ESRC National Centre for Research Methods

BIAS

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

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

Nicky Best[1], Alexina Mason[1],
Sylvia Richardson[1] and Lawrence McCandless[2]

[1] Department of Epidemiology and Biostatistics, Imperial College London, UK
[2] Faculty of Health Sciences, Simon Fraser University, Canada

4th International IMS/ISBA Joint Meeting, 2011

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

Unmeasured Confounding in Epidemiological Studies

- 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 (e.g. disease registers, hospital admissions,...)
- 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

Dealing with Unmeasured Confounding

Problem:

- Administrative databases are important source of data on health and socioeconomic outcomes on large populations, but they typically lack detailed information on potential confounding variables

Possible solutions to adjust for unmeasured confounders:

- Sensitivity analysis
  - requires prior information (fully elicited, ‘plausible bounds’, previous studies) about effects of unmeasured confounding
- Treat as missing data problem
  - requires use of additional data that contains more detailed information

In this talk, I will discuss the missing data perspective

Information About Unmeasured Confounders

Use of Supplementary (enriched) Datasets

- We consider the situation where
  - Confounders are identified but not measured in main database
  - 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 (compatibility) assumptions between different data sources....

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 (LBW; <2.5kg)
- Information was collected for 8780 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 THM water concentrations using postcode of residence at birth and a model to estimate THM concentration from the water company monitored samples.

The Primary Data: HES

- The primary data have the advantage of capturing information on all hospital births in the population under study.
  → Good statistical power, fully representative
- However, they contain only limited information on the mother and infant characteristics which impact birth weight.
  → Potentially biased
- 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.
  → How to account for unmeasured confounders?

Sources of Supplementary (enriched) Data

- The Millennium Cohort Study (MCS) contains survey information (stratified sample) on mothers and infants born during 2000-2001.
- 824 cohort births in study region — can be matched to the hospital data using postcode, sex, DOB
- Contains detailed information on maternal 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.
  → treat unmeasured confounders as 'missing data'
  → approx 90% births have missing data

Summary of Data Sources

	Primary data (*n*=7956)		Supplementary data (*n*=824)	
	THM$>60_{\mu\text{g/L}}$	THM$\leq60_{\mu\text{g/L}}$	THM$>60_{\mu\text{g/L}}$	THM$\leq60_{\mu\text{g/L}}$
LBW	144 (3.8%)	130 (3.1%)	14 (4.0%)	9 (1.9%)
Maternal age	27.9 ± 6.1	27.3 ± 6.0	27.8 ± 6.2	28.1 ± 5.9
Male baby	1956 (52%)	2076 (50%)	176 (51%)	254 (53%)
Deprivation index	4.1 ± 1.3	4.3 ± 1.2	4.0 ± 1.1	4.0 ± 1.2
Maternal smoking			126 (36%)	181 (38%)
Non-white ethnicity			77 (22%)	48 (10%)

Analysis results using a single data source

	Odds ratio (95% interval estimate)		
	HES only (n=8969)	MCS only (n=824)	MCS only (n=824)
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)
$25 - 29^{\star}$	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

Biased from unmeasured confounders?

Analysis results using a single data source

	Odds ratio (95% interval estimate)		
	HES only (n=8969)	MCS only (n=824)	MCS only (n=824)
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)
$25 - 29^{\star}$	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

Lacks power to detect an association

Analysis results using a single data source

	Odds ratio (95% interval estimate)		
	HES only (n=8969)	MCS only (n=824)	MCS only (n=824)
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)
$25-29^{\star}$	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

Some evidence of confounding

Modelling the Unmeasured Confounders

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
- We compare
  - Fully Bayesian joint model for the outcomes and missing confounders
  - Alternative two-stage imputation strategies
  - Bayesian propensity score adjustment for missing confounders

Bayesian graphical models

- 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 decompositions 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

Variables and Notation

Introducing some notation:

