Bayesian analysis of pair-matched case-control studies subject to outcome misclassification

Tanja Högg 1  Paul Gustafson 1  Yinshan Zhao 2,3
John Petkau 1  Helen Tremlett 2

1Department of Statistics, University of British Columbia
2Department of Medicine, University of British Columbia
3BC Centre for Improved Cardiovascular Health

Banff International Research Station
18.08.2013
Prodromal Multiple Sclerosis: the ProMS study

- Ongoing Canada-wide study (BC, NS, MA, SK) investigating the existence of a prodrome in multiple sclerosis (MS).
- Prevalence in Canada about .3%, one of the highest in the world.
- No definite diagnostic test and highly heterogeneous symptoms lead to diagnostic delays.
- Focus lies on five years prior to the first recognized symptom of MS.
- Among others, presence of 14 morbidities in prodromal phase (e.g. hypertension, depression).
- Study data extracted from provincial administrative health databases.
Health administrative databases of British Columbia

- **Medical Services Plan (MSP) Database**
  - claim information of fee-for-service practitioners in BC
  - since 1991, includes one to five ICD codes for reason of visit (e.g. 340 for MS)

- **Canadian Discharge Abstract Database**
  - captures administrative records for all hospital discharges
  - includes a maximum of 25 ICD codes per discharge

- **PharmaNet**
  - prescription medication dispensed by pharmacies across BC
  - includes information on drug type, quantity, directions for use

*Databases are linkable, giving near-universal coverage of healthcare contacts for British Columbians.*
ProMS study design

- Matched case-control study

- MS cases identified from admin data using case definition of $\geq 3$ MS-specific records, i.e.
  - ICD 340 in MSP or hospital discharge files
  - MS-specific prescription drugs in PharmaNet

- Date of first MS-specific claim (index date) marks end of five-year prodromal phase.

- Matched controls selected from peers without MS-related records.

- Matching variables are sex, postal code and age at index date.

- Linkage with British Columbia Multiple Sclerosis (BC MS) database.
Quality issues for administrative data

- ICD codes do not guarantee presence of a disease
  - ICD coding errors
  - Lack of specificity (e.g. ICD 780 - general symptoms)
  - High misdiagnosis rate for multiple sclerosis (false positive rate of 35% reported by Poser [3])
- Possibility of misclassified disease status in ProMS, leading to
  - apparent cases that are in fact controls
  - apparent controls that are in fact MS cases
- Analysis must take potentially imperfect MS status of study subjects into account
Suppose interest lies in the odds ratio $OR$ between a binary exposure $E$ and outcome $D$.

$D$ is unobserved and only available via surrogate $D^*$ produced by a non-differential classifier.

“Apparent” cases with $D^* = 1$ are matched to “apparent” controls with $D^* = 0$ on a set of confounders.

Let $(E_{1k}, E_{2k}), (D_{1k}, D_{2k})$ and $(D^*_{1k}, D^*_{2k})$ denote the exposure, true and observed outcome of the apparent case and control in the $k$th of $n$ pairs.

Cell counts (probabilities):

<table>
<thead>
<tr>
<th></th>
<th>$E_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>$n_{11} (\theta_{11})$</td>
</tr>
<tr>
<td>0</td>
<td>$n_{01} (\theta_{01})$</td>
</tr>
</tbody>
</table>
Consider the exposure risk model

$$\logit(P(E_{ik} = 1)) = \beta_k + \delta I(i = 1), \quad i = 1, 2$$

where $\beta_k$ is a pair-specific random effect.

Assuming $E_{1k}$ and $E_{2k}$ are independent given $\beta_k$, Prescott et al. (2005) show that

$$OR = \exp(\delta) = \frac{P(E_1 = 1, E_2 = 0)}{P(E_1 = 0, E_2 = 1)} = \frac{\theta_{10}}{\theta_{01}}$$

This gives

$$\hat{OR} = \frac{n_{10}}{n_{01}}$$

How do $\theta_{10}/\theta_{01}$ and $OR$ relate under outcome misclassification?
Bias under outcome misclassification

- Denote

\[ \theta_{lm|ij} = P(E_1 = l, E_2 = m|D_1 = i, D_2 = j), \quad i, j, l, m = 0, 1 \]

- Under non-differential misclassification, the numerator of (1) is

\[ \theta_{10} = \sum_{i,j \in \{0,1\}} \theta_{10|ij} P(D_1 = i, D_2 = j|D_1^* = 1, D_2^* = 0) \]

\[ = \sum_{i,j \in \{0,1\}} \theta_{10|ij} P(D_1 = i|D_1^* = 1)P(D_2 = j|D_2^* = 0) \]

where

\[ pp = P(D_1 = 1|D_1^* = 1) \quad \text{and} \quad np = P(D_2 = 0|D_2^* = 0) \]

