Statistics Canada Statistique Canada

Logistic Regression with Linked Data

Telling Canada's story in numbers

Abdelnasser Saïdi
Kennett Chu
Abel Dasylva
Felix Labrecque-Synnott

March. 9 2018
FCSM Conference Washington

Canada

Overview

1. Purpose
2. Logistic regression
3. Evaluation of the FP adjustment- Simulation study
4. Incorrect links & Unlinked records
5. Application: Modeling the probability of a record remaining unlinked
6. Challenges

1. Purpose

- **Analysis of association**
  - A focus on the relation between selected variables. Underlying model parameters (e.g., logistic regression coefficients) are secondary and so is the issue of their potential bias.
  - Example: cohort mortality study by linking a census and a mortality database (Wilkins et al., 2008)
- **Methods for Adjusting Statistical analyses for Linkage Errors**

1.1. Neter et al. (1965)

- “Neter and al. showed even small errors may greatly affect the results of the estimation procedures that do not tackle the errors in linkage” (Di. Consiglio and al. 2015)
- Linkage errors will generally lead to biased estimates , increased variability and can cause problems in analysing the relationships between variables across the linked files.
- The question of how to incorporate linkage errors in analysis is an open area of research.

2. Logistic regression

- Chambers et al. (2009)
  - Secondary analysis with Estimating Equations (GEE)
- Chipperfield et al. (2011)
  - Primary analysis with E-M algorithm and clerical-reviews

2.2. Notation

- Two unduplicated files
  - X with a fixed binary covariate X
  - Y with a binary response Y and linkage variables $\boldsymbol{Z}$
- Consider a one-to-one linkage.
- Linked pairs: $[(Y_i^*, x_i)]_{1 \le i \le n}$
- True response: Y_i for the i–th X record, s.t. $Y_i^* = Y_i$ in a correct link.
- Clerical decision: $D_i = 1$ if link i is correct, else $D_i = 0$.

2.2. Notation

- Regression model

$$P(Y_i = 1; x_i) = \mu_i(\boldsymbol{\beta}) = e^{\beta_0+\beta_1 x_i} / \left(1 + e^{\beta_0+\beta_1 x_i}\right)$$

- Clerical-review: assumed error-free
  - An SRS sample of pairs s for which we determine

    $P(D_i = 1 | Y_i^*, X_i)$ for all pairs x and y* associated with s.
- Observed data: denoted O_i by for the pair (Y_i^*, x_i) where
  - O_i comprises of Y_i^*, D_i for a pair in the review sample s.
  - O_i comprises of Y_i^* for the remaining pairs.

2.3. Incorrect links

Assumptions:

- No unlinked records in the linked file

- Links are incorrect at random (IAR)
  - Y_i and D_i are conditionally independent given X_i.

2.3. Incorrect links

When the response Y_i is unknown, use:

$$\sum_i \boldsymbol{X}_i^T \left(E\left[Y_i \middle| O_i\right] - \mu_i(\boldsymbol{\beta}) \right) \quad = \quad 0$$

$$\sum_{i \in s} \boldsymbol{X}_i^T \left(E\left[Y_i \middle| Y_i^*, D_i\right] - \mu_i(\boldsymbol{\beta}) \right) + \sum_{i \notin s} \boldsymbol{X}_i^T \left(E\left[Y_i \middle| Y_i^*\right] - \mu_i(\boldsymbol{\beta}) \right) \quad = \quad 0$$

where

$$E[Y_i|Y_i^*, D_i; \boldsymbol{\beta}] \quad = \quad D_i Y_i^* + (1 - D_i)\mu_i(\boldsymbol{\beta})$$

$$\begin{aligned} E[Y_i|Y_i^*; \boldsymbol{\beta}] &= E\left[E\left[Y_i \middle| Y_i^*, D_i\right] \middle| Y_i^*\right] \\ &= E\left[D_i Y_i^* + (1 - D_i)\mu_i \middle| Y_i^*\right] \\ &= P(D_i = 1|Y_i^*, X_i) Y_i^* + P\left(D_i = 0 \middle| Y_i^*, X_i\right) \mu_i(\boldsymbol{\beta}) \end{aligned}$$

2.3. Incorrect links

Estimation of $P(D_i = 1|Y_i^*, X_i)$:

- Estimated by the proportion of correct links in the clerical sample for each combination of x and y* (binary variables)

Consistency

The solution of the proposed method is consistent because:

