Detecting Anomalous Patterns of Care Using Health Insurance Claims

Sriram Somanchi
Mendoza College of Business
University of Notre Dame

Edward McFowland III
Carlson School of Management
University of Minnesota

Daniel B. Neill
H.J. Heinz III College
Carnegie Mellon University

Partially funded by National Science Foundation grants IIS-0916345, IIS-0911032, and IIS-0953330, and funding from Disruptive Health Technology Institute. We are also grateful to Highmark Health for providing data.
Agenda

- Introduction
  - Research Question
  - Motivating Example
  - Literature and Contribution

- Methods- Anomalous Patterns of Care (APC) Scan
  - Problem Formulation
  - Algorithm
  - Modeling the scoring function

- Empirical Analysis on Highmark Claims Data
  - Data
  - Results
  - Validation using regression analysis
Introduction

- Challenges the US healthcare system faces\(^1,2\)
  - Instances of over-treatment and under-treatment
  - Inconsistencies in execution of care

Introduction

- Huge opportunity to discover novel patterns of care that are potentially effective due to availability of:
  - Electronic Health Records
  - Documentation of patient care through health insurance claims

- Analyze patterns across patients and provide actionable insights
Research Question

- Given health insurance claims data, we wish to identify a **treatment** and a corresponding **sub-population** for whom that treatment corresponds to significantly better or worse outcomes.
  - Observational data
  - Multiple treatments
  - Population characteristics varying in multiple dimensions
  - Identify **most significant** combination of treatment and sub-population.
Motivating Example

Health Insurance Claims Data

Healthcare Analyst Patrick

Congestive Heart Failure Patients
1. Males
2. Age above 50
3. Similar co-morbidity (atrial fibrillation, on anticoagulant)

Taking Carvidilol is associated with longer stay in hospital

Can we automate the process of producing these interesting hypotheses?
Literature and Contribution

- Heterogeneous Treatments Effects with a given treatment
  - Randomized Control Trials
    - Imai and Ratkovic (2013)
    - McFowland et al. (2015) – see previous talk in this session
  - Observational Studies
    - Athey and Imbens (2015 arXiv)
    - Wager and Athey (2015 arXiv)

- Our Contributions
  - Given multiple treatments, identify combination of treatment and sub-population associated with anomalous outcomes.
  - Computationally efficient algorithm instead of evaluating exponentially many sub-populations
  - Observational studies
    Effectively use observational data to design future randomized control trials
Agenda

- Introduction
  - Research Question
  - Motivating Example
  - Literature and Contribution

- Methods- Anomalous Patterns of Care (APC) Scan
  - Problem Formulation
  - Algorithm
  - Modeling the scoring function

- Empirical Analysis on Highmark Claims Data
  - Data
  - Results
  - Validation using regression analysis
Problem Formulation

- Let $X = (X_1, X_2, ..., X_N)$ be the set of observed covariates for a patient (demographics, diagnoses, etc.)

- Let $T_1, T_2, ..., T_M$ be the set of available treatments

- Let $Y$ be the scalar outcome of interest (for example, total length of hospital stay in following 12 months).
We want to estimate the distribution of potential outcomes for treatment assignments $T_j = 1$, for a given sub-population, $S$

$$f_{j1,s} = f(y^{(1)} | x \in S)$$

Similarly, we want to estimate

$$f_{j0,s} = f(y^{(0)} | x \in S)$$
**Our Goal**

- Identify the combination of treatment and sub-population for which outcomes are most divergent between treated and untreated groups.

\[
\max_s \max_j Div(f_{j1,s}, f_{j0,s})
\]
**Anomalous Patterns of Care Scan**

1. Start with a random sub-population $S$
2. For each $T_j$
   a. Compute the propensity scores
   b. Reweight outcome distributions
   c. Compute Divergence $F_{j,S}$
3. $j^* = \arg\max_j F_{j,S}$
4. Reweight entire population outcomes based on $T_{j^*}$
5. Use MD-Scan to identify $S^* = \arg\max_S F_{j^*,S}$
6. Set $S = S^*$ and repeat steps 2 to 5 until score stops increasing
7. Repeat steps 1-6 for $R$ times
8. Compute statistical significance by randomization testing

Iterative Ascent algorithm between sub-populations and treatments
Inverse Propensity Score Weighting

- We use inverse propensity score weighting to estimate the outcome distribution from observational data

\[
f_{j1,S} = f(y^{(1)} | x \in S) \approx \sum_{x \in S} \frac{f(y, T_j=1, X=x)}{P(T_j=1 | X=x)}
\]

\[
f_{j0,S} = f(y^{(0)} | x \in S) \approx \sum_{x \in S} \frac{f(y, T_j=0, X=x)}{P(T_j=0 | X=x)}
\]
Efficiently Optimizing for Divergence

- **Parametric form**
  - Compute the sufficient statistic
  - Expectation-based Subset Scan framework

- In order to efficiently optimize, the divergence score needs to satisfy the **Linear Time Subset Scanning (LTSS)** property.

