Adjusting for selection bias in case control studies using Bayesian post-stratification

S.Geneletti, N.Best, S.Richardson

Imperial College Department of Epidemiology and Public Health

28/08/2008
Outline

• Problem of selection bias
• EMF and childhood Leukaemia data
• Simple example of SB
• Bias breaking model
• Potential sources of data
• Post-stratification method
• Adjustments
• Results
• Discussion
**Problem of selection bias**

**Basic problem**

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**
**EMF and childhood Leukaemia**

**Data**

- Case control study to investigate association of Electro-magnetic field exposure (via power lines) and childhood acute lymphoblastic Leukemia (ALL) in North Eastern US
- Eligible cases were diagnosed between 1987-1994 and registered at Childhood Cancer Registries
- Controls were contacted via random digit dialling
- No significant effect was found in the study

*The odds ratio for ALL was 1.24 (95 percent confidence interval, 0.86 to 1.79) at exposures of 0.200 μT or greater as compared with less than 0.065 μT.* (1)
Selection bias?

- Follow-up paper (2) raised issues about potential selection bias
- Partial participants - those who agreed to participate and gave demographic details BUT did not allow indoor measurements of exposure
- were systematically different w.r.t full participants (i.e. those who allowed indoor exposure measurement)
- In particular, they had different proportions of cases, white race, higher income (over $20k) and more urban dwellers
- As expected, all SES indicators
Selection bias?

**Full participants, N=1092**

- Case: 53%
- White: 94%
- >$20k: 73%
- Urb: 33%

**Partial participants, N=427**

- Case: 44%
- White: 87%
- >$20k: 53%
- Urb: 27%

**Full−Partial difference in percent**

- Case: 8%
- White: 7%
- >$20k: 20%
- Urb: 5%
Simple Example - Inheritance

- Male and female are independent $M \perp F$
Simple Example - Inheritance

- Male and female are independent $M \perp \perp F$
- Then they meet and have a child
Simple Example - Inheritance

- Male and female are independent $M \perp \perp F$
- Then they meet and have a child
- Now they are dependent through child $M \not\perp F | C$
Selection bias DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don’t know distribution of child given parents

- $Y$ is the outcome of interest, $W$ the exposure, $S$ the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don’t know $p(S|Y)$
Odds ratio

**True Odds ratio**

\[
\psi = \frac{p(Y = 1 \mid W = 1)p(Y = 0 \mid W = 0)}{p(Y = 0 \mid W = 1)p(Y = 1 \mid W = 0)}
= \frac{p(Y = 1, W = 1)p(Y = 0, W = 0)}{p(Y = 0, W = 1)p(Y = 1, W = 0)} \tag{1}
\]

**Observed Odds ratio**

\[
\psi^o = \frac{p(Y = 1, W = 1\mid S = 1)p(Y = 0, W = 0\mid S = 1)}{p(Y = 0, W = 1\mid S = 1)p(Y = 1, W = 0\mid S = 1)} \tag{2}
\]
Bias Breaking model

- The problem can be addressed if we can find a bias breaking variable $B$
- s.t. we can separate exposure $W$ from selection $S$

\[ A1 \quad W \perp S|(Y,B) \]  

- This means we can separate the exposure-disease process of interest from the nuisance of the selection process

**A2** Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

An assumption for simplicity:

**S1** Stratify $B$ if it is not discrete
Idea of Separation

The conditional independence $A1 \ W \perp \!\!\!\!\!\!\perp S| (Y, B)$ allows us to

1. separate the exposure disease mechanism of inferential interest
Idea of Separation

The conditional independence $A1 \ W \perp \perp S | (Y, B)$ allows us to

1. separate the exposure disease mechanism of inferential interest
2. from the nuisance selection bias mechanism
Idea of Separation

The conditional independence $A1 \ W \perp S \mid (Y, B)$ allows us to

1. separate the exposure disease mechanism of inferential interest
2. from the nuisance selection bias mechanism
3. by using $B$ to separate these mechanisms
Bias Breaking model

Now for example for controls we can estimate $p(W = 1 | Y = 0)$ as

$$p(W | Y = 0, S = 1, B) = p(W | Y = 0, B)$$

$$\sum_B p(W | Y = 0, B)p(B | Y = 0) = p(W | Y = 0)$$

- Focus is on finding estimates of $p(B | Y)$ as $p(W | Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias
Bias Breakers

Variables

- race - \textit{RAC} (white or other)
- urban status - \textit{URB} (urban or suburban vs other)
- income - \textit{INC} (family income of above $20k or lower)
- So $B = \{RAC, URB, INC\}$

