

# Challenges and Strategies in Analysis of Administrative Health Data

**X. Joan Hu**

Department of Statistics and Actuarial Science  
Simon Fraser University, Canada  
*joanh@stat.sfu.ca*

*Joint work with Rosychuk and Xiong*

Presentation at BIRS Workshop 22w5010, UBC - Okanagan  
May 24, 2022

# Outline

## 1. Introduction

## 2. Large Administrative Health Data Analysis

## 3. Final Remarks

# 1.1 Introduction: Large Administrative Health Data

## ▶ Readily Available

- ▶ Canadian Provincial Medical Insurance Databases  
(*Canadian Health Care System: Universally Accessible, Government-Sponsored*)
- ▶ Disease/Patient Registries: e.g. BC Cancer Registry
- ▶ Vital Statistics: e.g. BC Vital Statistics
- ▶ Various EHRs (electronic health records): e.g. Lab Test Results
- ▶ Recent Administrative Records of Covid-19 Infection

# 1.1 Introduction: Large Administrative Health Data

## ▶ Readily Available

- ▶ Canadian Provincial Medical Insurance Databases  
(*Canadian Health Care System: Universally Accessible, Government-Sponsored*)
- ▶ Disease/Patient Registries: e.g. BC Cancer Registry
- ▶ Vital Statistics: e.g. BC Vital Statistics
- ▶ Various EHRs (electronic health records): e.g. Lab Test Results
- ▶ Recent Administrative Records of Covid-19 Infection

## ▶ Lots of Information

- ▶ Affordable to be looked at (analyzed) from *different* perspectives, and answer *various* questions

⇒ **Many attempts to use such data to achieve particular aims**

# 1.1 Introduction: Large Administrative Health Data

## Challenges?

- ▶ *Large!* BigData issues  
e.g. large in size and/or many variables

# 1.1 Introduction: Large Administrative Health Data

## Challenges?

- ▶ *Large!* BigData issues  
e.g. large in size and/or many variables
- ▶ *Administrative*
  - ▶ **not collected/recorded to serve a particular research purpose:** Required to plan carefully for data extraction and analysis formulation  
e.g. what time origin to use?
  - ▶ **incomplete/imperfect data:** Available data are not in the ideal format  
e.g. information only available on a target population within a time window: *left- and/or right-censoring; truncation; measurement error/misclassification*

## 1.2. Examples of Large Administrative Health Data Based Research Programs

**Example 1.** Cancer Survivorship Research Program, BC Cancer Agency (BCCA), Canada (Phase I: 2004 - 2018, PI: *Mary McBride*; Phase II: 2019 - present, PI: *Stuart Peacock*)

- ▶ Started from the Childhood, Adolescent, Young Adult Cancer (CAYACS) Survivorship Program: risk classification, assessment, and prediction
- ▶ A longitudinal cohort study to examine long-term effect resulted from cancer, using linked data from the *BC Health Insurance (MSP)* and the *BC Cancer Registry*: e.g. hospitalization records, physician claims, diagnoses of other diseases

## 1.2. Examples of Large Administrative Health Data Based Research Programs

**Example 2.** Health Service Provided by Emergency Department (ED), University of Alberta, Canada (2009 - present; PI: *Rhonda Rosychuk*)

- ▶ Started with records of mental health (MH) related ED presentations by Alberta residents younger than 18 year-old, from the *Ambulatory Care Classification System* (ACCS, Alberta Health and Wellness)
- ▶ Having been expanded to include data from the *Dynamic Cohort of Complex, High System Users* developed by the *Canadian Institute for Health Information* (CIHI)



## 1.2. Examples of Large Administrative Health Data Based Research Programs

**Example 3.** Clinical Management of Opioid Use Disorder (OUD), BC Centre for Excellence in HIV/AIDS, Canada (2018 - present; PI: *Bohdan Nosyk*)

- ▶ Using seven linked population-level administrative databases: *BC Health Insurance, Discharge Abstract Database, PharmaNet, BC Corrections, Vital Statistics, National Ambulatory Care Reporting System (NACRS), and Perinatal Database*
- ▶ Aiming to emulate target trials to strengthen the evidence base for clinical management

**Example 1.** Cancer Survivorship Research Program, BC Cancer Agency (BCCA), Canada (Phase I: 2004 - 2018, PI: Mary McBride; Phase II: 2019 - present, PI: Stuart Peacock)

