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.

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

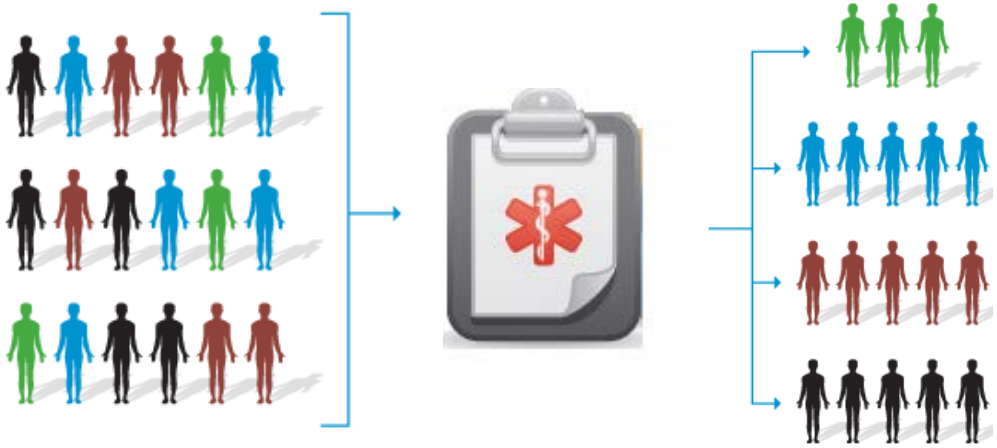
Daniel B. Neill
H.J. Heinz III College
Carnegie Mellon University

Carnegie Mellon University
EPD Lab
EVENT AND PATTERN DETECTION LABORATORY

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

1. N.C. Lallemand, “Health Policy Brief: Reducing Waste in Health Care,” Health Affairs, 13 Dec. 2012.
2. L.T. Kohn, et al., To Err Is Human: Building a Safer Health System, Inst. of Medicine/Nat’l Academy Press, 1999.

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 $\mathrm{X} = (X_1, X_2, \ldots, X_N)$ be the set of observed covariates for a patient (demographics, diagnoses, etc.)

- Let $T_1, T_2, \ldots, T_M$ 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).

Estimating Potential Outcome Distributions

- We want to estimate the distribution of potential outcomes for treatment assignments $T_j = 1$, for a given sub-population, S