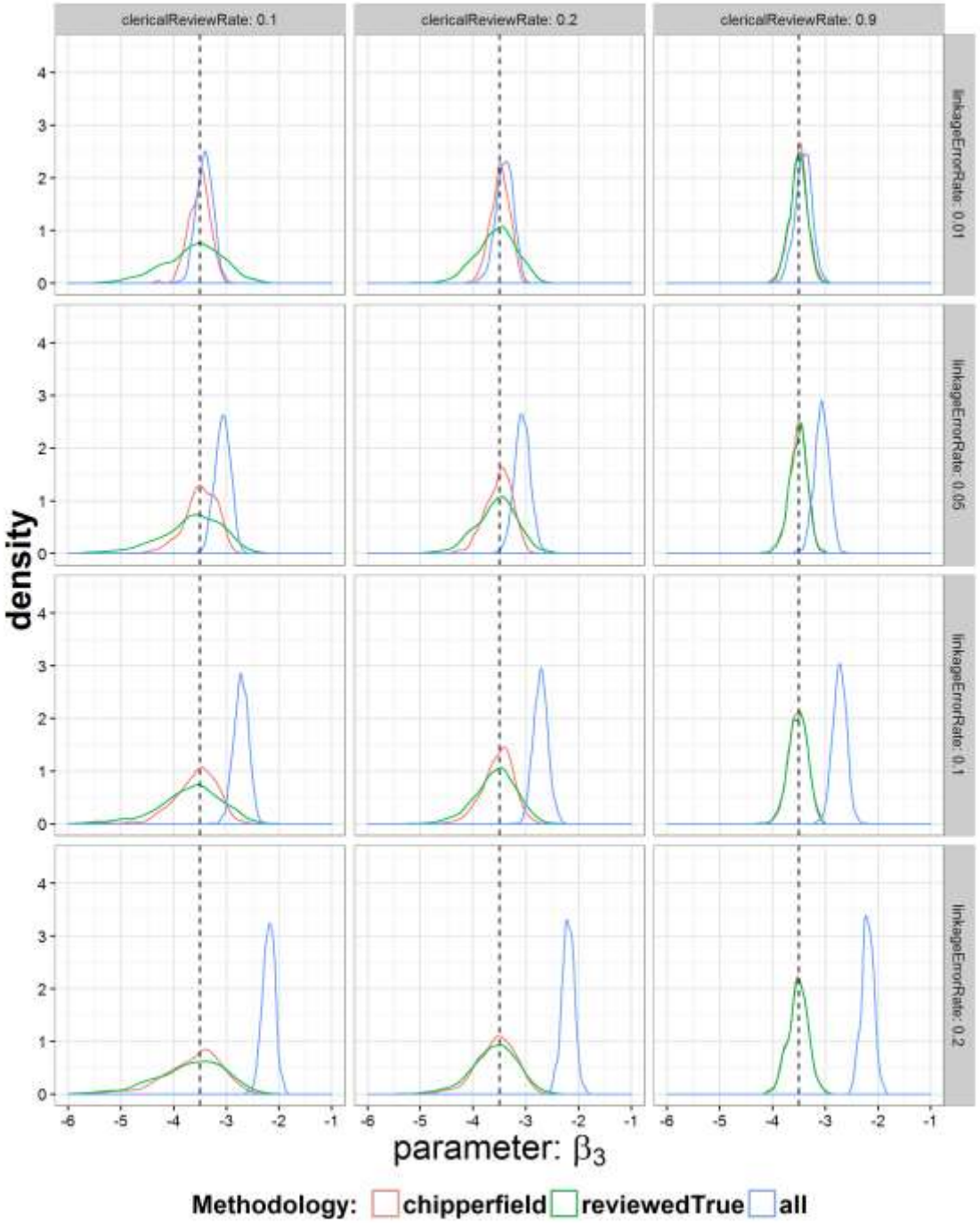
$$E[E[Y_i|Y_i^*, D_i; \boldsymbol{\beta}] - \mu_i(\boldsymbol{\beta})] \quad = \quad E[E[Y_i|Y_i^*; \boldsymbol{\beta}] - \mu_i(\boldsymbol{\beta})] \quad = \quad 0$$