*Ying (MSc 2006); Ma (MSc 2009); Wang (MSc 2012, PhD 2015); Zhao (PhD 2009); Li (PhD 2019)*

**Example 2.** Health Service Provided by Emergency Department (ED), University of Alberta, Canada (2009 - present; PI: Rhonda Rosychuk)

*Wang (MSc 2014); Thiessen (MSc 2020); Cui (Postdoc 2020); Xiong (Postdoc 2021); Chen (PhD 202x)*

**Example 3.** Clinical Management of Opioid Use Disorder (OUD), BC Centre for Excellence in HIV/AIDS, Canada (2018 - present; PI: Bohdan Nosyk)

*Kurz (MSc 2020); Thomson (PhD 202x)*

⇒ **The administrative health data have generated many interesting statistical problems.**

## 1.2. Examples of Large Administrative Health Data Based Research Programs

**Example 1.** Cancer Survivorship Research Program, BC Cancer Agency (BCCA), Canada (Phase I: 2004 - 2018, PI: Mary McBride; Phase II: 2019 - present, PI: Stuart Peacock)

**Example 2. Health Service Provided by Emergency Department (ED), University of Alberta, Canada (2009 - present; PI: Rhonda Rosychuk)**

**Example 3.** Clinical Management of Opioid Use Disorder (OUD), BC Centre for Excellence in HIV/AIDS, Canada (2018 - present; PI: Bohdan Nosyk)

⇒ **To motivate and illustrate the following presentation**

## 2. An Example for Analyses of Pediatric Mental Health Emergency Department (MHED) Visit Records

### 2.1 Formulation: Recurrent Events Data

### 2.2 Regression Analysis with Doubly-censored Data

### 2.3 Regression Analysis with Truncated Data

### 2.4 Analysis Outcomes

*PMHC Program Based on Emergency Department Visits (2009 - present; PI: Rhonda Rosychuk)*

**Goal.** To assess the need for pediatric mental health care and to improve the care system.

**Data.** Extracted from the Ambulatory Care Classification System (ACCS), a large, population-based database in Alberta, Canada, initiated in 1999 by the provincial health ministry.

**Specific Aims.**

- ▶ ... ..
- ▶ to evaluate frequency of children and youth's MHED visits
- ▶ to examine effects of risk factors/exposures, and identify important covariates and temporal changes
- ▶ ... ..

*PMHC Program Based on Emergency Department Visits (2009 - present; PI: Rhonda Rosychuk)*

**Goal.** To assess the need for pediatric mental health care and to improve the care system.

**Data.** Extracted from the Ambulatory Care Classification System (ACCS), a large, population-based database in Alberta, Canada, initiated in 1999 by the provincial health ministry.

**Specific Aims.**

- ▶ ... ..
- ▶ to evaluate frequency of children and youth's MHED visits
- ▶ to examine effects of risk factors/exposures, and identify important covariates and temporal changes
- ▶ ... ..

⇒ **Quick reaction: to conduct analysis of recurrent events**

# Available information to the PMHC Program?

The MHED data maintained at Rhonda Rosychuk's lab were extracted from *Alberta Health Care Insurance Plan*, including

- ▶ all the records of MHED visits from Alberta residents aged younger than 18 year-old during (i) April 1 2002 - March 31 2011, and (ii) April 1 2010 - March 31 2017.

# Available information to the PMHC Program?

The MHED data maintained at Rhonda Rosychuk's lab were extracted from *Alberta Health Care Insurance Plan*, including

- ▶ all the records of MHED visits from Alberta residents aged younger than 18 year-old during (i) April 1 2002 - March 31 2011, and (ii) April 1 2010 - March 31 2017.
- ▶ Each record includes
  - ▶ starting/ending date, time; age (in year) at service
  - ▶ triage level; discharge disposition
  - ▶ sex; pSES; residence region
  - ▶ birthdate if from April 1 2010 - March 31 2017



# Available information to the PMHC Program?

The MHED data maintained at Rhonda Rosychuk's lab were extracted from *Alberta Health Care Insurance Plan*, including

- ▶ all the records of MHED visits from Alberta residents aged younger than 18 year-old during (i) April 1 2002 - March 31 2011, and (ii) April 1 2010 - March 31 2017.
- ▶ Each record includes
  - ▶ starting/ending date, time; age (in year) at service
  - ▶ triage level; discharge disposition
  - ▶ sex; pSES; residence region
  - ▶ birthdate if from April 1 2010 - March 31 2017

**How to summarize/process the data to achieve the aims?**

## 2.1 Formulation of Alberta Pediatric MHED Visits

Consider the individuals with MHED records as *study units* ... ..