- Similarly for the denominator,

\[ \theta_{01} = \sum_{i,j \in \{0,1\}} \theta_{01|ij} P(D_1 = i|D_1^* = 1)P(D_2 = j|D_2^* = 0) \]
Bias under outcome misclassification (continued)

▶ Using

\[
\begin{align*}
\theta_{01 \mid 10} &= \theta_{10 \mid 01}, & \theta_{01 \mid 01} &= OR \ 	heta_{10 \mid 01} \\
\theta_{01 \mid 00} &= \theta_{10 \mid 00}, & \theta_{01 \mid 11} &= \theta_{10 \mid 11},
\end{align*}
\]

(3)

manipulations yield

\[
\frac{\theta_{10}}{\theta_{01}} = OR \frac{1 + \left( \frac{(1-np)}{np} a + \frac{(1-pp)}{pp} c \right) + \frac{(1-pp)(1-np)}{pp np} b}{1 + OR \left( \frac{(1-np)}{np} a + \frac{(1-pp)}{pp} c \right) + OR^2 \frac{(1-pp)(1-np)}{pp np} b}
\]

where

\[
a = \frac{\theta_{10 \mid 11}}{\theta_{10 \mid 10}}, \quad b = \frac{\theta_{10 \mid 01}}{\theta_{10 \mid 10}}, \quad c = \frac{\theta_{10 \mid 00}}{\theta_{10 \mid 10}}.
\]

▶ Therefore,

\[
\frac{\theta_{10}}{\theta_{01}} \leq OR \quad \text{if} \quad OR \geq 1 \quad \text{and} \quad \frac{\theta_{10}}{\theta_{01}} > OR \quad \text{if} \quad OR < 1.
\]
A Bayesian model for matched studies under outcome misclassification

- Assuming independence between pairs,

\[ (n_{11} + n_{00}, n_{10}, n_{01}) \sim \text{Multinomial} \left(n, (\theta_{11} + \theta_{00}, \theta_{10}, \theta_{01}) \right) \]

where

\[
\begin{align*}
\theta_{10} &= pp \, np \, \theta_{01|10} \, OR + (1 - pp)(1 - np) \, \theta_{01|10} + pp(1 - np)\theta_{10|00} + (1 - pp)np \, \theta_{10|11} \\
\theta_{01} &= pp \, np \, \theta_{01|10} + (1 - pp)(1 - np)\theta_{01|10} \, OR + pp(1 - np)\theta_{10|00} + (1 - pp)np \, \theta_{10|11}
\end{align*}
\]

- Taking the difference between cell probabilities,

\[
\theta_{10} - \theta_{01} = \theta_{01|10}(OR - 1)(pp \, np - (1 - pp)(1 - np))
\]

- Problem is non-identifiable when \( pp, np \) or \( \theta_{01|10} \) are unknown.

- **Needed**: prior input to inform prior distributions of \( pp, np \) and \( \theta_{01|10} \).
Prior distributions

- Six model parameters: \((pp, np, OR, \theta_{01|10}, \theta_{01|00}, \theta_{01|11})\)

- Choose informed, independent priors for \(pp, np\) and \(\theta_{01|10}\)

\[
pp \sim \text{Beta}(\alpha_1, \alpha_2) \\
np \sim \text{Beta}(\beta_1, \beta_2) \\
\theta_{01|10} \sim \text{Beta}(\gamma_1, \gamma_2)
\]

- Determine \(\alpha_j\) and \(\beta_j, j=1,2\) from previous estimates \(\hat{pp}, \hat{np}\) and \(se(\hat{pp}), se(\hat{np})\).

- Determine \(\gamma_j\) from validation data via

\[
\begin{align*}
m_{01} \mid \theta_{01|10} &\sim \text{Bin}(n_{val}, \theta_{01|10}) \\
\theta_{01|10} &\sim \text{Unif}(0, 1)
\end{align*}
\]

where \(m_{01}\) is the number of case-control pairs with \((E_1 = 0, E_2 = 1)\).