- Y - outcome, e.g. low birthweight
- X - exposure of interest, e.g. THM concentrations
- $\boldsymbol{C}$ - vector of fully measured confounders, e.g. mother's age, baby gender, deprivation index
- $\boldsymbol{U}$ - vector of partially measured confounders, e.g. smoking, ethnicity. *Note that covariates in U are identified but might be missing.*
- Objective: estimate the association between X and Y while controlling for $(\boldsymbol{C}, \boldsymbol{U})$

We now compare three Bayesian approaches

Approach 1: Bayesian 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 $\boldsymbol{U}$)
- 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

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

This strategy requires modelling distributional assumptions about $\boldsymbol{U}$ given $(X, \boldsymbol{C})$.

Specification of Bayesian joint model for case study

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

$$Y_i \sim Bernoulli(p_i), \quad \text{baby } i$$
$$logit(p_i) = \beta_0 + \beta_X X_i + \boldsymbol{\beta}_C^T \boldsymbol{C}_i + \boldsymbol{\beta}_U^T \boldsymbol{U}_i$$

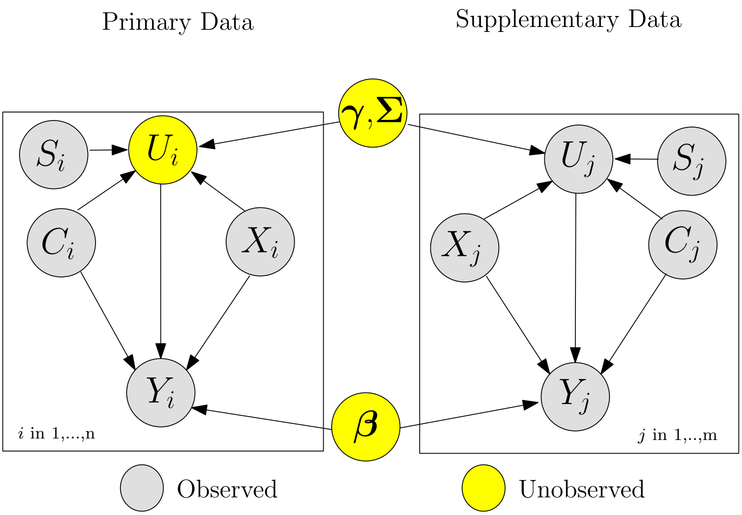
- Imputation model: Multivariate Probit for $P(U|X, \boldsymbol{C})$

$$\boldsymbol{U}_i^\star \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$
$$\boldsymbol{\mu}_i = \gamma_0 + \gamma_X X_i + \gamma_C^T \boldsymbol{C_i}$$
$$U_{iq} = I(U_{iq}^\star > 0), \; q = 1, 2$$
$$\boldsymbol{U}_i^\star = \begin{pmatrix} U_{i1}^\star \\ U_{i2}^\star \end{pmatrix}, \; \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \; \Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

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_j in the MCS cohort is associated with a stratum S_j and a sampling weight w_j
- 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 γ_0 with γ_{s_j})
  2. perform weighted imputation (i.e. replace Σ with $\Sigma_j = \frac{1}{w_j}\Sigma$)
- For clustered sampling designs, could include cluster random effects in imputation model

Graphical representation of Joint Bayesian Model



Analysis results using Approach 1

	Odds ratio (95% interval estimate)		
	HES only	HES+MCS (stratum adjusted)	HES+MCS (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

Approach 2: Two-stage imputation strategies

- Many imputation strategies for missing data do not use a fully Bayesian formulation, but a variety of two-stage procedures to approximate a fully Bayesian model
- Can be useful when full joint analysis is difficult computationally, but some bias can be expected