Subject  $i$ :

- ▶  $N_i(t)$  – the cumulative count of the MHED visits upto time  $t$
- ▶  $Z_i(t)$  – potential covariates at  $t$

## 2.1 Formulation of Alberta Pediatric MHED Visits

Consider the individuals with MHED records as *study units* ... ..

Subject  $i$ :

- ▶  $N_i(t)$  – the cumulative count of the MHED visits upto time  $t$
- ▶  $Z_i(t)$  – potential covariates at  $t$

*Questions* before a meaningful statistical analysis ...

- ▶ What is an appropriate time?
- ▶ What is  $Z_i(\cdot)$  to consider?
- ▶ What model to assume: it's appropriate; inference on it is meaningful and feasible to make with the current data?

## 2.1A Formulation of Alberta Pediatric MHED Visits

To *begin* with, **what information is available?**

- ▶ Focus on the individuals with MHED records,  $i = 1, \dots, n$
- ▶ Assume the individuals are independent
- ▶ Consider only external time-independent covariates

## 2.1A Formulation of Alberta Pediatric MHED Visits

To *begin* with, **what information is available?**

- ▶ Focus on the individuals with MHED records,  $i = 1, \dots, n$
- ▶ Assume the individuals are independent
- ▶ Consider only external time-independent covariates

For Subject  $i$ , consider *age* as the individual time:

- ▶  $N_i(a)$ ,  $a > 0$ : the cumulative count of Subject  $i$ 's MHED visits since birth at age  $a > 0$
- ▶  $Z_i$ : all the time-indtpt covariates of interest

## 2.1A Formulation of Alberta Pediatric MHED Visits

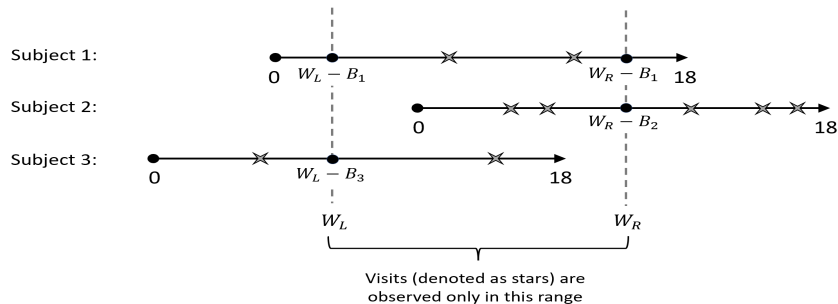
Further ...

- ▶ The data extraction window in the calendar time:  $[W_L, W_R]$
- ▶ The information available on Subject  $i$  is over the window in age:  $(C_{Li}, C_{Ri})$ 
  - ▶  $C_{Li} = \max(0, W_L - B_i), C_{Ri} = \min(18, W_R - B_i)$ , birthdate  $B_i$
- ▶ Available is the calendar time of Subject  $i$ 's  $j$ th recorded ED visit ( $T_{ij}$ ) but the age  $A_{ij} = T_{ij} - B_i$  is recorded in years.

## 2.1A Formulation of Alberta Pediatric MHED Visits

Further ...

- ▶ The data extraction window in the calendar time:  $[W_L, W_R]$
- ▶ The information available on Subject  $i$  is over the window in age:  $(C_{Li}, C_{Ri})$ 
  - ▶  $C_{Li} = \max(0, W_L - B_i), C_{Ri} = \min(18, W_R - B_i)$ , birthdate  $B_i$
- ▶ Available is the calendar time of Subject  $i$ 's  $j$ th recorded ED visit ( $T_{ij}$ ) but the age  $A_{ij} = T_{ij} - B_i$  is recorded in years.



## 2.1A Formulation of Alberta Pediatric MHED Visits

Further ...

