

Bayesian methods for combining multiple Individual and Aggregate data Sources in observational studies

Sara Geneletti

Department of Epidemiology and Public Health

Imperial College, London

s.geneletti@imperial.ac.uk

Based on work by S. Geneletti and L.McCandless

<http://www.bias-project.org.uk>

Overall goals

- To develop a set of statistical frameworks for combining data from multiple sources
- To improve the capacity of social science methods to handle the intricacies of observational data.
- Key statistical tools: Bayesian hierarchical models and ideas from graphical models will be used to formulate the basic building blocks for these developments

Outline

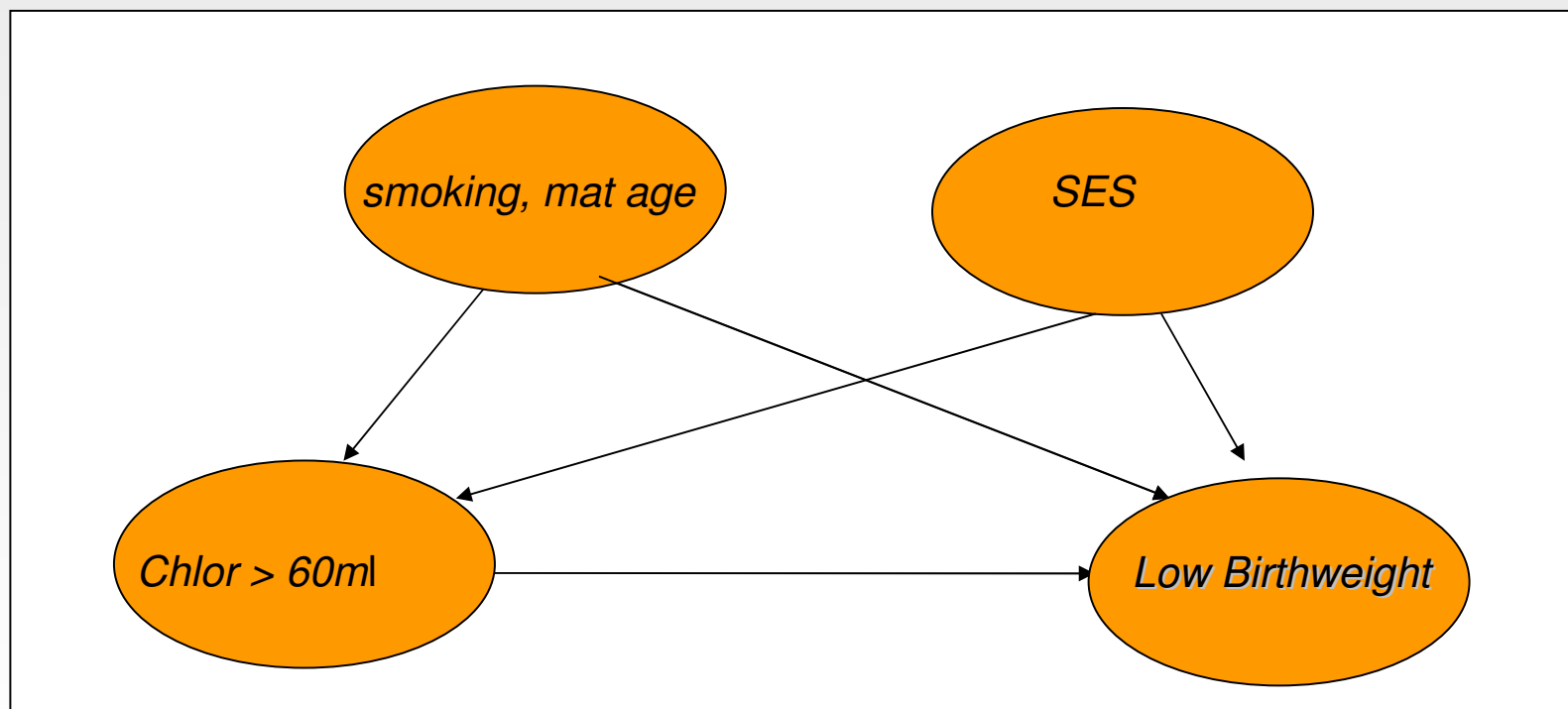
- Focus on two particular projects:
 - ◆ Combining multiple data sources with many confounders effectively using a *propensity score*
 - ◆ Modelling of selection bias in observational studies
- Both projects deal with **adjusting for bias**
- and use **multiple data sets** to adjust for it

