Bayesian graphical models for combining multiple data sources, with applications in environmental epidemiology

Sylvia Richardson[1]

sylvia.richardson@imperial.co.uk

Joint work with:

Alexina Mason[1], **Lawrence McCandless**[2] **& Nicky Best**[1]

Department of Epidemiology and Biostatistics, Imperial College London
Faculty of Health Sciences, Simon Fraser University, Canada

March 2010

Background

- The study of the influence of environmental risk factors on health is typically based on observational data
- Due to the nature of the research question, existing environmental contrasts (e.g. related to air pollution, water quality, ...) are commonly exploited in designs that link environmental measures with routinely collected administrative data
- Such data sources will typically have a limited number of variables for a large population, and might miss important confounders
- Exposure effect estimates will be biased without proper adjustment for confounders

The Problem of Unmeasured Confounding

Background:

- Environmental studies using large administrative databases and registries are commonly faced with confounding from unmeasured background variables

Possible solutions to adjust for unmeasured confounders:

- A: Source of prior information about unmeasured confounding?
  - Fully elicited versus use of additional data that contains more detailed information
- B: Possible analysis strategies?
  - Sensitivity analysis
  - Use of Bayesian hierarchical models to build a joint analysis of all data sources
  - Model-based versus semi-parametric

Information About Unmeasured Confounders

Use of Supplementary (enriched) Datasets

- We consider the situation where
  - Confounders are identified
  - Information about the unmeasured confounders may be available from additional datasets (e.g. surveys or cohort samples)
- We distinguish between the primary data versus the supplementary (enriched) data, which provide information about unmeasured confounders
- Analysis involves synthesis of multiple sources of empirical evidence
- This will require exchangeability assumptions ....
- Bayesian graphical models can be useful...

Case study: Water Disinfection By-Products and Risk of Low Birthweight

- Objective: To estimate the association between trihalomethane (THM) concentrations, a by-product of chlorine water disinfection potentially harmful for reproductive outcomes, and risk of full term low birthweight (<2.5kg)(Toledano, 2005).
- Information was collected for 8969 births between 2000 and 2001 in North West England, serviced by the United Utilities Water Company.
- Birth records obtained from the Hospital Episode Statistics (HES) data base were linked to estimated trihalomethane water concentrations using residence at birth and a model to estimate THM concentration from the water company monitored samples.
- First analysis in Molitor et al (2009)

The Primary Data: HES

- The primary data have the advantage of capturing information on all hospital births in the population under study.
  $\longrightarrow$ Increased power, fully representative
- However, they contain only limited information on the mother and infant characteristics which impact birth weight.
  $\longrightarrow$ Increased bias
- They contain data on mother's age, baby gender, gestational age and an index of deprivation, but no data on on maternal smoking or ethnicity.
  $\longrightarrow$ How to account for these?

Sources of Supplementary (enriched) Data

- The Millennium Cohort Study (MCS) contains survey information (stratified sample) on mothers and infants born during 2000-2001.
- Cohort births can be matched to the hospital data
- Contains detailed information on ethnicity, smoking, and other covariates, such as alcohol consumption, education, BMI.
- We combine information from the survey data with the hospital data using Bayesian hierarchical models.
  $\rightarrow$ treat unmeasured confounders as 'missing data'

Naive Analysis Results: Primary data (n=8969)

No adjustment for mother's smoking and ethnicity status

	Odds ratio (95% interval estimate) NAIVE
Trihalomethanes	
$> 60 \mu g/L$	1.39 (1.10,1.76)
Mother's age	
≤ 25	1.14 (0.86,1.52)
$25 - 29^{\star}$	1
$30 - 34$	0.81 (0.57,1.15)
≥ 35	1.10 (0.73,1.65)
Male baby	0.76 (0.60,0.96)
Deprivation index	1.37 (1.20,1.56)

* Reference group

$\longrightarrow$ Biased from unmeasured confounding?