3. Evaluation of the method

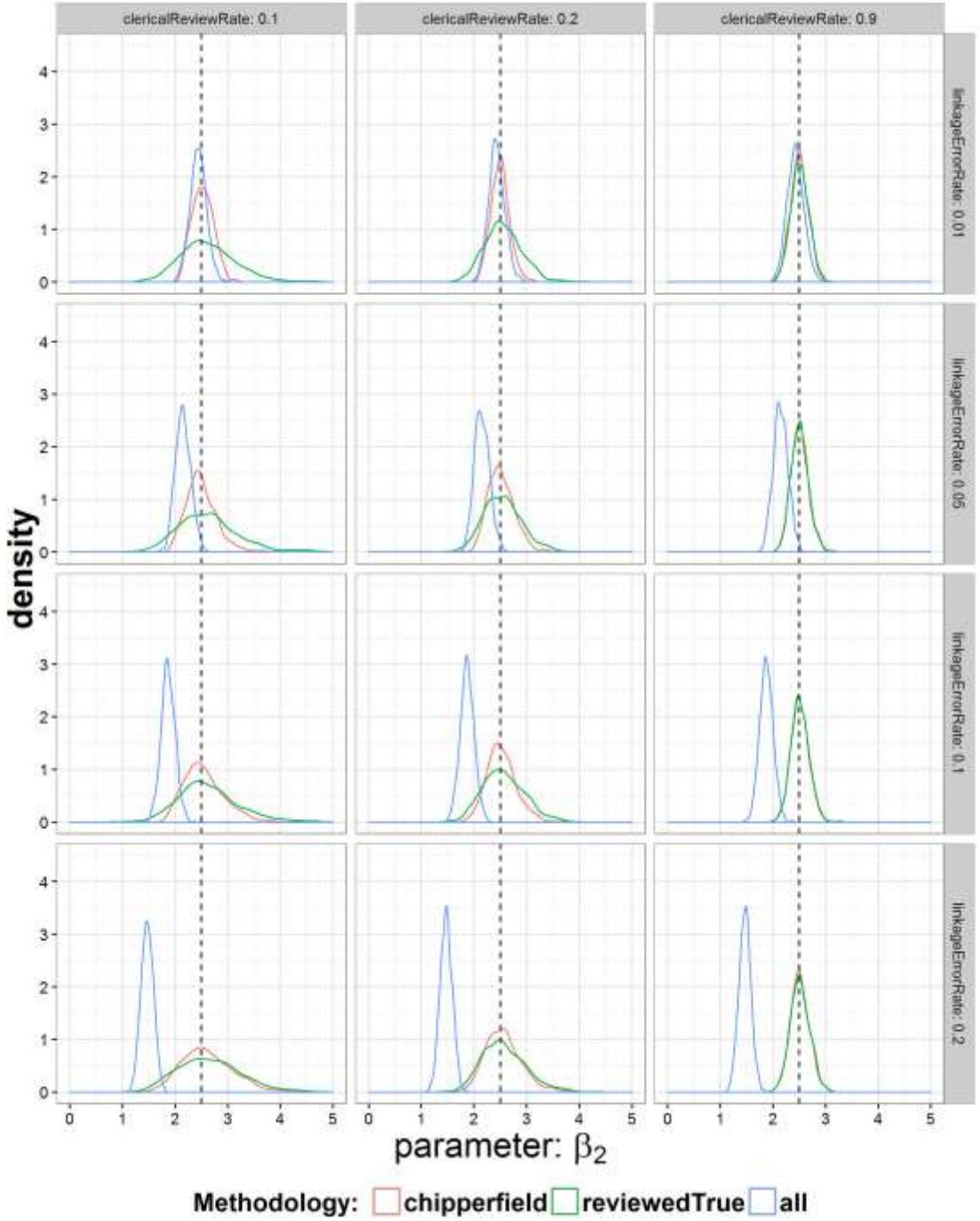
- each scenario corresponds to a combination of:
  - a linkage error rate of 0.01, 0.05, 0.1 or 0.2, and
  - a clerical review rate of 0.1, 0.2, and 0.9.
- for each scenario, 1,000 independent synthetic data sets were generated
  - Each contained 2000 linked pairs
  - Three independent binary predictor variables
  - True parameters (-0.5; 1.5; 2.5; -3.5) obeying

$$P(Y=1|X_1=x_1, X_2=x_2, X_3=x_3)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}$$