I– Project: Combining multiple data sources and confounders using propensity scores

- RA: Lawrence McCandless
- Methodology for adjusting for many measured and unmeasured confounders
- Case study
 - ◆ Assessing environmental exposures on birth weight (e.g. air pollution, water chlorination)

Water chlorination and low birthweight

- Does exposure to water chlorination during pregnancy increase the risk of low birth weight?
- must control for ethnicity, smoking, maternal age,... etc
- is there confounding perhaps by SES?



Water chlorination and low birthweight

Data sources

- Combine datasets with different strengths:
 - ◆ Survey data (Millennium Cohort Study - MCS)
 - ◆ Small, great individual detail (n=1115).
 - ◆ Confounders: alcohol, smoking, income, education, ethnicity ...
 - ◆ Administrative data (national births register -NBR)
 - ◆ Large, but little individual detail (n=7945).
 - ◆ Confounders: Mothers age and baby gender only

Water chlorination and low birthweight

Analysis approaches

- Analyse NBR data alone
 - ◆ **High power**, but biased from unmeasured **confounding**.
- Analyse MCS data alone
 - ◆ **Low power**, but estimates largely **unconfounded**.
- An alternative:
 - ◆ Treat unmeasured confounders in NBR as missing data and build model from MCS to generate confounders for NBR (Jassy Molitor)
 - ◆ Similar to two-stage sampling and regression with missing covariates.

Water chlorination and low birthweight

Analysis approaches

■ A challenge:

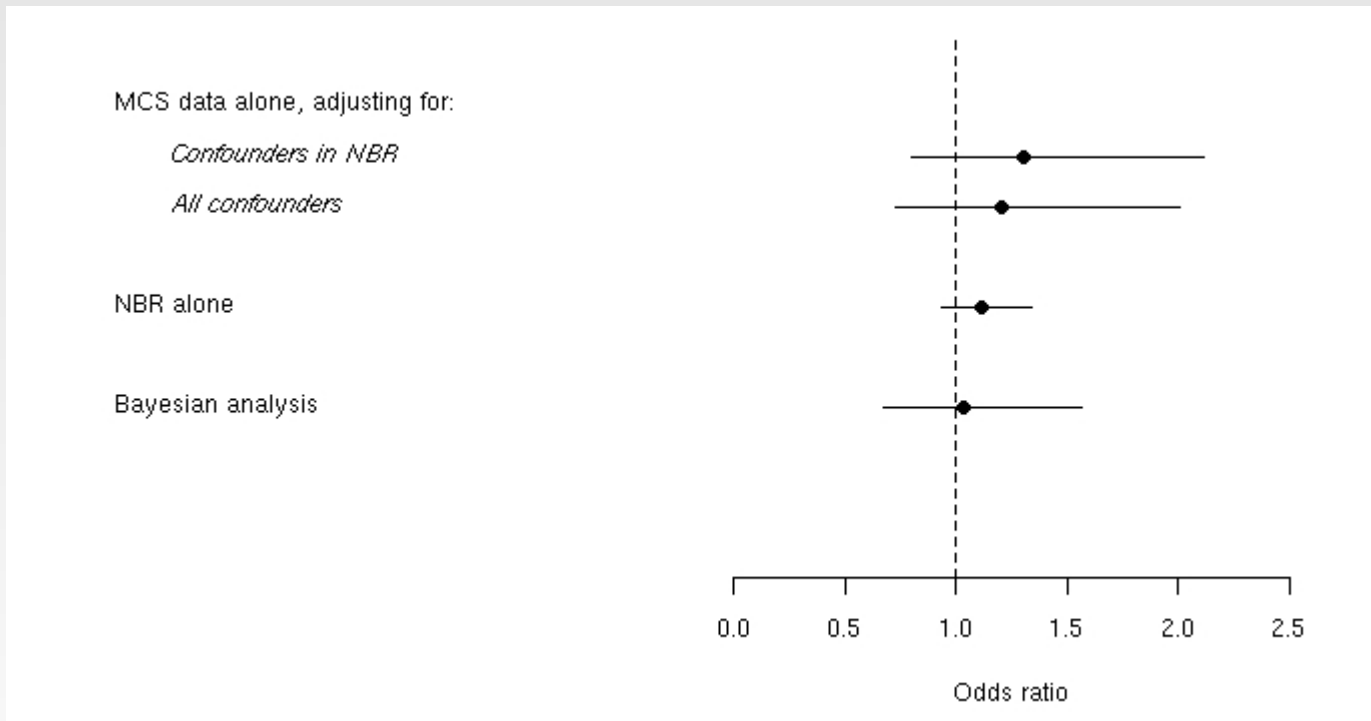
- ◆ Imputing a complex pattern of missing covariates is challenging – i.e. generating from a model
- ◆ Very many confounders and covariates (over 10)
- ◆ Also computationally expensive