- We assumed that $B$ are potential bias breakers as different distros of $B$ between full and partial participants
- All variables are dichotomized for simplicity
- How to estimate $p(B|Y)$ w/out bias?
- From data “outside” full participant study data
### Potential sources of data

#### Internal

- Pool full and partial participant data from study
- 1519 total, 1092 full participants and 427 partial participants (exclude 13 that have no measurements for one of the 3 BBs)
- Assume this represents target population
- Estimate **conditional** distribution $p(B|Y = y)$ as we know case/control status
- Estimate **marginal** distribution $p(B)$ if $p(B|Y = y) \approx p(B)$
- This is often plausible especially for controls
Potential sources of data

External

- Current Population Survey (CPS)
- The CPS is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics
- Download from US census website http://www.census.gov/cps/
- Can access geographic, demographic, health etc. data
- We used the Basic CPS from January 1995 for the 9 states in the study as this coincided roughly with the study period
- Variables were coded in the same way/similar as the variables in the study
- From CPS can only estimate $p(B)$ marginal
Recap

- We’ve assumed that $W \perp S|(Y, B)$
- We’ve identified a potential $B = \{RAC, INC, URB\}$
- We’ve got sources of data to estimate $p(B|Y)$ or $p(B)$ w/out selection bias
- Now we need to estimate $p(W|Y, B)$
- combine it with $p(B)$ to estimate $p(W|Y)$
- Use this to estimate marginal OR (as opposed to conditional OR we get from the logistic regression coefficient)
Poststratification Method

Idea

• Take a simple Bayesian logistic regression
• Estimate the parameters using WinBUGS and the full participant study data (i.e. those who have exposure data)
• Then for every possible (not necessarily observed) combination of the variables (in this case $Y$ and $B$) estimate the expected logit($p(W|Y, B)$)
• Inverse logit that quantity to get an estimate of $p(W|Y, B)$ for each of these combinations
• From pooled/external data get an empirical distribution for $p(B|Y)$ or $p(B)$
• Multiply this by appropriate $p(W|Y, B)$ and sum over to get an estimate of $p(W|Y)$ which you use to estimate OR
• Follows method in (3)
### Bayesian logistic regression model

- To explain, we focus on the simple model here.
- Variables: $w$ exposure, $y$ case/control status, $b$ single dichotomous variable.

#### Simple

<table>
<thead>
<tr>
<th>Simple</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>logit($p(w</td>
<td>y, b)) = \alpha + \beta y + \gamma b + \epsilon$</td>
</tr>
</tbody>
</table>

#### Why Bayesian?

- We have done this using a simple non-Bayesian logistic regression (4).
- The advantage with the Bayesian approach is that we do not have to worry about estimating the variance of the estimates in closed form.
- As we get a draw of 1000 from the posterior distribution of the ORs, the variance “comes out in the wash.”
**Poststratification method details**

Simple Bayesian logistic regression model with only one B participant data

<table>
<thead>
<tr>
<th>W</th>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Poststratification method details

Simple Bayesian logistic regression model with only one B participant data

<table>
<thead>
<tr>
<th>W</th>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

pooled internal data

<table>
<thead>
<tr>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

external data

<table>
<thead>
<tr>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Poststratification method details

Simple Bayesian logistic regression model with only one B participant data

<table>
<thead>
<tr>
<th>W</th>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

pooled internal data

<table>
<thead>
<tr>
<th>Y</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

external data

<table>
<thead>
<tr>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>


Poststratification method details

- From the full participant data we get estimates $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$
- $L\mathbf{W}$ is the vector of possible values of 
  $\text{logit}(p(W = 1|Y = y, B = b))$, $y, b \in \{0, 1\}$
  
  $L\mathbf{W} = (\hat{\alpha} \ \hat{\beta} \ \hat{\gamma}) \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$

\[ (5) \]

- The rows of the matrix are 1 for the intercept, $Y$ and $B$
- There are only 4 possible combinations for 2 binary variables
- So $L\mathbf{W}$ is a vector of 4 values
Poststratification method details

- From the full participant data we get estimates $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$
- $LW$ is the vector of possible values of $\logit(p(W = 1|Y = y, B = b))$, $y, b \in \{0, 1\}$

$$LW = (\hat{\alpha} \hat{\beta} \hat{\gamma}) \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

- The rows of the matrix are 1 for the intercept, $Y$ and $B$
- There are only 4 possible combinations for 2 binary variables
- So $LW$ is a vector of 4 values
Poststratification method details