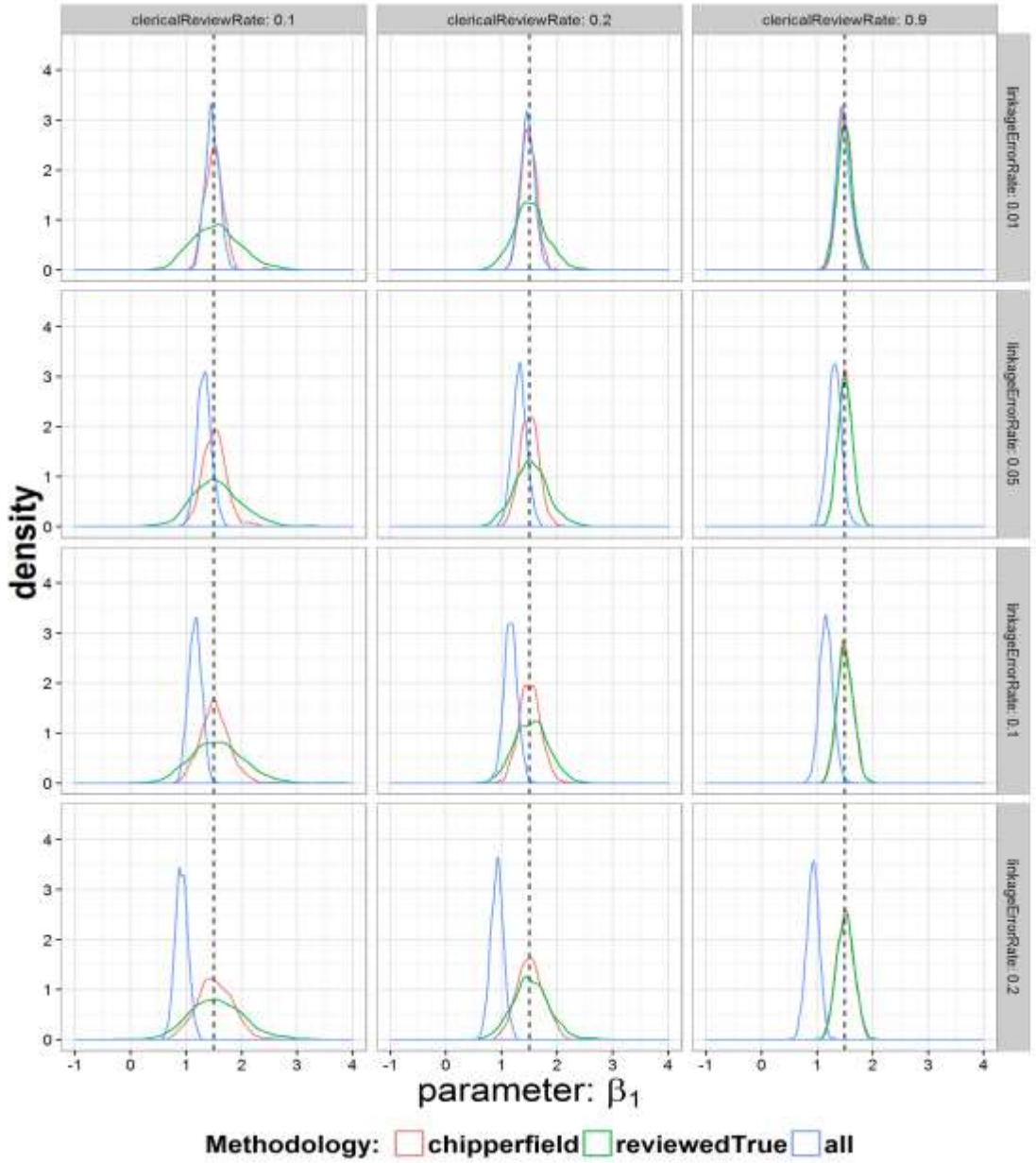
Results


clericalReviewRate: 0.1
clericalReviewRate: 0.2
clericalReviewRate: 0.9
linkageErrorRate: 0.01
linkageErrorRate: 0.05
linkageErrorRate: 0.1
linkageErrorRate: 0.2
density
parameter: β_3
Methodology: chipperfield reviewedTrue all