■ Our approach:

- ◆ Summarize the confounders with a **single score** called the **propensity score**
- ◆ Impute this quantity within a Bayesian framework.
- ◆ Use MCS to impute a single score for NBR
- ◆ Different from standard propensity score literature as we use **external data/informative prior** to generate score

Water chlorination and low birthweight

Initial findings



- The Bayesian propensity score pulls OR towards 1 as it adjusts for confounding

Initial findings

- **Adjusting for unmeasured confounding reduces the strength of the association.**
 - ◆ Perhaps mothers with other risk factors also more likely to live in regions with chlorination levels higher than 60mu/L?
 - ◆ Could this be due to SES?
- Data synthesis also reduces precision of estimate.
 - ◆ Using NBR alone gives inferences which are falsely precise because they ignore uncertainty due to bias

II–Project: Adjusting for selection bias in case–control studies

- RA: Sara Geneletti
- Methodology : Bias breaking variable
- uses graphical models
- Case study
 - ◆ Risk factors for Hypospadias (congenital anomaly affecting male babies)

Adjusting for selection bias (SB) in case-control studies

- **Case-control** studies in particular suffer from SB
- Hypospadias vs lifestyle factors
 - ◆ participant controls had higher SES (measured by Carstairs score) than participant cases
 - ◆ is there selection bias via differential participation of cases and controls due to SES?
- To understand use **graphical models**

Simple example of graphical model

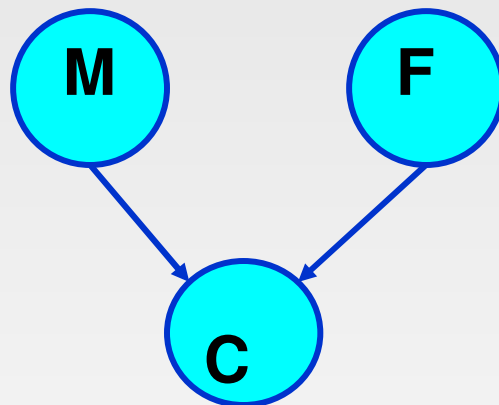
Mendelian inheritance



- M, F = genotype of a couple
- The genotypes of the couple are independent, just two random sets out in the world
- They meet and...

Simple example of graphical model

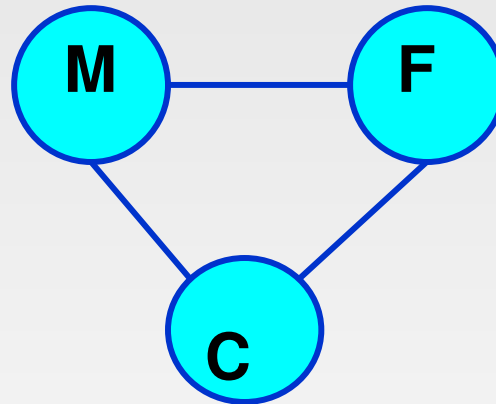
Mendelian inheritance



- C = genotype of child
- Once the couple have a child and become parents
- their genotypes become associated through the child – e.g. paternity testing