- From the full participant data we get estimates $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$
- $LW$ is the vector of possible values of $\logit(p(W = 1|Y = y, B = b))$, $y, b \in \{0, 1\}$

\[
LW = (\hat{\alpha} \ \hat{\beta} \ \hat{\gamma}) \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}
\] (7)

- The rows of the matrix are 1 for the intercept, $Y$ and $B$
- There are only 4 possible combinations for 2 binary variables
- So $LW$ is a vector of 4 values
Poststratification method details

• From the pooled internal data we get the vectors
  \[ ICA = (\hat{p}_{01}, \hat{p}_{11}) \]
  \[ ICN = (\hat{p}_{00}, \hat{p}_{10}) \]
  where \( \hat{p}_{by} = \hat{p}(B = b \mid Y = y) \) the internal conditional empirical distribution estimates
• From the external data we get the vector
  \[ EM = (\hat{p}_0, \hat{p}_1) \]
  where \( \hat{p}_b = \hat{p}(B = b) \) the external marginal empirical distribution estimate

\[
\hat{p}(W = 1 \mid Y = 1)_{ICA} = LW(2, 4) \times ICA \tag{8}
\]
\[
\hat{p}(W = 1 \mid Y = 0)_{ICN} = LW(1, 3) \times ICN \tag{9}
\]
\[
\hat{p}(W = 1 \mid Y = 1)_{EM} = LW(2, 4) \times EM \tag{10}
\]
\[
\hat{p}(W = 1 \mid Y = 0)_{EM} = LW(1, 3) \times EM \tag{11}
\]
Poststratification method details

- Finally, we combine (6) and (7)

\[
OR_{IC} = \frac{\hat{p}(W = 1|Y = 1)_{ICA} \times [1 - \hat{p}(W = 1|Y = 0)_{ICN}]}{\hat{p}(W = 1|Y = 0)_{ICN} \times [1 - \hat{p}(W = 1|Y = 1)_{ICA}]}
\] (12)

- And we combine (8) and (9)

\[
OR_{EM} = \frac{\hat{p}(W = 1|Y = 1)_{EM} \times [1 - \hat{p}(W = 1|Y = 0)_{EM}]}{\hat{p}(W = 1|Y = 0)_{EM} \times [1 - \hat{p}(W = 1|Y = 1)_{EM}]}
\] (13)

- Note that when we are using the marginal adjustment, we use the same distribution to adjust both cases and controls

- We look at whether this should be done in the discussion

- We can also pool the internal data to get marginal estimates for \(p(B)\) - this we term IM estimates
Analysis

- We estimated the parameters using WinBUGS
- We ran 20000 iterations and kept the last 1000
- Convergence was good
- We ran a number of different models including interaction terms but results were similar
- We also ran hierarchical models using the state information but results were similar
- We show density plots for the 1000 estimates of the ORs for one chain
**Results for Simple model with** $B = \{ RAC, INC, URB \}$

We began by looking at the Bayesian regression coefficient OR.
Results for Simple model with $B = \{RAC, INC, URB\}$

Then we included the both case/control adjusted estimates:

Densities of ORs

- **bayes lg**: OR = 1.17 (0.84, 1.61)
- **ICA+ICN**: OR = 1.28 (0.94, 1.75)
- **IM**: OR = 1.16 (0.84, 1.58)
- **EM**: OR = 1.16 (0.85, 1.57)
No selection bias?

- Generally there appears to be little reason to suspect selection bias
- The point estimates are close
- The variances are similar
- In fact, although the partial participants are atypical w.r.t some characteristics, it would appear that they are too few to sway the results
Discussion

- We estimate the marginal OR as opposed to the conditional OR in the logistic regression.
- We used Empirical distributions for the weights.
- For the external data it probably does not make sense to model the weights as there are over 50,000 data points.
- For the pooled internal data however, there might be some scope for modelling, especially when the number of partial participants is low.
- In particular, we can impose constraints on the weights if we believe that the OR point estimate must be above 1 etc.
Further work

- Develop a simulation study based on EMF data with selection bias and see what happens with adjustments.
- Replace $p(W|Y = y)$ with the 2x2 table estimate and see what happens if we assume there is no selection bias in either cases or controls.
- Is it possible to come up with bounds to determine just how bad selection bias has to be in order to significantly change results?
- Develop a model for the weights using Bayesian contingency table methods and use this to constrain the OR.
- More generally, relate to other weighting procedures such as inverse probability weighting, and imputation.