- If so, each conditional optimization step becomes linear rather than exponential in the arity of that attribute.
Multi-Dimensional Scan (MD-Scan)

$S^* = \arg\max_S F_{j,S}$

<table>
<thead>
<tr>
<th>Age</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30</td>
<td>$Y_{2M}$</td>
<td>$Y_{2F}$</td>
</tr>
<tr>
<td>30-40</td>
<td>$Y_{3M}$</td>
<td>$Y_{3F}$</td>
</tr>
<tr>
<td>40-50</td>
<td>$Y_{4M}$</td>
<td>$Y_{4F}$</td>
</tr>
<tr>
<td>&gt;50</td>
<td>$Y_{5M}$</td>
<td>$Y_{5F}$</td>
</tr>
</tbody>
</table>

Each step is computationally efficient if divergence function satisfies LTSS property.
Modeling the Scoring Function

- We model the scoring function as generalized log-likelihood ratio statistic

- We assume a parametric distribution for the outcome and compute the sufficient statistics of the expected distribution from the untreated group \((T_j = 0)\)
  - Expectation Based Poisson
  - Expectation Based Gaussian
  - Exponential family distributions
Expectation Based Poisson statistic for potential outcomes

\( H_0 \) : \( Y_i^{(1)} \mid X_i \in X_s \sim \text{Poisson}(\lambda_s) \quad \forall X_s \)

\[ \lambda_s = E[Y^{(0)} \mid X \in X_s] \]

\( H_1(S, q) \) : \( Y_i^{(1)} \mid X_i \in X_s \sim \text{Poisson}(q \ast \lambda_s) \quad X_s \in S \)
\( H_1(S, q) \) : \( Y_i^{(1)} \mid X_i \in X_s \sim \text{Poisson}(\lambda_s) \quad X_s \notin S \)

\[ F(S | q) = \log \frac{P(\text{Data} \mid H_1(S, q))}{P(\text{Data} \mid H_0)} \]

\[ F(S) = \max_q F(S | q) \quad S^* = \max_S F(S) \]
Agenda

- Introduction
  - Research Question
  - Motivating Example
  - Literature and Contribution

- Methods- Anomalous Patterns of Care (APC) Scan
  - Problem Formulation
  - Algorithm
  - Modeling the scoring function

- Empirical Analysis on Highmark Claims Data
  - Data
  - Results
  - Validation using regression analysis
Highmark Claims Data

- Patients with primary or admission diagnosis as ‘diseases of the circulatory system’ from the year 2008 to 2014
  - ~125K patients

![Diagram](chart.png)
Highmark Claims Data

- **Covariates** \((X)\) were built based on:
  - Demographics
  - Median income at patient’s zip code level
  - Diagnosis (primary and secondary)
  - Charlson Comorbidity Index \(^1\)
  - Length of current stay
  - Previous outpatient visits

- **Treatments** \((T_j)\)
  - Drug Therapeutic Class

- **Outcome** \((Y)\)
  - Number of hospitalizations, Total length of stay

---

## Descriptive Statistics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
<th>Percentage of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire Population</td>
<td></td>
<td>100% (124,146)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>53.0%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>47.0%</td>
</tr>
<tr>
<td>Age</td>
<td>Below40</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>40to60</td>
<td>19.8%</td>
</tr>
<tr>
<td></td>
<td>60to80</td>
<td>43.5%</td>
</tr>
<tr>
<td></td>
<td>Above80</td>
<td>33.9%</td>
</tr>
<tr>
<td>Hypertensive</td>
<td>Yes</td>
<td>53.9%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>46.1%</td>
</tr>
<tr>
<td>Diabetic</td>
<td>Yes</td>
<td>29.2%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>70.8%</td>
</tr>
<tr>
<td>Obese</td>
<td>Yes</td>
<td>11.1%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>88.9%</td>
</tr>
<tr>
<td>Primary Diagnosis</td>
<td>Rheumatic (390-398)</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>Hypertensive (401-405)</td>
<td>3.5%</td>
</tr>
<tr>
<td></td>
<td>Ischemic (410-414)</td>
<td>24.5%</td>
</tr>
<tr>
<td></td>
<td>Pulmonary (415-417)</td>
<td>3.7%</td>
</tr>
<tr>
<td></td>
<td>Heart Failure (420-429)</td>
<td>33.0%</td>
</tr>
<tr>
<td></td>
<td>Cerebrovascular (430-438)</td>
<td>16.6%</td>
</tr>
<tr>
<td></td>
<td>Arteries (440-448)</td>
<td>5.0%</td>
</tr>
<tr>
<td></td>
<td>Veins and lymphatics (451-459)</td>
<td>13.2%</td>
</tr>
</tbody>
</table>
Results