Simple example of graphical model

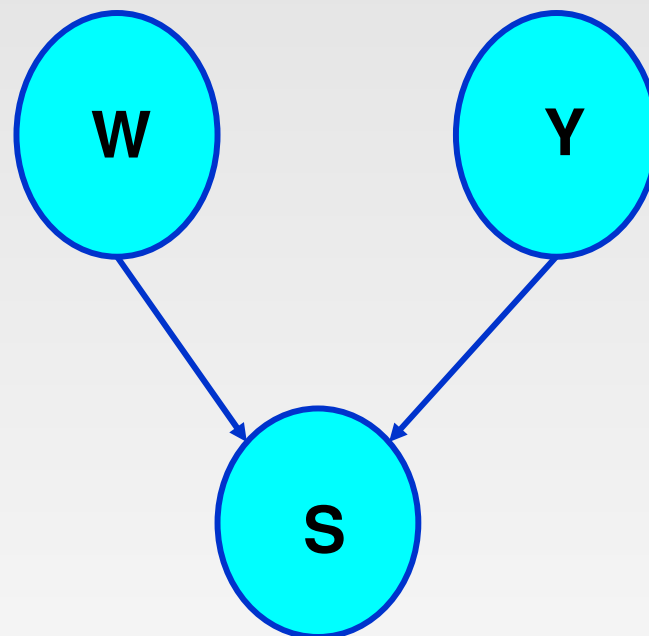
Mendelian inheritance



- In graphical terms, this connects the parents in a dependence relationship
- indicated by the line joining them
- graph with arrows called Directed Acyclic Graph (DAG)

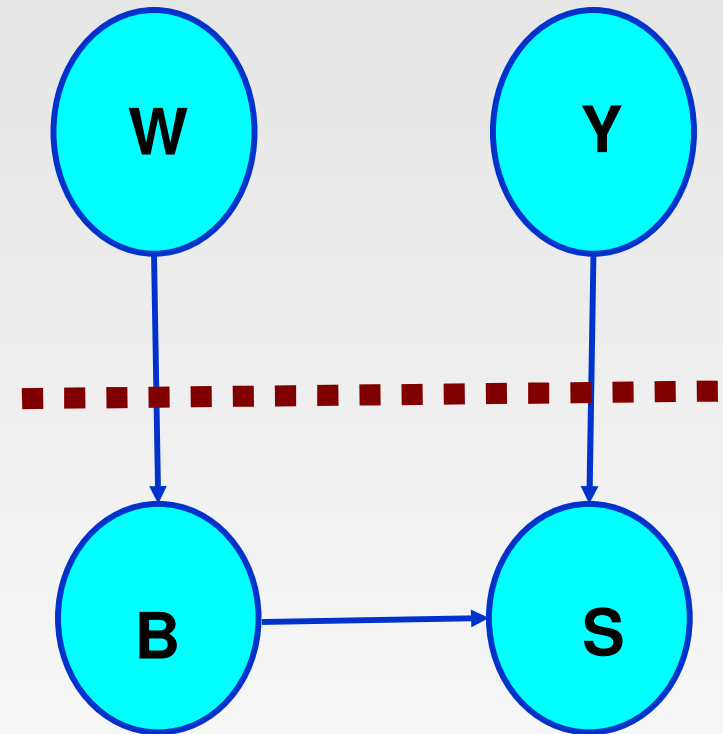
Using DAGs to understand

- W – smoking
- Y – hypospadias
- S – in the study?
- S is a “child” of smoking and Hypospadias
- even if W & Y not related in the general population, if there is selection bias, they are associated in the study



