

# Bayesian approaches for combining multiple data sources to adjust for missing confounders

Alexina Mason<sup>1</sup>, Sylvia Richardson<sup>1</sup>,  
Lawrence McCandless<sup>2</sup> and Nicky Best<sup>1</sup>

<sup>1</sup>Department of Epidemiology and Biostatistics, Imperial College London, UK

<sup>2</sup>Faculty of Health Sciences, Simon Fraser University, Canada

EAM-SMABS 2010, July 21-23, 2010 in Potsdam/Berlin

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

# Outline

## Introduction

The problem of unmeasured confounding  
Case study

Approach 1: Building a Bayesian joint model

Approach 2: Bayesian propensity score adjustment

# The problem of unmeasured confounding

- Data for the social, behavioural and health sciences typically come from observational studies
- Some data sources, e.g. routinely collected administrative data,
  - have a limited number of variables for a large population
  - but typically miss important confounders ⇒ **bias**
- Information about missing confounders may be available from other data sources,
  - e.g. surveys or cohort studies
  - containing detailed information on a small sample of individuals
- Problem of bias can be mitigated by combining multiple sources of data

# Reframing as a non standard missing data problem

- We consider the situation where
  - the confounders are identified
  - the **primary** data source misses important confounders
  - information on these unmeasured confounders is available from a **supplementary** data source
- If records from the two data sources can be matched,
  - we have a missing data type problem
  - where all the observed values of the variables with missing data come from the supplementary data

# Reframing as a non standard missing data problem

- We consider the situation where
  - the confounders are identified
  - the **primary** data source misses important confounders
  - information on these unmeasured confounders is available from a **supplementary** data source
- If records from the two data sources can be matched,
  - we have a missing data type problem
  - where all the observed values of the variables with missing data come from the supplementary data
- We compare 2 Bayesian approaches for addressing this problem
  1. build a joint model incorporating an analysis model and an imputation model for the missing confounders (Molitor et al, 2009)
  2. semi-parametric modelling approach using propensity score ideas (McCandless et al, 2009)

# Case study: water disinfection by-products and risk of low birth weight

- Objective:
  - estimate the association between trihalomethane (THM) concentrations and the risk of full term low birth weight (<2.5kg)

# Case study: water disinfection by-products and risk of low birth weight

- Objective:
  - estimate the association between trihalomethane (THM) concentrations and the risk of full term low birth weight (<2.5kg)
- Primary data:
  - 8969 birth records between 2000 and 2001 in North West England from the Hospital Episode Statistics (HES) database
  - linked to estimated trihalomethane water concentrations
  - contains data on mother's age, baby gender and an index of deprivation
  - but no data on maternal smoking and ethnicity

# Case study: water disinfection by-products and risk of low birth weight

- Objective:
  - estimate the association between trihalomethane (THM) concentrations and the risk of full term low birth weight (<2.5kg)
- Primary data:
  - 8969 birth records between 2000 and 2001 in North West England from the Hospital Episode Statistics (HES) database
  - linked to estimated trihalomethane water concentrations
  - contains data on mother's age, baby gender and an index of deprivation
  - but no data on maternal smoking and ethnicity
- Supplementary data:
  - survey information on mothers and infants from the Millennium Cohort Study (MCS)
  - contains detailed information on smoking and ethnicity
  - 824 cohort births matched to primary data

## Analysis results using a single data source

Odds ratio (95% interval estimate)

	HES only (n=8969) (excludes $U^\dagger$ )	MCS only (n=824) (excludes $U^\dagger$ )	MCS only (n=824) (includes $U^\dagger$ )
Trihalomethanes			
> 60 $\mu$ g/L	1.39 (1.10,1.76)	2.06 (0.85,4.98)	1.87 (0.76, 4.62)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.10 (0.73,1.65)	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.76 (0.60,0.96)	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	1.37 (1.20,1.56)	1.54 (0.78,3.02)	1.44 (0.73, 2.85)
Smoking			3.39 (1.26, 9.12)
Non-white ethnicity			2.66 (0.69,10.31)

\* Reference group; †  $U$  denotes smoking and ethnicity

Biased from unmeasured confounders?

