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]

[1]Department of Statistics, University of British Columbia

[2]Department of Medicine, University of British Columbia

[3]BC 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 ≥ 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

Preliminaries

- 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 kth of n pairs.
- Cell counts (probabilities):

		E_2	
		1	0
E_1	1	n_{11} (θ_{11})	n_{10} (θ_{10})
	0	n_{01} (θ_{01})	n_{00} (θ_{00})

Analysis of matched case-control data under perfect outcome classification

- Consider the exposure risk model

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

  where β_k is a pair-specific random effect.

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

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

- This gives

  $$\widehat{OR} = \frac{n_{10}}{n_{01}} \tag{2}$$

- How do θ_{10}/θ_{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

$$\begin{aligned}\theta_{10} &= \sum_{i,j \in \{0,1\}} \theta_{10|ij} \; P(D_1 = i, D_2 = j | D_1^* = 1, D_2^* = 0) \\ &= \sum_{i,j \in \{0,1\}} \theta_{10|ij} \; P(D_1 = i | D_1^* = 1) P(D_2 = j | D_2^* = 0)\end{aligned}$$

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

$$\begin{aligned} \theta_{01|10} &= \theta_{10|01}, \quad \theta_{01|01} = OR\,\theta_{10|01} \\ \theta_{01|00} &= \theta_{10|00}, \quad \theta_{01|11} = \theta_{10|11}, \end{aligned} \tag{3}$$

  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

$$a = \frac{\theta_{10|11}}{\theta_{10|10}}, \quad b = \frac{\theta_{10|01}}{\theta_{10|10}}, \quad c = \frac{\theta_{10|00}}{\theta_{10|10}}.$$

- Therefore,

$$\frac{\theta_{10}}{\theta_{01}} \leq OR \quad \text{if} \quad OR \geq 1 \qquad \text{and} \qquad \frac{\theta_{10}}{\theta_{01}} > OR \quad \text{if} \quad OR < 1.$$

A Bayesian model for matched studies under outcome misclassification

- Assuming independence between pairs,

$$(n_{11} + n_{00}, n_{10}, n_{01}) \sim \mathit{Multinomial}\Big(n, (\theta_{11} + \theta_{00}, \theta_{10}, \theta_{01})\Big)$$

  where

$$\theta_{10} = pp\,np\,\theta_{01|10} OR + (1 - pp)(1 - np)\,\theta_{01|10} + pp(1 - np)\theta_{10|00} + (1 - pp)np\,\theta_{10|11}$$
$$\theta_{01} = pp\,np\;\theta_{01|10} + (1 - pp)(1 - np)\theta_{01|10} OR + pp(1 - np)\theta_{10|00} + (1 - pp)np\,\theta_{10|11}$$

- Taking the difference between cell probabilities,

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

- 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 $\theta_{01|10}$

$$
\begin{aligned}
pp &\sim Beta(\alpha_1, \alpha_2) \\
np &\sim Beta(\beta_1, \beta_2) \\
\theta_{01|10} &\sim Beta(\gamma_1, \gamma_2)
\end{aligned}
$$

- Determine α_j and $\beta_{j,\ j=1,2}$ from previous estimates $\widehat{pp}$, $\widehat{np}$ and $se(\widehat{pp})$, $se(\widehat{np})$.
- Determine γ_j from validation data via

$$
\begin{aligned}
m_{01} \mid \theta_{01|10} &\sim Bin(n_{val}, \theta_{01|10}) \\
\theta_{01|10} &\sim Unif(0, 1)
\end{aligned}
$$

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

- Implies $\gamma_1 = m_{01} + 1$ and $\gamma_2 = n_{val} - m_{01} + 1$.

Prior distributions (continued)

Choose uniform priors for OR, $\theta_{01|00}$ and $\theta_{01|11}$ as

$$
\begin{gathered}
OR \mid pp, np, \theta_{01|10} \sim Unif(0, t_1) \\
\theta_{01|00} \mid OR, pp, np, \theta_{01|10} \sim Unif(0, t_2) \\
\theta_{01|11} \mid OR, pp, np, \theta_{01|10}, \theta_{01|00} \sim Unif(0, t_3)
\end{gathered}
$$

where

$$
\begin{aligned}
t_1 &= \min\left(\frac{1}{\theta_{01|10}}, \frac{1}{\theta_{01|10}(pp\ np + (1-pp)(1-np))} - 1\right) \\
t_2 &= \min\left(1, \frac{1-(OR+1)\theta_{01|10}(pp\ np + (1-pp)(1-np))}{2pp(1-np)}\right) \\
t_3 &= \min\left(1, \frac{1-(OR+1)\theta_{01|10}(pp\ np + (1-pp)(1-np)) - 2pp(1-np)\theta_{01|00}}{2(1-pp)np}\right)
\end{aligned}
$$