Multiple Imputation

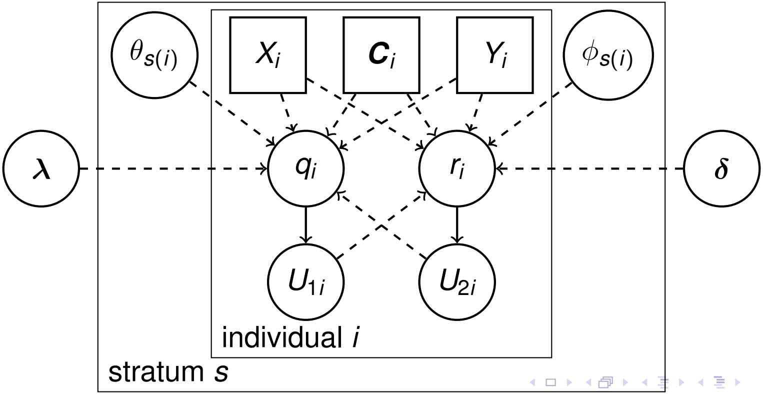
- Multiple Imputation, MI (Rubin, 1978, 1987) is a widely used two-stage procedure for imputing missing data
  - first impute the missing data, $P(\boldsymbol{U}|X, \boldsymbol{C}, Y, S)$
  - then analyse the completed datasets, $P(Y|X, \boldsymbol{C}, \hat{\boldsymbol{U}}_k), \quad k = 1, ..., K$, and pool results
- Rubin justifies MI as an approximate Bayesian procedure if the imputations ($\hat{\boldsymbol{U}}_k$) are draws from a posterior predictive distribution for the missing data given the observed data (and a suitable model)
- Notice that imputation model needs to include all variables related to the missing variables (including response, Y) and stratum variables S related to missingness
- Most of the practical issues with MI concern the choice of, and draws from, the imputation model

Multiple Imputation

- When $\boldsymbol{U}$ is multivariate and includes categorical variables, drawing from fully defined joint distribution, $P(\boldsymbol{U}|X, \boldsymbol{C}, Y, S)$, can be difficult in practice
    - One alternative is to iterate between a set of univariate conditional distributions $P(U_q|X, \boldsymbol{C}, Y, S, \boldsymbol{U}_{\setminus q}), \quad q = 1, ...Q,$
    - Implemented in, e.g. MICE (van Buuren)
    - Convergence to valid joint posterior not guaranteed

Graphical representation for MICE approach

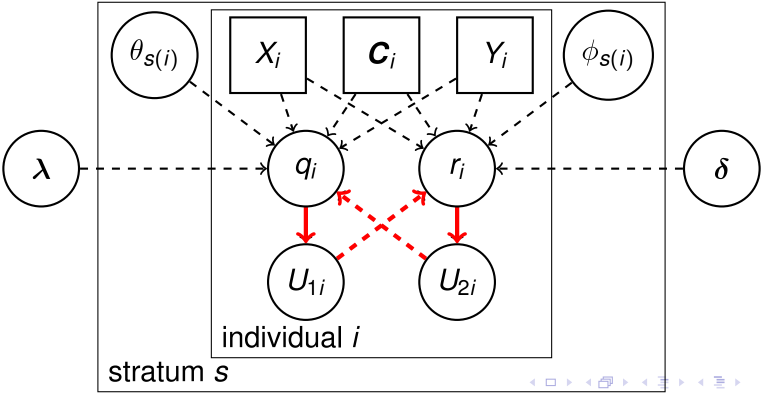
$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$



Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

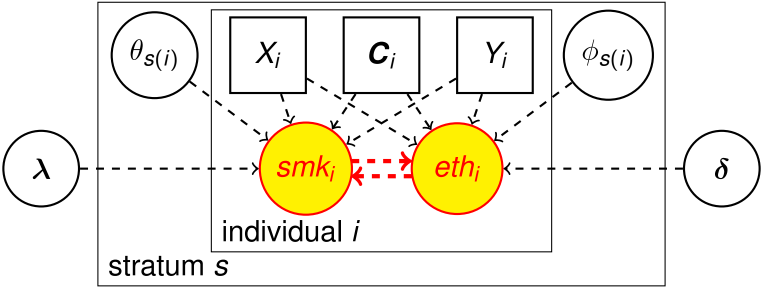
We have a cycle so diagram is NOT a DAG!