- ▶ The data extraction window in the calendar time:  $[W_L, W_R]$
- ▶ The information potentially available on Subject  $i$  is over the window in age:  $(C_{Li}, C_{Ri})$ 
  - ▶  $C_{Li} = \max(0, W_L - B_i), C_{Ri} = \min(18, W_R - B_i)$ , birthdate  $B_i$
- ▶ Available is the calendar time of Subject  $i$ 's  $j$ th recorded ED visit ( $T_{ij}$ ) but the age  $A_{ij} = T_{ij} - B_i$  is recorded in years.

**Provided all  $B_i$  available**, the available data are *doubly-censored*:

$$\bigcup_{i \in \mathcal{O}_1} \left[ \{dN_i(a) : a \in (C_{Li}, C_{Ri}]\}; Z_i \right]$$



## 2.1A Formulation of Alberta Pediatric MHED Visits

Further ...

- ▶ The data extraction window in the calendar time:  $[W_L, W_R]$
- ▶ The information potentially available on Subject  $i$  is over the window in age:  $(C_{Li}, C_{Ri})$ 
  - ▶  $C_{Li} = \max(0, W_L - B_i), C_{Ri} = \min(18, W_R - B_i)$ , birthdate  $B_i$
- ▶ Available is the calendar time of Subject  $i$ 's  $j$ th recorded ED visit ( $T_{ij}$ ) but the age  $A_{ij} = T_{ij} - B_i$  is recorded in years.

**Provided all  $B_i$  available**, the available data are *doubly-censored*:

$$\bigcup_{i \in \mathcal{O}_1} \left[ \{dN_i(a) : a \in (C_{Li}, C_{Ri})\}; Z_i \right]$$

**If  $B_i$  unavailable?** the available data are *doubly-censored with missing/coarsened censoring time*:

- ▶  $C_{Li} = \max(0, W_L - ???), C_{Ri} = \min(18, W_R - ???)$

## 2.1B Formulation of Alberta Pediatric MHED Visits

Note that study subjects associated with the PMHC database form a sample ( $\mathcal{O}_1$ ) from a *biased* sub-population ( $\mathcal{P}_1$ ) of the general population ( $\mathcal{P}$ ), a sample of which the Alberta residents under 18-year-old during 2002-2017 form ( $\mathcal{O}$ ).

- ▶ The available MHED visits are *truncated data* when  $\mathcal{P}$  is targeted.

*How to make inference with the data?*

## 2.1B Formulation of Alberta Pediatric MHED Visits

Note that study subjects associated with the PMHC database form a sample ( $\mathcal{O}_1$ ) from a *biased* sub-population ( $\mathcal{P}_1$ ) of the general population ( $\mathcal{P}$ ), a sample of which the Alberta residents under 18-year-old during 2002-2017 form ( $\mathcal{O}$ ).

- ▶ The available MHED visits are *truncated data* when  $\mathcal{P}$  is targeted.

*How to make inference with the data?*

Instead of modeling the truncation, following Hu and Lawless (1996),

- ▶ to use the available information on  $\mathcal{O}_0 = \mathcal{O} \setminus \mathcal{O}_1$ : subjects in  $\mathcal{O}_0$  did not experience any MHED during 2002-2017

$$\left[ \bigcup_{i \in \mathcal{O}_1} [\{dN_i(a) : a \in (C_{Li}, C_{Ri})\}; Z_i] \right] \cup \left[ \bigcup_{i \in \mathcal{O}_0} \{dN_i(a) = 0 : a \in (C_{Li}, C_{Ri})\} \right]$$

## 2.1B Formulation of Alberta Pediatric MHED Visits

Note that study subjects associated with the PMHC database form a sample ( $\mathcal{O}_1$ ) from a *biased* sub-population ( $\mathcal{P}_1$ ) of the general population ( $\mathcal{P}$ ), a sample of which the Alberta residents under 18-year-old during 2002-2017 form ( $\mathcal{O}$ ).

- ▶ The available MHED visits are *truncated data* when  $\mathcal{P}$  is targeted.