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 $\widehat{pp}$, $\widehat{np}$ 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:

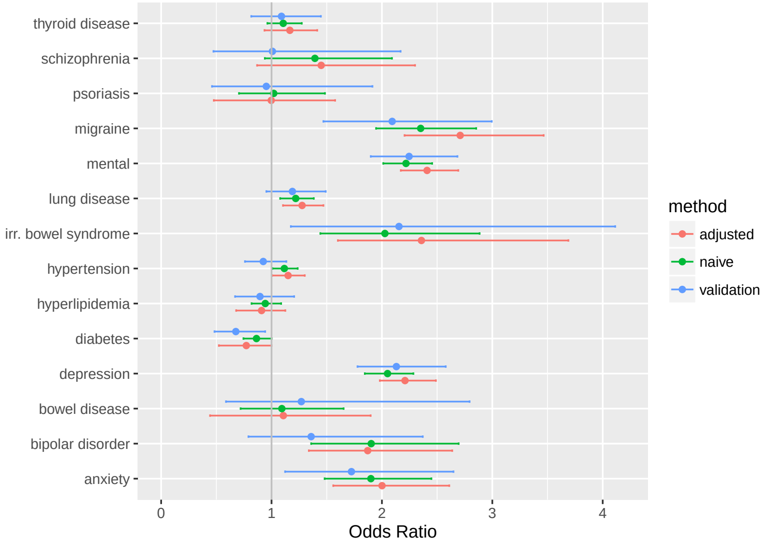
		naive			**adjusted**		
SN	SP	median	coverage	length	median	coverage	length
0.7	0.7	1.29	0.00	0.47	2.00	0.96	1.72
	0.9	1.53	0.15	0.56	1.97	0.95	1.11
	1.0	1.84	0.83	0.68	2.05	0.96	0.92
0.9	0.7	1.44	0.06	0.53	2.00	0.96	1.26
	0.9	1.70	0.53	0.63	2.03	0.97	1.01
	1.0	1.94	0.92	0.73	2.04	0.97	0.82
1	0.7	1.53	0.16	0.57	1.99	0.96	1.10
	0.9	1.78	0.71	0.67	2.01	0.96	0.93
	1.0	2.00	0.95	0.75	2.00	0.95	0.70

$OR = 2,\ n = 1000,\ n_{val} = 200,\ se(\widehat{pp}) = se(\widehat{np}) = 0.02.$

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 ≥ 4 disease-related records within 2 years.
- Assume $np = 1$ and use $\widehat{pp} = 0.83$, $se(\widehat{pp}) = 0.02$ based on Marrie et al. [2] for prior input on pp.
- Validation cohort defined as subset with ≥ 20 MS-specific ICD codes ($n_{val} = 929$).

Results



thyroid disease
schizophrenia
psoriasis
migraine
mental
lung disease
irr. bowel syndrome
hypertension
hyperlipidemia
diabetes
depression
bowel disease
bipolar disorder
anxiety
0
1
2
3
4
Odds Ratio
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.

[1] Marrie, R. A., J. D. Fisk, K. J. Stadnyk, H. Tremlett, C. Wolfson, S. Warren, V. Bhan, B. N. Yu, and CIHR Team in the Epidemiology and Impact of Comorbidity on Multiple Sclerosis (2014). Performance of administrative case definitions for comorbidity in multiple sclerosis in Manitoba and Nova Scotia. *Chronic Diseases and Injuries in Canada 34*(2-3), 145–53.

[2] Marrie, R. A., J. D. Fisk, K. J. Stadnyk, B. N. Yu, H. Tremlett, C. Wolfson, S. Warren, and V. Bhan (2013). The incidence and prevalence of multiple sclerosis in Nova Scotia, Canada. *The Canadian Journal of Neurological Sciences 40*(06), 824–831.

[3] Poser, C. M. (1997). Misdiagnosis of multiple sclerosis and β-interferon. *The Lancet 349*(9069), 1916.

[4] Prescott, G. J. and P. H. Garthwaite (2005). Bayesian analysis of misclassified binary data from a matched case–control study with a validation sub-study. *Statistics in Medicine 24*(3), 379–401.