## Analysis results using a single data source

Odds ratio (95% interval estimate)

	HES only (n=8969) (excludes $U^\dagger$ )	MCS only (n=824) (excludes $U^\dagger$ )	MCS only (n=824) (includes $U^\dagger$ )
Trihalomethanes			
> 60 $\mu$ g/L	1.39 (1.10,1.76)	2.06 (0.85,4.98)	1.87 (0.76, 4.62)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.10 (0.73,1.65)	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.76 (0.60,0.96)	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	1.37 (1.20,1.56)	1.54 (0.78,3.02)	1.44 (0.73, 2.85)
Smoking			3.39 (1.26, 9.12)
Non-white ethnicity			2.66 (0.69,10.31)

\* Reference group; †  $U$  denotes smoking and ethnicity

Lacks power to detect an association

## Analysis results using a single data source

Odds ratio (95% interval estimate)

	HES only (n=8969) (excludes $U^\dagger$ )	MCS only (n=824) (excludes $U^\dagger$ )	MCS only (n=824) (includes $U^\dagger$ )
Trihalomethanes			
> 60 $\mu$ g/L	1.39 (1.10,1.76)	<b>2.06 (0.85,4.98)</b>	<b>1.87 (0.76, 4.62)</b>
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.10 (0.73,1.65)	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.76 (0.60,0.96)	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	1.37 (1.20,1.56)	1.54 (0.78,3.02)	1.44 (0.73, 2.85)
Smoking			<b>3.39 (1.26, 9.12)</b>
Non-white ethnicity			<b>2.66 (0.69,10.31)</b>

\* Reference group; †  $U$  denotes smoking and ethnicity

**Some evidence of confounding**

# Outline

## Introduction

The problem of unmeasured confounding

Case study

## Approach 1: Building a Bayesian joint model

## Approach 2: Bayesian propensity score adjustment

## Notation and Objective

- Introducing some notation:
  - $Y$  - outcome, e.g. low birthweight
  - $X$  - exposure of interest, e.g. trihalomethane concentrations (THM)
  - $\mathbf{C}$  - vector of measured confounders, e.g. mother's age, baby gender, deprivation index
  - $\mathbf{U}$  - vector of partially measured confounders, e.g. smoking, ethnicity
- **Objective:** estimate the association between  $X$  and  $Y$  while controlling for  $(\mathbf{C}, \mathbf{U})$

We now compare two Bayesian approaches

## Approach 1: building a joint model

- Build a Bayesian joint model (BJM) consisting of
  - an analysis sub-model (to answer question of interest)
  - an imputation sub-model (to impute missing  $U$ )
- This is a single stage process in which the unknown parameters and missing data are estimated simultaneously
  - ensures consistency
  - all sources of uncertainty are automatically propagated
- By contrast, multiple imputation (MI) is a two stage process
  - first imputes the missing data
  - then analyses the completed datasets and pools results

## Specification of Bayesian joint model for case study

- Analysis model: Logit for  $P(Y|X, \mathbf{C}, \mathbf{U})$

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \mathbf{C}_i + \beta_U^T \mathbf{U}_i$$

- Imputation model: Multivariate Probit for  $P(U|X, \mathbf{C})$

$$\mathbf{U}_i^* \sim \text{MVN}(\boldsymbol{\mu}_i, \Sigma)$$

$$\boldsymbol{\mu}_i = \gamma_0 + \gamma_X X_i + \boldsymbol{\gamma}_C^T \mathbf{C}_i$$

$$U_{ij} = I(U_{ij}^* > 0), \quad j = 1, 2$$

$$\mathbf{U}_i^* = \begin{pmatrix} U_{i1}^* \\ U_{i2}^* \end{pmatrix}, \quad \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

## Accounting for the sampling bias in the MCS

- The supplementary data (MCS) is not a random sample from the primary data (HES)
- The MCS cohort is a stratified sample (oversamples low socio-economic and ethnic categories)
- Each outcome  $Y_i$  in the MCS cohort is associated with a stratum  $S(i)$  and a sampling weight  $w_i$
- We have implemented two approaches to account for this sampling bias
  1. include the stratum in the imputation model as stratum specific intercepts (i.e. replace  $\gamma_0$  with  $\gamma_{S(i)}$ )
  2. perform weighted imputation (i.e. replace  $\Sigma$  with  $\Sigma_i = \frac{1}{w_i} \Sigma$ )

## Analysis results using Approach 1

	Odds ratio (95% interval estimate)		
	HES only	HES+MCS (stratum adjusted)	HES+MCS (weight adjusted)
<b>Trihalomethanes</b>			
> 60 $\mu$ g/L	<b>1.39 (1.10,1.76)</b>	<b>1.17 (0.88,1.53)</b>	<b>1.20 (0.87,1.59)</b>
<b>Mother's age</b>			
≤ 25	1.14 (0.86,1.52)	1.02 (0.71,1.38)	0.99 (0.71,1.35)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.85 (0.57,1.21)	0.85 (0.57,1.20)
≥ 35	1.10 (0.73,1.65)	1.43 (0.88,2.21)	1.40 (0.86,2.16)
Male baby	0.76 (0.60,0.96)	0.76 (0.59,0.97)	0.76 (0.58,0.97)
Deprivation index	1.37 (1.20,1.56)	1.19 (1.01,1.38)	1.27 (1.10,1.47)
Smoking		3.91 (1.35,9.92)	3.97 (1.35,9.53)
Non-white ethnicity		3.56 (1.75,6.82)	4.11 (1.23,9.74)