Analysis of Supplementary MCS data only (n=824)

	Odds ratio (95% interval estimate)	
	MCS data	MCS data
Trihalomethanes		
$> 60 \mu g/L$	2.06 (0.85,4.98)	1.87 (0.76, 4.62)
Mother's age		
≤ 25	0.65 (0.23,1.79)	0.57 (0.20, 1.61)
$25 - 29^{\star}$	1	1
$30 - 34$	0.13 (0.02,1.11)	0.13 (0.02, 1.11)
≥ 35	1.57 (0.49,5.08)	1.82 (0.55, 5.99)
Male baby	0.59 (0.25,1.43)	0.62 (0.25, 1.49)
Deprivation index	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

Overall Objectives

- Building models that can link various sources of data containing different sets of covariates
  - to fit a common regression model
  - and to account adequately for uncertainty arising from missing or partially observed confounders in large data bases
- Investigating alternative formulations of imputation and adjustment for unknown confounders

Bayesian hierarchical models (BHM)

- Bayesian graphical models provide a coherent way to connect local sub-models based on different datasets into a global unified analysis.
- BHM allow propagation of information between the model components following the graph
- In the case of missing confounders, several decomposition of the marginal likelihood can be used, as well as different imputation strategies
  - Lead to different ways for information propagation or feedback between the model components
- Modularity helps our understanding of assumptions made when adjusting for missing confounders

Adjustment for Multiple Unmeasured Confounders

Variables and Notation

Introducing some notation:

- Let Y denote an outcome, e.g. low birthweight
- Let X denote the exposure of interest, e.g. THM
- Let C denote a vector of measured confounders, e.g. mother's age, baby gender, deprivation
- Let U denote a vector of partially measured confounders, e.g. smoking, ethnicity. *Note that covariates in U are identified but might be missing.*

- 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

- Usual approach (1): 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 given (X, C).

- Alternative approach (2):

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

  This follows *propensity score ideas* for assessing the 'causal' effect of X on Y.

Adjustment for Multiple Unmeasured Confounders

Modelling $P(U \mid X, C)$ – Approach (1)

- Outcome model:

$$\text{Logit}[P(Y = 1|X, C, U)] \quad = \quad \alpha + \beta_X X + \xi_C^T \boldsymbol{C} + \Psi_U^T \boldsymbol{U}$$

- Imputation model: Multivariate Probit for $P(U \mid X, C)$

$$\boldsymbol{U}^\star \sim MVN(\boldsymbol{\mu}, \Sigma)$$

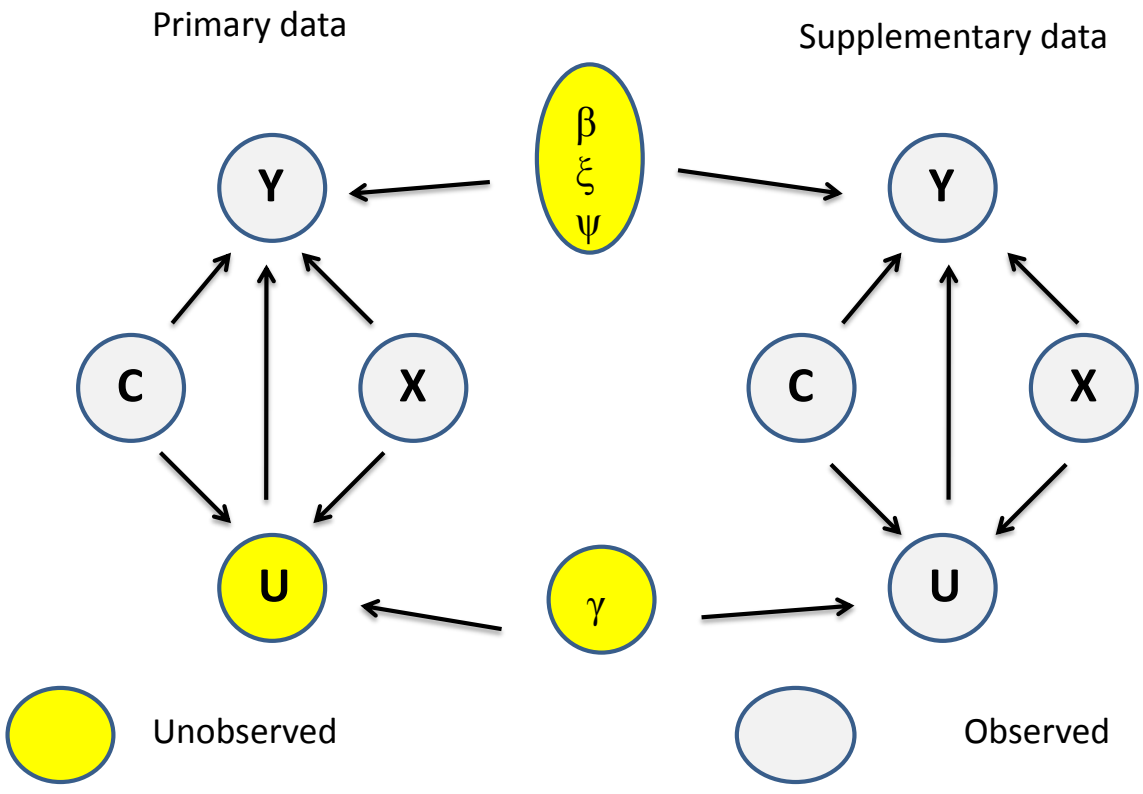
$$\boldsymbol{\mu} = \gamma_0 + \gamma_X X + \boldsymbol{\gamma_C}^T \boldsymbol{C}$$