Graphical representation for MICE approach

$$U_{1i} \sim Bernoulli(q_i)$$
$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$
$$U_{2i} \sim Bernoulli(r_i)$$
$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

We have a cycle so diagram is NOT a DAG!



We iterate between 2 parts of imputation model, then fit analysis model

Graphical representation for MICE approach

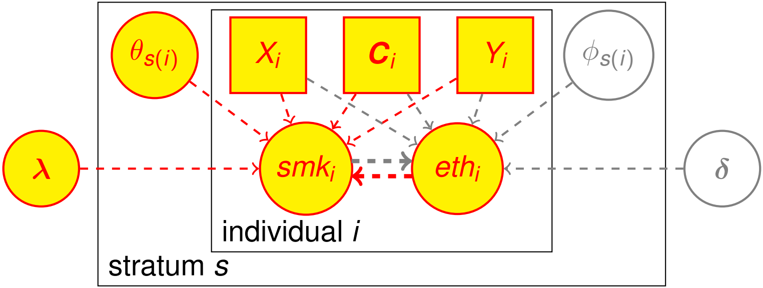
$$U_{1i} \sim Bernoulli(q_i)$$

$$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \boldsymbol{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$$

$$U_{2i} \sim Bernoulli(r_i)$$

$$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \boldsymbol{C_i} + \delta_U U_{1i} + \delta_Y Y_i$$

We have a cycle so diagram is NOT a DAG!



We iterate between 2 parts of imputation model, then fit analysis model

Graphical representation for MICE approach

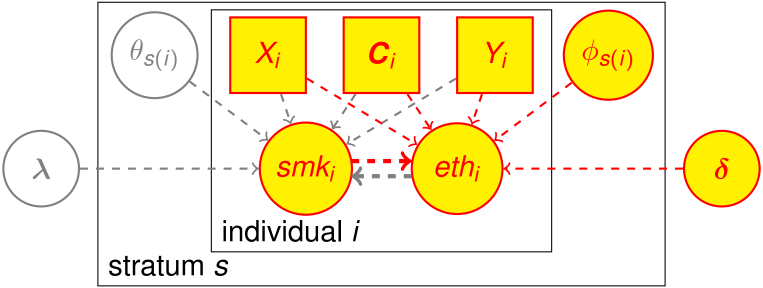
$U_{1i} \sim Bernoulli(q_i)$

$logit(q_i) = \theta_{s(i)} + \lambda_X X_i + \boldsymbol{\lambda}_C^T \mathbf{C_i} + \lambda_U U_{2i} + \lambda_Y Y_i$

$U_{2i} \sim Bernoulli(r_i)$

$logit(r_i) = \phi_{s(i)} + \delta_X X_i + \boldsymbol{\delta}_C^T \mathbf{C_i} + \delta_U U_{1i} + \delta_Y Y_i$

We have a cycle so diagram is NOT a DAG!