- Implies \(\gamma_1 = m_{01} + 1\) and \(\gamma_2 = n_{val} - m_{01} + 1\).
Choose uniform priors for $OR$, $\theta_{01|00}$ and $\theta_{01|11}$ as

$$
OR \mid pp, np, \theta_{01|10} \sim Unif(0, t_1)
$$

$$
\theta_{01|00} \mid OR, pp, np, \theta_{01|10} \sim Unif(0, t_2)
$$

$$
\theta_{01|11} \mid OR, pp, np, \theta_{01|10}, \theta_{01|00} \sim Unif(0, t_3)
$$

where

$$
t_1 = \min \left( \frac{1}{\theta_{01|10}}, \frac{1}{\theta_{01|10}(pp np + (1 - pp)(1 - np))} - 1 \right)
$$

$$
t_2 = \min \left( 1, \frac{1 - (OR + 1)\theta_{01|10}(pp np + (1 - pp)(1 - np))}{2pp(1 - np)} \right)
$$

$$
t_3 = \min \left( 1, \frac{1 - (OR + 1)\theta_{01|10}(pp np + (1 - pp)(1 - np)) - 2pp(1 - np)\theta_{01|00}}{2(1 - pp)np} \right)
$$

to ensure that $\theta_{10} + \theta_{01} \leq 1$ and $\theta_{ij|lm} \leq 1$. 
Simulation study

- Generate
  - \( n \) apparent case-control pairs
  - \( n_{val} \) true case-control pairs,
  matched on a binary confounder \( U \) with \( D - E \) association \( OR \).

- Evaluate
  1. posterior median of \( OR \),
  2. length and coverage of 95\% posterior credible interval of \( OR \),
  3. empirical size and power of the hypothesis test \( H_0 : OR = 1 \)
for naive and proposed analysis.

- Examine different settings of
  - disease-exposure association \( OR \),
  - cohort sizes \( n \) and \( n_{val} \),
  - misclassification (\( SN, SP \)),
  - prior uncertainty about \( pp \) and \( np \).
  - deviations of \( \hat{pp}, \hat{np} \) from true \( pp, np \)
Results: Median, length and coverage

Median of posterior distribution of OR, coverage and length of 95% posterior credible interval, averaged over 1000 runs:

<table>
<thead>
<tr>
<th>$SN$</th>
<th>$SP$</th>
<th>naive median</th>
<th>naive coverage</th>
<th>naive length</th>
<th>adjusted median</th>
<th>adjusted coverage</th>
<th>adjusted length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>1.29</td>
<td>0.00</td>
<td>0.47</td>
<td>2.00</td>
<td>0.96</td>
<td>1.72</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7</td>
<td>1.44</td>
<td>0.06</td>
<td>0.53</td>
<td>2.00</td>
<td>0.96</td>
<td>1.26</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>1.60</td>
<td>0.04</td>
<td>0.58</td>
<td>2.00</td>
<td>0.96</td>
<td>1.13</td>
</tr>
<tr>
<td>1.0</td>
<td>0.9</td>
<td>1.85</td>
<td>0.07</td>
<td>0.63</td>
<td>2.03</td>
<td>0.97</td>
<td>1.01</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>2.00</td>
<td>0.08</td>
<td>0.75</td>
<td>2.00</td>
<td>0.95</td>
<td>0.70</td>
</tr>
</tbody>
</table>

$OR = 2, n = 1000, n_{val} = 200, se(\hat{pp}) = se(\hat{np}) = 0.02.$
Application - Morbidities in MS prodrome

- Estimate odds ratio of MS and presence of 14 morbidities in the prodromal phase.
- Study cohort of 7250 apparent case-control pairs.
- Determine presence of morbidities via case definitions of Marrie et al. [1].
- E.g. hypertension is considered prevalent if \( \geq 4 \) disease-related records within 2 years.
- Assume \( np = 1 \) and use \( \hat{pp} = 0.83, \ se(\hat{pp}) = 0.02 \) based on Marrie et al. [2] for prior input on \( pp \).
- Validation cohort defined as subset with \( \geq 20 \) MS-specific ICD codes \( (n_{val} = 929) \).
### Results

<table>
<thead>
<tr>
<th>Condition</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>thyroid disease</td>
<td></td>
</tr>
<tr>
<td>schizophrenia</td>
<td></td>
</tr>
<tr>
<td>psoriasis</td>
<td></td>
</tr>
<tr>
<td>migraine</td>
<td></td>
</tr>
<tr>
<td>mental</td>
<td></td>
</tr>
<tr>
<td>lung disease</td>
<td></td>
</tr>
<tr>
<td>irr. bowel syndrome</td>
<td></td>
</tr>
<tr>
<td>hypertension</td>
<td></td>
</tr>
<tr>
<td>hyperlipidemia</td>
<td></td>
</tr>
<tr>
<td>diabetes</td>
<td></td>
</tr>
<tr>
<td>depression</td>
<td></td>
</tr>
<tr>
<td>bowel disease</td>
<td></td>
</tr>
<tr>
<td>bipolar disorder</td>
<td></td>
</tr>
<tr>
<td>anxiety</td>
<td></td>
</tr>
</tbody>
</table>

**Method**
- **adjusted**
- **naive**
- **validation**
Acknowledgements

All inferences, opinions, and conclusions drawn in this presentation are those of the authors, and do not reflect the opinions or policies of the Data Steward(s).
References

Thank you.