*How to make inference with the data?*

Instead of modeling the truncation, following Hu and Lawless (1996),

- ▶ to use the available information on  $\mathcal{O}_0 = \mathcal{O} \setminus \mathcal{O}_1$ : subjects in  $\mathcal{O}_0$  did not experience any MHED during 2002-2017

$$\left[ \bigcup_{i \in \mathcal{O}_1} [\{dN_i(a) : a \in (C_{Li}, C_{Ri})\}; Z_i] \right] \cup \left[ \bigcup_{i \in \mathcal{O}_0} \{dN_i(a) = 0 : a \in (C_{Li}, C_{Ri})\} \right]$$

- ▶ to integrate with the available demographic information (on  $Z$ ) of  $\mathcal{O}$ .

## 2.1C Formulation: A Two-sample Problem

To explore temporal changes in frequency of children and youth's MHED visits and the associated effects of risk factors/exposures,  $\implies$

**Focus on comparisons of MHED visits 2002-2010 (early decade) vs MHED visits 2010-2017 (late decade) via**

- ▶ (i) the *doubly-censored* recurrent events data in §2.2, targeting on  $\mathcal{P}_1$ ;
- ▶ (ii) the *truncated* recurrent events data integrated with population based demographic information in §2.3, targeting on  $\mathcal{P}$ ;

## 2.1C Formulation: A Two-sample Problem


To explore temporal changes in frequency of children and youth's MHED visits and the associated effects of risk factors/exposures,  $\implies$

**Focus on comparisons of MHED visits 2002-2010 (early decade) vs MHED visits 2010-2017 (late decade) via**

- ▶ (i) the *doubly-censored* recurrent events data in §2.2, targeting on  $\mathcal{P}_1$ ;
- ▶ (ii) the *truncated* recurrent events data integrated with population based demographic information in §2.3, targeting on  $\mathcal{P}$ ;

For Subject  $i \in \mathcal{P}^*$  (the target population), assume a marginal model (cf: Hu and Rosychuk, 2016):

$$E[dN_i(a)|Z_i, X_i] = \exp\{\alpha(a)X_i + \beta'(a)Z_i + \gamma'(a)Z_iX_i\}d\Lambda_0(a),$$

$X_i$  is the indicator of Subject  $i$  from late decade. 

For Subject  $i \in \mathcal{P}^*$  (the target population),

$$E[dN_i(a)|Z_i, X_i] = \exp\{\alpha(a)X_i + \beta'(a)Z_i + \gamma'(a)Z_iX_i\}d\Lambda_0(a), \quad (1)$$

$X_i$  is the indicator of Subject  $i$  from late decade.

### Remarks:

- ▶ Model (1) is equivalent to

$$E[dN_i(a)|Z_i, X_i] = \begin{cases} \exp\{\beta'_E(a)Z_i\}d\Lambda_{0E}(a), & i \text{ with early decade} \\ \exp\{\beta'_L(a)Z_i\}d\Lambda_{0L}(a), & i \text{ with late decade} \end{cases}$$

$$\beta_E(a) = \beta(a), \Lambda_{0E}(a) = \Lambda_0(a), \beta_L(a) = \beta(a) + \gamma(a), \text{ and} \\ d\Lambda_{0L}(a) = \exp(\alpha(a))d\Lambda_0(a).$$

For Subject  $i \in \mathcal{P}^*$  (the target population),

$$E[dN_i(a)|Z_i, X_i] = \exp\{\alpha(a)X_i + \beta'(a)Z_i + \gamma'(a)Z_iX_i\}d\Lambda_0(a), \quad (1)$$

$X_i$  is the indicator of Subject  $i$  from late decade.

### Remarks:

- ▶ Model (1) is equivalent to

$$E[dN_i(a)|Z_i, X_i] = \begin{cases} \exp\{\beta'_E(a)Z_i\}d\Lambda_{0E}(a), & i \text{ with early decade} \\ \exp\{\beta'_L(a)Z_i\}d\Lambda_{0L}(a), & i \text{ with late decade} \end{cases}$$

$$\beta_E(a) = \beta(a), \Lambda_{0E}(a) = \Lambda_0(a), \beta_L(a) = \beta(a) + \gamma(a), \text{ and} \\ d\Lambda_{0L}(a) = \exp(\alpha(a))d\Lambda_0(a).$$

- ▶ We may choose to consider different specifications of Model (1): e.g. Model CCC is  $\alpha(a) \equiv \alpha$ ,  $\beta(a) \equiv \beta$ ,  $\gamma(a) \equiv \gamma \implies$  the proportional mean/rate (Andersen and Gill) model.