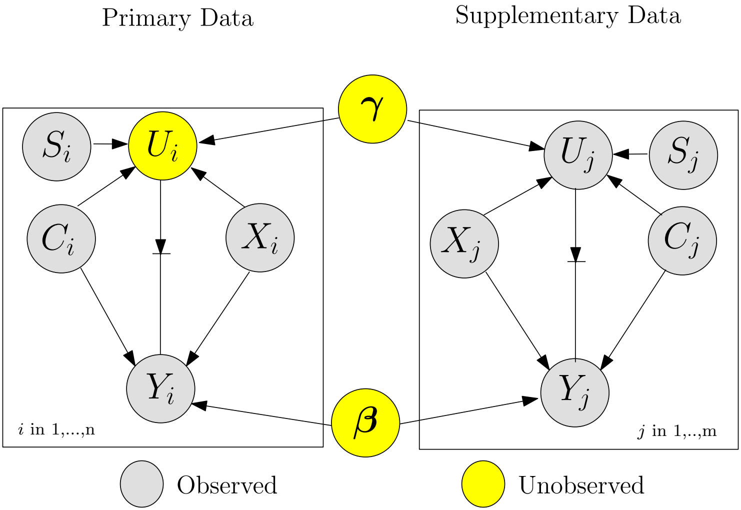


We iterate between 2 parts of imputation model, then fit analysis model

‘Feedforward’ imputation strategy

- A related idea with flavours of MI and JBM is to fit an approximate JBM using a ‘Feedforward’ strategy
    - performs successively $P(\boldsymbol{U}|X, \boldsymbol{C}, S)$ then $P(Y|X, \boldsymbol{C}, \boldsymbol{U})$ within same MCMC run
    - can be thought of as cutting feedback from Y to $\boldsymbol{U}$ (and implemented using e.g. the cut-function in Winbugs)

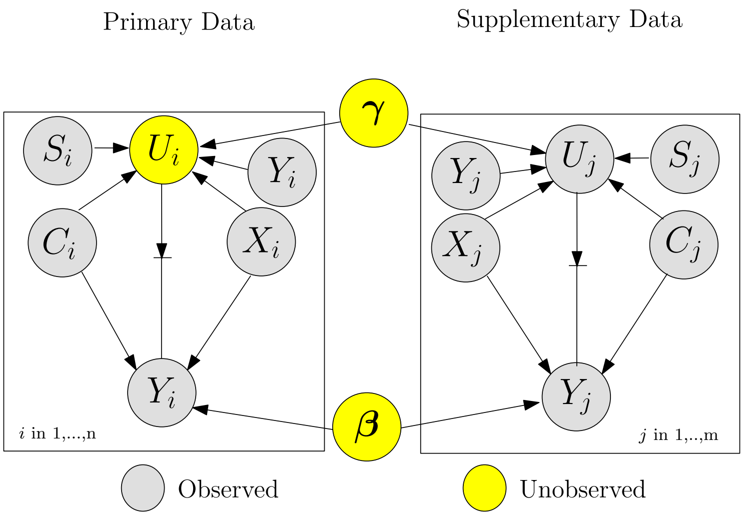
Graphical representation of ‘Feedforward’ Model



‘Feedforward’ imputation strategy

- A related idea with flavours of MI and JBM is to fit an approximate JBM using a ‘Feedforward’ strategy
  - performs successively $P(\boldsymbol{U}|X, \boldsymbol{C}, S)$ then $P(Y|X, \boldsymbol{C}, \boldsymbol{U})$ within same MCMC run
  - can be thought of as cutting feedback from Y to $\boldsymbol{U}$ (and implemented using e.g. the cut-function in Winbugs)
- Should modify the sampling distribution of $\boldsymbol{U}$ to include Y
  - performs successively $P(\boldsymbol{U}|X, \boldsymbol{C}, S, Y)$ then $P(Y|X, \boldsymbol{C}, \boldsymbol{U})$

Graphical representation of ‘Feedforward’ Model



‘Feedforward’ imputation strategy

- A related idea with flavours of MI and JBM is to fit an approximate JBM using a ‘Feedforward’ strategy
  - performs successively $P(\boldsymbol{U}|X, \boldsymbol{C}, S)$ then $P(Y|X, \boldsymbol{C}, \boldsymbol{U})$ within same MCMC run
  - can be thought of as cutting feedback from Y to $\boldsymbol{U}$ (and implemented using e.g. the cut-function in Winbugs)
- Should modify the sampling distribution of $\boldsymbol{U}$ to include Y
  - performs successively $P(\boldsymbol{U}|X, \boldsymbol{C}, S, Y)$ then $P(Y|X, \boldsymbol{C}, \boldsymbol{U})$
- Imputation model: multivariate imputation of $\boldsymbol{U}$
- No need for MI combining rules as sampled U’s are fed automatically into analysis model at each MCMC iteration
- For normal linear analysis model with vague priors, feedforward model is equivalent to JBM
- Main advantage is that cutting feedback can improve computational efficiency and robustness