$$\boldsymbol{U^*} = \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

$$U_j = I(U_j^* > 0), j = 1, 2$$

A graphical representation of the fully Bayesian model

- Joint estimation in primary and supplementary data
- The supplementary data informs the imputation model
- The uncertainty on *U* is propagated coherently



Exchangeability assumptions

Accounting for sampling bias

- It is often the case that the supplementary data is not a random sample from the primary data
- Assumptions of exchangeability that underline the BHM model synthesis will not hold
- Need to include additional modelling of the sampling of supplementary data to render both sources of data exchangeable
- In our case study, the MCS cohort sampling was stratified in order to oversample in the UK low socio-economic categories

Accounting for sampling bias in MCS cohort

- Each outcome Y_i in the MCS cohort is associated with a stratum S_i as well as a sampling weight.
- We have implemented two approaches to account for the stratified sampling
  - Include the stratum S in the imputation model equation:

$$P(U \mid X, C) \longrightarrow P(U \mid X, C, S)$$

  - Perform weighted imputation, i.e. replace Σ by

$$\Sigma_i = w_i \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix} = \begin{pmatrix} w_i & w_i\kappa \\ w_i\kappa & w_i \end{pmatrix}$$

where $w_i = \frac{1}{\text{weight}_i}$

Comparison of naive and fully Bayesian analysis

	Odds ratio (95% interval estimate)		
	NAIVE	Fully Bayesian (stratum adjusted)	Fully Bayesian (weight adjusted)
Trihalomethanes			
$> 60 \mu g/L$	1.39 (1.10,1.76)	1.17 (0.88,1.53)	1.20 (0.87,1.59)
Mother's age			
≤ 25	1.14 (0.86,1.52)	1.02 (0.71,1.38)	0.99 (0.71,1.35)
$25 - 29^{\star}$	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

Approximations to approach (1)

- 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
- In a 'Feedforward' strategy, perform successively
  - $P(U \mid X, C)$, then
  - $P(Y \mid X, C, U)$
  - This can be thought of as cutting feedback from Y to U (and implemented using e.g. the cut-function in Winbugs)
- Alternatively, modify the sampling distribution of U to include Y to perform Bayesian multiple imputation
  - $P(U \mid X, C, Y)$, then
  - $P(Y \mid X, C, U)$

Comparison of full Bayes with alternative strategies

	Odds ratio (95% interval estimate)		
	Fully Bayesian	Feedforward only (no response)	Feedforward only (with response)
Trihalomethanes			
$> 60 \mu g/L$	1.20 (0.87,1.59)	1.38 (1.06,1.78)	1.28 (0.77,2.03)
Mother's age			
≤ 25	0.99 (0.71,1.35)	1.14 (0.85,1.51)	0.98 (0.66,1.40)
$25 - 29^{\star}$	1	1	1
$30 - 34$	0.85 (0.57,1.20)	0.82 (0.57,1.15)	0.86 (0.55,1.29)
≥ 35	1.40 (0.86,2.16)	1.13 (0.73,1.66)	1.45 (0.84,2.35)
Male baby	0.76 (0.58,0.97)	0.76 (0.59,0.97)	0.77 (0.53,1.07)
Deprivation index	1.27 (1.10,1.47)	1.37 (1.20,1.56)	1.28 (1.10,1.50)
Smoking	3.97 (1.35,9.53)	1.10 (0.79,1.47)	4.83 (1.94,10.10)
Non-white ethnicity	4.11 (1.23,9.74)	1.14 (0.65,1.78)	3.41 (0.63,9.15)

* Reference group

Simple Feedforward provides inadequate adjustment.

