

# Bayesian Approaches to Adjustment for Unmeasured Confounders

Sylvia Richardson & Lawrence McCandless

`sylvia.richardson@imperial.co.uk`

Joint work with:

Jassy Molitor, Nicky Best

Department of Epidemiology and Public Health, Imperial College London

March 2008

# The Problem of Unmeasured Confounding

## Background

- Confounding from unmeasured background variables is a common problem in observational studies.
- Exposure effect estimates will be biased without proper adjustment.
  
- This is a common problem in studies using large administrative databases and registries.
- Such databases are often missing important information and have measurement error.

# The Problem of Unmeasured Confounding

## Possible solutions

- Solution is to try to adjust for unmeasured confounders.
- A **framework** for existing methodologies:
- **A:** Source of prior information about unmeasured confounding?
  - Fully elicited versus use of external data
- **B:** Possible modelling strategies?
  - Model-based versus semi-parametric

# Prior Information About Unmeasured Confounders

## A Fully Elicited Approach

- At one extreme, we have the **fully elicited** approach.
- Information is elicited from prior beliefs.
- For example Greenland (2003), *J Am Stat Assoc*:
  - Greenland studies the association between magnetic field exposure and child leukemia assuming that there is a binary unmeasured confounder.
  - Prior information about possible confounders is elicited from the literature.
  - Greenland concludes that published associations may be spurious artifacts of bias.

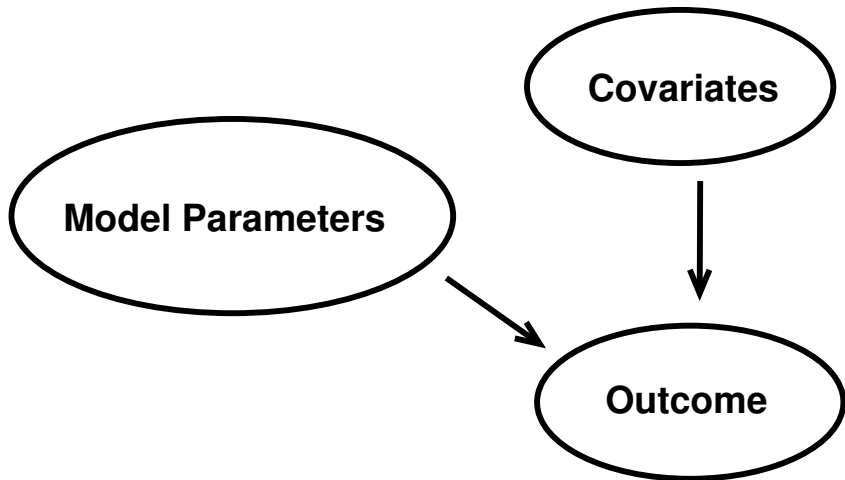
# Prior Information About Unmeasured Confounding

## Use of External Datasets

- Alternatively, information about unmeasured confounding may be available from external datasets. (e.g. Surveys or secondary samples).
- We distinguish between the **primary data** versus the **external data**, which provide information about unmeasured confounding.
- Analysis involves synthesis of multiple sources of empirical evidence.
- This requires complex models and exchangeability assumptions.
- **Bayesian graphical models** can be useful...

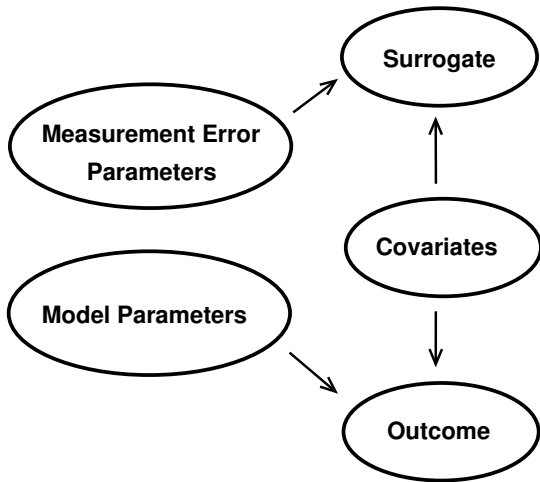
# Bayesian Graphical Models for Complex Data Structures

Example: A regression analysis



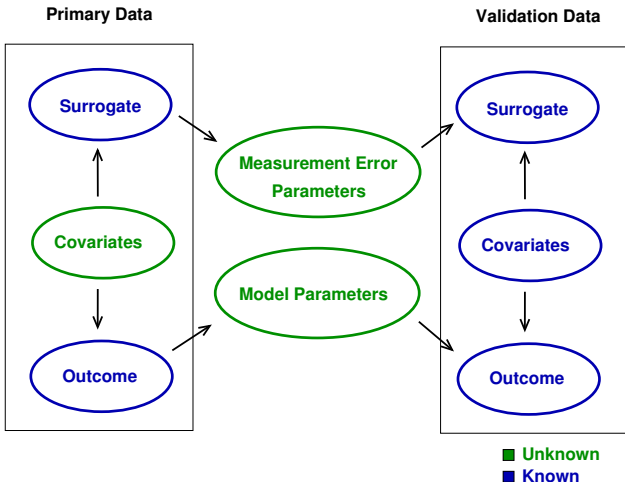
# Bayesian Graphical Models for Complex Data Structures