$$f_{j1,S} = f\left(y^{(1)} \mid x \in S\right)$$

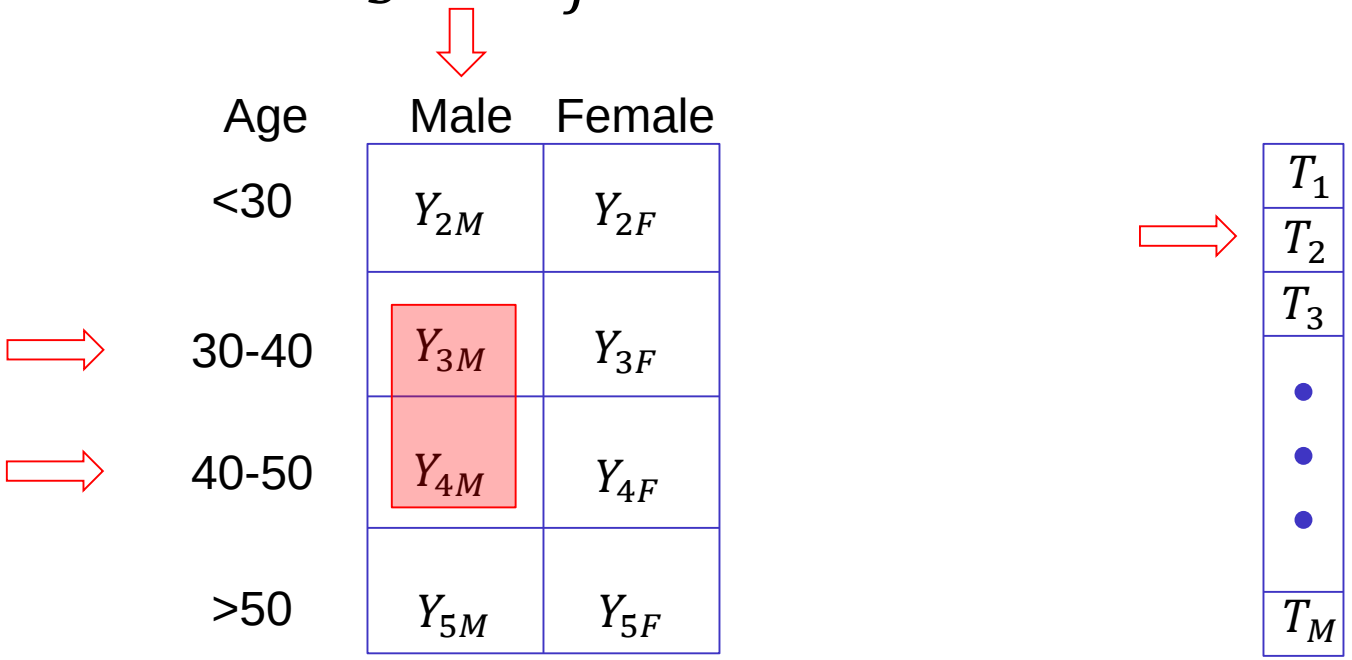
- Similarly, we want to estimate

$$f_{j0,S} = f\left(y^{(0)} \mid x \in S\right)$$

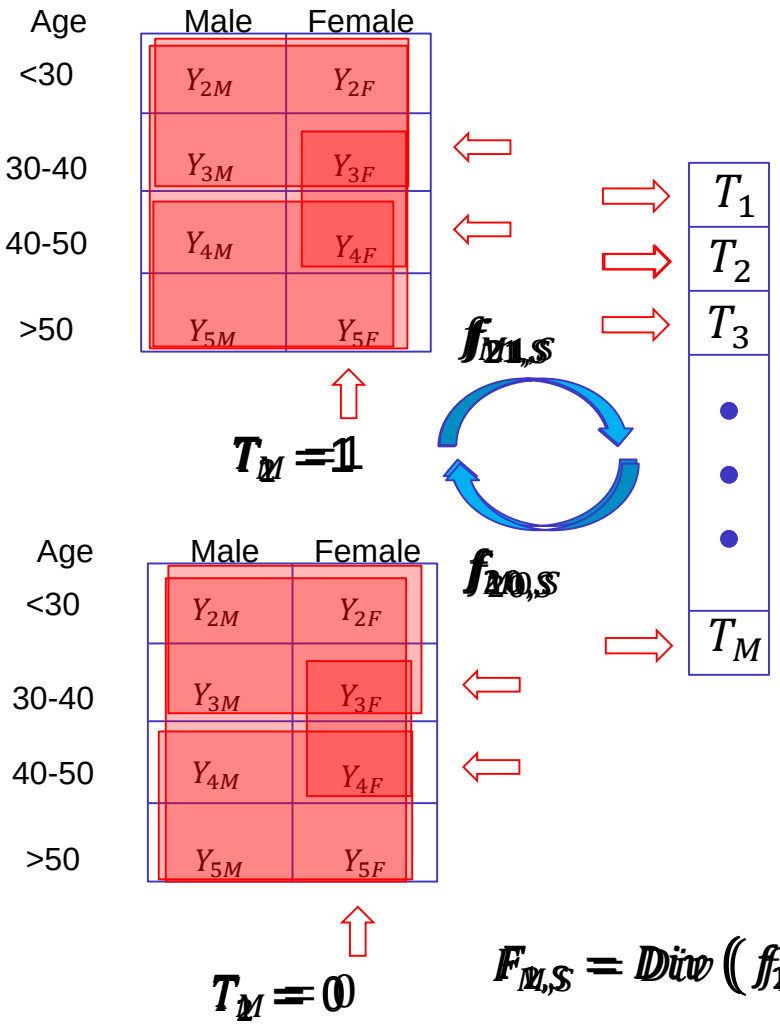
Our Goal

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

$$\max_{S} \max_{j} \; Div\left(f_{j1,S}, f_{j0,S}\right)$$



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. $j^* = \text{argmax}_j F_{j,S}$
4. Reweight entire population outcomes based on T_{j^*}
5. Use MD-Scan to identify $S^* = argmax_S F_{j^*,S}$
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

$$f_{j1,S} = f\left(y^{(1)} \mid x \in S\right) \approx \sum_{x \in S} \frac{f(y, T_j=1, X=x)}{P\left(T_j=1 \mid X=x\right)}$$