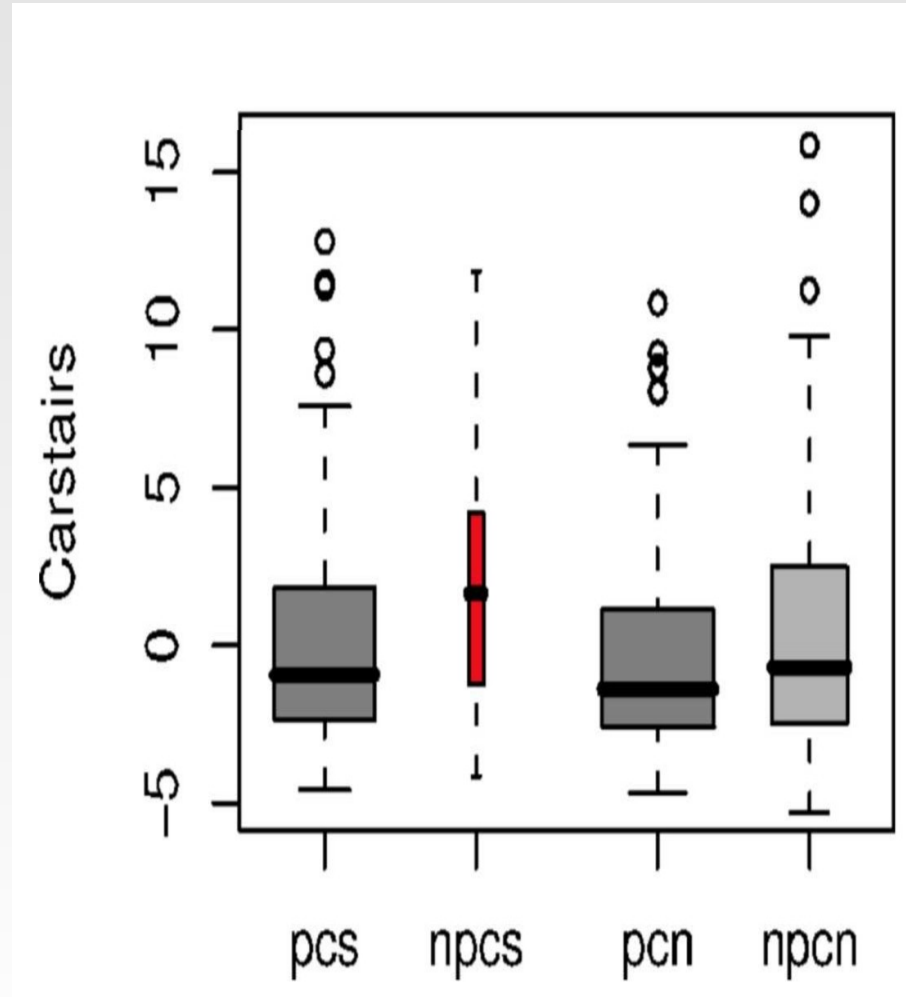
The Bias breaking variable

- To adjust for SB find a variable that **separates** the **exposure-disease mechanism** from the **biasing mechanism** – maybe **SES**?
- If we can estimate distribution of SES (B) w/out bias then can adjust



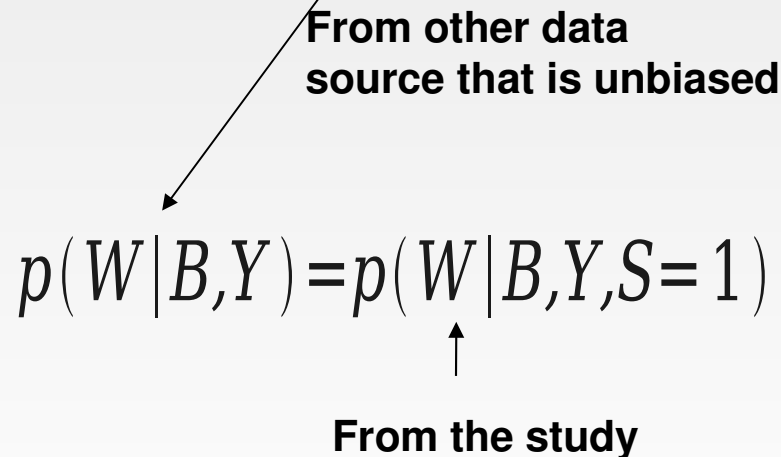
The Bias breaking variable

- All potential participants contacted via GP
- So even when declined to participate we know their ward and Carstairs
- **Non-participants appear more deprived**
- Is that a problem?



