Using the Literature to Identify Confounding Variables for Performing Pharmacovigilance with Clinical Notes

Scott Malec, Assaf Gottlieb, Elmer Bernstam, Trevor Cohen
Presented By: Scott Malec, MLIS, MSIT
PhD Candidate, Pre-Doctoral NLM Fellow, & W.M. Keck Center Trainee
University of Texas Health Science Center at Houston
School of Biomedical Informatics
2017-11-04
Overview

- Motivation
- Background
  - Confounding
  - Causal Modeling
  - Literature-Based Discovery
- Data and Methods
- Results and Discussion
- Current/Future Work
Motivation

- **Adverse drug events (ADEs):** 807K events and 124K deaths in 2014 (FAERS)
- **Pharmacovigilance:** *FDA approval is not the end*
- **Spontaneous Reporting Systems (SRSs) (e.g., FAERS, EudraVigilance):**
  - Underreporting, lack of context, inaccuracy, Botsis, 2015; Hersh et al., 2013
- **Electronic Health Records (EHR):**
- **Confounding** introduces bias in between predictor and outcome of interest
- **Prior Approaches**
- **Why causal modeling and discovery?**
  - We wish to know the direction of influence, not merely correlation
  - \( X \rightarrow Y : X \text{ “CAUSES” } Y \)
Literature-Based Confounder Discovery

- Literature Based Discovery ("LBD")
  - Swanson et al., 1986 (Raynaud’s syndrome and fish oil)
  - Hristovski et al., 2006 (Discovery Patterns [“DPs”])
  - Kilicoglu et al., 2012 (SemMedDB)
  - Shang et al., 2014 (DPs for Pharmacovigilance)

**DPs:** relational constraints that help discover meaningful implicit relationships.

---

lithium  \[\text{INHIBITS}\] sodium channel  \[\text{ASSOCIATED WITH}\] Brugada Syndrome
Hypothesis

**Hypothesis:** Predictions from literature-informed causal models will be more accurate than those from unadjusted statistical correlation.

**Goal:** Improve detection of drug safety signals in clinical notes.
Hypothesis

**Hypothesis**: Predictions from literature-informed causal models will be more accurate than those from unadjusted statistical correlation.

**Goal**: Improve detection of drug safety signals in clinical notes.
Data and Knowledge

- **Data**
  - Curated reference set - 399 drug-ADE pairs, *Ryan et al., 2013*
  - UT Clinical Data Warehouse subset ~2.2M outpatient notes (~364k patients)
  - Processed notes with MedLEE, *Friedman et al., 1995* → ~ 33K UMLS concepts

- **Knowledge**
  - SemMedDB: ~70M predications extracted from MEDLINE as triple stores
  - EpiphaNet LBD system ([http://epiphanet.uth.tmc.edu](http://epiphanet.uth.tmc.edu)), *Cohen et al., 2010*

- **Reference standard**
  - 165 positive and 234 negative examples of drug ADE pairs, *Ryan et al., 2013*
  - Acute kidney injury; acute liver injury; gastrointestinal bleed; myocardial infarction
Method

- Identify confounders - query EpiphaNet backend
  Discovery Pattern: drug TREATS x; x CAUSES ADE
- Test confounders using observational clinical data using FGeS in TETRAD, Ramsey, 2015
- Combinatorial expansion of all unique sets of confounders
Method

- Identify confounders - query EpiphaNet backend
  Discovery Pattern: drug TREATS x; x CAUSES ADE

- Test confounders using observational clinical data using FGeS in TETRAD, Ramsey, 2015

- Combinatorial expansion of all unique sets of confounders
Combinatorial Expansion

- A continuous value is required in order to calculate AUC using “ground truth” in reference dataset, *Ryan et al., 2013*
- Say we have a set of confounders \{ A, B, C \} that we wish to include in a model. We identify all unique permutations (order does not matter).
  \{ [A], [B], [C], [A, B], [A, C], [B, C], [A, B, C] \}
- To calculate score for this set of confounders given 4 directed edges, the score would be \( \frac{4}{7} = 0.5714 \)
<table>
<thead>
<tr>
<th>ADE</th>
<th>Baseline AUC (unadjusted logistic regression)</th>
<th>Causal model AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gastrointestinal Bleeding</td>
<td>0.5643</td>
<td>0.6912</td>
</tr>
<tr>
<td>Acute Kidney Injury</td>
<td>0.5547</td>
<td>0.6598</td>
</tr>
<tr>
<td>Acute Liver Injury</td>
<td>0.4957</td>
<td>0.5449</td>
</tr>
<tr>
<td>Acute Myocardial Infarction</td>
<td>0.4946</td>
<td>0.56</td>
</tr>
<tr>
<td>Overall</td>
<td>0.504</td>
<td>0.5704</td>
</tr>
</tbody>
</table>
Discussion

- **Findings**
  - LBD was able to identify confounders in the EHR
  - Improvements of 0.10-0.13 when baseline is above noise

- **Limitations**
  - Improvements only evident when there is already decent “signal”
  - Co-medications would have been helpful
  - Performance is not on par with meta-analysis, Li et al., 2015, Harpaz et al., 2017
What is next?

- **Future Work**
  - Replace combinatorial expansion procedure with *parameter estimation*
  - Incorporate more procedures to reject unhelpful covariates
  - Identify and incorporate *co-medications* confounders
Thank you!

The End

This research was generously supported by US National Library of Medicine grant R01LM011563, NIH/BD2K supplement R01LM011563-02S1, and by the NLM Training Program in Biomedical Informatics (NLM Grant No. T15 LM007093).

Acknowledgements
Carnegie Mellon University
* Clark Glymour, PhD
University of Pittsburgh
* Harry Hochheiser, PhD
* Richard Boyce, PhD
University of Borås
* Sándor Darányi
UTHealth SBMI:
* Swaroop Gantela, MD
University of Michigan
* Frank Manion, PhD
Anonymous Reviewers