- We ran our methodology on this dataset to identify patterns of interest.
- We have ranked order of the highest scoring combination of subpopulation and treatments.
- As a case study, here we discuss the highest scoring subpopulation and treatment pair.
Highest Scoring Subpopulation-Treatment Combination

Subpopulation Characteristics Identified
- Gender
  - Male
- Medical condition
  - Hypertension
  - Obese or Overweight
- Age
  - 40 to 80
- Primary diagnosis
  - Ischemic Heart disease (ICD9 410 – 414)
  - Heart Failure (ICD9 420 – 429)
  - Cerebrovascular heart disease (ICD9 430 – 439)
- Secondary diagnosis
  - No respiratory (ICD9 460 – 519)
  - Endocrine and Immunity disorders (ICD9 240 – 279)

Drug therapeutic class
- Glucocorticoids

Outcome
- More number of hospitalizations

<table>
<thead>
<tr>
<th>Glucocorticoids</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>264</td>
<td>1713</td>
</tr>
<tr>
<td>Mean Number of Hospitalizations</td>
<td>0.606 (0.069)</td>
<td>0.280 (0.016)</td>
</tr>
</tbody>
</table>
Validation of our results

There is huge literature in the medical community on Glucocorticoids and Cardiovascular issues:

- Association using 10 years of observational data (Heart, 2004)
- Metabolic and tissue level effects in heart (European Journal of Endocrinology, 2007)
- Experiments at micro level analysis of glucocorticoids signaling certain receptors in heart for mice (J of Biochemical and Molecular Biology, 2015)
Confirming the results using regression analysis

- We randomly split the data into:
  - 60% for running our APC Scan
  - 40% for running the regression analysis

- Regression with outcome $Y$ as number of hospitalizations with Glucocorticoids as one of independent variable $X$, for
  - The entire population
  - The entire population with a dummy for subpopulation identified by APC Scan
  - The subpopulation identified by APC Scan
  - The complementary subpopulation
Regression analysis (Poisson) on a Hold-Out set

<table>
<thead>
<tr>
<th></th>
<th>Number of Hospitalizations (1)</th>
<th>(2)</th>
<th>Number of Hospitalizations (3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucocorticoids</td>
<td>0.101*** (0.007)</td>
<td></td>
<td>0.410*** (0.089)</td>
<td></td>
</tr>
<tr>
<td>Glucocorticoids Subpopulation</td>
<td></td>
<td>0.265*** (0.088)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subpopulation</td>
<td></td>
<td>-0.313*** (0.068)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.079*** (0.004)</td>
<td></td>
<td>-0.040 (0.079)</td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>0.116*** (0.008)</td>
<td></td>
<td>0.113*** (0.008)</td>
<td></td>
</tr>
<tr>
<td>Hypertensive</td>
<td>-0.163*** (0.008)</td>
<td></td>
<td>-0.161*** (0.008)</td>
<td></td>
</tr>
<tr>
<td>Diabetic</td>
<td>0.286*** (0.008)</td>
<td></td>
<td>0.193*** (0.089)</td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>0.007 (0.013)</td>
<td></td>
<td>0.020 (0.013)</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.773*** (0.044)</td>
<td></td>
<td>-1.634*** (0.120)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>49,658</td>
<td></td>
<td>796</td>
<td></td>
</tr>
</tbody>
</table>

We have included all input characteristics $X$ for our regression

Note: *p<0.1; **p<0.05; ***p<0.01
Sensitivity analysis

- Modifications to the identified subpopulation dramatically reduce the effect.
Robustness checks

- Typical diseases treated using Glucocorticoids
  - Rheumatic Arthritis
  - Chronic Obstructive Pulmonary Disease
  - Cushing’s syndrome

- Ruled out hospital level biases in propensity to treat with Glucocorticoids
  - Overlap coefficient between two groups is 0.78
Ongoing work

- Better estimation of treated and non-treated outcome distributions given sparse data.
- Moving beyond categorical input attributes and binary treatments → incorporate BMI, lab results, etc.
- Using other scoring functions (both parametric and non-parametric).
Summary of our contributions

- Developed a general framework for detecting combinations of treatment and subpopulation that have large deviations in their observed outcomes

- Used multidimensional constraints to scan a large number of subpopulation and treatment combinations in a computationally efficient manner

- Theoretical analysis:
  - Showed that our scoring functions with propensity reweighted outcomes removes the bias from the observed characteristics

- Empirical evaluation:
  - Generated interesting hypothesis related to heart disease by analyzing large, complex and observational health care claims data