Identifying Causal Effects of Medication Exposures from Electronic Health Record Data with i2b2

Victor M. Castro, Michael C. Ollendieck, Shawn N. Murphy

4th Workshop on Data Mining for Medical Informatics: Causal Inference for Health Data Analytics

November 4, 2017
Aims

• Build a tool to identify causal drug effects in large EHR datasets

• Integrate with i2b2 framework to enable reproducibility and generalizability across many sites

• Control for confounding typically encountered in observation EHR data:
  • Confounding by unmeasured observations (the open system problem)
  • Confounding by severity of disease/utilization:
    • Sicker patients visit the doctor/get admitted more often and are more likely to diagnoses
    • Onset of disease may not be well documented

• Applications include:
  • Identifying unknown drug side effects
  • Repurposing existing drugs
  • Identifying treatment-resistant subgroups
Matched Case-Controls Design

- Case-control studies are a staple of observational data analysis that can be applied to EHR data for large-scale drug screening
  - Cases = patients with a disease
  - Controls = patients without a disease
  - Exposure window = time preceding onset of a disease

- Defining the cohorts based on the disease (outcome) allows efficiently screening many drugs
Confounding by Severity of Illness/Utilization

• Using matching and adjusting for utilization measures, we can mitigate confounding effects

Castro, et al, 2014 - JBI
Coarsened Exact Matching

- Coarsen matching variables based on heuristics or common histogram binning techniques

- Perform exact matching on the coarsened variables

- Group observations into strata

- Prune any stratum with 0 case or 0 controls

- “Sacrifice some data to avoid bias” - Blackwell
Coarsened Exact Matching

Examples of Coarsened Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coarsened to</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at index = 38.5 years</td>
<td>coarsened to 35-39</td>
<td>35-39</td>
</tr>
<tr>
<td>Year of index event = 2009</td>
<td>coarsened to 2009-2012</td>
<td>2009-2012</td>
</tr>
<tr>
<td>Gender = F</td>
<td>not coarsened</td>
<td>F</td>
</tr>
<tr>
<td>Race = Asian</td>
<td>coarsened to Non-Caucasian</td>
<td>Non-Caucasian</td>
</tr>
</tbody>
</table>

Implemented using cem R package
Analysis Pipeline

Source EMR Data

- i2b2 Star Schema

SQL stored procedure extracts data to CSV

Analysis Files

- CSV
- Demographics
- Concept Dict.
- Analysis Facts

Run method

- Load Data
- Data Filter/Clean
- Matching
- Case-control Analysis

Results / Visualizations
Experimental Setting

• Population: 69,121 patients consented in the Partners Healthcare Biobank

• Cases
  • Patients with a diagnosis of Osteoporosis (ICD-10: M80.*; M81.*; ICD-9 733.0*)

• Controls
  • No lifetime history of Osteoporosis
  • Matched using CEM on index_year, age_at_index, gender and race

• Exposures
  • All RxNorm drug ingredients/combinations) prescribed to at least 100 patients (901 drugs)

• Effect Estimates
  • Unadjusted Risk Ratio (riskratio.boot from R epitools package)
  • logit: Logistic Regression (adjusted for index_year and number of visits in window (log-adjusted)
  • clogit: Conditional logistic regression (adjusted for index_year, number of visits and matching stratum from R survival package)
Time Parameters

• **Index Date:**
  • Cases: First record diagnosis of Osteoporosis
  • Controls: Random visit date selected on the same year of the matched case

• **Exposure Window:**
  • 730 days (2 years) to 30 days prior to index date
  • Patients with no visits in the exposure window are excluded from both cases and controls
Effect Estimates

<table>
<thead>
<tr>
<th>Osteoporosis</th>
<th>Unexposed</th>
<th>Exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO Osteoporosis</td>
<td>(a) NO Osteoporosis + Drug NOT prescribed in exposure window</td>
<td>(b) Osteoporosis + Drug NOT prescribed in exposure window</td>
</tr>
<tr>
<td>NO Osteoporosis</td>
<td>(c) NO Osteoporosis + Drug prescribed in Exposure Window</td>
<td></td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>(d) Osteoporosis + Drug prescribed in exposure window</td>
<td></td>
</tr>
</tbody>
</table>