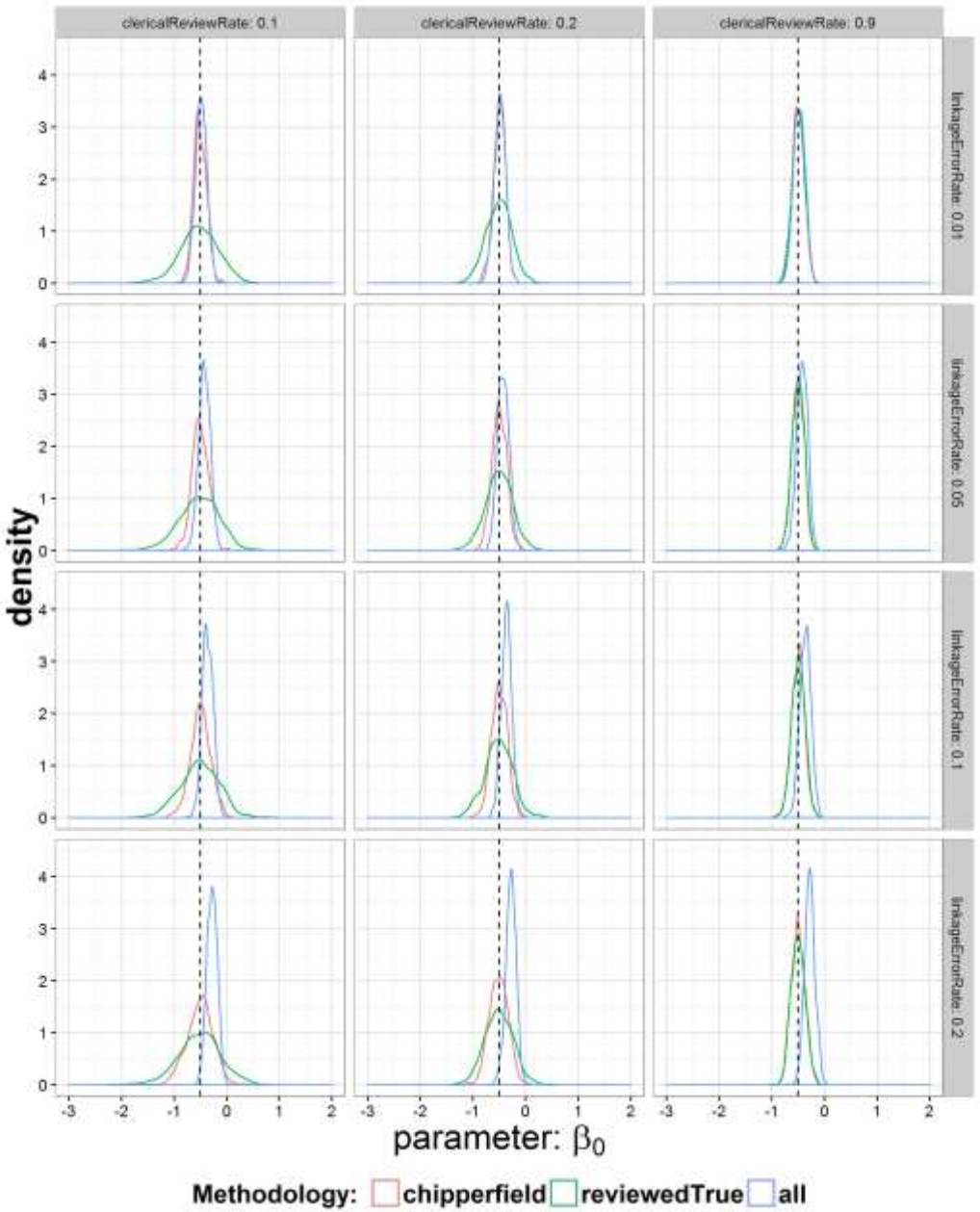
Results


clericalReviewRate: 0.1
clericalReviewRate: 0.2
clericalReviewRate: 0.9
linkageErrorRate: 0.01
linkageErrorRate: 0.05
linkageErrorRate: 0.1
linkageErrorRate: 0.2
density
parameter: β_2
Methodology: chipperfield reviewedTrue all

Results


clericalReviewRate: 0.1
clericalReviewRate: 0.2
clericalReviewRate: 0.9
linkageErrorRate: 0.01
linkageErrorRate: 0.05
linkageErrorRate: 0.1
linkageErrorRate: 0.2
density
parameter: β_1
Methodology: chipperfield reviewedTrue all

Results


clericalReviewRate: 0.1
clericalReviewRate: 0.2
clericalReviewRate: 0.9
linkageErrorRate: 0.01
linkageErrorRate: 0.05
linkageErrorRate: 0.1
linkageErrorRate: 0.2
density
parameter: β_0
Methodology: chipperfield reviewedTrue all