## 2.2 Analysis of Doubly-censored Recurrent Events

### Data with $\mathcal{P}^* = \mathcal{P}_1$

Provided available birthdates and independent subjects, consider the estimating functions under Model (1): for  $a \in [\tau_L, \tau_R]$ ,

$$U(\phi(a); a | \mathbf{B}) = \int_0^{18} K_h(u - a) \sum_{i \in \mathcal{O}_1} Y_i(u | B_i) \{ V_i^*(u, a) - \bar{V}^*(\phi(a); u, a) \} dN_i(u)$$

$$\begin{aligned} \phi(a) &= \left( \theta(a)', \dot{\theta}(a)' \right)' \text{ with } \theta(\cdot) = (\alpha(\cdot), \beta(\cdot)', \gamma(\cdot)')'; \text{ a kernel function } K(\cdot); \\ V_i^*(u, a) &= (V_i', (u - a)V_i')' \text{ with } V_i = (X_i, Z_i', X_i Z_i')'; \\ \bar{V}^*(\phi(a); u, a) &= S^{(1)}(\phi(a); u, a) / S^{(0)}(\phi(a); u, a) \text{ with} \\ S^{(q)}(\phi(a); u, a) &= \sum_{i \in \mathcal{O}_1} Y_i(u) [V_i^*(u, a)]^q \exp\{\phi(a) V_i^*(u, a)\} \end{aligned}$$

### Remarks:

- ▶ following Hu and Rosychuk (2016) to accommodate the situations without birthdates by assuming birthdates uniformly distributed over a year.
- ▶ checking for the required independence assumption for the two groups.

## 2.3 Analysis of Truncated Recurrent Events Data with $\mathcal{P}^* = \mathcal{P}$

By the available Alberta census information during 2022-2017, consider the estimating functions under Model (1): for  $a \in [\tau_L, \tau_R]$ ,

$$U(\phi(a); a | \mathbf{B}) = \int_0^{18} K_h(u - a) \sum_{i \in \mathcal{O}} Y_i(u | B_i) \left\{ V_i^*(u, a) - \tilde{V}^*(\phi(a); u, a) \right\} dN_i(u),$$

where  $\tilde{V}^*(\phi(a); u, a)$  is an approximation to

$$\bar{V}^*(\phi(a); u, a) = \frac{\sum_{m \in \mathcal{O}} Y_m(u | B_m) V_m^*(u, a) \exp\{\phi'(a) V_m^*(u, a)\}}{\sum_{m \in \mathcal{O}} Y_m(u | B_m) \exp\{\phi'(a) V_m^*(u, a)\}}.$$

$\phi(a) = (\theta(a)', \dot{\theta}(a)')$  with  $\theta(\cdot) = (\alpha(\cdot), \beta(\cdot)', \gamma(\cdot)')$ , and

$V_i^*(u, a) = (V_i', (u - a)V_i')$  with  $V_i = (X_i, Z_i', X_i Z_i')$ .

### Remarks:

- ▶ the approximation by the aggregated population information?
- ▶ asymptotic properties of the estimator?

## 2.4 Analysis Outcomes

### ► Estimates for Time-independent Coefficients

**TABLE:** Parameter estimates and estimated standard errors of time-independent covariate effects with Model CCC<sup>a</sup> with PMHC data

| Covariates |                              | wrt $\mathcal{P}_1$ | wrt $\mathcal{P}$  |
|------------|------------------------------|---------------------|--------------------|
| $\alpha$   | Period                       | <b>.282(.020)</b>   | <b>.605(.020)</b>  |
|            | Sex (vs. female)             | <b>-.049(.016)</b>  | <b>-.414(.016)</b> |
| $\beta$    | Edmonton (vs. other regions) | <b>.050(.021)</b>   | <b>-.225(.019)</b> |
|            | Calgary (vs. other regions)  | .003 (.019)         | <b>-.387(.019)</b> |
|            | PeriodSex                    | <b>-.058(.023)</b>  | <b>-.124(.023)</b> |
| $\gamma$   | PeriodEdmonton               | .003(.029)          | <b>-.080(.029)</b> |
|            | PeriodCalgary                | .015(.026)          | <b>.196(.026)</b>  |

Model CCC<sup>a</sup>:  $E[dN_i(a)|Z_i, X_i] = \exp(\alpha X_i + \beta' Z_i + \gamma' Z_i X_i) d\Lambda_0(a)$ .