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)
$25 - 29^\star$	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/inefficiency seems to remain

Comparison of Bayesian models with MICE

	Odds ratio (95% interval estimate)		
	Fully Bayesian joint model	Feedforward only (with response)	MICE: 20 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)
$25 - 29^{\star}$	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,
but with narrower intervals

Approach 3: Bayesian propensity score adjustment

- JBM and MI become more difficult computationally as the dimension of $\boldsymbol{U}$ increases
- JBM and MI require parametric assumptions about $\boldsymbol{U}$ that can be difficult to verify
- In approach 3, we attempt to overcome these difficulties using a propensity score approach
- In approaches 1 and 2, we model

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

- By contrast, in approach 3 we model

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

or, more precisely $\int P(Y|X,\boldsymbol{C},Z(\boldsymbol{U}))P(X|Z(\boldsymbol{U}),\boldsymbol{C})P(Z(\boldsymbol{U})|\boldsymbol{C})dZ(\boldsymbol{U})$

Specification of the propensity score model

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

$$
\begin{aligned}
logit[P(Y = 1|X, \boldsymbol{C}, Z(\boldsymbol{U}))] &= \beta_0 + \beta_X X + \beta_C^T \boldsymbol{C} + \beta_U^T g\{Z(\boldsymbol{U})\} \\
logit[P(X = 1|\boldsymbol{C}, Z(\boldsymbol{U}))] &= \gamma_0 + \gamma_C^T \boldsymbol{C} + \gamma_U^T \boldsymbol{U}
\end{aligned}
$$

- The scalar quantity $Z(\boldsymbol{U}) = \gamma_U^T \boldsymbol{U}$ is called the conditional propensity score (conditional on $\boldsymbol{C}$)
- Can show that there is no unmeasured confounding of the $Y - X$ association conditional on $\boldsymbol{C}, Z(\boldsymbol{U})$
- In general, the quantity $g\{Z(\boldsymbol{U})\}$ is a semi-parametric linear predictor with regression coefficients β_U. Its link to Y has to be modelled flexibly, e.g using natural splines.

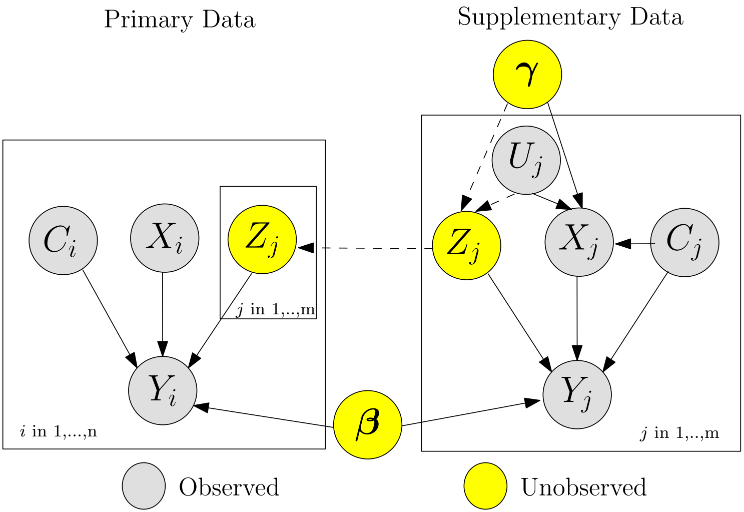
Bayesian propensity score with missing data