Unadjusted Risk Ratio (RR)

**Logistic regression:**
Drug Prescribed in Window (0/1) ~ Osteoporosis (0/1) + Year of Index Event + log(visits in exposure window)

**Conditional logistic regression:**
Drug Prescribed in Window (0/1) ~ Osteoporosis (0/1) + Year of Index Event + log(visits in exposure window) + strata(match_strata)
Results

Osteoporosis Case vs Matched Controls - All Drug Exposures
Results – No control for utilization

Osteoporosis Case vs Matched Controls - All Drug Exposures
Results

Psychotic Disorder Case vs Matched Controls - All Drug Exposures

- Valproic acid
- Clonazepam
- Risperidone
- Quetiapine
- Lithium
- Aripiprazole
- Trazodone
- Mirtazapine
- Hydroxyzine
- Benztpine
- Oxcarbazepine
- Bupropion
- Fluoxetine
- Duloxetine
- Haloperidol
- Prazosin
- Gabapentin
- Topiramate
- Citalopram
- Phenytoin
- Levetiracetam
- Buspirone
- Paroxetine
- Midazolam
- Carbamazepine
- Sertraline
- Fentanyl
- Nicotine
- Nicotine polacrilex
- Thiamine

Legend:
- Antidotes, detergents and poison control
- Antihistamines
- Antimicrobials
- Antineoplastics
- Antiparasitics
- Autonomic medications
- Blood products/modifiers/volume expanders
- Cardiovascular medications
- Central nervous system medications
- Dental and oral agents, topical
- Dermatological agents
- Diagnostic agents
- Gastrointestinal medications
- Genitourinary medications
- Herbs/alternative therapies
- Hormones/synthetics/modifiers
- Immunological agents
- Musculoskeletal medications
- Nasal and throat agents, topical
- Ophthalmic agents
- Otic agents
- Pharmaceutical aids/reagents
- Respiratory tract medications
- Therapeutic nutrients/minerals/electrolytes
Conclusion

• A case-control approach applied to large EHR datasets can identify true causal effects of drug exposures with the aim of monitoring the safety of medications and identifying candidates for drug repurposing.

• The Ri2b2casecontrol tools implements the methods in an i2b2 framework.

• Future efforts will be aimed at minimizing bias due to missing data and inaccurate outcome definitions.

• Longer-term goal of incorporating genomic and environmental data into methods.

• Improve data workflow with UI within the i2b2 webclient.
Ri2b2casecontrol R package

https://github.com/vcastro/Ri2b2casecontrol

case_control

Description
The main case control function to run a case control analysis from i2b2-generated data files

Usage

```
case_control(p, d, dict, exposure_cds, visit_cd = "V", outcome_cd = "O",
             riskWindow_daysPreIndex_start = 365, riskWindow_daysPreIndex_end = 1,
             controls_num = 3)
```

Arguments

- `p` A data frame of patients demographics
- `d` A data frame of patient data in a tall file
- `dict` A data frame of descriptions for each concept_cd in the data_tbl
- `exposure_cds` A string vector of exposures codes to test
- `visit_cd` The code used for visits (default is V)
- `outcome_cd` The code used for the outcome in the data_tbl (default is O)
- `riskWindow_daysPreIndex_start` The number of days prior to the index date to start the risk window (default is 365)
- `riskWindow_daysPreIndex_end` The number of days prior to the index date to end the risk window (default is 1)
- `controls_num` Number of controls to match to each case

Value
Resources

• Standalone R package
  • Ri2b2casecontrol
    https://github.com/vcastro/Ri2b2casecontrol
  • Ri2b2matchcontrols (implements CEM based on R cem package)
    https://github.com/vcastro/Ri2b2matchcontrols

• https://i2b2.org/

• Questions

• vcastro@partners.org