Including Y is beneficial but some bias seems to remain.

Adjustement for multiple confounders through Bayesian propensity score

Approach (2)

- In this approach we model the conditional density $P(Y, X|C, U)$ using a pair of equations:

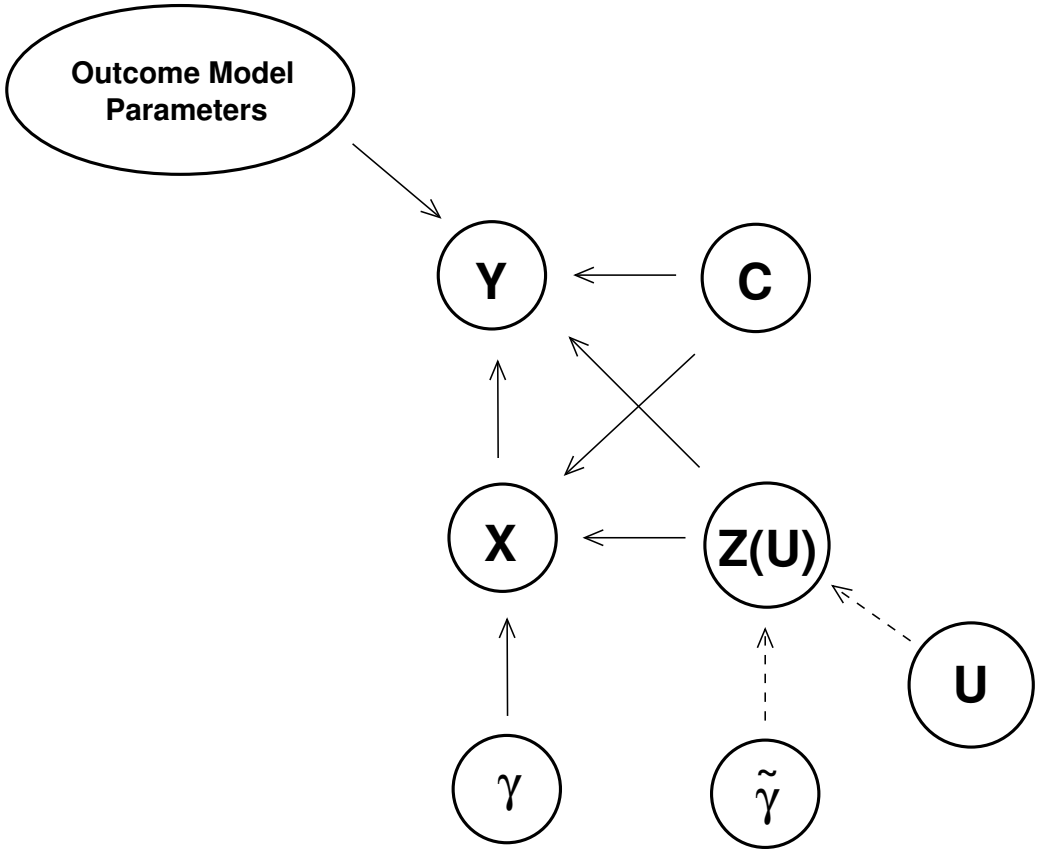
$$\begin{aligned} \text{Logit}[P(Y = 1|X, C, U)] &= \alpha + \beta X + \xi^T C + \tilde{\Psi}^T g\{Z(U)\}^T \\ \text{Logit}[P(X = 1|C, U)] &= \gamma_0 + \gamma^T C + \tilde{\gamma}^T U \end{aligned}$$

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

- One can show that there is no unmeasured confounding of the $Y - X$ association conditional on $C, Z(U)$.
- In general, the quantity $g\{Z(U)\}$ is a semi-parametric linear predictor with regression coefficients $\tilde{\Psi}$. Its link to Y has to be modelled flexibly, e.g using natural splines.