- Recall $P(Y,X|\boldsymbol{C}) = \int P(Y|X,\boldsymbol{C},Z)P(X|Z,\boldsymbol{C})P(Z|\boldsymbol{C})dZ$ where, for notational convenience, $Z = Z(\boldsymbol{U})$
- To complete specification, require a model for $P(Z|\boldsymbol{C})$
- Assumption 1: $\boldsymbol{U}$ and $\boldsymbol{C}$ marginally independent
  - We may approximate:

    $$P(Y,X|\boldsymbol{C}) = E_Z\{P(Y,X,Z|\boldsymbol{C})\} \approx \frac{1}{m}\sum_{j=1}^{m} P(Y|X,\boldsymbol{C},Z_j)P(X|Z_j,\boldsymbol{C})$$

    where $\{Z_j|j=1,\ldots,m\}$ is the empirical distribution of the estimated propensity score in the Supplementary data
  - Weighting can be included in the summation to account for the stratified sampling
  - Requires no parametric assumptions about distribution of $\boldsymbol{U}$ or Z

Graphical representation of Propensity Score Model



Bayesian propensity score with missing data

- Assumption 2: ***U*** and ***C*** *not* marginally independent
  - Now require $P(Y, X|\boldsymbol{C}) = E_{Z|\boldsymbol{C}}\{P(Y, X, Z|\boldsymbol{C})\}$
  - If sample size of Supplemental data is sufficiently large and ***C*** low dimensional:
    - could stratify empirical distribution of $\{Z_j|j = 1, \ldots, m\}$ by ***C***-strata
    - then estimate expectation by empirical summation as before

Bayesian propensity score with missing data

- Supplementary data sample typically not large enough to stratify

→ Fit simple univariate parametric model to estimate conditional distribution of $Z|\boldsymbol{C}$ in Supplementary data

$$Z|\boldsymbol{C} \sim N(\hat{\boldsymbol{\theta}}^T\boldsymbol{C}, \hat{\sigma}^2)$$

where $\hat{\boldsymbol{\theta}}^T\boldsymbol{C} = \hat{\theta}_0 + \hat{\theta}_1 C_1 + ... + \hat{\theta}_p C_p$ is the estimated mean propensity score conditional on $\boldsymbol{C}$

- $(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ are ML point estimates obtained from preliminary regression analysis of Supplementary data
- We then approximate

$$E_{Z|\boldsymbol{C}}\{P(Y, X, Z|\boldsymbol{C})\} \approx \frac{1}{m}\sum_{k=1}^{K}\omega_k P(Y|X, \boldsymbol{C}, Z_{\boldsymbol{C},k})P(X|Z_{\boldsymbol{C},k}, \boldsymbol{C})$$

where $(Z_{\boldsymbol{C},k}, \omega_k)$ is a histogram approximation to the Normal distribution $N(\hat{\boldsymbol{\theta}}^T\boldsymbol{C}, \hat{\sigma}^2)$

Contrasting propensity score conditioning with imputation

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

⇒ Uncertainty in estimation of propensity score coefficient in the supplementary data is propagated into the primary analysis

  - but, ignores parameter uncertainty in conditional distribution of propensity score in case of non-independence of $\boldsymbol{U}$ and $\boldsymbol{C}$

- 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 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)
$25 - 29^\star$	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; † $P(\boldsymbol{U}|\boldsymbol{C}) = P(\boldsymbol{U})$

Both approaches reduce OR of THM

Comparison of full Bayes imputation v Bayesian propensity score adjustment

	Odds ratio (95% interval estimate)		
	HES only	Fully Bayesian	Bayesian propensity
		(weight adjusted)	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)
$25-29^{\star}$	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; † $P(\boldsymbol{U}|\boldsymbol{C}) = P(\boldsymbol{U})$

Other OR unchanged if they are correlated with $\boldsymbol{U}$

Bayesian propensity score adjustment: Comparison of assumptions about $P(\boldsymbol{U}|\boldsymbol{C})$