3.1. Simulation results

- Estimator bias can be large when the presence of linkage errors is completely ignored, even if the linkage error rate is as low as 0.05.
- Classical logistic regression deployed on only the reviewed-to-be-correctly-linked records appeared unbiased, as expected.
- For small linkage error and clerical review rates, the method of Chipperfield *et al*. (2011) appeared to "outperform" the method of applying classical logistic regression on only the reviewed-to-be-correct linked records, in the sense that the former appeared to remain unbiased but exhibit a smaller variance than the latter.
- The Chipperfield method did not appear to underperform in any of the twelve scenarios we examined.

4. Incorrect links & unlinked records

Pseudo ML and Reweighting

- Solving weighted versions of score function where a record's weight is the inverse of the probability that a record remains unlinked: $P(U_i = 1 | X_i, Z_i)$
- For adjusting (sample) weights for linkage, Proc WTADJUST in Sudaan is useful (Judson and al. 2013)
  - Flexible enough for implementing different approaches to re-weighting
  - Weight adjustments are created using a model-based calibration approach

5. Application: Modeling the probability of a record remaining unlinked

- Applied to the linkage (Census cohort -Tax)
- Census cohort 2001 aged 19+ vs Tax T1 2000-2001
- Manual review showed that 99.8% of links to tax were true. So we may ignore the false positive linkage error
- Of the in-scope 2001 Census respondents linkage rate was 78.6% (no names in the census at that time)
- Linkage rate was lower for person of younger age [19-24]
- For persons of any Aboriginal identity (64% linked)
- For persons who have moved in the past year (63% LR)

5.1 Estimation Results

Age-standardized mortality rate (ASMR) per 100,000 person-years at risk, for women aged 20 years or greater at baseline (weighted and unweighted), by selected socioeconomic characteristics, Canada, 2001 - 2011.

		Unweighted				Weighted			
			95% CI				95% CI		
Variable	Number respondents	ASMR	from	to	width	ASMR	from	to	width
Employment									
Employed	1,049,085	693	665	723	58	680	652	708	55
Unemployed	72,310	964	870	1069	199	992	897	1087	189
Not in labour force	672,045	1027	1020	1033	13	1025	1016	1033	17
Educational attainment									
University degree	287,370	723	704	743	39	731	712	749	37
Postsecondary non-university	398,395	780	768	793	25	785	770	799	29
High school with/without trades certificate	611,275	872	862	881	19	871	861	880	19
Less than secondary school graduation	496,395	1055	1047	1062	15	*1045*	1037	1054	17

6. Challenges

- **Accurate models of linkage errors without training data**
  - Difficulty into practical applications is that nobody have developed suitably accurate methods for estimating all false match rates for all pairs when no training data is available (Winkler 2017)
- **How to get rid of manual review in FP and FN adjustment?**
- **Difficulty in determining unlinked records. How to discard unlinked records and otherwise records?**
- **How to measure RL errors and develop adjustment methods for multi-linkage?**

References

- Chambers, R., Chipperfield, J.O, Davis, W. and Kovacevic, M. (2009). *Regression Inference Based on Estimating equations and Probability-Linked Data*. Submitted to Publication
- Chipperfield, J.O, Bishop, G. R., and Campbell, P. (2011). *Maximum Likelihood estimation for contingency tables and logistic regression with incorrectly linked data*, Survey Methodology, Vol. 37, No. 1, pp 13-24
- Judson DH, Parker JD, Larsen,MD. (2013) *Adjusting sample weights for linkage-eligibility using SUDAAN.* National center for Health statistics, Hyattsville Maryland. May 2013. http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudan.pdf
- Neter, J., Maynes, E. S., and Ramanathan, R. (1965), *The effect of Mismatching on the Measurement of Response Errors*, Journal of the American statistical Association, 60, 1005-1027
- Wilkins, R., M. Tjepkema, C. Mustard, and R. Choinière. *2008 "The Canadian census mortality follow-up study, 1991 through 2001"*, Health Reports 19 (3): 25-43. Statistics Canada Catalogue no. 82-003-XPE
- Winkler, W. E., *Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation* (2017). Submitted to Publication

Questions? Comments? Ideas

Thank you for attending!
Merci de votre intérêt !

Abdelnasser.saidi@Canada.ca