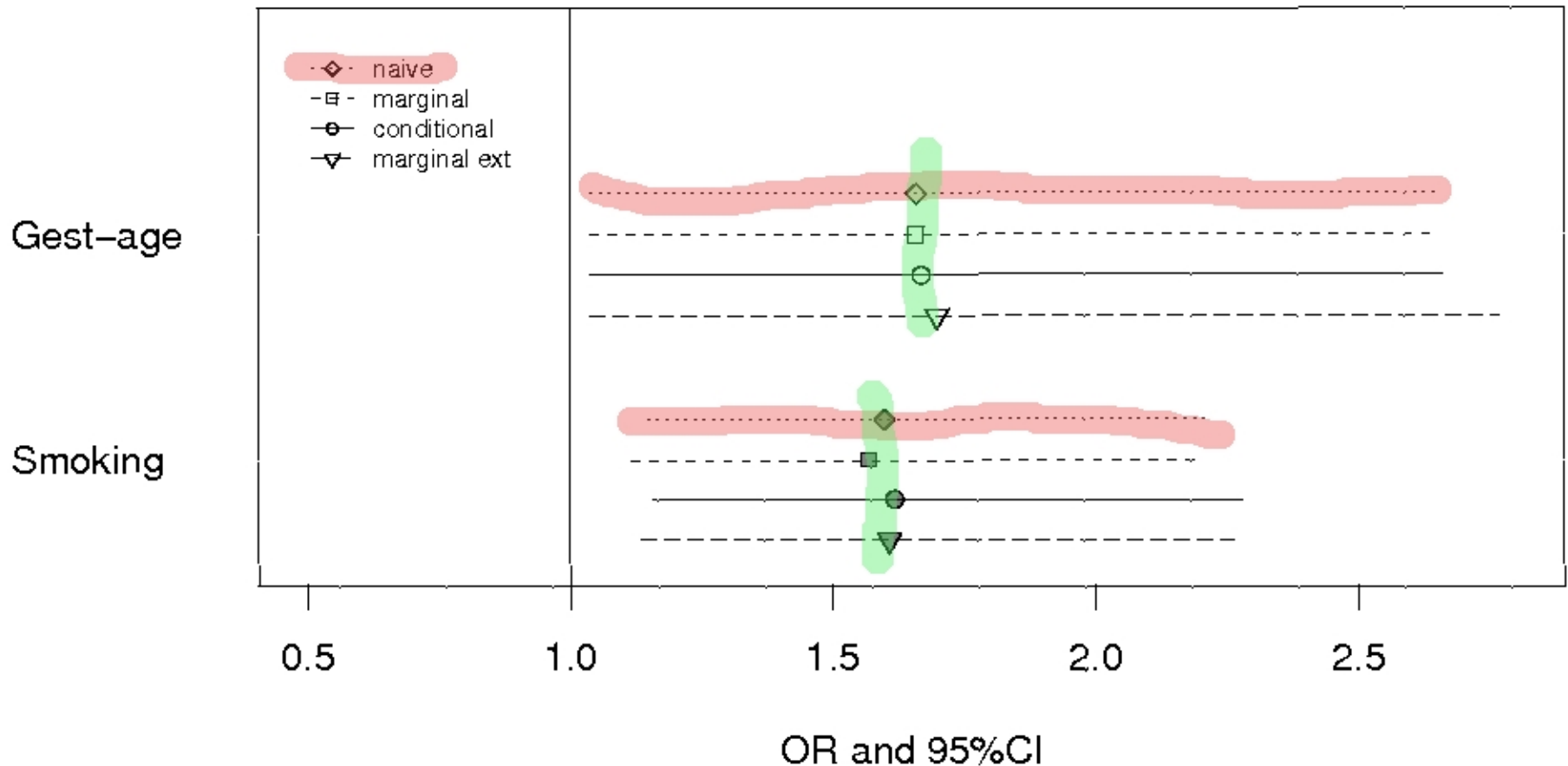
Some equations

- **Observed** odds ratio a function of _____ $p(W|Y, S=1)$
- **True** odds ratio a function of _____ $p(W|Y)$
- Law of total probability _____ $p(W|Y) = \sum p(W|B, Y) p(B|Y)$
 - We can get $p(W|Y, B)$ from study by conditional independence assumption and stratifying over B
 - If we can get $p(B|Y)$ from unbiased sources then we can estimate $p(W|Y)$ and so OR



Adjusted estimates (3 types)