\* Reference group

**Accounting for missing confounders has reduced OR of THM**

## Alternative imputation strategies

- Many imputation strategies for missing data do not use a fully Bayesian formulation but a variety of two-stage procedures
- Can be useful when full joint analysis difficult, but some bias can be expected
- For example, a ‘Feedforward’ strategy
  - performs successively  $P(U|X, \mathbf{C})$  then  $P(Y|X, \mathbf{C}, \mathbf{U})$
  - can be thought of as **cutting feedback from  $Y$  to  $\mathbf{U}$**
- Should modify the sampling distribution of  $\mathbf{U}$  to include  $Y$ 
  - performs successively  $P(U|X, \mathbf{C}, Y)$  then  $P(Y|X, \mathbf{C}, \mathbf{U})$

## Comparison with alternative imputation strategies

	Odds ratio (95% interval estimate)		
	Fully Bayesian joint model	Feedforward only (no response)	Feedforward only (with response)
Trihalomethanes			
> 60 $\mu$ g/L	1.17 (0.88,1.53)	1.33 (1.02,1.72)	1.24 (0.76,1.93)
Mother's age			
≤ 25	1.02 (0.71,1.38)	1.15 (0.85,1.52)	1.03 (0.71,1.45)
25 – 29*	1	1	1
30 – 34	0.85 (0.57,1.21)	0.82 (0.57,1.15)	0.85 (0.55,1.25)
≥ 35	1.43 (0.88,2.21)	1.16 (0.74,1.68)	1.36 (0.81,2.17)
Male baby	0.76 (0.59,0.97)	0.76 (0.59,0.96)	0.77 (0.55,1.05)
Deprivation index	1.19 (1.01,1.38)	1.34 (1.17,1.53)	1.22 (1.02,1.45)
Smoking	3.91 (1.35,9.92)	1.09 (0.78,1.48)	3.33 (1.40,6.49)
Non-white ethnicity	3.56 (1.75,6.82)	1.34 (0.92,1.87)	2.84 (1.01,6.27)

\* Reference group

Simple Feedforward provides inadequate adjustment  
Including Y is beneficial but some bias seems to remain

# Comparison of Bayesian models with MICE

	Odds ratio (95% interval estimate)		
	Fully Bayesian joint model	Feedforward only (with response)	MICE: 5 imputations (with response)
Trihalomethanes			
> 60 $\mu$ g/L	1.17 (0.88,1.53)	1.24 (0.76,1.93)	1.22 (0.91, 1.62)
Mother's age			
≤ 25	1.02 (0.71,1.38)	1.03 (0.71,1.45)	0.98 (0.69, 1.38)
25 – 29*	1	1	1
30 – 34	0.85 (0.57,1.21)	0.85 (0.55,1.25)	0.84 (0.58, 1.22)
≥ 35	1.43 (0.88,2.21)	1.36 (0.81,2.17)	1.32 (0.86, 2.03)
Male baby	0.76 (0.59,0.97)	0.77 (0.55,1.05)	0.73 (0.58, 0.93)
Deprivation index	1.19 (1.01,1.38)	1.22 (1.02,1.45)	1.23 (1.05, 1.44)
Smoking	3.91 (1.35,9.92)	3.33 (1.40,6.49)	4.01 (1.32,12.15)
Non-white ethnicity	3.56 (1.75,6.82)	2.84 (1.01,6.27)	2.73 (1.83, 4.09)

\* Reference group

MICE provides similar adjustment to Feedforward only

# Outline

## Introduction

The problem of unmeasured confounding

Case study

Approach 1: Building a Bayesian joint model

Approach 2: Bayesian propensity score adjustment

## Why consider Approach 2?

- Approach 1 becomes more difficult computationally as the dimension of  $\mathbf{U}$  increases
- Whereas approach 1 requires parametric assumptions about individual  $U$ , approach 2 does not
- So approach 2 is easier to extend to high numbers of missing confounders
- In approach 1 we model

$$P(Y|X, \mathbf{C}) = \int P(Y|X, \mathbf{C}, \mathbf{U})P(\mathbf{U}|X, \mathbf{C})d\mathbf{U}$$

- By contrast, in approach 2 we model

$$P(Y, X|\mathbf{C}) = \int P(Y|X, \mathbf{C}, \mathbf{U})P(X|\mathbf{U}, \mathbf{C})P(\mathbf{U}|\mathbf{C})d\mathbf{U}$$