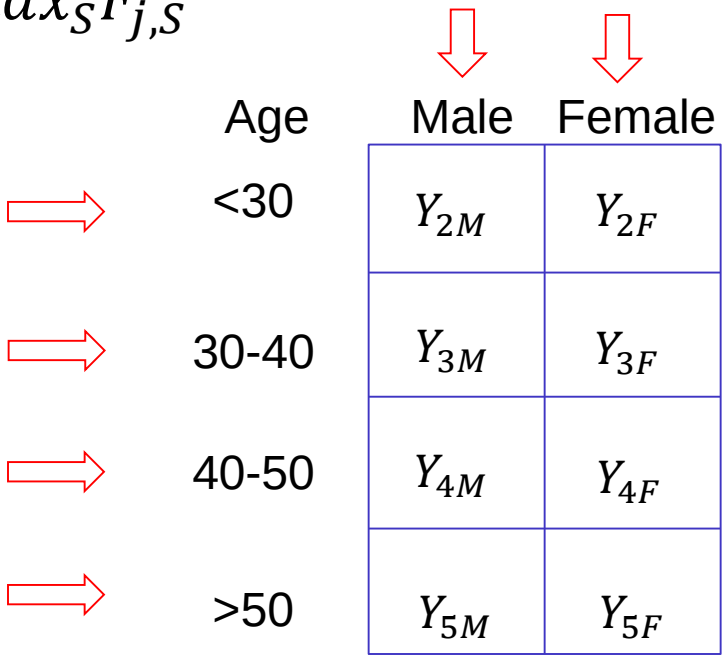
$$f_{j0,S} = f\left(y^{(0)} \mid x \in S\right) \approx \sum_{x \in S} \frac{f(y, T_j=0, X=x)}{P\left(T_j=0 \mid X=x\right)}$$

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)

$$S^* = argmax_S F_{j,S}$$



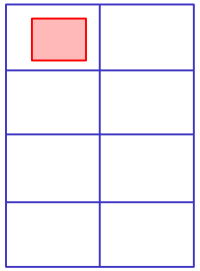
Each step is computationally efficient if divergence function satisfies LTSS property

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



$$\boldsymbol{H_0} \quad : \quad Y_i^{(1)} \mid X_i \in X_s \sim Poisson(\lambda_s) \qquad \forall\, X_s$$

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

$$\boldsymbol{H_1(S,q)} : \quad Y_i^{(1)} \mid X_i \in X_s \sim Poisson(q * \lambda_s) \qquad X_s \in S$$

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

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

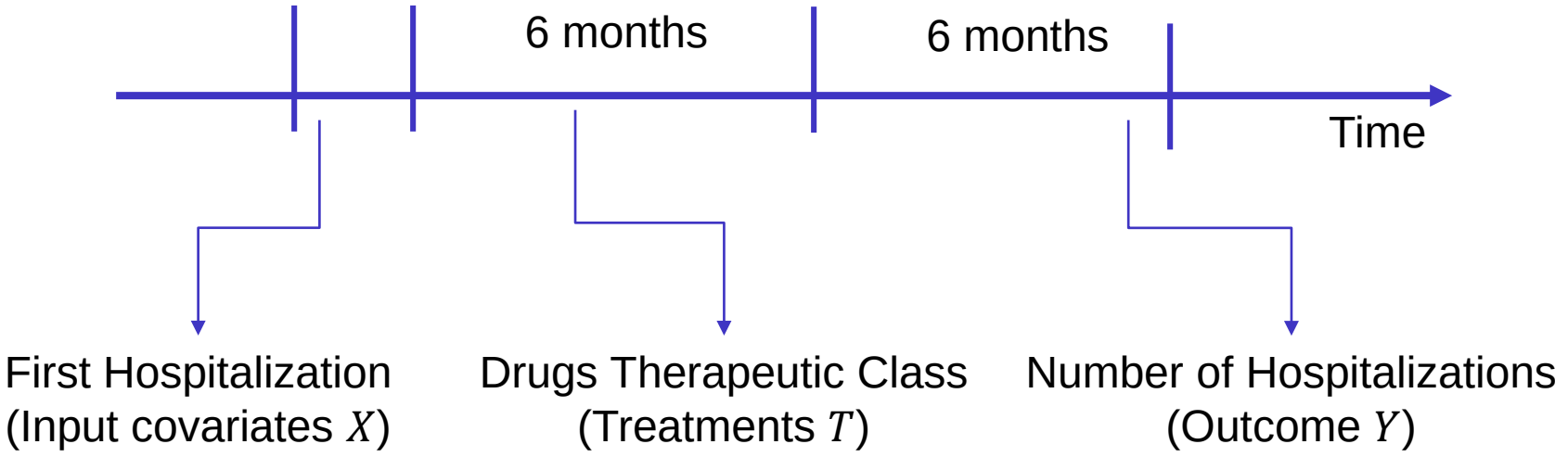
$$F(S) = \max_q F(S|q) \qquad S^* = \max_S F(S)$$

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



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

Bronchial Dilators
Glucocorticoids
Thyroid Preparations
Diabetic Therapy
Lipotropics
Hypotensives
Vasodilators
Digitalis Preparations
Cardiovascular Preparations
Anticoagulants
Diuretics

1. Quan et al (2005). Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. Medical Care, 43(11):1130–1139

Descriptive Statistics

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

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

	Glucocorticoids	
	Yes	No
Number of Patients	264	1713
Mean Number of Hospitalizations	0.606 (0.069)	0.280 (0.016)