## 2.4 Analysis Outcomes

### ► Estimates for Time-independent Coefficients (cont'd)

**TABLE:** Parameter estimates and standard errors for time-independent covariate effects with Model VCC<sup>b</sup> with data from the two time periods

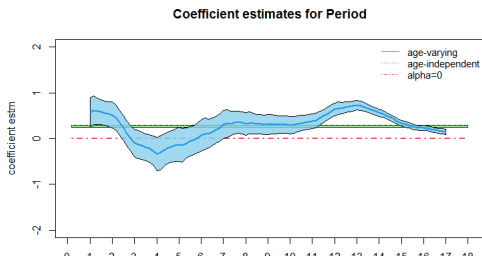
| Covariates |                              | wrt $\mathcal{P}_1$ | wrt $\mathcal{P}$  |
|------------|------------------------------|---------------------|--------------------|
| $\alpha^*$ | Period                       | –                   | –                  |
|            | Sex (vs. female)             | <b>-.049(.016)</b>  | <b>-.410(.015)</b> |
| $\beta$    | Edmonton (vs. other regions) | <b>.051(.021)</b>   | <b>-.218(.019)</b> |
|            | Calgary (vs. other regions)  | .004(.019)          | <b>-.385(.018)</b> |
|            | PeriodSex                    | <b>-.055(.020)</b>  | <b>-.130(.020)</b> |
| $\gamma$   | PeriodEdmonton               | .001(.029)          | <b>-.089(.025)</b> |
|            | PeriodCalgary                | .015(.026)          | <b>.198(.022)</b>  |

Model VCC<sup>b</sup>:  $E[dN_i(a)|Z_i, X_i] = \exp(\alpha(a)X_i + \beta'Z_i + \gamma'Z_iX_i)d\Lambda_0(a)$ .

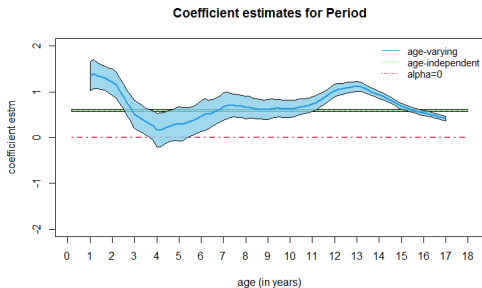
\*: time-varying model components.

## 2.4 Analysis Outcomes: $\alpha(\cdot)$ Estimates with Models CCC, VVV

(a)



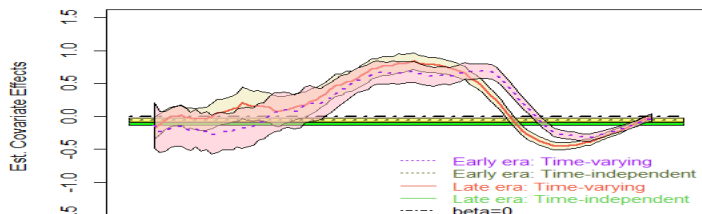
(b)



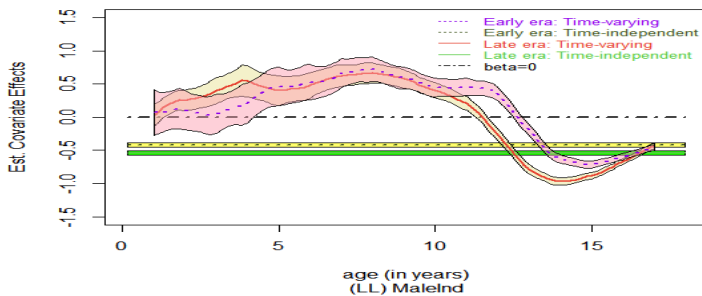
(a) the MHEd cohort ( $\mathcal{P}^* = \mathcal{P}_1$ ); (b) the Alberta residents ( $\mathcal{P}^* = \mathcal{P}$ )

## 2.4 Analysis Outcomes: Estimates of $\beta_E(\cdot)$ and $\beta_L(\cdot)$ (Sex) with Models CCC, VVV – (a) MHED cohort ( $\mathcal{P}^* = \mathcal{P}_1$ ); (b) Alberta residents ( $\mathcal{P}^* = \mathcal{P}$ )

(a)



(b)



### 3. Final Remarks: What have we done for the PMHC program?

This project has conducted

- ▶ useful summary of the MHED data (from  $\mathcal{O}_1$ ), and then  $\mathcal{O}$ .
- ▶ statistical inference on  $\mathcal{P}_1$ , and then  $\mathcal{P}$  if  $\mathcal{O}_1/\mathcal{O}$  forms a random sample.
- ▶ What if  $\mathcal{O}_1$  or  $\mathcal{O}$  is not a random sample of the target population?