## Specification of propensity score adjustment models

$$P(Y, X|\mathbf{C}) = \int P(Y|X, \mathbf{C}, \mathbf{U})P(X|\mathbf{U}, \mathbf{C})P(\mathbf{U}|\mathbf{C})d\mathbf{U}$$

- $P(Y, X|\mathbf{C}, \mathbf{U})$  is modelled using a pair of equations:

$$\text{logit}[P(Y = 1|X, \mathbf{C}, \mathbf{U})] = \beta_0 + \beta X + \xi_C^T \mathbf{C} + \xi_{UG}^T \{Z(\mathbf{U})\}$$

$$\text{logit}[P(X = 1|\mathbf{C}, \mathbf{U})] = \gamma_0 + \gamma_C^T \mathbf{C} + Z(\mathbf{U})$$

- $Z(\mathbf{U})$  is a scalar summary called the **conditional propensity score**
- $P(\mathbf{U}|\mathbf{C})$  is estimated using information from the supplementary data based on
  - either the empirical distribution of  $\mathbf{U}$
  - or by making simple parametric assumptions about the distribution of the propensity score
- see McCandless et al, 2009 for further details

# Comments on propensity score adjustment models

- The missing confounders are not imputed individually
- But summarised with a single summary score informed by the supplementary data
- This score breaks the dependence between the exposure and the missing confounders within levels of covariates
- Uncertainty in estimation of the coefficient of the propensity score on the supplementary data is propagated into the primary analysis

## Comparison of results from two approaches: full Bayes imputation v Bayesian propensity score adjustment

	Odds ratio (95% interval estimate)		
	HES only	Fully Bayesian (weight adjusted)	Bayesian propensity score adjustment
Trihalomethanes			
> 60 $\mu$ g/L	1.39 (1.10,1.76)	1.20 (0.87,1.59)	1.21 (0.91,1.60)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.99 (0.71,1.35)	1.14 (0.86,1.52)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.85 (0.57,1.20)	0.80 (0.58,1.14)
≥ 35	1.10 (0.73,1.65)	1.40 (0.86,2.16)	1.11 (0.74,1.67)
Male baby	0.76 (0.60,0.96)	0.76 (0.58,0.97)	0.76 (0.60,0.95)
Deprivation index	1.37 (1.20,1.56)	1.27 (1.10,1.47)	1.35 (1.19,1.55)
Smoking		3.97 (1.35,9.53)	
Non-white ethnicity		4.11 (1.23,9.74)	

\* Reference group

Both approaches reduce OR of THM

## Comparison of results from two approaches: full Bayes imputation v Bayesian propensity score adjustment

	Odds ratio (95% interval estimate)		
	HES only	Fully Bayesian (weight adjusted)	Bayesian propensity score adjustment
Trihalomethanes			
> 60 $\mu$ g/L	1.39 (1.10,1.76)	1.20 (0.87,1.59)	1.21 (0.91,1.60)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.99 (0.71,1.35)	1.14 (0.86,1.52)
25 – 29*	1	1	1
30 – 34	0.81 (0.57,1.15)	0.85 (0.57,1.20)	0.80 (0.58,1.14)
≥ 35	1.10 (0.73,1.65)	1.40 (0.86,2.16)	1.11 (0.74,1.67)
Male baby	0.76 (0.60,0.96)	0.76 (0.58,0.97)	0.76 (0.60,0.95)
Deprivation index	1.37 (1.20,1.56)	1.27 (1.10,1.47)	1.35 (1.19,1.55)
Smoking		3.97 (1.35,9.53)	
Non-white ethnicity		4.11 (1.23,9.74)	

\* Reference group

Other OR unchanged if they are correlated with **U**

# Summary

	Fully Bayesian	Propensity Score	Multiple Imputation
$X - Y$ relationship	✓	✓	✓
$\mathbf{C} - Y$ relationship	✓	✗	✓
coherency	✓	✓	✗
high dimension $\mathbf{U}$	✗	✓	✗

## Further Information and Acknowledgements

- Coming soon: paper on comparisons of different imputation strategies - see BIAS web site ([www.bias-project.org.uk](http://www.bias-project.org.uk))
  - Funding by ESRC: the BIAS project (PI N Best), based at Imperial College, London, is a node of the Economic and Social Research Council's National Centre for Research Methods (NCRM)
- ▶ [McCandless, L. C., Richardson, S., and Best, N. \(2009\).](#)  
Adjustment for Missing Confounders Using External Validation Data and Propensity Scores.  
[under revision for JASA \(available at \[www.bias-project.org.uk\]\(http://www.bias-project.org.uk\)\).](#)
- ▶ [Molitor, N.-T., Best, N., Jackson, C., and Richardson, S. \(2009\).](#)  
Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: Application to low birth-weight and water disinfection by-products.  
*Journal of the Royal Statistical Society, Series A*, **172**, (3), 615–37.