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

	Number of Hospitalizations (1)	(2)	Number of Hospitalizations (3)	(4)
Glucocorticoids	0.101*** (0.007)	0.099*** (0.007)	0.410*** (0.089)	0.099*** (0.007)
Glucocorticoids* Subpopulation		0.265*** (0.088)		
Subpopulation		-0.313*** (0.068)		
Age	0.079*** (0.004)	0.079*** (0.004)	-0.040 (0.079)	0.080*** (0.004)
Females	0.116*** (0.008)	0.113*** (0.008)		0.113*** (0.008)
Hypertensive	-0.163*** (0.008)	-0.161*** (0.008)		-0.161*** (0.008)
Diabetic	0.286*** (0.008)	0.286*** (0.008)	0.193*** (0.089)	0.287*** (0.008)
Obesity	0.007 (0.013)	0.020 (0.013)		0.020 (0.013)
…	…	…	…	…
Constant	-0.773*** (0.044)	-0.772*** (0.044)	-1.634*** (0.120)	-0.772*** (0.044)
Observations	49,658	49,658	796	48,862

10.6%

50.6%

(1) Entire Population

(2) Entire Population with dummy for the subpopulation

(3) Only with the subpopulation identified by APC-Scan

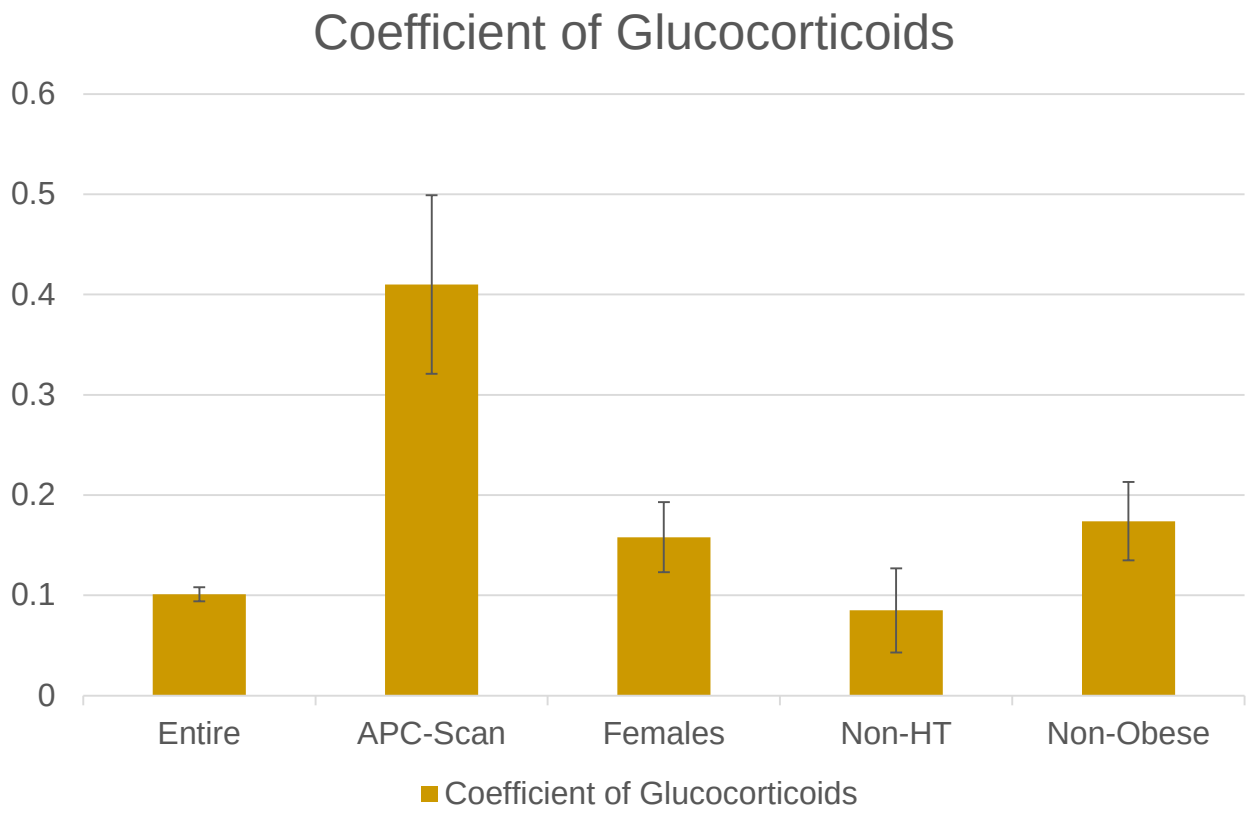
(4) Only with the complementary subpopulation

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