OR estimates: naive and adjusted

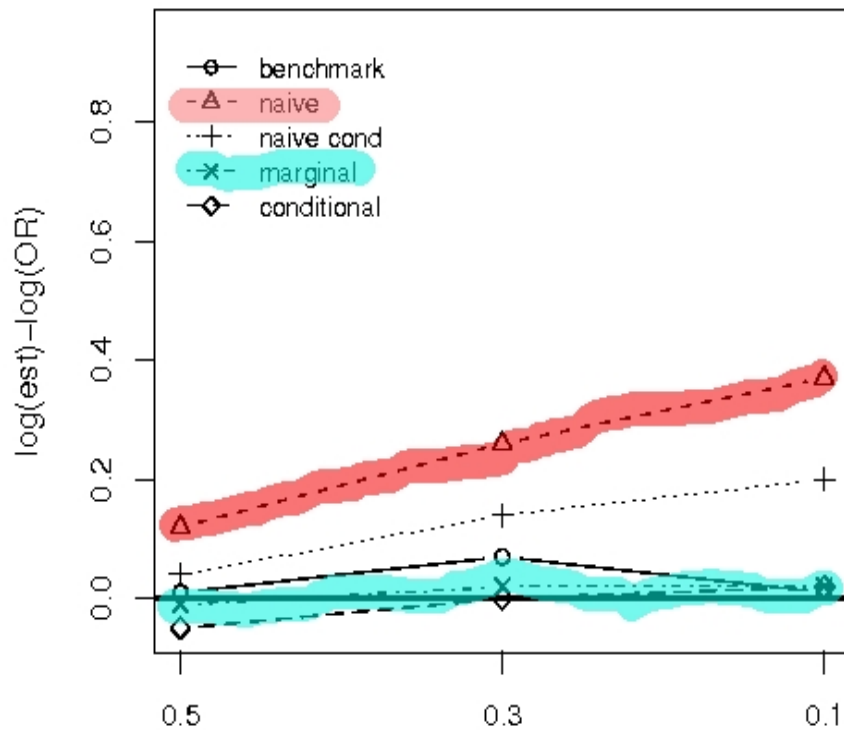


The Bias breaking variable

- The different estimates correspond to the **naïve** OR estimate – i.e. standard (highlighted in pink)
- Others are variations on our methods
- From the analysis on the Hypospadias data set there seems to be **no selection bias** mediated by SES as all the estimates including the naïve are very similar (green highlight)

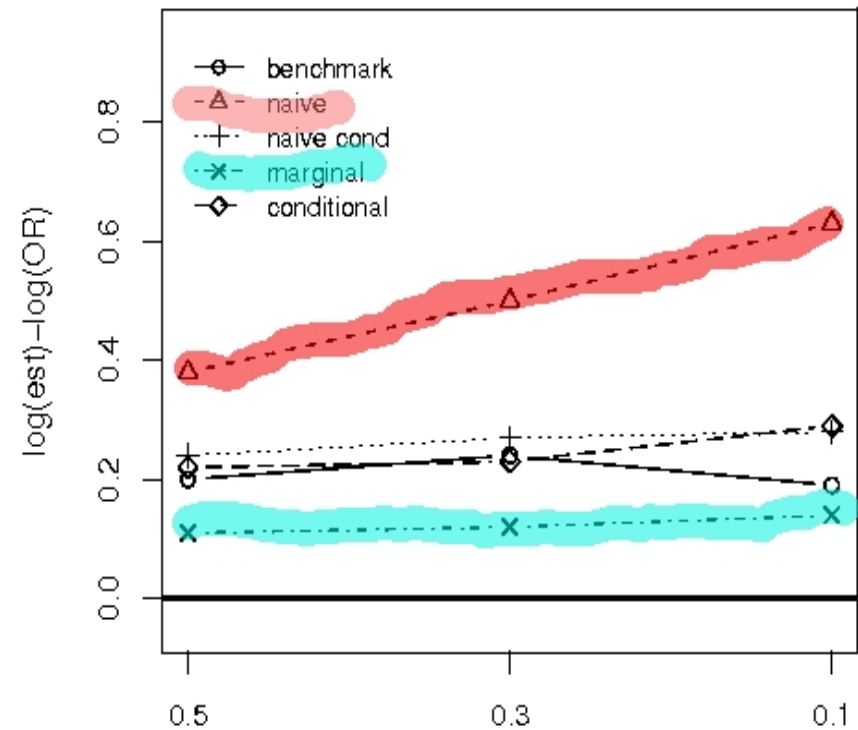
Simulations: when there IS selection bias

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

Conclusions about Selection bias

- The method we propose **adjusts for selection bias** when it is there...
- AND It does **not** introduce bias when this is not present!
- It is based on the use of external data:
assumption that a **bias breaking** variable exists
and we have **data to estimate its distribution** in
addition to the data in the study

Further work and comments

- ◆ Running **simulations** to confirm that the propensity score method works
- ◆ Extending selection bias model to a **Bayesian** framework.
- ◆ Are looking to apply it to a Leukaemia and EMF exposure dataset that **does appear to suffer** from selection bias
- ◆ Lawrence has a poster manned by Nicky and Jassy