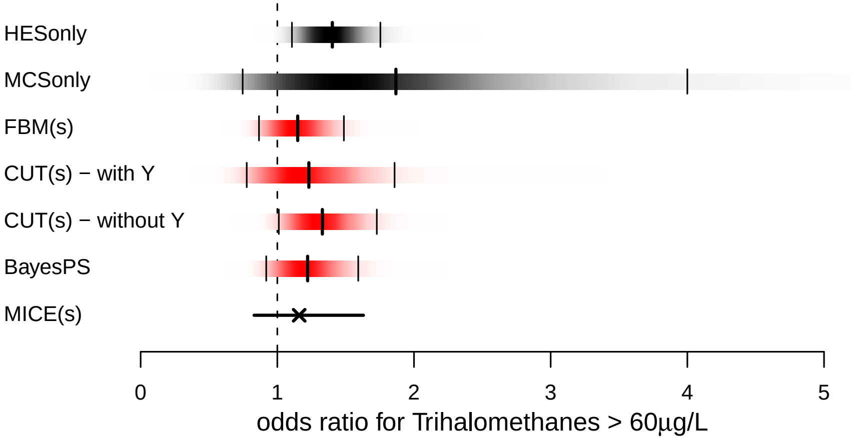
	Odds ratio (95% interval estimate)		
	Fully Bayesian (weight adjusted)	Bayesian propensity score adjustment $P(\boldsymbol{U}\mid\boldsymbol{C}) = P(\boldsymbol{U})$	Bayesian propensity score adjustment $P(\boldsymbol{U}\mid\boldsymbol{C}) \neq P(\boldsymbol{U})$
Trihalomethanes			
$> 60\mu g/L$	1.20 (0.87,1.59)	1.21 (0.91,1.60)	1.23 (0.94, 1.64)
Mother's age			
≤ 25	0.99 (0.71,1.35)	1.14 (0.86,1.52)	1.13 (0.85, 1.46)
$25-29^{\star}$	1	1	1
$30-34$	0.85 (0.57,1.20)	0.80 (0.58,1.14)	0.82 (0.59, 1.13)
≥ 35	1.40 (0.86,2.16)	1.11 (0.74,1.67)	1.18 (0.79, 1.75)
Male baby	0.76 (0.58,0.97)	0.76 (0.60,0.95)	0.75 (0.58, 0.94)
Deprivation index	1.27 (1.10,1.47)	1.35 (1.19,1.55)	1.28 (1.07, 1.56)
Smoking	3.97 (1.35,9.53)		
Non-white ethnicity	4.11 (1.23,9.74)		

Assuming dependence between $\boldsymbol{U}$ and $\boldsymbol{C}$ has little impact

Case Study: Conclusions

- Adjustment for unmeasured confounding in environmental studies is feasible through the use of additional data sources (e.g. surveys, cohorts, validation subgroups ...)
- In our case study, exposure effect estimate is driven towards the null once the important confounding effects of mother's smoking and ethnicity are taken into account
- 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

Comparison of Methods (1)



Comparison of Methods (2)

	Joint Model	Feedforward	MI (MICE)	Propensity Score
X - Y relationship	✓	✓	✓	✓
$\boldsymbol{C}$ - Y relationship	✓	✓	✓	✗
Coherency	✓	(✓)	✗	✓
$\boldsymbol{U}$ high dimension	(✗)	(✗)	(✓)	✓
Bayesian analysis model	✓	✓	✗	✓

Comparison of Methods (3)

Model	Software	Burn-in	Sample size	ESS	Run time
Joint model	WinBUGS	20000	2×20000	3268	12 hrs
Feed forward	WinBUGS	10000	2×10000	6467	5 hrs
MICE	R package	19	20×1	20	2 mins
Propensity score	R code	40000	2×100000	?	27 hrs

Concluding remarks

- Conceptually, Bayesian models offer an elegant tool for synthesizing information across data sources to handle problems of bias in observational data
- BUT, MCMC still imposes practical constraints and can make Bayesian methods a “hard sell”
- Might expect greater differences between approaches in situations with
    - more complex hierarchical structure
    - many more unmeasured confounders
    - informative priors
    - model mis-specification

...... work in progress

Acknowledgments

- Funding by UK Economic and Social Research Council (ESRC)
- The BIAS project, 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