### 3. Final Remarks: What have we done for the PMHC program?

This project has conducted

- ▶ useful summary of the MHED data (from  $\mathcal{O}_1$ ), and then  $\mathcal{O}$ .
- ▶ statistical inference on  $\mathcal{P}_1$ , and then  $\mathcal{P}$  if  $\mathcal{O}_1/\mathcal{O}$  forms a random sample.
- ▶ What if  $\mathcal{O}_1$  or  $\mathcal{O}$  is not a random sample of the target population?

With  $\mathcal{D} = \{X_1, \dots, X_n\}$ ,

- ▶ to calculate the sample mean  $\bar{X} = \sum_{i=1}^n X_i/n$  to summarize  $\mathcal{D}$ .
- ▶ If  $\mathcal{D}$  is a random sample of the target population  $\mathcal{P}$  with mean  $\mu$ , one may use  $\bar{X}$  to estimate  $\mu$  and form  $\mu$ 's (approximate) confidence interval.
- ▶ What if not?



### 3. Final Remarks: What have we done for the PMHC program?

This project has conducted

- ▶ useful summary of the MHED data (from  $\mathcal{O}_1$ ), and then  $\mathcal{O}$ .
- ▶ statistical inference on  $\mathcal{P}_1$ , and then  $\mathcal{P}$  if  $\mathcal{O}_1/\mathcal{O}$  forms a random sample.
- ▶ What if  $\mathcal{O}_1$  or  $\mathcal{O}$  is not a random sample of the target population?

With  $\mathcal{D} = \{X_1, \dots, X_n\}$ ,

- ▶ to calculate the sample mean  $\bar{X} = \sum_{i=1}^n X_i/n$  to summarize  $\mathcal{D}$ .
- ▶ If  $\mathcal{D}$  is a random sample of the target population  $\mathcal{P}$  with mean  $\mu$ , one may use  $\bar{X}$  to estimate  $\mu$  and form  $\mu$ 's (approximate) confidence interval.
- ▶ What if not? e.g. *inverse probability weighted estimator*

## 3. Final Remarks: Whatelse to do for the PMHC program?

*Many interesting issues to further explore:*

- ▶ What if  $B_i \not\sim$  the uniform distn?
- ▶ Use alternative models to the proportional type regression models? e.g. to consider intensity-based modeling
- ▶ How to explore and account for spatio-temporal correlation underlying the MHED visits?
- ▶ Consider different study units, such as health regions/sub-regions, and calendar time, to explore the visits spatio-temporally?
- ▶ ... ..

## 3. Final Remarks: Large Administrative Health Data

- ▶ **Readily Available**
- ▶ **Lots of Information**
  - ▶ Affordable to be looked at/analyzed from *different* perspectives, to answer *many* interesting questions
- ▶ **Challenges = Interesting/Not Boring**
  - ▶ *Large!* BigData issues
  - ▶ *Administrative!* **not collected/recorded to serve a particular research purpose**
- ▶ **To Use Administrative Data**
  - ▶ Careful design/plan is required for extraction.
  - ▶ When analyzing administrative data,
    - ▶ account for their special features appropriately, and
    - ▶ **consider to use available additional information**  
⇒ efficient, robust, easy-to-implement procedures

# Acknowledgements

- ▶ The key part of the presentation is based on a recent joint work with *R. Rosychuk* (Univ of Alberta) and *Y. Xiong* (Fred Hutchinson Cancer Research Center), sponsored by NSERC and CIHR: Xiong, Hu, and Rosychuk (2022).
- ▶ A comment from *Jerry Lawless* on a previous version of the presentation has led to a big modification of the work.

**Thank-you for the attention!**