Regression with measurement error



# Bayesian Graphical Models for Complex Data Structures

Model for multiple data sources



# An Example Using External Data

## Water Disinfection By-Products and Risk of Low Birthweight

- Example from Molitor et al. (2008)
- **Objective:** To estimate the association between trihalomethane concentrations and risk of full term low birthweight.
- Information was collected for 9278 births between 2000 and 2001 in North West England, serviced by the United Utilities Water Company.
- Birth records obtained from the **National Birth Registry** and were linked to trihalomethane water concentrations using residence at birth.

## The Primary Data: National Birth Registry

- The primary data have the advantage of capturing information on all births in the population under study.  
→ Increased power
- However, they contain only limited information on the mother and infant characteristics which impact birth weight.  
→ Increased bias
- They contain data on mother's age and baby gender, but not smoking, ethnicity.
- Also, gestational age is unknown.

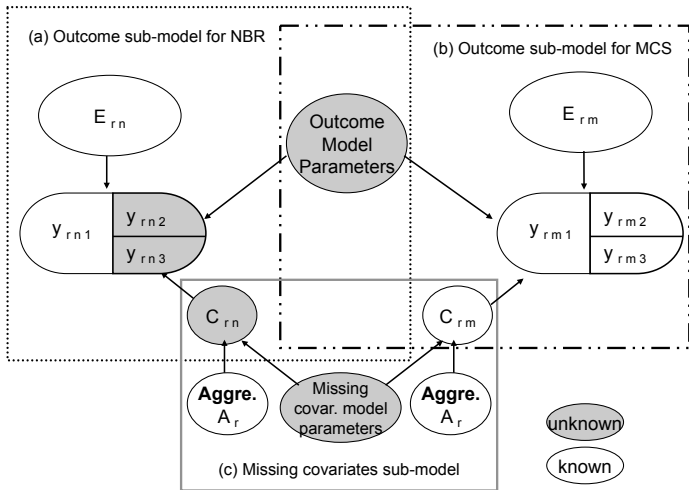
# Adjustment for Unmeasured Confounders

## Sources of External Data

- The **Millennium Cohort Study** contains survey information on mothers and infants born during 2000-2001.
- Contains information on ethnicity, smoking and gestational age.
- We link the survey data with that of the National Birth Registry using Bayesian hierarchical models.
  
- Aggregate census data contains information on unmeasured confounders.

# Adjustment for Unmeasured Confounders

## A Graphical Model for Trihalomethane Data



## Analysis Results

Description	MCS data alone Odds ratio (95% interval estimate)	Full Bayesian Model MCS + NBR + Aggr. Odds ratio (95% interval estimate)
Trihalomethanes > 60 $\mu$ g/L	1.51 (0.76, 3.01)	2.26 (1.09, 4.17)
Smoker	2.45 (1.18, 5.07)	2.73 (1.25, 5.31)
Non-White	5.25 (2.17, 12.72)	7.41 (3.34, 14.49)

We see a positive association between trihalomethane exposure and full-term low birthweight.

# Adjustment for Multiple Unmeasured Confounders

## Overview

- Model based approaches to adjustment for unmeasured confounders work well when there are only few covariates.
- We can use parameteric approaches.
- For example, in previous investigation the variables ethnicity and smoking are dichotomous.
  
- But for multiple unmeasured confounders, model-based adjustment is more difficult.
- The variables may be correlated, continuous or categorical.
- We now present a **semi-parametric** approach.

# Adjustment for Multiple Unmeasured Confounders

## Variables and Notation

Introducing some notation:

- Let  $Y$  denote a dichotomous outcome
  - Let  $X$  denote a dichotomous exposure
  - Let  $C$  denote a vector of measured confounders
  - Let  $U$  denote a vector of unmeasured confounders.
- 
- The objective is to estimate the association between  $X$  and  $Y$  while controlling for  $(C, U)$

# Adjustment for Multiple Unmeasured Confounders

Modelling  $U$  as a Latent Variable

- **One approach:** Model  $P(Y|X, C)$  as

$$P(Y|X, C) = \int P(Y|X, C, U)P(U|X, C)dU$$

This strategy requires modelling distributional assumptions about  $U$ .