A graphical representation of role of the propensity score

- We include the scalar summary $Z(U)$ in the outcome model



Bayesian propensity score with missing data

- Recall:

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

- To complete specification, require a model for $P(U|C)$:
  - Assume U and C marginally independent and use the empirical distribution of U from the supplementary data.
  - This gives a model for $P(Y, X|C) = E\{P(Y, U, X|C)\}$
- We may approximate:

$$P(Y, X|C) \approx \frac{1}{m}\sum_{j=1}^{m} P(Y|X, C, U_j)P(X|U_j, C).$$

for $j = 1, \ldots, m$ in the Supplementary data
[Weighting can be included in the summation to account for the stratified sampling]

Contrasting propensity score conditioning with imputation

- A major benefit of this approach is that it can easily extend when dim(U)>2, whereas a multivariate probit imputation model will become difficult to implement with a high dimensional U
- The Us are not imputed but their empirical distribution in the supplementary data is used
- As before, a joint model of the primary and supplementary data is used

$\Rightarrow$ Uncertainty in estimation of the coefficient of the propensity score on the supplementary data is propagated into the primary analysis

- Note that effects of other covariates on Y will not necessarily be estimated without bias if they are correlated with U.

Comparison of full Bayes imputation with Bayesian propensity score adjustment

	Odds ratio (95% interval estimate)		
	Bayesian propensity score adjustment	Fully Bayesian	Feedforward only no response
Trihalomethanes			
$> 60 \mu g/L$	1.21 (0.91,1.60)	1.20 (0.87,1.59)	1.38 (1.06,1.78)
Mother's age			
≤ 25	1.14 (0.86,1.52)	0.99 (0.71,1.35)	1.14 (0.85,1.51)
$25 - 29^{\star}$	1	1	1
$30 - 34$	0.80 (0.58,1.14)	0.85 (0.57,1.20)	0.82 (0.57,1.15)
≥ 35	1.11 (0.74,1.67)	1.40 (0.86,2.16)	1.13 (0.73,1.66)
Male baby	0.76 (0.60,0.95)	0.76 (0.58,0.97)	0.76 (0.59,0.97)
Deprivation index	1.35 (1.19,1.55)	1.27 (1.10,1.47)	1.37 (1.20,1.56)
Smoking		3.97 (1.35,9.53)	1.10 (0.79,1.47)
Non-white ethnicity		4.11 (1.23,9.74)	1.14 (0.65,1.78)

Implementation of Bayesian propensity score adjustment with an extended set of confounders

Smoking, ethnicity + Education, lone parent, alcohol consumption, BMI

	Odds ratio (95% interval estimate) Bayesian propensity score adjustment	
	2 missing confounders	6 missing confounders
Trihalomethanes		
$> 60 \mu g/L$	1.21 (0.91,1.60)	1.23 (0.92,1.60)
Mother's age		
≤ 25	1.14 (0.86,1.52)	1.13 (0.85,1.53)
$25 - 29^{\star}$	1	1
$30 - 34$	0.80 (0.58,1.14)	0.80 (0.56,1.15)
≥ 35	1.11 (0.74,1.67)	1.11 (0.73,1.64)
Male baby	0.76 (0.60,0.95)	0.75 (0.59,0.95)
Deprivation score	1.35 (1.19,1.55)	1.35 (1.19,1.53)

* Reference group

Adjustment for Multiple Unmeasured Confounders

Summary

- Exposure effect estimate is driven towards zero once the important confounding effects of mother's smoking and ethnicity are taken into account
- The Bayesian propensity score approach can effectively adjust the effect of X for multiple unmeasured confounders
- We found very good concordance between the two approaches for the $X - Y$ relationship
- The confounding effect of U (smoking and ethnicity) on the $C - Y$ relationship (mother's age, deprivation) is only captured adequately by the fully Bayesian imputation
- Cutting feedback creates some bias, including the response in the imputation model partially mitigates this.

Conclusion

- Adjustment for unmeasured confounding in environmental studies is feasible through the use of additional data sources (e.g. surveys, cohorts, validation subgroup ...)
- Bayesian methods can be flexibly adapted to synthesize information across of range of additional data sources, e.g. can incorporate additional sources of data such as area-level census variables in the imputation model
- One precaution is that we must be careful to study exchangeability assumptions between data sets and account for any sampling weights or stratification in the imputation model
- Among the methods available, propensity score presents a computationally feasible alternative when faced with many unmeasured confounders.

Acknowledgments

- 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).
- For papers and technical reports, see our web site www.bias-project.org.uk