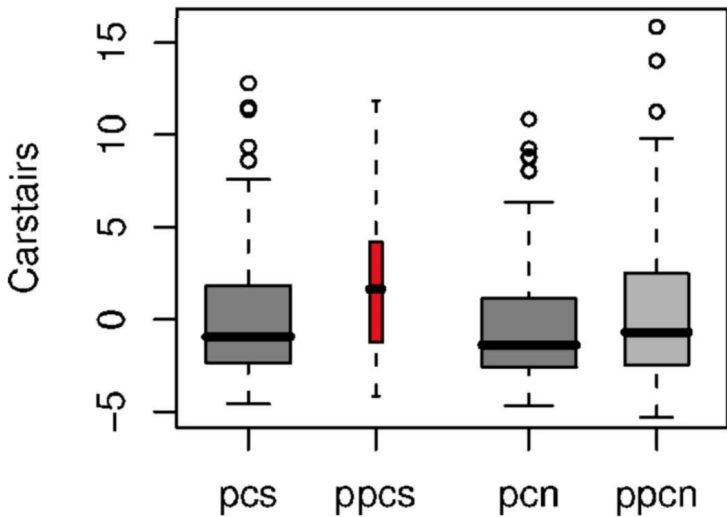
1. Conditional estimate - based on  $p(B|Y)$  OR
2. Marginal estimate - based on  $p(B)$  OR

## Hypospadias case control study

### Story

1. Hypospadias a congenital malformation - is it associated to gestational age or smoking?
2. Concern that **controls have a higher SES than cases** using Carstairs score - **selection bias?**
3. Data collection was such that we had ward data on people who had declined to participate - call these *partial* participants
4. Ward data means we have their Carstairs score.
5. Noticed that partial participant cases have lower SES (red in boxplot)
6. Decision to adjust for selection bias with SES as bias breaker in **both cases and controls**

## Boxplot



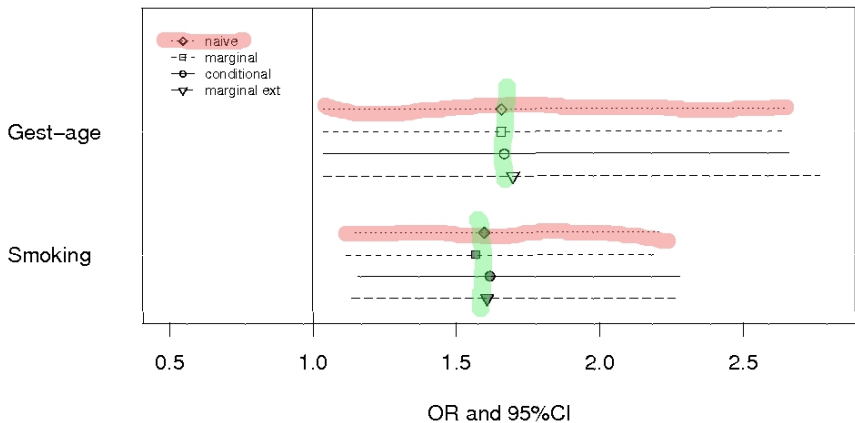
## Hypospadias case control study

### *Data*

1. Carstairs score of people who were asked to participate but declined (partial participants)
  2. Carstairs score of people who participated
  3. Carstairs score of people who lived in the region the study was conducted from census
- ▶ Pooling 1 and 2 and assuming this is a representative sample can perform internal adjustment and estimate  $p(SES|Y)$  **conditional** as well as  $p(SES)$  **marginal**
  - ▶ Using data from 3 can do an external adjustment based on just  $p(SES)$  **marginal ext.**

## Results

### OR estimates: naive and adjusted



## Hypospadias case control study

### *Conclusions*

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

## Simulations

### *Set-up*

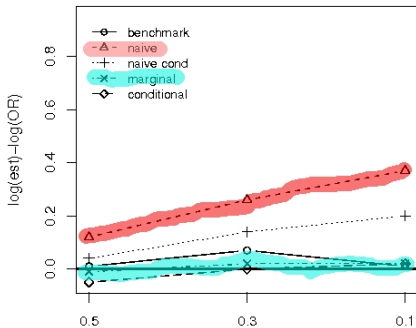
1. True OR = 1 and 2.41 (and  $B$  is a confounder)
2.  $B$  has 3 levels
3. Introduce bias by changing the probability of being selected into study if in 3rd level
4. for different probabilities of being in 3rd level.

### *Monitor*

1. Benchmark estimate (Logistic regression coefficient with  $B$  as covariate in data that is not biased)
2. Naive estimate - 2x2 table
3. Logistic regression coefficient with  $B$  as covariate
4. Marginal internal estimator based on pooled study data on  $B$
5. Marginal external estimator based only on census data on

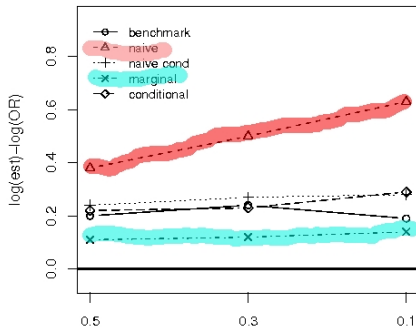
## Results

true OR=1



$p(S=1|B=3, Y=0)$ , selection bias increases  $\rightarrow$

true OR=2.41



$p(S=1|B=3, Y=0)$ , selection bias increases  $\rightarrow$

## 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
6. Should come out soon in Biostatistics

### *Further work*

1. Have developed Bayesian version
2. Will apply to EMF data from the US