# Adjustment for Multiple Unmeasured Confounders

## Illustration in Trihalomethane Data

- Primary Data: National Births Registry
  - Sample size  $m = 7945$
  - Contains only baby gender, mother's age
- Validation Data: Millenium Cohort Study
  - Sample size  $m = 1115$
  - We let  $U$  be a vector of 9 unmeasured confounders. Including: smoking and ethnicity. Also income, education, alcohol ...

## Analysis of **only** the Primary Data While Ignoring Unmeasured Confounding

Description	Odds ratio (95% interval estimate) NAIVE
Trihalomethanes > 60 $\mu$ g/L	<b>2.27 (1.68, 3.06)</b>
Mother's age	
$\leq 20$	0.35 (0.16, 0.77)
20 - 24	1.86 (1.29, 2.66)
25 - 29*	1.0
30 - 34	0.67 (0.42, 1.05)
$\geq 35$	0.89 (0.53, 1.47)
Male baby	0.87 (0.65, 1.16)

\* Reference group

→ **Biased from unmeasured confounding?**

# Analysis of **only** the Millenium Cohort Study

Description	Odds ratio (95% interval estimate)	
	Adjusting for $C_j$ only	Adjusting for $C_j$ and $U_j$
Trihalomethane > 60 $\mu$ g/L	2.52 (1.03, 6.15)	2.43 (0.95, 6.24)
Mother's age		
$\leq$ 20	0.54 (0.08, 3.47)	0.29 (0.04, 2.05)
20 - 24	1.81 (0.66, 4.97)	1.02 (0.34, 3.04)
25 - 29*	0.0	0.0
30 - 34	0.37 (0.09, 1.48)	0.58 (0.14, 2.45)
$\geq$ 35	0.79 (0.19, 3.19)	1.37 (0.31, 6.05)
Male baby	0.92 (0.39, 2.13)	0.75 (0.31, 1.82)
Lone parent family	.	1.40 (0.43, 4.58)
Number of live children	.	0.90 (0.71, 1.13)
Smoking during pregnancy	.	5.79 (1.54, 21.69)
Ethnicity		
White/Other*	.	0.0
Asian	.	8.98 (1.54, 52.36)
Black	.	2.58 (0.10, 63.36)
Alcohol during pregnancy	.	1.84 (1.01, 3.37)
Body mass index <sup>†</sup> (Kg/m <sup>2</sup> )	.	0.76 (0.38, 1.50)
Income (1=Low, 2=Med, 3=High)	.	0.49 (0.18, 1.32)
High school diploma	.	1.04 (0.41, 2.59)

\* Reference group

† Measured prior to pregnancy

→ Unmeasured confounders appear to be important?

# Adjustment for Multiple Unmeasured Confounders

## A Semi-parametric Modelling Approach

- We now present a semi-parametric method for adjusting for a vector of unmeasured confounders.
- Our method builds on idea of the propensity score, introduced by Rosenbaum and Rubin (1983), *Biometrika*.
- The propensity score is defined as the probability of treatment given confounders.
- They showed that to control confounding, it suffices to adjust for the propensity score in a regression model for the outcome.

# Adjustment for Multiple Unmeasured Confounders

## A Semi-parametric Modelling Approach

- Our model

$$\text{Logit}[P(Y = 1|X, C, U)] = \alpha + \beta X + \xi^T C + \tilde{\xi}^T g\{Z(U)\}^T$$

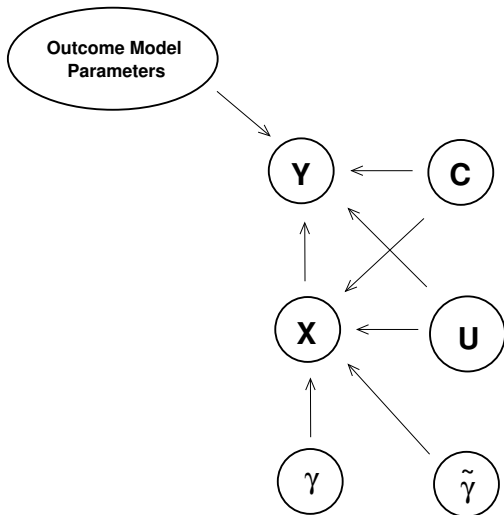
$$\text{Logit}[P(X = 1|C, U)] = \gamma_0 + \gamma^T C + \tilde{\gamma}^T U$$

where  $Z(U) = \tilde{\gamma}^T U$  is called the **conditional propensity score**.

- One can show that for  $Z(U)$  as defined, we have

$$X \perp\!\!\!\perp U | C, Z(U)$$

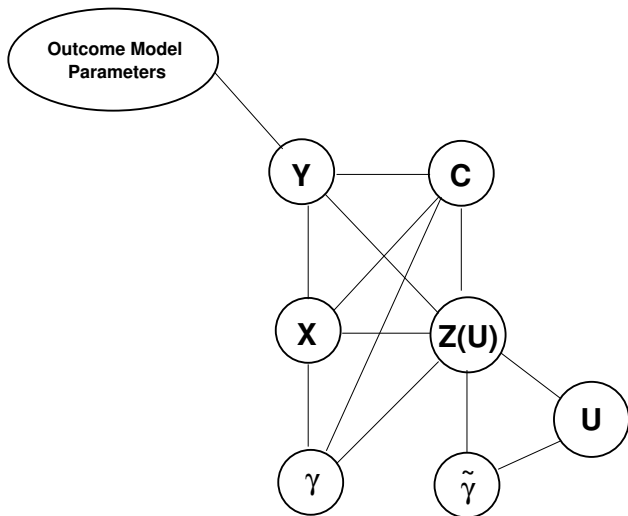
# Adjustment for Multiple Unmeasured Confounders



- We see the role of the unmeasured confounders  $U$ .

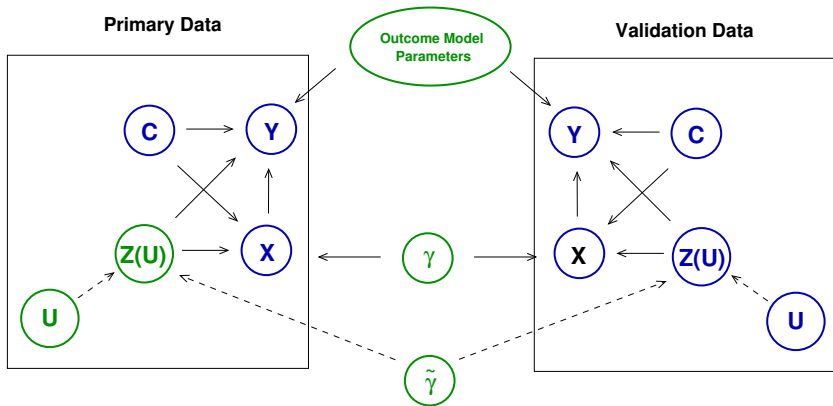


# Adjustment for Multiple Unmeasured Confounders



- Moral graph and we see that  $X \perp\!\!\!\perp U | C, Z(U)$

# Adjustment for Multiple Unmeasured Confounders



- Model for primary and validation data.

# Adjustment for Multiple Unmeasured Confounders

## A Semi-parametric Modelling Approach

- Conditional on  $Z(U)$ , the variable  $U$  is not a confounder.
- Can substitute a scalar covariate in place of a vector.
- The advantage is that to adjust for unmeasured confounding, we need only model a scalar summary score.

# Adjustment for Multiple Unmeasured Confounders

## A Semi-parametric Modelling Approach

- We model

$$P(Y, X|C) = \int P(Y|X, C, U)P(X|U, C)P(U|C)dU$$

instead of the usual

$$P(Y|X, C) = \int P(Y|X, C, U)P(U|X, C)dU$$

- **Easier?** The approach is appealing because propensity score approach already requires a model for  $P(X|U, C)$ .

# Adjustment for Multiple Unmeasured Confounders

A semi-parametric modelling approach

- To complete the specification, we require a model for  $P(U|C)$ .
- We assume that  $U$  and  $C$  are marginally independent and use the empirical distribution of  $U$  from the validation data.
- This gives a model for  $P(Y, X|C)$ .

# Analysis of the Primary Data Adjusting for Unmeasured Confounders

## Analysis of NBR Data

Description	Odds ratio (95% interval estimate)	
	NAIVE	BAYES
Trihalomethanes > 60 $\mu$ g/L	2.27 (1.68, 3.06)	1.96 (1.32, 2.92)
Mother's age		
$\leq$ 20	0.35 (0.16, 0.77)	0.32 (0.16, 0.68)
20 - 24	1.86 (1.29, 2.66)	1.77 (1.25, 2.48)
25 - 29*	0.0	0.0
30 - 34	0.67 (0.42, 1.05)	0.61 (0.39, 0.94)
$\geq$ 35	0.89 (0.53, 1.47)	0.86 (0.53, 1.37)
Male baby	0.87 (0.65, 1.16)	0.84 (0.64, 1.12)

\* Reference group

# Adjustment for Multiple Unmeasured Confounders

## Summary

- Exposure effect estimate is driven towards zero. This is consistent with analysis of the MCS data.
- 95% credible interval is 30% longer on log odds scale (15% longer on odds ratio scale). A consequence of modelling uncertainty from unmeasured confounding.
- Separate simulations show that the propensity score method can effectively adjust for multiple unmeasured confounders.

## Summary and Conclusion

- Methods for adjustment for unmeasured confounding are available and feasible.
- One precaution is that we must be careful to study exchangeability assumptions between datasets.